PRECISION ENVIRONMENTAL HEALTH — AN OMICS-BASED WHOLE-MIXTURE APPROACH

by

XIAOJING LI

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF BIRMINGHAM

School of Bioscience

College of Life and Environmental Sciences

University of Birmingham

June 2021

Dedication

To my grandparents and parents

"It was the best of times, it was the worst of times,

it was the age of wisdom, it was the age of foolishness,

it was the epoch of belief, it was the epoch of incredulity,

it was the season of light, it was the season of darkness,

it was the spring of hope, it was the winter of despair."

— Charles Dickens

Abstract

In the natural waters, hundreds to thousands of chemicals co-exist as complex mixtures, which needs a holistic assessment of their health effects. Identifying and testing each individual chemicals in the environment is undoubtably an insurmountable challenge to ecotoxicological studies and an unrealistic approach to reveal mixture effect at environmental relevant concentration, which may require insight from toxicogenomic studies. In this thesis, a new way of understanding and potentially discovering solutions to the mixture effect problem of safeguarding the health of human populations and the environment from the unknown effects of real-world chemical mixtures, specifically targeting pollutants of inland waters.

In Chapter 1, the current status of environmental monitoring, its challenges and limitations by highlighting environmental sample classification and harmful chemical component prioritisation are described and discussed as the major issues.

The conceptual framework of Precision Environmental Health is then proposed in Chapter 2, emphasising the importance of chemical mixture modes of action in the view of multi-omics. The Precision Environmental Health framework applies an omics-based bioassay approach to comprehensively characterise the effect of environmental chemical mixtures. The core of this framework focuses on the identification and interpretation of the molecular key event (mKE), which is responsive of foreign chemical exposure and indicative of potential adverse outcome. The mKEs are subsequently applied to classify the mixture effect and identify associated chemical components. This conceptual framework aims at

integrating the data-driven biological signatures generated by omics profiles and prior knowledge of gene functions and pathways of counterpart genetic model species.

Chapter 3 explains and verifies the mathematical basis of the framework, which relies on multi-block correlation analysis. Two case studies are included to demonstrate this framework in action, and two chemical components (caffeine and carbamazepine) are selected as prove-of-concept. The Data-driven biological features are compared with prior knowledge and compared between two case, in order to prove the effectiveness and robustness of the mathematical assumption behind this framework.

Derived from PEH framework, the mKE was used to group and classify the mixture effects of chemicals at environmentally relevant concentrations in two case studies, as gene clusters of highly variable genes in the transcriptomic profiles were identified and grouping pattern of gene clusters associated with chemical responses in Chapter 4 and further identify chemical component associated signatures that may reflect the chemicals' modes of action in Chapter 5. In Chapter 4, expression-based clustering analysis of five gene clusters revealed that the environmental chemical mixture of a single site (M16) induced relatively higher expression levels in stress response and cellular homeostasis, and these differences are significantly related to Dibenz[a,h]anthracene, Erythromycin and Trimethoprim in the Chaobai case study. In Chapter 5, similarity analysis of chemical profiles and transcriptomic profiles reveal similar grouping pattern, as expression-based clustering analysis of gene clusters revealed that distinctive transcriptomic profiles of two sites (D11 and

D12) reveal down-regulation of xenobiotic biodegradation and antioxidative response pathways.

This thesis ends by highlighting in Chapter 6 the promise of Precision Environmental Health to address harm caused by real world chemical pollutants based on my findings and discusses need for future verification.

Acknowledgements

Throughout my four years at the University of Birmingham (UoB), I have received a great deal of support and assistance.

First and foremost, I would like to express my most profound appreciation to my committee. As my supervisor, Prof. John Colbourne inspires me with novel ideas and supports me with invaluable expertise in formulating the research topic. The completion of my thesis would not have been possible without the constant support and nurturing of John. I am also extremely grateful to my co-supervisor Prof. Ben Brown, who always provides me insightful suggestions on methodology and practical suggestions of machine learning modelling and biostatistics testing. I am deeply indebted to Prof. Mark Viant, who supports me in conducting my experiments and training in the metabolomic technique. I would like to extend my deepest gratitude to my co-supervisor Prof. Liang-hong Guo, who helps me in research and convenience in life when I was conducting my first case study at the Research Centre of Eco-Environmental Science (RCEES) and constructive advice on my thesis. I would like to thank Prof. Stuart Harrod sincerely. He provides expert advice on non-targeted chemical analysis and gives me opportunities to join the Network POPs Conference and build connections with researchers in this field.

Secondly, I would like to acknowledge my colleagues from Environmental Genomics groups for their tremendous support and collaboration. Special thanks to Luisa, who always encourages me, shares her extraordinary experiences in experimental design with me, and invites me to work on the manuscript related to Daphnia gut

microbiota. Many thanks to Tim, who supports me greatly with his experiences and expertise, helps me with my presentation and progress report, and especially provides the precious achieved data and samples from the Solutions project. I am grateful to Vignesh, who is always helpful and patient with my problems related to the Bluebear. Thanks also to Steve for his professional techniques in the molecular lab, which assures my progress. And thanks should also go to Caroline for his indispensable work and support in the Daphnia wet lab. And I am also grateful to Prof. Jan-Ulrich Kreft, who inspires me with valuable suggestions on my project and is willing to discuss with me during lunch break. Moreover, I would like to extend my sincere thanks to my collaborators at RCEES, namely Prof. Yao-hui Bai, Dr. Bin Wan, Dr. Wei-wei Ben, Dr.Meng Qiao, Dr. Wei Xiong, Dr. Guo-rui Liu, for their supports in scheduling fielding sampling, sharing data and opinions, and supporting organic extractions.

I also had the great pleasure of working with Rosie, Marie, Hollie, Niamh, and Muhammad. Especially helpful to me during the first few weeks that Rosie and Marie kindly invited me to camp and helped me quickly integrate into the group. And Niamh always kindly explain the British culture and common sense of life to me. I'd also like to gratefully acknowledge the help of Cate, Julia, Judith, Jelena, Hanna, Julia, Sophia, Rolf and Martin for guiding me through metabolome training and data collection.

最后的最后，我想要感谢我的父母一直以来给予我最大的鼓励和支持，让我能够有机会完成自己的梦想，心无旁骛的从事科研工作。我也要感谢我的亲朋好友一直以来给我的关心和帮助，尤其是在我不自信的时候。

于我而言，这四年来最最幸运的是遇到了我挚爱的家锐，不管是在科研还是生活中他都是我最最坚实的后盾。即便在我最彷徨无助的时候，他也依然会提醒我我所不自知的优秀之处。感谢他为我做的所有事！

Table of Contents

List of Figures

Chapter 4

Chapter 5

List of Tables

Chapter 3

Chapter 4

Chapter 5

List of Abbreviations

AAA	n-Acetyl-4-aminoantipyrine
ABC	ATP-binding cassette (transporter)
Ace	Acenaphthene
ACF	Acesulfame
Acy	Acenaphthylene
Ant	Anthracene
AOP	Adverse Outcome Pathway
ATD	Atrazine-desethyl
ATE	Atenolol
ATR	Atrazine
AZN	Azithromycin
BaA	Benzo[a]anthrancene
BaP	Benzo[a]pyrene
BbF	Benzo[b]fluoranthene
BEN	Bentazone
BF	Bezafibrate
BghiP	Benzo[g,h,i]perylene
BkF	Benzo[k]fluoranthene
BZT	1H-Benzotriazole
CA	Concentration Addition
CAF	Caffeine

CBZ	Carbamazepine
Chry	Chrysene
CIP	Ciprofloxacin
CLA	Clarithromycin
COT	Cotinine
CYP	Cytochrome P450
DBA	Dibenz[a,h]anthracene
DDC	10,11-Dihydro-10,11-dihydroxycarbamazepine
DIMS	Direct Infusion Mass Spectrometry
DOX	Doxycycline
EC50	Half maximal effective concentration
EDA	Effect-directed analysis
ENR	Enrofloxacin
ERY	Erythromycin
Flua	Fluoranthene
Fluo	Fluorene
GC	Gas chromatography
GPX	Glutathione peroxidase
GST	Glutathione S-transferase
HCA	Hierarchical clustering analysis
HPLC	Ultra-performance liquid chromatography
HRMS	High-resolution mass spectrometry
IA	Independent Action
IATA	Integrated Approaches of Testing and Assessment

IncdP	Indeno[1,2,3-cd]pyrene
JDS	Joint Danube Survey
km	kilometer
LC	Liquid chromatography
LOM	Lomefloxacin
LOQ	Limit of quantification
m/z	mass-to-charge ratio
MBZ	5-Methyl-1H-benzotriazole
MET	Metoprolol
mKE	Molecular key event
MoA	Mode of action
MS	Mass spectrometer
MTZ	Metazachlor
Nap	Naphthalene
nESI	Nano-electrospray ionisation assembly
NOEL	No Observed Effect Level
NOR	Norfloxacin
NTS	Non-targeted screening
OG	Ortholog group
OTC	Oxytetracycline
PAHs	Polycyclic aromatic hydrocarbons
PEH	Precision Environmental Health
PFOS	Perfluorooctanesulfonic acid
Phe	Phenanthrene

PROP	Propranolol
Pyr	Pyrene
RCEES	Research Centre of Eco-Environmental Science
ROX	Roxithromycin
rpm	revolutions per minute
sCCA	Sparse Canonical Correlation Analysis
SDZ	Sulfadiazine
SMR	Sulfamerazine
SMX	Sulfamethoxazole
SOD	Superoxide dismutase
SPE	Solid-phase extraction
TDS	Total dissolved solids
TER	Terbuthylazine
TET	Tetracycline
TMP	Trimethoprim
TOC	Total organic carbon
TRA	Tramadol
UFZ	Helmholtz Centre for Environmental Research
WWTP	Wastewater treatment plant

1 Effect Assessment of Environmental Chemical Mixtures

1.1 Abstract

Chemical substances in the environment pose potential health effect on humans and the environment. Effect assessment of chemical substance is the basis of prioritising chemical substances for regulatory monitoring and management. This process includes obtaining and processing environmental samples (field sampling), describing the environmental chemical substances (chemical analysis), toxicological testing with bioanalytical approaches (toxicological bioassay), and identifying chemical substances that drive the overall toxicity (chemical prioritisation). In this chapter, conventional methods of effect assessment are discussed, and solutions that combine whole mixture approach and omics-based bioassays are proposed to facilitate the effect characterisation of the environmental chemical mixture (Figure 1.1).

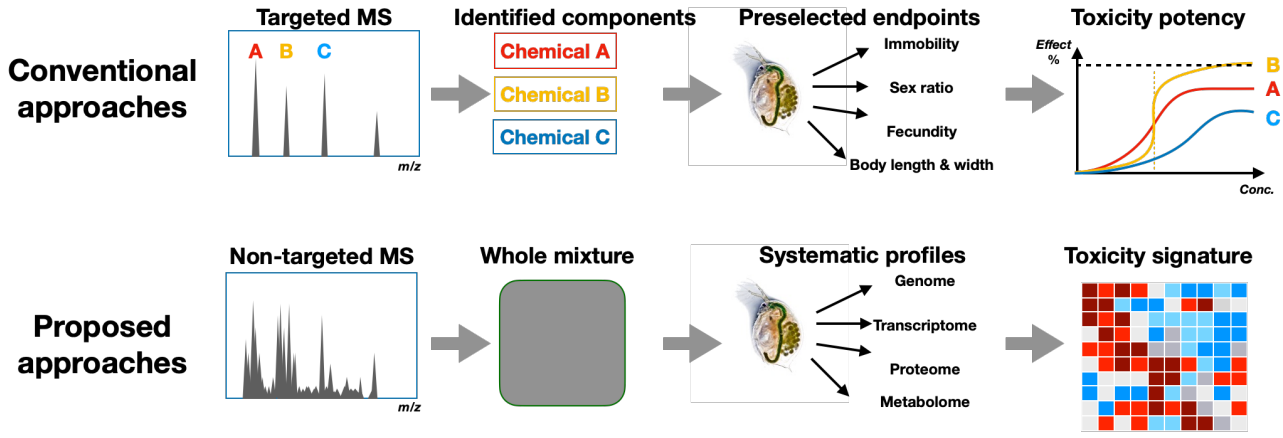

Figure 1. 1 Conventional and proposed approaches of chemical mixture effect assessment.

1.2 Environmental chemical mixture problem

A growing number of chemicals are introduced into the aquatic environment year upon year (Rüdel et al., 2020). To date, there are more than 350,000 chemicals registered for commercial usage globally (Wang et al., 2020). A large number of these chemicals are making their way into the aquatic environment through emission and discharge (Villanueva et al., 2014), resulting in numerous possible combinations of those environmental chemicals co-occurring as complex mixtures (B. I. Escher et al., 2020). Environmental scientists and environmental policymakers realise that such environmental chemical mixtures pose potential threats to aquatic organisms and humans and aquatic ecosystems (Posthuma et al., 2019). According to the Lancet Commission report on pollution and health, chemical pollutants are associated with approximately 9 million premature deaths (Landrigan et al., 2018). Pollution has other severe health effects on human that are known, yet challenging to quantify, including neurotoxicity, developmental toxicity, reproductive toxicity and endocrine disruption (Silva et al., 2015). The environmental chemical mixtures are also disrupting ecosystems by causing growth inhibition, delayed or reduced reproduction and changes in biodiversity affecting various trophic level dynamics (Amoatey and Baawain, 2019), which may further impair the sustainability of natural resources (Rhind, 2009). However, most of the chemicals are regulated at the individual compound level by the regulatory agencies for the protection of water quality and aquatic ecosystems. For example, only 76 priority substances and 17 substances within the watch list are under routine monitoring based on the 2013/39/EU Water Framework Directive in European countries (Kern, 2014); a total of 126 priority pollutants are under regulation of Clean Water Act in the United States (EPA 1991). Regulation on chemical mixtures is only

implemented for specified groups of chemicals. For example, the EU Water Framework Directive specifies thresholds for chemical groups like cyclodiene pesticides, polyaromatic hydrocarbons (PAHs), perfluorooctane sulfonic acid and its derivates, and dioxin and dioxin-like compounds (Kern, 2014). In contrast, there is no current regulation on the mixture effect generated by multiple chemical substances of different chemical classes (Kortenkamp and Faust, 2018). Thus, to better understand and prevent the environmental chemical mixture problem, an exposure assessment of potential health effects should be considered in new policymaking to assure a “good state” of the aquatic environment.

1.3 Conventional methods and their limitations

Conventional methods for effect assessment of environmental chemical substances are mainly targeted, represented by targeted chemical measurement and pre-selected apical endpoints for toxicological bioassay (Brack et al., 2018).

1.3.1 Chemical analysis

Chemical substances in the environment exhibit various physiochemical structures and a wide range of concentrations and polarities; it is challenging to develop a holistic profiling method of all the chemical components in the environment by using a single chemical analytical platform. The gas chromatography (GC) is commonly applied to separate volatile and thermally stable compounds, while the liquid chromatography (LC) can separate non-volatile, semi-polar, and polar compounds, such as pesticides (e.g., organophosphorus, organochlorine and carbamate), pharmaceuticals, illicit drugs, and personal care products (Brack et al., 2016). Coupling the efforts of multiple technical

platforms targeting various characteristics may better resemble the complete picture of the environmental chemical mixture (Ulrich et al., 2019).

Current environmental monitoring relies on targeted analysis, which aims at quantifying the pre-selected chemicals with *a priori* well-defined reference data (e.g., retention time, mass spectrum, and tandem mass spectrum). This procedure typically accounts for tens to hundreds of detectable substances and has been well accepted in many countries and practised for decades, such as the U.S.A surface water quality criteria (Stephan *et al.* 1985), the European framework for water policy 2000/60/EC (Todo and Sata 2002), and the UK water framework directive (DEFRA 2014). It must be noted that environmental chemical monitoring typified by the chosen substances, also known as prioritised chemicals, can only represent a tip of a chemical cocktail iceberg (Brack et al., 2018). Current chemical-based monitoring is limited to about 1 % of detectable substances, while a large number of chemical substances with potential toxicity may leak through regular monitoring as they are not on the priority list. Chemicals may be sometimes below the detection levels, leading to incomplete profiling of the environmental chemicals. Limited information of the total amounts, composition, and individual identities of the environmental chemical mixtures in the waters impedes an understanding of the scope and impact of the aquatic pollution problem, with timely and sufficient evidence to intervene (Altenburger et al., 2015).

1.3.2 Toxicological bioassay for single chemical substance

Toxicological bioassay depicts the levels of biological responses as a function of the concentration of the selected chemical substance. It aims at revealing the monotonic relationship between detected chemical substance and the toxicity potency within the

concentration range or dose frequency defined in the testing settings (Tsatsakis et al., 2018). The prediction of toxicity level can thereby rely on measuring the concentration levels of the tested chemical substance. It is critical to address chemical effect assessment and predict the risk level of chemical substances at the environmental levels. However, the concentration range chosen in any toxicological bioassay may be limited from 5 to 7 concentration levels per study. Most of the testing is done to observe acute toxicity at relatively higher concentration levels far beyond environmental levels of exposure. Such a dose-response model might not have enough observations within the concentration range reflecting environmental levels, leading to an unreliable prediction of effects in the environment (Knillmann et al., 2018).

The adverse effect of a chemical substance is assessed and characterised by the design of the bioassay. Such toxicity relevant bioassays are based on the variation in a specified biomarker (indicative gene, functional protein or metabolite) or a phenotypic trait. Related to xenobiotic metabolism, the biomarkers at receptor levels may include aryl hydrocarbon receptor activation (Brennan et al., 2015), pregnane X receptor activation (Lemaire et al., 2006), hormone receptor (like androgen and estrogen; Wilson, 2002), metabolism homeostasis (PPAR-γ nuclear peroxisome proliferator-activated receptor-γ, Neale et al., 2017) (oxidative stress,(Farmen et al., 2010), photosynthesis inhibition (Muller et al., 2008), enzyme inhibition (like acetylcholinesterase inhibition, Ellman et al., 1961), DNA damage (Lee and Steinert, 2003), protein depletion and lipid peroxidation (Baryla et al., 2000). While at the organismal level, the phenotypic traits may consist of embryo development (Zhang et al., 2003), population growth (de Almeida et al., 2017), and motility and feeding behaviour (Barata et al., 2008). The systemic response would be further summarised

as neurotoxicity (Tohyama, 2016), genotoxicity (Tabrez et al., 2011), mutagenicity (Hakura et al., 2021), immunotoxicity (Germolec et al., 2017), carcinogenicity (Cohen et al., 2019), endocrine disruption (Kwak et al., 2018), reproductive and development impairment (Sidorkiewicz et al., 2017). The harmful effects of chemicals detected in the environment are mainly assessed by toxicity testing in a single-compound-at-one-time manner, and the bioassays are anchored and prescribed by a finite number of toxicological endpoints (Serra et al., 2020).

1.3.3 Effect assessment of chemical mixture using a component-based approach

Since the known chemical components in the environment are limited by the efforts of chemical analysis, the effects of detected chemical components are crucial for assessing the effect of environmental chemical mixture, which is known as the component-based approach. The component-based approach can be regarded as a mixture-wise extension of toxicological effect assessment. It is based on investigating the effects of combinations of a limited number of the detected chemical component in the environment chosen by their occurrences or concentration levels (Kumari and Kumar, 2020). Three assumptions of the component-based approach are:

(1) The mixture of selected chemical compounds can approximate the effect of the chemical mixture in the environment, when the mixture effect is mainly contributed by the effects of bioactive components in a dose or effect addition way (Kumari and Kumar, 2020).

(2) The dose-response relationships of selected chemical components can be established by full factorial design, and the mathematical prediction can be derived from individual chemical toxicity data (Altenburger and Greco, 2009).

(3) The combined effect of selected chemical components can be detectable and characterised as an array of modes of action contributed by each component (McCarty and Borgert, 2006).

The selected chemical components then form an artificial chemical mixture. The toxicity potency of this artificial chemical mixture relies on two models, the Concentration Addition model (CA; Loewe and Muischnek 1926; Loewe S. 1927) and the Independent Action model (IA; Bliss C.I. 1939) (B. Escher et al., 2020). The prediction of the combined effect is calculated by the weighted sum of concentrations of individual chemical components that share a similar mode of action in the CA model or arithmetic sum of probabilities of the response of individual chemical components under a different mode of action in the IA model (Jonker et al., 2005). It is worth noting that these two models could approximate the real-world scenarios only if the following three premises are satisfied:

(1) Joint effects are additive—all the components in the mixture collectively contribute to the mixture effect, and the combined effect can be formulated in an additive way, based on their concentrations (Kortenkamp, 2007).

(2) Chemical interactions can be neglected—interactions between components, either by direct chemical-chemical interaction in the environmental media or toxicokinetic and toxicodynamic phases (Gao et al., 2020), do not affect the overall toxicity significantly.

(3) The mode of action (MoA) of each chemical component is well defined and acknowledged—the knowledge of chemical-related biological responses with the indication of a potential adverse outcome (Meek et al., 2014) is available for all the components in the mixture; and the similarity of the MoA can be

determined by either shared MoA terms or similar responsive features (Spurgeon et al., 2010).

Both models are widely applied and accepted (Altenburger *et al.* 2000, 2004, 2009, 2012, 2013; Backhaus *et al.* 2003; Backhous and Faust 2012; Belden *et al.* 2007; Nina Cedergreen 2014); some studies even regard CA as the default model for mixture toxicity assessment (Backhaus *et al.* 2004; Syberg *et al.* 2009). Currently, the component-based approach is applied to study the mixture effect of selected chemical components with well-defined toxicity delivered by individual chemical toxicological bioassay and mathematical modelling. (Groten *et al.* 2001; Backhaus and Karlsson 2014; Bopp *et al.* 2018). For instance, twelve well-defined chemicals are combined as two artificial mixtures with carefully selected concentration combinations. These 12 chemical substances were widely detected in the Danube River and selected based on their MoAs in a European Union funded SOLUTIONs research project (Busch *et al.* 2016). Twenty-one bioassays were then applied on these two artificial mixtures across multiple test species, from invertebrate (*Daphnia magna*) to vertebrate (*Danio rerio* and *Oryzias latipes*) and mammalian cell lines, to generate toxicological signatures of the mixture effects (Altenburger *et al.* 2018). This study was performed as a proof-of-concept case study that the CA/IA model is able to describe the effects of active chemical component and their mixtures (Hashmi *et al.* 2018). The component-based approach may be used to estimate the independent, additive effect of selected chemical compounds and reveal the combined mixture effect drawn from the knowledge of MoAs.

However, the component-based approach is woefully impractical at predicting natural environmental chemical mixtures. The number of chemical components included in a

component-based mixture study is usually limited to 2 - 20 compounds. It is impossible to assess every possible combination (binary, trinary or even higher-order) of all components (covering several concentration levels) in the environmental chemical mixtures; the total number of samples in a full-factorial design would be challenging and unrealistic to achieve. Moreover, chemicals may either be chosen based on regulatory concern or be limited to chemicals with well-defined MoAs. Even though the well-studied substances may be proven to be hazardous to human health, such a small set of substances may only represent a tiny portion of the actual environmental chemical mixture and may not characterise the overall complexity, thereby underestimating the combined effect of the corresponding environmental chemical mixtures (Cizmas *et al.* 2004). Finally, the potential low dose effects of compounds create another challenging issue for component-based studies. The artificial chemical mixtures at their environmental concentrations are unlikely to trigger an observable toxicological effect during acute exposures. To address this issue, artificial chemical mixtures may be enriched by increasing the environmentally relevant concentrations of their components (100- to 1000-fold) to trigger observable toxicity potency changes (Altenburger *et al.* 2018). However, enrichment of mixture concentrations faces the same dilemma of individual chemical-based toxicity testing by failing to represent toxicological effects under real world environmental scenarios. Based on individual chemical-based toxicity testing, the concentration level of a specific chemical that induces no observable effect may be identified, known as the No Observed Effect Level (NOEL). However, the nature of NOEL is a statistical term where the effect size is not significantly different from the control level. The combined effect of chemicals at low dose levels might deviate from CA/IA model or even be enhanced, known as synergism

(Jonker *et al.* 2005). A few studies have reported significant synergistic effects of a mixture consisting of chemicals at individual NOEL or low effect level (Walter *et al.* 2002; Reffstrup *et al.* 2010; Hass *et al.* 2012; Orton *et al.* 2013; Kortenkamp A. 2008, 2014), suggesting that by extrapolating mixture effect from higher-dose-range, substance-based concentration-effect data may underestimate the overall toxicity effect.

Therefore, although effect assessment of chemical mixture with the component-based approach is easier to apply and verify, the mixture effect can be

(1) highly biased by selected chemical components,

(2) significantly underestimated if low dose synergic effects exist, and

(3) cannot account for a large amount of unknown (without toxicity data) chemical components that exist in the environment nor cover potential harmful effects that are unintentionally neglected because of observations of a limited set of targeted endpoints.

1.3.4 Effect assessment of chemical mixture using effect-directed analysis

The effect-directed analysis (EDA) was proposed in the early 1980s to combine the efforts of chemical analysis, chemical mixture fractionation, and *in vitro* bioassays in a tiered approach in order to reduce mixture complexity and provide evidence of mixture toxicity and unknown causative toxic chemical substance in the environmental chemical mixture (Brack W. 2003; Brack and Burgess, 2011).

For water samples, the sequential removal of the non-toxic chemical components (fractionation) is achieved by solid-phase extraction (SPE) and LC with specified physicochemical characteristics (Brack W. 2003). Several fractionation steps may be

included until the toxic components are isolated and verified by bioassays (Brack *et al.* 2008; Brack *et al.* 2016). The associations between active components and toxic effects can thus be carefully established (Thomas *et al.* 2004), and the toxic chemical components can be identified and confirmed by further chemical analyses. For identifying the toxic driver at site or basin scale, EDA can feasibly deliver evidence of effects caused by a few individual chemicals or specific chemical classes, especially for a receptor-related effect like enzyme inhibition, metabolic failure, or endocrine disruption (Brack *et al.* 2018). It has been applied to analyse an organic extract from a Danube River water sample affected by untreated wastewater. The results revealed that selective fractionation contained compounds that may cause androgenic and estrogenic responses (Hashmi *et al.* 2018). Using a parallel fractionation approach, Muschket *et al.* (2018) also linked the likely cause of antiandrogenic activity to be 4-methyl-7-diethylaminocoumarin and two derivatives. Latest development in high-throughput bioassays for mutagenicity and endocrine disruption activities allow efficient identification of mutagen and androgen in the surface water and wastewater treatment plant effluents (Houtman *et al.* 2020; Zwart *et al.* 2020).

As chemical mixture partitioning is the critical step in the EDA approach, issues related to extraction and effect confirmation have a considerable impact on the approach's success (Brack *et al.* 2016). For extracting multiple environmental samples in parallel, the recovery rate at extracting may differ between chemical compounds (Zhou *et al.* 2017) and the co-eluting chemical compounds might be affected by the matrix effect (Benijts *et al.* 2004). Optimising the recovery rate during extraction is key to assure the recovery of the active toxic components in the fractions. Effect confirmation can only be performed after the substance identities have been confirmed by targeted analysis,

which requires further efforts and time for structure confirmation with the aid of a standard compound (Vughs *et al.* 2018). However, if the standard compound is not available (*e.g.*, the substances to be confirmed is a transformed product due to sample processing), it remains impossible to confirm the main driver of the observed toxicity.

Hence current methods for effect assessment of environmental chemical substances are limited by the capacity of targeted analysis and chemical identity confirmation and the choice of toxicological bioassays (limited apical endpoints).

1.4 Preferred approaches (clarify the novelties proposed in the thesis)

Given the limitations of current methods outlined above, a fundamentally novel solution should include a whole-mixture approach that would fully reveal the reality and complexity of environmental chemical mixture, combined with a non-targeted bioassay that would comprehensively capture the subtle biological responses that precisely represent biomolecular differences in the underlying mechanism of toxicity.

1.4.1 The whole-mixture approach

A whole-mixture approach to chemical effect assessment measures the environmental chemical mixtures at real world concentration levels and composition. It treats the whole chemical mixture in its entirety. The composition of the whole mixture could be the totality of the chemical substances in the environment treating used for toxicity testing or a sub-total of chemical components that possess similar physicochemical characteristics based on the selectivity of an extraction/enrichment method (Brack *et al.* 2016; Mount and Hockett, 2000; Burgess *et al.* 2013). Yet, the relative abundance or

concentration of each component could also be incomplete. Advancements in analytical chemistry has enabled a wider analytical window.

1.4.1.1 The non-targeted analysis of chemical mixture

The state-of-art ionisation technique and high-resolution mass spectrometry (HRMS) promise greater performance in sensitivity, mass accuracy, mass resolution and mass range (Krauss *et al.* 2010; Hollender *et al.* 2017). The non-targeted analysis based on HRMS aims at describing the presence and composition of chemical substances without reference to prior information. It provides an opportunity of identifying unknown chemical substances that are not regularly included in any targeted analysis (Schymanski *et al.* 2014; Hernández *et al.* 2005; Aceña *et al.* 2015). The chemical profile generated by non-targeted analysis may be interpreted as a chemical fingerprint that is unique to a sample by virtue of its unique chemical composition. With non-targeted analysis of micro-pollutants in water, comprehensive profiles of 546 pesticides and 1212 pharmaceuticals (Masiá *et al*. 2014), 760 petroleum metabolites (Mohler *et al*., 2013), or more than 1880 organic compounds (Hernández *et al*. 2015) in surface waters can be obtained by using advanced HRMS. Such screening method would assist in describing and understanding of the whole mixture. The whole-mixture approach may benefit from holistic chemical fingerprinting by combining target and non-target screening techniques (Gago-Ferrero *et al.* 2015; Ruff *et al.* 2015; Postigo *et al.* 2021) with multiple analytical platforms (Kortenkamp *et al.* 2019). Gago-Ferrero *et al.* (2015) conducted suspect screening on surfactants and pharmaceutical

metabolites, accompanied with non-targeted screening (NTS) on the unknown mass peaks in the wastewater. Ruff *et al.* (2015) performed target and non-targeted analysis of water samples from the Rhine River, which revealed quantitative measurements of 302 substances and existence of two substances (Tizanidine and 1,3-Dimethyl-2-imidazolidinone) that were never reported before. Postigo *et al.* (2021) described the quantities of 47 disinfection by-products (DBPs) in the drinking water with targeted analysis and tentatively identified 86 DBPs with NTS.

Although the non-targeted analysis is promising in generating comprehensive chemical fingerprinting of the environmental chemical mixtures, costs in instruments and needs for expertise hinder larger-scale deployment of this technique. It may be difficult to define and describe the whole mixture of an environmental sample, which may be unstable over time when tests are performed. Factors such as time, pH, temperature, and sunlight may have a significant impact on the stability, solubility and volatility of the chemical components of a mixture. The chemical mixture may include components that are unidentifiable and unquantifiable. Some of the chemical substances could be temporary transformed products in the environment or metabolites during bio-mediated metabolism, which may not be described or detected before (La Farre *et al.* 2008; Celiz *et al.* 2009). Assays may be unable to detect chemical components that fall below detection levels or are loss during the sampling or extraction processes.

1.4.1.2 The whole-mixture bioassays

The whole-mixture approach can be also implemented in the bioassays. For example, Escher *et al.* (2014) combined 103 *in vitro* bioassays across multiple model species (human cell line, zebrafish, yeast, algae, *etc.*) to study the potential effects of 10 water

samples, revealing the feasibility of evaluating biological responses of the chemical mixture by applying a battery of *in vitro* bioassays; Blackwell *et al.* (2019) applied 69 assays in a multiplexed way targeting a variety of metabolic pathways to study the potential effect of surface waters in the United States, indicating the potentially of high-throughput screening bioassay. The diversity of readouts for toxicological testing in these two examples are exceptional (compared to other cases, citations), which indicates assessing the environmental water samples as a whole with a wider spectrum of toxicological bioassays can deliver rich information in toxicological signatures for discriminating subtle differences in the mixture effects of water samples.

1.4.2 The omics-based bioassay

Developments in the DNA sequencing platforms, high-resolution mass spectrometers, and computational capability contribute to the development of omics or multi-omics techniques that could reveal in-depth toxicological responses at the biomolecular level (Canzler *et al.* 2020; Sun *et al.* 2019). The in-depth measurements of gene composition (genome), gene expression (transcriptome), protein composition (proteome), metabolic activity (metabolome), and DNA methylation and histone modification (epigenome), may altogether offer simultaneously comprehensive measurements of the toxicological responses to chemical mixtures from a singular sample. Such non-targeted, hypothesis-free approaches can thereby provide

(1) systemic and holistic description of the response mechanism, which may further facilitate the discovery of critical molecular events related to chemical exposure (Stegeman *et al.* 2018);

(2) co-responsive or biologically related metabolic events linked to the same phenotype (Spurgeon *et al.* 2010), which sheds light on response mechanism of toxicologically relevant molecular processes (Groten *et al.* 2004).

1.4.2.1 Transcriptome

Transcriptome can reveal all the messenger RNAs in a biological test system to identify expressional genetic variations linked to exposed chemical compounds (Joseph P. 2017). Take the transcriptome of ecotoxicological model species *Daphnia magna* as example. The transcriptomic profiles of *Daphnia magna* has been extensively used in characterising the biological responses to metals (Antczak *et al.* 2013; Brun *et al.* 2019), endocrine disruptors (Antczak *et al.* 2013; Jeong *et al.* 2013), pharmaceuticals (Antczak *et al.* 2013; Russo *et al.* 2018, Fuertes *et al.* 2019b), flame retardants (Scanlan *et al.* 2015), benzotriazoles (Giraudo *et al.* 2017), pesticide (Fuertes *et al.* 2019a), herbicide (Suppa *et al.* 2020), and their simple mixtures (Garcia-Reyero *et al.* 2012; Fuertes *et al.* 2019b; Brun *et al.* 2019). Only a few studies used transcriptomic profiles to depict the biological responses of environmental chemical mixtures (Garcia-Reyero *et al.* 2012; Kim *et al.* 2017).

1.4.2.2 Metabolome

The metabolome can detect and profile metabolites in the biological system so as to describe xenobiotic-driven variations (Perhar and Arhonditsis 2015; Viant *et al.* 2019; Pomfret *et al.* 2020). Since 2010s, the metabolomic assay has been applied to reveal the mode of action of PAHs (Vandenbrouck *et al.* 2010), metals (Taylor *et al.* 2010; Poynton *et al.* 2011; Nagato *et al.* 2013; Li *et al.* 2015;), insecticides (Taylor *et al.* 2010;

Kovacevic *et al.* 2019), pharmaceuticals (Taylor *et al.* 2010; Wagner *et al.* 2017; Kovacevic *et al.* 2018; Wagner *et al.* 2018), fungicides (Kovacevic *et al.* 2016; Wagner *et al.* 2017; Kovacevic *et al.* 2019), perfluorooctanesulfonic acid (PFOS, Wagner *et al.* 2017; Kariuki *et al.* 2017; Wagner *et al.* 2018; Kovacevic *et al.* 2019), organophosphates (Nagato *et al.* 2016; Kovacevic *et al.* 2018), bisphenol-A (Nagato *et al.* 2016; Garreta-Lara *et al.* 2021), flame retardant (Kovacevic *et al.* 2019), halogenated acetic acids (Labine and Simpson 2021), *etc.* A few studies reported unique metabolomic profiles of *Daphnia magna* under chemical mixture exposure, which were related to energy disruption represented by decrease in glucose (Wagner *et al.* 2018; Kovacevic *et al.* 2019). Wagner *et al.* (2019) tried to display wastewater effluents' (before and after chlorination) impacts on *Daphnia magna* metabolome after 48 hours exposure and revealed that exposure to chlorinated effluent may induce decreases in energy molecules, suggested that metabolome is sensitive enough to depict the mixture effect of environmental chemical mixtures that facilitates environmental biomonitoring.

As omics-based bioassay provides unprecedented details of biological responses at the molecular level, such molecular profiles could be used to reveal the subtle differences in chemical mixture effects that improves our understanding of the effects of the environmental chemical mixture (Seeger *et al.* 2019).

1.4.3 Chemical prioritisation based on the chemical mode of action

Currently, the individual chemical toxicity testing provides information of toxic potency at certain concentration levels, which are used to identify potential key driver of the mixture effect of environmental chemical cocktails. The identified chemical

components are prioritised as the toxic candidates of concern, and thresholds may be subsequently set for regulatory monitoring and strict management, as part of water quality criteria and wastewater outlet standards (Daginnus *et al.* 2011; von der Ohe *et al.* 2011). The NORMAN network developed a prioritisation scheme for emerging environmental substances in European surface and drinking water with EDA approach at assessing European surface and drinking waters (Brack *et al.* 2012). The latest NORMAN list of emerging substances includes 967 compounds (http://www.norman-network.net/?q=node/19), while with the newly developed non-targeted screening methods, this list will continuously grow (Dulio *et al.* 2018). But the knowledge of prioritised chemical substances generated by these conventional methods is quite limited and cannot meet the need for assessing the vast majority of environmental chemicals.

With whole-mixture approach and omics-based bioassays, the resulting molecular profiles provide comprehensive mechanistic characterisation of the effect of chemical components and their mixture (Martins *et al.* 2019). The biological signatures generated by omics-bioassays with whole-mixture approach can be assembled as the MoA of mixture effect. The MoA of mixture effect can be applied in generating categorisation of the chemical mixtures (Sparks *et al.* 2015), proposing hypothesis of MoA of chemical component and interactions between chemical components (Ge *et al.* 2015), and identifying robust biomarkers for environmental monitoring and risk assessment (Borgert *et al.* 2004).

The integration of multiple omics (multi-omics) may help generalise the biological findings across multiple omics platform. It substantiates cohesive analysis of target pathways responding to foreign compounds and provides an opportunity to identify

incidental pathways that revealed the cellular processes from multiple perspectives (Norris *et al.* 2017). The integration of multiple omics can further clarify the relationships between the factor of interest (environmental factors or chemical pollutants or diseases) and genotype for discovering molecular mechanisms of biological responses towards a certain chemical component (Hasin *et al.* 2017; Canzler *et al.* 2019). As the integration of transcriptome and metabolome can facilitate the portray of post-transcription activities, it provides evidence in both gene expressional activity and functional metabolite activity under the same exposure condition (Kumar *et al.* 2016) for constructing gene-metabolite regulatory network.

Hence, the MoA seems to be a promising approach to describe the differences in the effect of environmental chemical mixtures comprehensively and effectively. It may further assist in identifying unknown or undetected chemical component with considerable contribution to the mixture effect.

1.5 Conclusion

Current environmental monitoring and effect assessment was limited by substance-wise adversity-based toxicological research. For assessing the effect of the environmental chemical mixture, a whole-mixture approach should be applied to establish a realistic exposure that represents the environmental scenarios for evaluating the dissimilarity among different mixtures, and an omics-based bioassay should be involved in order to reveal the systemic biological responses as the mixture effect.

1.6 Reference

Altenburger, R., Ait-Aissa, S., Antczak, P., Backhaus, T., Barceló, D., Seiler, T.-B., Brion, F., Busch, W., Chipman, K., de Alda, M.L., de Aragão Umbuzeiro, G., Escher, B.I., Falciani, F., Faust, M., Focks, A., Hilscherova, K., Hollender, J., Hollert, H., Jäger, F., Jahnke, A., Kortenkamp, A., Krauss, M., Lemkine, G.F., Munthe, J., Neumann, S., Schymanski, E.L., Scrimshaw, M., Segner, H., Slobodnik, J., Smedes, F., Kughathas, S., Teodorovic, I., Tindall, A.J., Tollefsen, K.E., Walz, K.-H., Williams, T.D., Van den Brink, P.J., van Gils, J., Vrana, B., Zhang, X., Brack, W., 2015. Future water quality monitoring — Adapting tools to deal with mixtures of pollutants in water resource management. Sci. Total Environ. 512–513, 540–551. https://doi.org/10.1016/j.scitotenv.2014.12.057

Altenburger, R., Greco, W.R., 2009. Extrapolation concepts for dealing with multiple contamination in environmental risk assessment. Integr Env. Assess Manag 7.

Amoatey, P., Baawain, M.S., 2019. Effects of pollution on freshwater aquatic organisms. Water Environ. Res. 91, 1272–1287. https://doi.org/10.1002/wer.1221

Barata, C., Alañon, P., Gutierrez-Alonso, S., Riva, M.C., Fernández, C., Tarazona, J.V., 2008. A Daphnia magna feeding bioassay as a cost effective and ecological relevant sublethal toxicity test for Environmental Risk Assessment of toxic effluents. Sci. Total Environ. 405, 78–86. https://doi.org/10.1016/j.scitotenv.2008.06.028

Baryla, A., Laborde, C., Montillet, J.-L., Triantaphylidès, C., Chagvardieff, P., 2000. Evaluation of lipid peroxidation as a toxicity bioassay for plants exposed to copper. Environ. Pollut. 109, 131–135. https://doi.org/10.1016/S0269-7491(99)00232-8

Brack, W., Ait-Aissa, S., Burgess, R.M., Busch, W., Creusot, N., Di Paolo, C., Escher, B.I., Mark Hewitt, L., Hilscherova, K., Hollender, J., Hollert, H., Jonker, W., Kool, J., Lamoree, M., Muschket, M., Neumann, S., Rostkowski, P., Ruttkies, C., Schollee, J., Schymanski, E.L., Schulze, T., Seiler, T.-B., Tindall, A.J., De Aragão Umbuzeiro, G., Vrana, B., Krauss, M., 2016. Effect-directed analysis supporting monitoring of aquatic environments — An in-depth overview. Sci. Total Environ. 544, 1073–1118. https://doi.org/10.1016/j.scitotenv.2015.11.102

Brack, W., Escher, B.I., Müller, E., Schmitt-Jansen, M., Schulze, T., Slobodnik, J., Hollert, H., 2018. Towards a holistic and solution-oriented monitoring of chemical status of European water bodies: how to support the EU strategy for a non-toxic environment? Environ. Sci. Eur. 30, 33. https://doi.org/10.1186/s12302-018-0161-1

Brennan, J.C., He, G., Tsutsumi, T., Zhao, J., Wirth, E., Fulton, M.H., Denison, M.S., 2015. Development of Species-Specific Ah Receptor-Responsive Third Generation CALUX Cell Lines with Enhanced Responsiveness and Improved Detection Limits. Environ. Sci. Technol. 49, 11903–11912. https://doi.org/10.1021/acs.est.5b02906

Cohen, S.M., Boobis, A.R., Dellarco, V.L., Doe, J.E., Fenner-Crisp, P.A., Moretto, A., Pastoor, T.P., Schoeny, R.S., Seed, J.G., Wolf, D.C., 2019. Chemical carcinogenicity revisited 3: Risk assessment of carcinogenic potential based on the current state of knowledge of carcinogenesis in humans. Regul. Toxicol. Pharmacol. 103, 100–105. https://doi.org/10.1016/j.yrtph.2019.01.017

de Almeida, A.C.G., Petersen, K., Langford, K., Thomas, K.V., Tollefsen, K.E., 2017. Mixture toxicity of five biocides with dissimilar modes of action on the growth and photosystem II efficiency of *Chlamydomonas reinhardtii*. J. Toxicol. Environ. Health A 80, 971–986. https://doi.org/10.1080/15287394.2017.1352176

Ellman, G.L., Courtney, K.D., Andres, V., Featherstone, R.M., 1961. A new and rapid colorimetric determination of acetylcholinesterase activity. Biochem. Pharmacol. 7, 88–95. https://doi.org/10.1016/0006-2952(61)90145-9

Escher, B., Braun, G., Zarfl, C., 2020. Exploring the Concepts of Concentration Addition and Independent Action Using a Linear Low-Effect Mixture Model. Environ. Toxicol. Chem. 39, 2552–2559. https://doi.org/10.1002/etc.4868

Escher, B.I., Stapleton, H.M., Schymanski, E.L., 2020. Tracking complex mixtures of chemicals in our changing environment. Science 367, 388–392. https://doi.org/10.1126/science.aay6636

Farmen, E., Olsvik, P.A., Berntssen, M.H.G., Hylland, K., Tollefsen, K.E., 2010. Oxidative stress responses in rainbow trout (Oncorhynchus mykiss) hepatocytes exposed to pro-oxidants and a complex environmental sample. Comp. Biochem. Physiol. Part C Toxicol. Pharmacol. 151, 431–438. https://doi.org/10.1016/j.cbpc.2010.01.008

Gao, Y., Xie, Z., Feng, M., Feng, J., Zhu, L., 2020. A biological characteristic extrapolation of compound toxicity for different developmental stage species with toxicokinetic-toxicodynamic model. Ecotoxicol. Environ. Saf. 203, 111043. https://doi.org/10.1016/j.ecoenv.2020.111043

Germolec, D., Luebke, R., Rooney, A., Shipkowski, K., Vandebriel, R., van Loveren, H., 2017. Immunotoxicology: A brief history, current status and strategies for future immunotoxicity assessment. Curr. Opin. Toxicol. 5, 55–59. https://doi.org/10.1016/j.cotox.2017.08.002

Hakura, A., Awogi, T., Shiragiku, T., Ohigashi, A., Yamamoto, M., Kanasaki, K., Oka, H., Dewa, Y., Ozawa, S., Sakamoto, K., Kato, T., Yamamura, E., 2021. Bacterial mutagenicity test data: collection by the task force of the Japan pharmaceutical manufacturers association. Genes Environ. 43, 41. https://doi.org/10.1186/s41021-021-00206-1

Jonker, M.J., Svendsen, C., Bedaux, J.J.M., Bongers, M., Kammenga, J.E., 2005. SIGNIFICANCE TESTING OF SYNERGISTIC/ANTAGONISTIC, DOSE LEVEL–DEPENDENT, OR DOSE RATIO–DEPENDENT EFFECTS IN MIXTURE DOSE–RESPONSE ANALYSIS. Environ. Toxicol. Chem. 24, 2701. https://doi.org/10.1897/04-431R.1

Kern, K., 2014. New Standards for the Chemical Quality of Water in Europe under the New Directive 2013/39/EU. J. Eur. Environ. Plan. Law 11, 31–48. https://doi.org/10.1163/18760104-01101002

Knillmann, S., Orlinskiy, P., Kaske, O., Foit, K., Liess, M., 2018. Indication of pesticide effects and recolonization in streams. Sci. Total Environ. 630, 1619–1627. https://doi.org/10.1016/j.scitotenv.2018.02.056

Kortenkamp, A., 2007. Ten Years of Mixing Cocktails: A Review of Combination Effects of Endocrine-Disrupting Chemicals. Environ. Health Perspect. 115, 98–105. https://doi.org/10.1289/ehp.9357

Kortenkamp, A., Faust, M., 2018. Regulate to reduce chemical mixture risk. Science 361, 224–226. https://doi.org/10.1126/science.aat9219

Kumari, M., Kumar, A., 2020. Identification of component-based approach for prediction of joint chemical mixture toxicity risk assessment with respect to human health: A critical review. Food Chem. Toxicol. 143, 111458. https://doi.org/10.1016/j.fct.2020.111458

Kwak, K., Ji, K., Kho, Y., Kim, P., Lee, J., Ryu, J., Choi, K., 2018. Chronic toxicity and endocrine disruption of naproxen in freshwater waterfleas and fish, and steroidogenic alteration using H295R cell assay. Chemosphere 204, 156–162. https://doi.org/10.1016/j.chemosphere.2018.04.035

Landrigan, P.J., Fuller, R., Acosta, N.J.R., Adeyi, O., Arnold, R., Basu, N. (Nil), Baldé, A.B., Bertollini, R., Bose-O'Reilly, S., Boufford, J.I., Breysse, P.N., Chiles, T., Mahidol, C., Coll-Seck, A.M., Cropper, M.L., Fobil, J., Fuster, V., Greenstone, M., Haines, A., Hanrahan, D., Hunter, D., Khare, M., Krupnick, A., Lanphear, B., Lohani, B., Martin, K., Mathiasen, K.V., McTeer, M.A., Murray, C.J.L., Ndahimananjara, J.D., Perera, F., Potočnik, J., Preker, A.S., Ramesh, J., Rockström, J., Salinas, C., Samson, L.D., Sandilya, K., Sly, P.D., Smith, K.R., Steiner, A., Stewart, R.B., Suk, W.A., van Schayck, O.C.P., Yadama, G.N., Yumkella, K., Zhong, M., 2018. The Lancet Commission on pollution and health. The Lancet 391, 462–512. https://doi.org/10.1016/S0140-6736(17)32345-0

Lee, R.F., Steinert, S., 2003. Use of the single cell gel electrophoresis/comet assay for detecting DNA damage in aquatic (marine and freshwater) animals. Mutat. Res. Mutat. Res. 544, 43–64. https://doi.org/10.1016/S1383-5742(03)00017-6

Lemaire, G., Mnif, W., Pascussi, J.-M., Pillon, A., Rabenoelina, F., Fenet, H., Gomez, E., Casellas, C., Nicolas, J.-C., Cavaillès, V., Duchesne, M.-J., Balaguer, P., 2006. Identification of New Human Pregnane X Receptor Ligands among Pesticides Using a Stable Reporter Cell System. Toxicol. Sci. 91, 501–509. https://doi.org/10.1093/toxsci/kfj173

McCarty, L.S., Borgert, C.J., 2006. Review of the toxicity of chemical mixtures: Theory, policy, and regulatory practice. Regul. Toxicol. Pharmacol. 45, 119–143. https://doi.org/10.1016/j.yrtph.2006.03.004

Meek, M.E. (Bette), Palermo, C.M., Bachman, A.N., North, C.M., Jeffrey Lewis, R., 2014. Mode of action human relevance (species concordance) framework: Evolution of the Bradford Hill considerations and comparative analysis of weight of evidence. J. Appl. Toxicol. 34, 595–606. https://doi.org/10.1002/jat.2984

Muller, R., Schreiber, U., Escher, B.I., Quayle, P., Bengtson Nash, S.M., Mueller, J.F., 2008. Rapid exposure assessment of PSII herbicides in surface water using a novel chlorophyll a fluorescence imaging assay. Sci. Total Environ. 401, 51–59. https://doi.org/10.1016/j.scitotenv.2008.02.062

Neale, P.A., Altenburger, R., Aït-Aïssa, S., Brion, F., Busch, W., de Aragão Umbuzeiro, G., Denison, M.S., Du Pasquier, D., Hilscherová, K., Hollert, H., Morales, D.A., Novák, J., Schlichting, R., Seiler, T.-B., Serra, H., Shao, Y., Tindall, A.J., Tollefsen, K.E., Williams, T.D., Escher, B.I., 2017. Development of a bioanalytical test battery for water quality monitoring: Fingerprinting identified micropollutants and their contribution to effects in surface water. Water Res. 123, 734–750. https://doi.org/10.1016/j.watres.2017.07.016

Posthuma, L., van Gils, J., Zijp, M.C., van de Meent, D., de Zwart, D., 2019. Species sensitivity distributions for use in environmental protection, assessment, and management of aquatic ecosystems for 12 386 chemicals. Environ. Toxicol. Chem. 38, 905–917. https://doi.org/10.1002/etc.4373

Rhind, S.M., 2009. Anthropogenic pollutants: a threat to ecosystem sustainability? Philos. Trans. R. Soc. B Biol. Sci. 364, 3391–3401. https://doi.org/10.1098/rstb.2009.0122

Rüdel, H., Körner, W., Letzel, T., Neumann, M., Nödler, K., Reemtsma, T., 2020. Persistent, mobile and toxic substances in the environment: a spotlight on current research and regulatory activities. Environ. Sci. Eur. 32, 5. https://doi.org/10.1186/s12302-019-0286-x

Serra, H., Brion, F., Chardon, C., Budzinski, H., Schulze, T., Brack, W., Aït-Aïssa, S., 2020. Estrogenic activity of surface waters using zebrafish- and human-based in vitro assays: The Danube as a case-study. Environ. Toxicol. Pharmacol. 78, 103401. https://doi.org/10.1016/j.etap.2020.103401

Sidorkiewicz, I., Zaręba, K., Wołczyński, S., Czerniecki, J., 2017. Endocrine-disrupting chemicals—Mechanisms of action on male reproductive system. Toxicol. Ind. Health 33, 601–609. https://doi.org/10.1177/0748233717695160

Silva, M., Pham, N., Lewis, C., Iyer, S., Kwok, E., Solomon, G., Zeise, L., 2015. A Comparison of ToxCast Test Results with In Vivo and Other In Vitro Endpoints for Neuro, Endocrine, and Developmental Toxicities: A Case Study Using Endosulfan and Methidathion: TOXCAST DATA AND IN VIVO ENDPOINTS COMPARED. Birth Defects Res. B. Dev. Reprod. Toxicol. 104, 71–89. https://doi.org/10.1002/bdrb.21140

Spurgeon, D.J., Jones, O.A.H., Dorne, J.-L.C.M., Svendsen, C., Swain, S., Stürzenbaum, S.R., 2010. Systems toxicology approaches for understanding the joint effects of environmental chemical mixtures. Sci. Total Environ. 408, 3725–3734. https://doi.org/10.1016/j.scitotenv.2010.02.038

Tabrez, S., Shakil, S., Urooj, M., Damanhouri, G.A., Abuzenadah, A.M., Ahmad, M., 2011. Genotoxicity Testing and Biomarker Studies on Surface Waters: An Overview of the Techniques and Their Efficacies. J. Environ. Sci. Health Part C 29, 250–275. https://doi.org/10.1080/10590501.2011.601849

Tohyama, C., 2016. Developmental neurotoxicity test guidelines: problems and perspectives. J. Toxicol. Sci. 41, SP69–SP79. https://doi.org/10.2131/jts.41.SP69

Tsatsakis, A.M., Vassilopoulou, L., Kovatsi, L., Tsitsimpikou, C., Karamanou, M., Leon, G., Liesivuori, J., Hayes, A.W., Spandidos, D.A., 2018. The dose response principle from philosophy to modern toxicology: The impact of ancient philosophy and medicine in modern toxicology science. Toxicol. Rep. 5, 1107–1113. https://doi.org/10.1016/j.toxrep.2018.10.001

Ulrich, E.M., Sobus, J.R., Grulke, C.M., Richard, A.M., Newton, S.R., Strynar, M.J., Mansouri, K., Williams, A.J., 2019. EPA's non-targeted analysis collaborative trial (ENTACT): genesis, design, and initial findings. Anal. Bioanal. Chem. 411, 853–866. https://doi.org/10.1007/s00216-018-1435-6

Villanueva, C.M., Kogevinas, M., Cordier, S., Templeton, M.R., Vermeulen, R., Nuckols, J.R., Nieuwenhuijsen, M.J., Levallois, P., 2014. Assessing Exposure and Health Consequences of Chemicals in Drinking Water: Current State of Knowledge and Research Needs. Environ. Health Perspect. 122, 213–221. https://doi.org/10.1289/ehp.1206229

Wang, Z., Walker, G.W., Muir, D.C.G., Nagatani-Yoshida, K., 2020. Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of

National and Regional Chemical Inventories. Environ. Sci. Technol. 54, 2575–2584. https://doi.org/10.1021/acs.est.9b06379

Wilson, V.S., 2002. A Novel Cell Line, MDA-kb2, That Stably Expresses an Androgen- and Glucocorticoid-Responsive Reporter for the Detection of Hormone Receptor Agonists and Antagonists. Toxicol. Sci. 66, 69–81. https://doi.org/10.1093/toxsci/66.1.69

Zhang, L., Gibble, R., Baer, K.N., 2003. The effects of 4-nonylphenol and ethanol on acute toxicity, embryo development, and reproduction in Daphnia magna. Ecotoxicol. Environ. Saf. 55, 330–337. https://doi.org/10.1016/S0147-6513(02)00081-7

2 Precision Environmental Health: a framework for effect assessment of environmental chemical mixtures

2.1 Abstract

Environmental chemical mixture problems pose potential health threats to human and dwelling organisms. For the need of better effect assessments of environmental chemical mixtures, a framework named Precision Environmental Health (PEH) serves the need of studying the relationships between the chemical mixtures in the environment and the biological responses of the testing objects in bioassays. The core of the PEH framework is to identify the molecular key events (mKEs) that are responsive to environmental chemical exposure and indicative of adversity. The conservation of mKEs across multiple species may facilitate cross-species extrapolation, so that the harmful impacts observed in the tested species can be translated to toxic potency in other non-tested species and further extend to ecotoxicity potency. This work describes the conceptual basis and key attributes of the PEH framework.

2.2 Introduction

In the aquatic environment, chemicals co-exist as complex mixtures and pose potential health threats to humans, natural resources, and ecosystems. To understand potential hazards of these complex chemical mixtures may need understanding of their composition and corresponding biological effects. Comprehensive profiling of the

environmental chemical mixtures requires chemical fingerprinting technique, such as non-targeted screening analysis with high resolution mass spectrometry (Hollender et al., 2017). However, it is still difficult and time-consuming to identify and clarify every single chemical peak in the profile (Pourchet et al., 2020). Chemical components like isomers or transformation products with limited structure information may cause uncertainty in peak annotation (Escher and Fenner, 2011). And it is also questionable to compare semi-quantified non-targeted data generated by different analytical pipelines and across multiple teams (Sobus et al., 2018). Even with a comprehensive profile of all the chemicals in the environment, effect assessment of the environmental chemical mixture heavily relies on toxicity testing records of individual chemical component (Heys et al., 2016; Syberg et al., 2009). Among over 191 million inorganic and organic chemical substances registered in the Chemical Abstracts Services (CAS; Chen et al., 2021), 26,147 chemicals were recorded as registered chemicals in the European REACH system with public available toxicity references (European Commission. Joint Research Centre. Institute for Health and Consumer Protection., 2010) and 17,242 chemical substances were filed with toxicity observations in ECOTOX database established by US EPA (Villeneuve et al., 2019), which suggests no more than 0.01 % of the registered chemical were ever tested for their potential toxicities. Moreover, only a small fraction of the chemical substances are studied when it compares to the chemical substances being discharged into the environment (Hernández et al., 2019) and 158 chemical compounds were listed as prioritised chemical for environmental monitoring under the Water Framework Directive (Brack et al., 2017). One of the commonly applied approaches, the component-based approach, estimates the toxicity of mixtures within environmental samples based on the toxicity

of known, bioactive, detectable, and well-defined chemical components (Posthuma et al., 2019). However, since a significant proportion of the co-existing chemical components are unknown or unidentified in the natural environment, effect assessment based on such component-based approach is likely to fail at delivering a reasonable evaluation of the joint effects.

An alternative way exploits the advantages of high-throughput omics technologies that assist systematic profiling of the biological effect at molecular level. Advanced techniques like high-throughput sequencing and high-resolution mass spectrometry are particularly promising at providing a cost-effective way to generate a multi-omics profile of biological responses from the same test system in favour of a mechanistic understanding of chemical effects (Dugourd et al., 2021; Larras et al., 2020). This has fuelled significant advancements in fields like system biology (Pinu et al., 2019) and precision medicine (Olivier et al., 2019); and it has also laid the foundation of the fast evolvement of toxicogenomic studies (Martins et al., 2019). The omics-based assays, including transcriptome and metabolome, aim at obtaining a holistic profile of all the biomolecular signals, such as DNA transcripts and metabolites in the biological system (Roede et al., 2014). Omics approaches are generally data-rich and unbiased, as they generate a global perspective of the molecular biological responses for further exploration without specifying potential responsive features to be interrogated beforehand (Martins et al., 2019). Each omics profile may represent an aspect of the complex biological responses that reveal a specified level of biological variations (Nguyen and Wang, 2020). Therefore, the omics technologies can characterise the chemical mixture effect as perturbation in the biological systems in greater details (Sturla et al., 2014). As increasing efforts focuses on improving acceptance of omics

bioassays in regulatory necessities (Mondou et al., 2020), research projects delivering omics-based signatures with broader content of hazard assessments undoubtably enrich the understanding of chemical toxicity (Escher et al., 2019).

The effect characterisation of chemical mixtures can be achieved by an array of bioassays, including molecular profiling (omics data) and biochemical endpoints (biomarkers), that are indicative of metabolic mechanism and downstream perturbation (Leung, 2018). A critical challenge lies in establishing the associations between the chemical profiles and the corresponding biological responses, and ultimately translating the knowledge into application in chemical mixture risk assessment. For chemical profiles, the number of chemical features will be far more than the number of available samples, suggesting that the samples may never be able to cover all possible combinations of all the concentration levels of individual chemical components in the mixture. General concentration series exposure test with full factorial design cannot work its magic under these circumstances. For biological profiles, the information of individual biological features is so complicated that describing the variation patterns of each biological feature is unrealistic. Summarising the scatter variation patterns of features may fail to convey the major functional changes (such as activation or down-regulation at the pathway level). Moreover, stablishing association between the chemical profiles and the corresponding biological profiles is a matrix-to-matrix correlation problem, where traditional univariate correlation analysis may fail to reveal the patterns when a set of chemical features intrigues variations in a set of biological features or pathways. Furthermore, it is difficult to translate the knowledge of resulting biological features of one tested species to another non-tested species. Thus, the Precision Environmental Health (PEH) framework was developed to address this

challenge (as Figure 2.1). Herein I describe the conceptual basis of the PEH framework and touch on several relevant topics to further substantiate the findings under this framework.

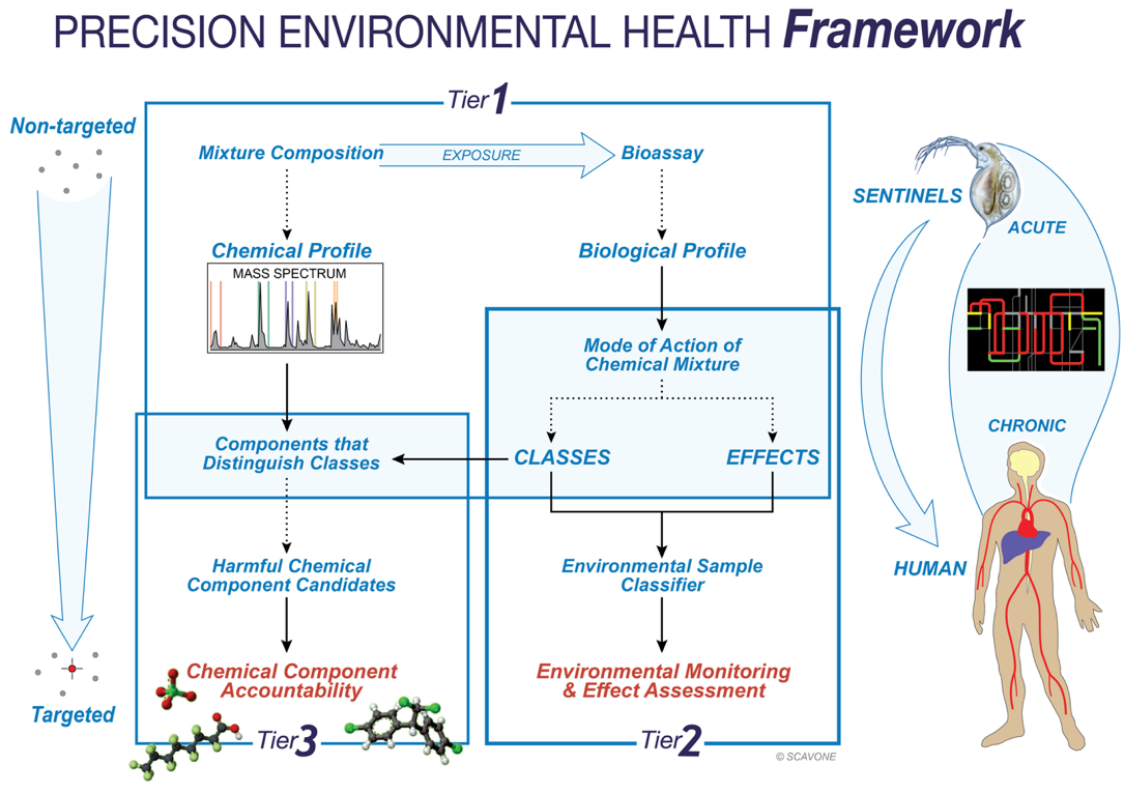

Figure 2. 1 The framework of Precision Environmental Health is conceptualised as a tiered approach. Tier 1 focuses on establish the exposure tests on chemical mixture with comprehensive profiling of both chemical profiles and biological responses. Tier 2 exploits the structure of the biological responses in order to deliver the class and effect of chemical mixtures, further summarised as mode of action. Tier 3 investigates the potential associations between chemcial components and classes of mixture effect, which provides a list of chemical candidates that might be accountable for differences in the mixture effects. It faciltates a zooming-in strategy, from non-targeted to targeted. It may further establish the probalistic relationships between molecular signatures in short-term exposure to long-term effects, and even enable cross-species extrapolation based on the evolutionary conservation of the molecular mechanism shared by multiple species.

2.3 Overview of the PEH framework

The PEH is a conceptual framework aiming at characterising and classifying the mixture effects of environmental chemical exposure (as Figure 2.1). It is a three-tier approach that allows in-depth investigation of the effects of the environmental chemical mixtures.

2.3.1 Tier one: exposure testing and profiles generation

An environmental monitoring and assessment case study provides environmental samples, including samples collected from the natural environment, pre-processed the environmental sample (extraction and/or enrichment) (Schulze et al., 2017), and surrogate chemical mixture constructed in the laboratory with known components and at pre-determined concentration ratios (Hashmi et al., 2018). The exposure test for characterising chemical mixtures effect consists of comprehensive chemical profiling and systematic biological response measurements. Chemical profiling can be achieved by the joint efforts of suspect, targeted and non-targeted screening analysis to retrieve the composition and relative quantification measurements (Brunner et al., 2020; Ccanccapa-Cartagena et al., 2019; Hollender et al., 2019). Suspect screening enables peak annotation against pre-established library of known chemical components (Gago-Ferrero et al., 2018), while nontargeted screening requires on-line chemical database for putative annotation of unknown chemical peaks (Hollender et al., 2017). Biological responses can be captured by various omics-based techniques at whole organismal level (Fuertes et al., 2019; Taylor et al., 2018) or at single cell level (Zhang et al., 2017). The omics assays for detecting perturbation in the testing organisms include transcriptomics, metabolomics, proteomics and epigenomics

(Canzler et al., 2020). Different omics assay can provide complementary view of the systematic responses (Tong et al., 2020).

2.3.2 Tier two: biological responses analysis and data integration

Biological response profiles can be analysed in two ways: at single feature level and as a set of multiple features. Analysis at single feature level clarifies the variation of specific feature, as it may behave in a concentration-dependent manner (Hou et al., 2017). A set of features with a similar variation pattern can be identified as co-responsive signals; for example, co-expression transcripts (Saha et al., 2017) and co-accumulation metabolites (Sakurai et al., 2011). The co-responsiveness of biological signals establishes the bridge between intra-omics regulation (e.g. co-regulation within a functional pathway; Josyula et al., 2020) and data-driven modelling (e.g. co-expression network analysis; Zhang and Horvath, 2005). Data integration of multi-omics data allow the identification of inter-omics relationships; for example, transcripts and metabolites involved in the same functional pathway (Subramanian et al., 2020). Integration between omics data and phenotypic traits variation can establish the probabilistic relationships between genotype and phenotype (Costanzo et al., 2019), which facilitates prediction of adverse outcome at individual/population levels based on molecular responses at biomolecular level.

The mixture effects are thereby characterised by systematic profiling of various omics-features that re-constructed as multiple co-responsive sets (or modules in the network). The mixture effects are therefore classified based on the homogeneity of the biological responses of these co-responsive modules, as the class of mixture effect depends on the type of omics assay (biomolecular type), the coverage of the omics assay (biomolecular entirety), and the biological functions revealed by co-responsive

modules (biomolecular function). The classifier generated based on the activities and compositions of co-responsive modules facilitates environmental sample classification. While the effect of each class relies on functional analysis of co-responsive modules. Gene products and metabolites with the same or similar functions refer to annotations from gene ontologies (e.g., GO; Gene Ontology Consortium, 2004) and gene orthologies (e.g., orthoDB; Zdobnov et al., 2021). A pathway-level profile is further summarised for individual co-responsive module with pathway databases like KEGG (Kanehisa et al., 2007), Reactome (Fabregat et al., 2018), and PANTHER (Mi et al., 2019).

2.3.3 Tier three: chemical component accountability

The associations between chemical components in the mixture and biological responses in the testing system are investigated for identifying the potential causal links between toxic components and corresponding perturbations in the bio-signals. The set of co-responsive features (either genes, their products, metabolites, or those combined) to chemical components (a single chemical substance, a set of chemical components, or a class of chemical compound) can be identified via multi-block correlation analysis (Tenenhaus and Tenenhaus, 2011). Suggested by correlation analysis, subsets of bio-features may be linearly correlated to the distribution of the selected chemical component (Mishra et al., 2021). Based on “Guilty-by-association” assumption (Girvan and Newman, 2002), features sharing similar response patterns (e.g. co-expression) may be co-regulated in the same sets of functional pathways, and the co-regulation can be suggested by network analysis generated in Tier two. Therefore, co-responsive features associated with a specific chemical component or class may use to characterise the corresponding biomolecular responses in the testing

organisms. As co-responsive features can be also used to predict adversities at higher biological levels, the chemical component or class (estimated to be) related to potent adverse outcome may be proposed in this step. Verification of biological impacts of these chemical component candidates substantiate the causal relationships between selected chemical components and corresponding effects suggested by the co-responsive biomolecular features.

The chemical component accountability can be established for both known and unknown chemical components in the mixtures. Accountability of unknown chemical components captured in the non-targeted analysis can be used as a data mining process that highlights potential toxicity drivers for downstream chemical annotation and bioassay verification.

2.4 Molecular key events as the core of PEH

The PEH focuses on identifying the mode of action (MoA) of chemical mixture, which is the knowledge term of the functional roles of the biological signatures representing a specific class of mixture effect. Such MoA is derived from empirical observations, which represents by a set of molecular key events (mKEs). An mKE is a set of co-responsive features that are associate with specific chemical components (or classes) and indicative of potential adverse outcome (as Figure 2.2).

2.4.1 System-biology perspective

The mKE provides a system-based insight supported by knowledge of system biology, as the mKEs serve as the basic functional units that assist in representing mechanistic constitution of chemical-related perturbation at the system level. The mKE sets are generated within data-driven network (e.g. co-expression network; Deng et al., 2010),

and further verified by knowledge-based network (e.g. Reatome; Fabregat et al., 2018). The identification of mKE requires incorporating both the intra-omics structure delivered by single omics assay and the inter-omics connections delivered by multi-omics assays. Namely, the mKE can be derived from single omics, for example, presenting as a set of co-regulating genes that might be functionally related; and it can be also derived from integration of multiple omics, for instance, presenting as a post-transcriptional signature that consists of co-responsive transcripts and metabolites linked by the same sets of functional pathways.

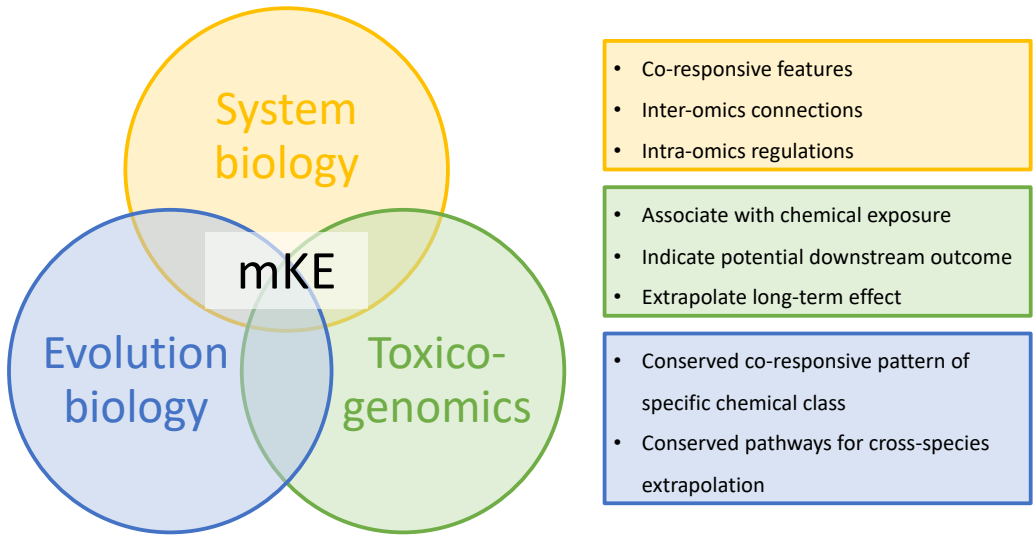

Figure 2. 2 The molecular key events (mKEs) in the PEH framework. The mKEs are co-responsive features that provide perspectives from system biology, toxicogenomics and evolution biology.

2.4.2 Toxicogenomics perspective

The mKE also provides the empirical observations supporting toxicogenomic studies. The features within a module of the co-expression network represent differentially expressed features with similar variation pattern under the chemical mixture exposure. The distinctive variation patterns of these features are closely related to differences among chemical mixture exposures, as the differentially expressed features may describe the perturbation in the biological system (Alexander-Dann et al., 2018). Such module consists of co-responsive features that may be induced by chemical mixture or involved in the metabolism of the chemical compounds. Thus, the mKE identified in the PEH framework are a set of co-responsive biological features that are differentially expressed under the chemical mixture exposure. Pathway analysis like overrepresentation analysis (Karp et al., 2021) and gene set enrichment analysis (Subramanian et al., 2005) may further assist biological interpretation of the functional roles of each mKE. For example, in a pilot project, the co-responsive modules associated with caffeine and carbamazepine in the river water samples were identified and functional pathways were summarised as pathway profiles (as revealed in Figure 2.3; Detail of this case study is described in Chapter 3).

2.4.3 Evolution-biology perspective

The features in each mKE can be re-annotated by its evolutionary conserved groups (e.g., ortholog groups for genetic models; Koonin, 2005) as a way to reveal its cross-species consensus. Previous observations of protein-protein interaction in five model species revealed that orthologous pairs of interacting proteins are more likely to be co-expressed (Tirosh and Barkai, 2005). Besides, the genes in a functional pathway may enhance their homogenous expression pattern (co-express or high correlation) under

specific conditions (Tegge et al., 2012). Theoretically, the co-responsive features of the mKE found in one testing species may be “conserved” across multiple species under similar exposure conditions. The “conservation” manner refers to features that are functionally conserved, with similar co-express pattern, and even under co-regulation of inter-connected pathways. Therefore, the conservation characteristic of mKEs provides the knowledge base of cross-species extrapolation.

2.4.4 Relation to Adverse Outcome Pathway (AOP)

The mKE term is closely related to the adverse outcome pathway (AOP). As the chemical mode of action in AOP can be depicted as a molecular initiating event (MIE) and a series of key event (KEs) that may lead to an adverse outcome induced by a chemical compound (Ankley and Edwards, 2018). The mKEs here focus on biological variation at the molecular level, where subtle differences in the biological signatures can be used to predict the adverse outcome at higher biological levels (e.g. individual or population level). Unlike AOP, the mKEs in the PEH framework integrate multiple omics perspectives in order to reveal the systematic responses.

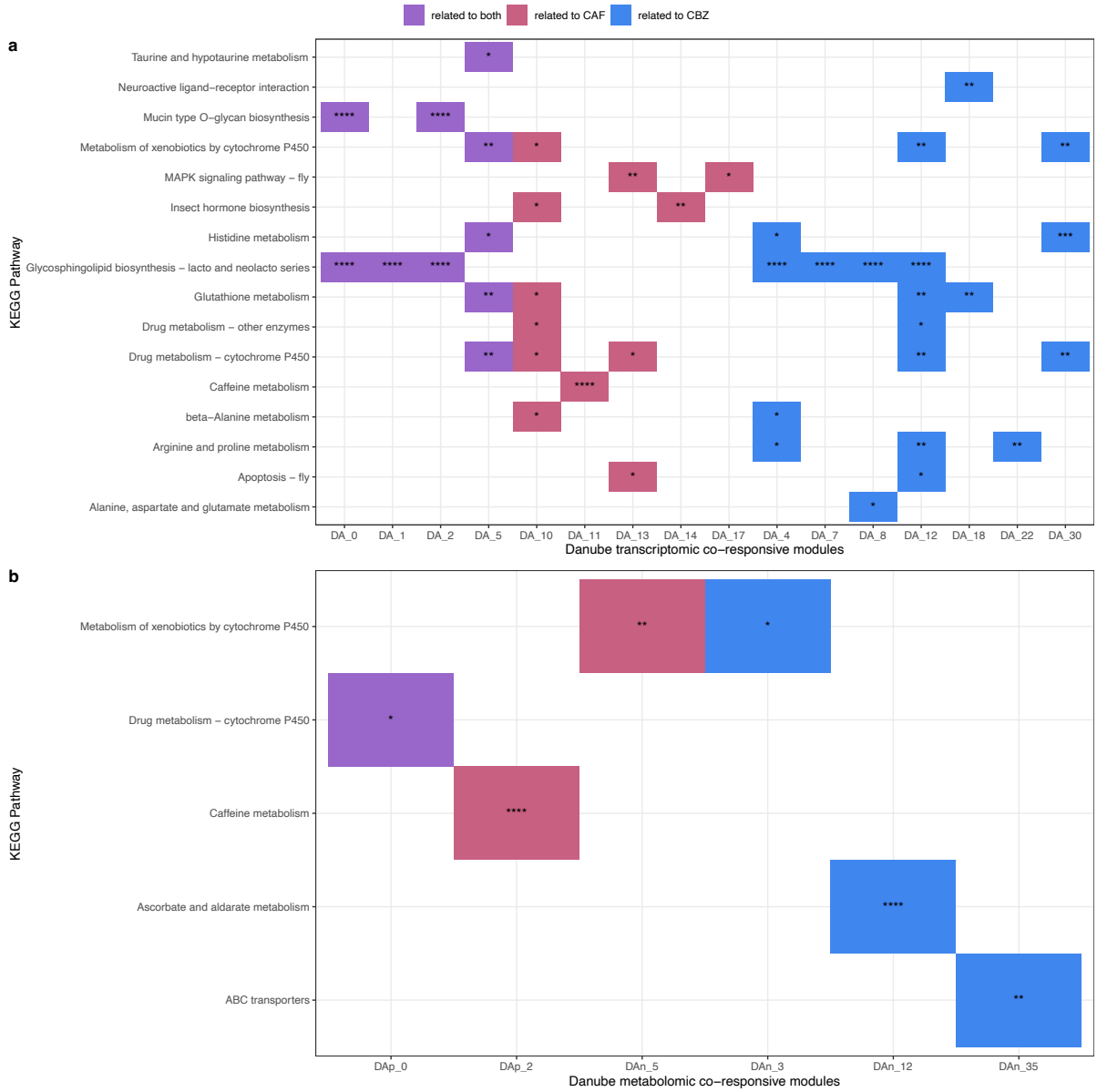

Figure 2. 3 Co-responsive modules associated with caffeine and carbamazepine. In the case study on Danube river water samples, 9 and 11 co-responsive modules were identified to be associated with caffeine and carbamazepine exposure, respectively. Pathways like xenobiotic metabolisms and caffeine metabolism are found in these selected modules, which suggested that the data-driven modelling can effectively identify pathways described in prior knowledge.

2.4.5 Challenges and technical considerations

(1) Whether the co-responsiveness within a specific omics profile is consistent across multiple case studies is still under study. The robustness of the co-responsive pattern

relate to a specific type of mixture effect may reflect the specificity and prediction capacity of the identified mKEs.

(2) The relationships between mKEs associated with the same mixture effect is undefined. The relationships between mKEs can be further interpreted as inter-connected sets of functional features, but whether there are subsequent co-regulation between mKEs is hard to prove.

(3) The prediction model for short-term adversity can be established based on omics profiles and phenotypic traits; however, whether the long-term adversity can be predicted by omics profiles capture at early stages or in short-term exposure is still questionable.

2.5 Conclusion

The Precision Environmental Health, proposed here as a conceptual framework, can provide a valuable and pragmatic roadmap upon which prior knowledge and case-based evidence can be integrated to systematically reveal the mode of action of mixture effect can be identified, and through which predictive and quantitative approaches to environmental monitoring can be improved and guide the design of local, or even national, the environmental regulatory policy. The PEH framework is developed to characterise and compare the joint effects of different environmental chemical mixtures, to establish the relationship between chemical mixtures and their modes of action, and to reveal the real-world exposure-related toxicological mechanisms of individual chemical components. This approach provides opportunities for optimising environmental monitoring and identifying harmful chemical component in the environment. The PEH framework is proposed to be: (1) a powerful approach for

locating pollution hot spots in field surveys as a means to group and classify environmental samples for further study; (2) a sufficient tool for characterising the effects of chemical mixtures by identifying the modes of action of mixture effects of the environmental samples; also (3) a promising IATA framework for pragmatical application of AOP scheme by identifying the MOAs of chemical components and potential toxicity driver in the chemical mixture.

2.6 Reference

Alexander-Dann, B., Pruteanu, L.L., Oerton, E., Sharma, N., Berindan-Neagoe, I., Módos, D., Bender, A., 2018. Developments in toxicogenomics: understanding and predicting compound-induced toxicity from gene expression data. Mol. Omics 14, 218–236. https://doi.org/10.1039/C8MO00042E

Ankley, G.T., Edwards, S.W., 2018. The adverse outcome pathway: A multifaceted framework supporting 21st century toxicology. Curr. Opin. Toxicol. 9, 1–7. https://doi.org/10.1016/j.cotox.2018.03.004

Brack, W., Dulio, V., Ågerstrand, M., Allan, I., Altenburger, R., Brinkmann, M., Bunke, D., Burgess, R.M., Cousins, I., Escher, B.I., Hernández, F.J., Hewitt, L.M., Hilscherová, K., Hollender, J., Hollert, H., Kase, R., Klauer, B., Lindim, C., Herráez, D.L., Miège, C., Munthe, J., O'Toole, S., Posthuma, L., Rüdel, H., Schäfer, R.B., Sengl, M., Smedes, F., van de Meent, D., van den Brink, P.J., van Gils, J., van Wezel, A.P., Vethaak, A.D., Vermeirssen, E., von der Ohe, P.C., Vrana, B., 2017. Towards the review of the European Union Water Framework Directive: Recommendations for more efficient assessment and management of chemical contamination in European surface water resources. Sci. Total Environ. 576, 720–737. https://doi.org/10.1016/j.scitotenv.2016.10.104

Brunner, A.M., Bertelkamp, C., Dingemans, M.M.L., Kolkman, A., Wols, B., Harmsen, D., Siegers, W., Martijn, B.J., Oorthuizen, W.A., ter Laak, T.L., 2020. Integration of target analyses, non-target screening and effect-based monitoring to assess OMP related water quality changes in drinking water treatment. Sci. Total Environ. 705, 135779. https://doi.org/10.1016/j.scitotenv.2019.135779

Canzler, S., Schor, J., Busch, W., Schubert, K., Rolle-Kampczyk, U.E., Seitz, H., Kamp, H., von Bergen, M., Buesen, R., Hackermüller, J., 2020. Prospects and challenges of multi-omics data integration in toxicology. Arch. Toxicol. 94, 371–388. https://doi.org/10.1007/s00204-020-02656-y

Ccanccapa-Cartagena, A., Pico, Y., Ortiz, X., Reiner, E.J., 2019. Suspect, non-target and target screening of emerging pollutants using data independent acquisition: Assessment of a Mediterranean River basin. Sci. Total Environ. 687, 355–368. https://doi.org/10.1016/j.scitotenv.2019.06.057

Chen, J., Zhang, S., Allen, D.T., Subramaniam, B., Licence, P., 2021. Expectations for Manuscripts Contributing to the Field on Management of Synthetic Chemicals in *ACS Sustainable Chemistry & Engineering*. ACS Sustain. Chem. Eng. 9, 3376–3378. https://doi.org/10.1021/acssuschemeng.1c01134

Costanzo, M., Kuzmin, E., van Leeuwen, J., Mair, B., Moffat, J., Boone, C., Andrews, B., 2019. Global Genetic Networks and the Genotype-to-Phenotype Relationship. Cell 177, 85–100. https://doi.org/10.1016/j.cell.2019.01.033

Deng, Y., Johnson, D.R., Guan, X., Ang, C.Y., Ai, J., Perkins, E.J., 2010. In vitro gene regulatory networks predict in vivo function of liver. BMC Syst. Biol. 4, 153. https://doi.org/10.1186/1752-0509-4-153

Dugourd, A., Kuppe, C., Sciacovelli, M., Gjerga, E., Gabor, A., Emdal, K.B., Vieira, V., Bekker-Jensen, D.B., Kranz, J., Bindels, Eric.M.J., Costa, A.S.H., Sousa, A., Beltrao, P., Rocha, M., Olsen, J.V., Frezza, C., Kramann, R., Saez-Rodriguez, J., 2021. Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. Mol. Syst. Biol. 17. https://doi.org/10.15252/msb.20209730

Escher, B.I., Fenner, K., 2011. Recent Advances in Environmental Risk Assessment of Transformation Products. Environ. Sci. Technol. 45, 3835–3847. https://doi.org/10.1021/es1030799

Escher, S.E., Kamp, H., Bennekou, S.H., Bitsch, A., Fisher, C., Graepel, R., Hengstler, J.G., Herzler, M., Knight, D., Leist, M., Norinder, U., Ouédraogo, G., Pastor, M., Stuard, S., White, A., Zdrazil, B., van de Water, B., Kroese, D., 2019. Towards grouping concepts based on new approach methodologies in chemical hazard assessment: the read-across approach of the EU-ToxRisk project. Arch. Toxicol. 93, 3643–3667. https://doi.org/10.1007/s00204-019-02591-7

European Commission. Joint Research Centre. Institute for Health and Consumer Protection., 2010. Characterisation of the REACH pre-registered substances list by chemical structure and physicochemical properties. Publications Office, LU.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C.D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P., 2018. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 46, D649–D655. https://doi.org/10.1093/nar/gkx1132

Fuertes, I., Jordão, R., Piña, B., Barata, C., 2019. Time-dependent transcriptomic responses of Daphnia magna exposed to metabolic disruptors that enhanced storage lipid accumulation. Environ. Pollut. 249, 99–108. https://doi.org/10.1016/j.envpol.2019.02.102

Gago-Ferrero, P., Krettek, A., Fischer, S., Wiberg, K., Ahrens, L., 2018. Suspect Screening and Regulatory Databases: A Powerful Combination To Identify Emerging Micropollutants. Environ. Sci. Technol. 52, 6881–6894. https://doi.org/10.1021/acs.est.7b06598

Gene Ontology Consortium, 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32, 258D – 261. https://doi.org/10.1093/nar/gkh036

Girvan, M., Newman, M.E.J., 2002. Community structure in social and biological networks. Proc. Natl. Acad. Sci. 99, 7821–7826. https://doi.org/10.1073/pnas.122653799

Hashmi, M.A.K., Escher, B.I., Krauss, M., Teodorovic, I., Brack, W., 2018. Effect-directed analysis (EDA) of Danube River water sample receiving untreated municipal wastewater from Novi Sad, Serbia. Sci. Total Environ. 624, 1072–1081. https://doi.org/10.1016/j.scitotenv.2017.12.187

Hernández, F., Bakker, J., Bijlsma, L., de Boer, J., Botero-Coy, A.M., Bruinen de Bruin, Y., Fischer, S., Hollender, J., Kasprzyk-Hordern, B., Lamoree, M., López, F.J., Laak, T.L. ter, van Leerdam, J.A., Sancho, J.V., Schymanski, E.L., de Voogt, P., Hogendoorn, E.A., 2019. The role of analytical chemistry in exposure science: Focus on the aquatic environment. Chemosphere 222, 564–583. https://doi.org/10.1016/j.chemosphere.2019.01.118

Heys, K.A., Shore, R.F., Pereira, M.G., Jones, K.C., Martin, F.L., 2016. Risk assessment of environmental mixture effects. RSC Adv. 6, 47844–47857. https://doi.org/10.1039/C6RA05406D

Hollender, J., Schymanski, E.L., Singer, H.P., Ferguson, P.L., 2017. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? Environ. Sci. Technol. 51, 11505–11512. https://doi.org/10.1021/acs.est.7b02184

Hollender, J., van Bavel, B., Dulio, V., Farmen, E., Furtmann, K., Koschorreck, J., Kunkel, U., Krauss, M., Munthe, J., Schlabach, M., Slobodnik, J., Stroomberg, G., Ternes, T., Thomaidis, N.S., Togola, A., Tornero, V., 2019. High resolution mass spectrometry-based non-target screening can support regulatory environmental monitoring and chemicals management. Environ. Sci. Eur. 31, 42. https://doi.org/10.1186/s12302-019-0225-x

Hou, J., Liu, X., Cui, B., Bai, J., Wang, X., 2017. Concentration-dependent alterations in gene expression induced by cadmium in Solanum lycopersicum. Environ. Sci. Pollut. Res. 24, 10528–10536. https://doi.org/10.1007/s11356-017-8748-4

Josyula, N., Andersen, M.E., Kaminski, N.E., Dere, E., Zacharewski, T.R., Bhattacharya, S., 2020. Gene co-regulation and co-expression in the aryl hydrocarbon receptor-mediated transcriptional

regulatory network in the mouse liver. Arch. Toxicol. 94, 113–126. https://doi.org/10.1007/s00204-019-02620-5
Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., Yamanishi, Y., 2007. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 36, D480–D484. https://doi.org/10.1093/nar/gkm882
Karp, P.D., Midford, P.E., Caspi, R., Khodursky, A., 2021. Pathway size matters: the influence of pathway granularity on over-representation (enrichment analysis) statistics. BMC Genomics 22, 191. https://doi.org/10.1186/s12864-021-07502-8
Koonin, E.V., 2005. Orthologs, Paralogs, and Evolutionary Genomics. Annu. Rev. Genet. 39, 309–338. https://doi.org/10.1146/annurev.genet.39.073003.114725
Larras, F., Billoir, E., Scholz, S., Tarkka, M., Wubet, T., Delignette-Muller, M.-L., Schmitt-Jansen, M., 2020. A multi-omics concentration-response framework uncovers novel understanding of triclosan effects in the chlorophyte Scenedesmus vacuolatus. J. Hazard. Mater. 397, 122727. https://doi.org/10.1016/j.jhazmat.2020.122727
Leung, K.M., 2018. Joining the dots between omics and environmental management: Omics and Environmental Management. Integr. Environ. Assess. Manag. 14, 169–173. https://doi.org/10.1002/ieam.2007
Martins, Dreij, Costa, 2019. The State-of-the Art of Environmental Toxicogenomics: Challenges and Perspectives of “Omics” Approaches Directed to Toxicant Mixtures. Int. J. Environ. Res. Public. Health 16, 4718. https://doi.org/10.3390/ijerph16234718
Mi, H., Muruganujan, A., Ebert, D., Huang, X., Thomas, P.D., 2019. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 47, D419–D426. https://doi.org/10.1093/nar/gky1038
Mishra, P., Roger, J.-M., Jouan-Rimbaud-Bouveresse, D., Biancolillo, A., Marini, F., Nordon, A., Rutledge, D.N., 2021. Recent trends in multi-block data analysis in chemometrics for multi-source data integration. TrAC Trends Anal. Chem. 137, 116206. https://doi.org/10.1016/j.trac.2021.116206
Mondou, M., Hickey, G.M., Rahman, H.T., Maguire, S., Pain, G., Crump, D., Hecker, M., Basu, N., 2020. Factors Affecting the Perception of New Approach Methodologies (NAMs) in the Ecotoxicology Community. Integr. Environ. Assess. Manag. 16, 269–281. https://doi.org/10.1002/ieam.4244
Nguyen, N.D., Wang, D., 2020. Multiview learning for understanding functional multiomics. PLOS Comput. Biol. 16, e1007677. https://doi.org/10.1371/journal.pcbi.1007677
Olivier, M., Asmis, R., Hawkins, G.A., Howard, T.D., Cox, L.A., 2019. The Need for Multi-Omics Biomarker Signatures in Precision Medicine. Int. J. Mol. Sci. 20, 4781. https://doi.org/10.3390/ijms20194781
Pinu, F.R., Beale, D.J., Paten, A.M., Kouremenos, K., Swarup, S., Schirra, H.J., Wishart, D., 2019. Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community. Metabolites 9, 76. https://doi.org/10.3390/metabo9040076
Posthuma, L., Altenburger, R., Backhaus, T., Kortenkamp, A., Müller, C., Focks, A., de Zwart, D., Brack, W., 2019. Improved component-based methods for mixture risk assessment are key to characterize complex chemical pollution in surface waters. Environ. Sci. Eur. 31, 70. https://doi.org/10.1186/s12302-019-0246-5
Pourchet, M., Debrauwer, L., Klanova, J., Price, E.J., Covaci, A., Caballero-Casero, N., Oberacher, H., Lamoree, M., Damont, A., Fenaille, F., Vlaanderen, J., Meijer, J., Krauss, M., Sarigiannis, D., Barouki, R., Le Bizec, B., Antignac, J.-P., 2020. Suspect and non-targeted screening of chemicals of emerging concern for human biomonitoring, environmental health studies and support to risk assessment: From promises to challenges and harmonisation issues. Environ. Int. 139, 105545. https://doi.org/10.1016/j.envint.2020.105545
Roede, J.R., Uppal, K., Park, Y., Tran, V., Jones, D.P., 2014. Transcriptome–metabolome wide association study (TMWAS) of maneb and paraquat neurotoxicity reveals network level interactions in toxicologic mechanism. Toxicol. Rep. 1, 435–444. https://doi.org/10.1016/j.toxrep.2014.07.006
Saha, A., Kim, Y., Gewirtz, A.D.H., Jo, B., Gao, C., McDowell, I.C., The GTEx Consortium, Engelhardt, B.E., Battle, A., 2017. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. Genome Res. 27, 1843–1858. https://doi.org/10.1101/gr.216721.116
Sakurai, N., Ara, T., Ogata, Y., Sano, R., Ohno, T., Sugiyama, K., Hiruta, A., Yamazaki, K., Yano, K., Aoki, K., Aharoni, A., Hamada, K., Yokoyama, K., Kawamura, S., Otsuka, H., Tokimatsu, T., Kanehisa, M., Suzuki, H., Saito, K., Shibata, D., 2011. KaPPA-View4: a metabolic pathway

database for representation and analysis of correlation networks of gene co-expression and metabolite co-accumulation and omics data. Nucleic Acids Res. 39, D677–D684. https://doi.org/10.1093/nar/gkq989

Schulze, T., Ahel, M., Ahlheim, J., Aït-Aïssa, S., Brion, F., Di Paolo, C., Froment, J., Hidasi, A.O., Hollender, J., Hollert, H., Hu, M., Kloß, A., Koprivica, S., Krauss, M., Muz, M., Oswald, P., Petre, M., Schollée, J.E., Seiler, T.-B., Shao, Y., Slobodnik, J., Sonavane, M., Suter, M.J.-F., Tollefsen, K.E., Tousova, Z., Walz, K.-H., Brack, W., 2017. Assessment of a novel device for onsite integrative large-volume solid phase extraction of water samples to enable a comprehensive chemical and effect-based analysis. Sci. Total Environ. 581–582, 350–358. https://doi.org/10.1016/j.scitotenv.2016.12.140

Sobus, J.R., Wambaugh, J.F., Isaacs, K.K., Williams, A.J., McEachran, A.D., Richard, A.M., Grulke, C.M., Ulrich, E.M., Rager, J.E., Strynar, M.J., Newton, S.R., 2018. Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA. J. Expo. Sci. Environ. Epidemiol. 28, 411–426. https://doi.org/10.1038/s41370-017-0012-y

Sturla, S.J., Boobis, A.R., FitzGerald, R.E., Hoeng, J., Kavlock, R.J., Schirmer, K., Whelan, M., Wilks, M.F., Peitsch, M.C., 2014. Systems Toxicology: From Basic Research to Risk Assessment. Chem. Res. Toxicol. 27, 314–329. https://doi.org/10.1021/tx400410s

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. 102, 15545–15550. https://doi.org/10.1073/pnas.0506580102

Subramanian, I., Verma, S., Kumar, S., Jere, A., Anamika, K., 2020. Multi-omics Data Integration, Interpretation, and Its Application. Bioinforma. Biol. Insights 14, 117793221989905. https://doi.org/10.1177/1177932219899051

Syberg, K., Jensen, T.S., Cedergreen, N., Rank, J., 2009. On the Use of Mixture Toxicity Assessment in REACH and the Water Framework Directive: A Review. Hum. Ecol. Risk Assess. Int. J. 15, 1257–1272. https://doi.org/10.1080/10807030903304922

Taylor, N., Gavin, A., Viant, M., 2018. Metabolomics Discovers Early-Response Metabolic Biomarkers that Can Predict Chronic Reproductive Fitness in Individual Daphnia magna. Metabolites 8, 42. https://doi.org/10.3390/metabo8030042

Tegge, A.N., Caldwell, C.W., Xu, D., 2012. Pathway Correlation Profile of Gene-Gene Co-Expression for Identifying Pathway Perturbation. PLoS ONE 7, e52127. https://doi.org/10.1371/journal.pone.0052127

Tenenhaus, A., Tenenhaus, M., 2011. Regularized Generalized Canonical Correlation Analysis. Psychometrika 76, 257–284. https://doi.org/10.1007/s11336-011-9206-8

Tirosh, I., Barkai, N., 2005. Computational verification of protein-protein interactions by orthologous co-expression. BMC Bioinformatics 6, 40. https://doi.org/10.1186/1471-2105-6-40

Tong, L., Mitchel, J., Chatlin, K., Wang, M.D., 2020. Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. BMC Med. Inform. Decis. Mak. 20, 225. https://doi.org/10.1186/s12911-020-01225-8

Villeneuve, D.L., Coady, K., Escher, B.I., Mihaich, E., Murphy, C.A., Schlekat, T., Garcia-Reyero, N., 2019. High-throughput screening and environmental risk assessment: State of the science and emerging applications: High-throughput screening and environmental risk assessment. Environ. Toxicol. Chem. 38, 12–26. https://doi.org/10.1002/etc.4315

Zdobnov, E.M., Kuznetsov, D., Tegenfeldt, F., Manni, M., Berkeley, M., Kriventseva, E.V., 2021. OrthoDB in 2020: evolutionary and functional annotations of orthologs. Nucleic Acids Res. 49, D389–D393. https://doi.org/10.1093/nar/gkaa1009

Zhang, B., Horvath, S., 2005. A General Framework for Weighted Gene Co-Expression Network Analysis. Stat. Appl. Genet. Mol. Biol. 4. https://doi.org/10.2202/1544-6115.1128

Zhang, B., Huang, K., Zhu, L., Luo, Y., Xu, W., 2017. Precision toxicology based on single cell sequencing: an evolving trend in toxicological evaluations and mechanism exploration. Arch. Toxicol. 91, 2539–2549. https://doi.org/10.1007/s00204-017-1971-4

3 Co-responsive biological features characterise chemical component associated effects in environmental mixtures

3.1 Abstract

In the aquatic environment, co-occurring chemicals leads to complex mixture effect that may pose health threats on living organisms and the ecosystem. Yet, for the design of mitigating strategies in the protection of human health and the environment, it is essential to reveal the effects of the chemical components within the mixture to identify potentially harmful substances. Here, I purpose a method to identify the co-responsive biological features associated with exposure to chemical components detected in the surface waters at their environmental concentration levels, based on transcriptomics and metabolomics profiling in the exposed *Daphnia magna*. The method includes constructing co-expression network of transcriptomic and metabolomic features and establishing multi-block correlation models to identify co-responsive features. Two well-studied chemical components, caffeine and carbamazepine, were selected for testing the effectiveness of this method to identify the features associated with the chemical component in quest. Datasets from two case studies, the Chaobai River and the Danube River, were included in this work and analysed in parallel, as a way to evaluate the robustness of the selected co-responsive features delivered by this method across two case studies. Overall, the co-responsive modules associated with one or both chemical components are biological plausibly associated with caffeine and carbamazepine, as both xenobiotic metabolism and

caffeine metabolism are enriched in the selected feature sets. The transcriptome-based case comparisons revealed consistent conclusions over the two studies, although dissimilar sets of pathways were found to be associated with caffeine. Thus, this work presents and verifies a novel method that can reasonably identify the co-responsive biological features associated with a chemical component in the environmental chemical mixtures, which may further assist the characterisation of chemical mode of action.

3.2 Introduction

Current approaches at identifying the chemical component effects mostly rely on individual chemical substance toxicity testing (Posthuma et al., 2019). The criteria for determining the potential hazards of the chemical components are anchored by a few selected toxicological endpoints. When bioassays are employed, these are typically assessing specific adverse outcome pathways (Neale et al., 2015), which can potentially overlook other potential, or even harmful, effects by such a screening process. It is unrealistic to test all the chemicals in the environment considering all possible combinations in a full-factorial design experiment, in order to reflect the relative contribution of each chemical component in the mixture to the joint mixture effect. The inherent challenge is to disentangle the environmental chemical mixture effect to detect and characterise the contribution of chemical components to the joint effect (Altenburger et al., 2019, 2015).

The global biological responses from exposure to the chemical mixtures within the environment can be effectively captured by omics-based bioassays corresponding to the whole mixture. To be specific, the gene expression profile (transcriptome) and the

metabolomic profile (metabolome) jointly describe the composition and dynamics of transcripts and resulting metabolites responding to different environmental chemical mixtures. The co-varying features that share similar variation patterns across different experimental conditions may be co-regulated genes and/or metabolites that share condition-specific expression patterns to the same external pressure (Carter et al., 2004; Stuart, 2003). Such co-varying features can be identified via co-varying network analysis, such as weighted co-expression gene network analysis (WGCNA; Langfelder and Horvath, 2007) and co-accumulation metabolite network analysis (Sakurai et al., 2011). The modules comprising co-varying features here may be putative functional units associated with a few functional pathways and can be further applied to distinguish the distinctiveness of systemic biological responses corresponding to different experimental conditions (Orsini et al., 2018). The co-responsive modules establish the basis of structuralising the omics data into multiple co-varying sets that links the mathematical modelling (pairwise correlation of biological features) with functional relationships (co-regulation of biological features) (Josyula et al., 2020; Kustatscher et al., 2019). Chemical components in a mixture may affect the biological systems in an independent and additive way, as the molecular features that are associated with one specific chemical component in single substance-based exposure testing may be observed in the mixture exposure testing. If the chemical substance is bioactive and the associated features reveal concentration-dependent response, the linear combination of responses of these associated features may be also correlated to the concentration levels of the chemical components in the mixture. Such chemical-associated features can be identified via multi-block correlation analysis, which identifies the linear combination of biomolecular features that are correlated with the

chemical components of interest within the chemical mixtures (Tenenhaus et al., 2014; Tenenhaus and Tenenhaus, 2011). The multi-block correlation analysis exemplified by the Canonical Correlation Analysis (CCA; Jun et al., 2018) studies the inter-connection between multiple data sources. The sparse version of Regularized Generalized Canonical Correlation Analysis (RGCCA/SGCCA) is particularly appropriate in this case as it may find the subset of biological features that are linearly correlated with the chemical component. In that case, those identified biomolecular features may be regarded as chemical-associated features, which provides the insight of association between biological responses and chemical factors. With the profiles of co-regulating features (co-varying modules) and chemical-associted features (sCCA selected features), modules with features that are identified as chemical-associated features are further regarded as co-responsive modules. Such co-responsive modules are significantly associated with chemical components of chemical mixtures and presented as functionally related features.

Biological interpretation of the gene clusters may require comprehensive information of gene function and pathway. Although pathway databases like KEGG (Kanehisa et al., 2007) and Reactome (Jassal et al., 2019) include a large amount of annotated information of multiple model species, the functional annotation of the genomes of a few ecotoxicological model species (like *Daphnia magna*) is quite poor. A cross-species extrapolation can be employed to annotate genes of the poorly defined species by referred to well-studied species based on their orthologs. The ortholog groups (OGs), which are evolutionarily conserved, can be useful in identifying functionally conserved proteins between any two species (Koonin, 2005). These phylogenetic-based OGs (like OGs in orthoDB database; Zdobnov et al., 2021) may be the

biomolecular targets of bioactive chemical substances in the aquatic environment (Gunnarsson et al., 2008). The basic assumption of cross-species extrapolation is that the pathways that are functionally conserved between two species may consist of OGs that are shared by those two species to fulfil their function. Previous studies suggested that the essential regulatory genes may interact the same way in the kernel of the gene regulatory networks of *Drosophila melanogaster* and other invertebrates (Davidson and Erwin, 2006). These kernels networks may control and maintain the general functions of any organisms (Kim et al., 2013). And the ortholog-based protein-protein interaction network suggested that the connections between orthologs may preserve the same in different species (Lee et al., 2008). Thus, it is reasonable to assume that the OGs that fulfil an evolutionarily conserved pathway may also be conserved across multiple species, which is the premise of cross-species extrapolation. Based on this assumption, the OGs-pathway associations establish on one well-defined model species may be transferrable to another species. It is reasonable to annotate the undefined *Daphnia* genes in respect of their OGs that shared with the well-studied model species like *Drosophila melanogaster*. *Daphnia magna* and *Drosophila melanogaster* are both belongs to the same phylum *Arthropoda*. Comparative genomics studies on the genomes of *Daphnia* and *Drosophila melanogaster* already revealed that the functional protein sequences were highly conserved in circadian proteins (Tilden et al., 2011), C2H2 zinc-finger proteins (Seetharam and Stuart, 2013), and DNA-binding proteins (Kato et al., 2008), which suggested that it is feasible to use *Drosophila melanogaster* as a counterpart to reveal the biological functions of protein sequences conserved between *Daphnia* and *Drosophila melanogaster*. The OGs-pathway associations in the *Daphnia magna* may be predicted by the OGs-pathway

associations defined in the well-studied genetic model species (*Drosophila melanogaster*). As the functions of unknown genes can be putatively annotated by the corresponding OGs' function, the gene-pathway association can be thereby transformed into an OGs-pathway association. If the OGs composition of every pathway is unique, the OGs-pathway association can be used to (1) distinguish different pathways and (2) applied as the reference data, similar to gene sets serving as background knowledge in the pathway overrepresentation analysis (Khatri et al., 2012).

In this work, I propose a method that combines the co-varying network analysis and the multi-block correlation analysis to identify chemical component associated co-responsive biological features. To test the effectiveness and rationale of this method, two chemical components, caffeine and carbamazepine, were selected as chemicals of interest. Two case studies were included to identify the associated effects of these two chemical components within the environmental chemical mixtures. The aim of this work is twofold: (1) to identify the co-responsive biological features associated with caffeine and carbamazepine in two case studies and compare the identified biological features by data-driven model (method proposed in this work) with prior knowledge (research papers and associated pathways); and (2) to compare results between two case studies in order to evaluate the robustness of the method.

3.3 Methods

3.3.1 Chemical component selection

Over 40,000 organic compounds are regarded as emerging threats in the aquatic ecosystem (Sun et al., 2018). Most of these organic compounds are widely spread at low concentrations (Barbosa et al., 2016). Still, some unregulated organic compounds

are detected with relatively higher concentrations (Sousa et al., 2018); for example, caffeine and carbamazepine. Both caffeine and carbamazepine are widely detected in freshwater rivers worldwide (Bean et al., 2018; Mutiyar et al., 2018; Su et al., 2020; Yang et al., 2018) that both are often used as anthropogenic markers of municipal outlets (Cunningham et al., 2010; Silva et al., 2014).

Caffeine is a xanthine alkaloid used as a stimulant within beverages (coffee and tea) or a psychoactive drug (Cappelletti et al., 2015). Caffeine is known to reduce oxidative stress and apoptosis by increasing the activity of antioxidants in human (Carelli-Alinovi et al., 2016; Kolahdouzan and Hamadeh, 2017). However, caffeine can cause reproduction inhibition and developmental delay in aquatic invertebrates (Rivetti et al., 2015). The caffeine metabolism in human and bacteria is documented in KEGG (map00915; Kanehisa et al., 2007) and PharmGKB (Thorn et al., 2012).

Carbamazepine is an antiepileptic drug applied in the medication of neuropathic pain (Tolou-Ghamari et al., 2013). It is persistent in the environment (Andreozzi, 2002) and may bioaccumulate in fish (Ramirez et al., 2009) and zooplankton (Nkoom et al., 2019). Carbamazepine is a medicine used to modulate the levels of neurotransmitters (Beutler et al., 2005). But carbamazepine at an environmental-relevant concentration level can cause oxidative stress (Nkoom et al., 2019) and act as an endocrine disruptor that may affect reproduction (Oropesa et al., 2016) and delay maturation (Dieterle et al., 2006). The carbamazepine metabolism in human and bacteria is also documented in KEGG (map00982, as part of “drug metabolism – cytochrome P450”; Kanehisa et al., 2007) and PharmGKB (Thorn et al., 2011).

3.3.2 Field sampling and targeted chemical analysis

Two case studies were included in this work. The Chaobai River case study included 30 surface water samples, and the Danube River case study included 12 organic extracts from river samples.

The concentration levels of caffeine and carbamazepine of the Chaobai River water samples and of organic extracts from the Danube River were measured by targeted chemical analytical methods. In the Chaobai River case, the organic substances were extracted by SPE with Oasis HLB cartridges (500mg, 6ml, Waters, U.S.A.), eluted with methanol, dried under nitrogen at room temperature, and dissolved in 40% methanol solvent (v:v). Targeted analysis of caffeine and carbamazepine was conducted on the Agilent 1290 ultra-performance liquid chromatography (UPLC) system equipped with the Agilent 6420 Triple Quad mass spectrometer (MS). Details of extraction method and instrumental settings were described in (Ben et al., 2018; Su et al., 2020), respectively.

In the Danube River case, over 500 L of surface water were pumped into the stainless-steel tank filled with sorbents for neutral, anionic, and cationic ions for each site, according to the description in (Schulze et al., 2017). The SPE was performed on-site with the large volume solid phase extraction (LVSPE) device by the JDS3 team (Schulze et al., 2017). The elutes were dried under nitrogen and stored at -20 °C. The dried extracts were then shipped to the University of Birmingham, maintained in methanol, and stored -20 °C. The targeted analysis of organic substances was performed by ultra-high pressure liquid chromatography tandem mass spectrometry (UHPLC-MS-MS) coupled with a hybrid triple quadrupole linear ion trap mass

spectrometer (QqQ-LIT-MS). The instrument settings are described in Liška *et al.* (2015).

3.3.3 Whole-mixture *in vitro* bioassay

The whole-mixture *in vitro* bioassay with *Daphnia magna* as the test system were applied to both case studies.

In the Chaobai River case study, 30 filtered water samples collected from the Chaobai River were used as exposure media. Each water sample treatment had three biological replicates, and the water sampled at site B01 (as reference level) had eight biological replicates. Each 5 ml glass vial was filled with 4.5 ml filtered water samples. Neonates of the same population of *Daphnia magna* (Bham2 strain) hatched within 24 hours were collected within 2 hours and attributed to each glass vial before the assay (5 neonates per vial).

In the Danube River case, 12 re-constructed borehole media injected by organic extracts from 12 sites of the Danube River were transferred to glass vials before the bioassay. Each treatment group had six biological replicates. And the negative control group had 24 biological replicates. The borehole media with 0.08 % methanol (methanol as the carrier of the organic extracts) was treated as the negative control in this case study. Each 20 ml glass vial was filled with 15 ml re-constructed borehole media. Neonates of the same population of *Daphnia magna* (Bham2 strain) hatched within 24 hours were collected within 2 hours and attributed to each glass vial before the bioassays (15 neonates per vial).

Following OECD test guideline 202 (OECD 202), 48 hours exposure tests were conducted in the laboratory. After 48 hours of exposure, the number of immobilised

neonates was recorded. The exposed neonates were flash-frozen within liquid nitrogen and stored at -80 ℃ prior to total RNA and/or metabolites extraction.

3.3.4 Multi-omics extraction

For the Chaobai River case study, all the frozen tissues were homogenised within lysis buffer (included in the Agencourt RNAdvance Tissue Total RNA kit) using the 2020 Genogrinder (SPEX SamplePrep, U.S.A.) at the speed of 1750 rpm for 45 seconds. Total RNA extraction was performed using the Agencourt RNAdvance Tissue Total RNA kit (Beckman Coulter, U.S.A.), following the manufacturer's instructions. RNA was absorbed by magnetic beads, washed twice for rinsing salts, and eluted in 100 μl RNAse-free H_2O. RNA concentrations were quantified by Nanodrop 8000 Spectrophotometer (Labtech Ltd., U.K.). RNA qualities (integrity and purity) were measured on TapeStation 2200 (Agilent Technologies, U.S.A.). These RNA samples were stored at -80°C until cDNA library construction.

For the Danube River case study, all the frozen tissues were homogenised within a methanol solution (640 μl methanol and 256 μl H_2O) using the Genogrinder at the same speed (1750 rpm) for 90 seconds. One-third of the homogenate was used for transcriptome profiling, and the rest was used for metabolome profiling. The RNAs was isolated and purified using the same method as described above. The polar metabolites were extracted using the methanol: chloroform solution (methanol: chloroform: H_2O = 2: 2: 1.8) as described previously (Wu *et al.* 2008). Briefly, the homogenate was added with 215 μl methanol, 640 μl chloroform and 405 μl H_2O, and vortexed at a maximum speed of Vortex-Genie 2 (Scientific Industries, Inc., U.S.A.) for 30 seconds. After 10 minutes of incubation on ice, the mixed solvents were centrifuged at 4000 rpm for 10 minutes. The upper layer supernatant containing polar metabolites

was transferred to two new 2 ml tubes, each with 400 µl. The polar aliquots were then dried in a Speed Vac Concentrator (Eppendorf, Germany) for 45-50 minutes and stored at -80°C until DIMS analysis.

3.3.5 Transcriptome sequencing and pre-processing

A cDNA library was generated for each sample from 150 ng RNA using NEBNext Ultra II Directional RNA Library Prep Kit for Illumina, following the manufacturer's instructions. All of the sample libraries were normalised to the same molecular weight and pooled together using the adapter indices supplied by the manufacturer. RNA-seq sequencing was performed on the Hiseq4000 (Illumina, U.S.A) and DNBseq (MGI, China) at BGI for the Chaobai and Danube case, respectively. All the samples were run in two lanes in parallel to avoid potential systemic bias. The reads from both lanes were merged into one. The nucleotide sequence reads were trimmed in Trimmomatic (version 0.32; Bolger et al., 2014) to remove sequencing adapters and obtain sequences with phred scores of at least 30. FastQC was used to screen the overall sequence quality (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Then high-quality reads were mapped to the *Daphnia magna* transcriptome reference (Orsini et al., 2016) using Salmon (version 0.8.2; Patro et al., 2017). The mapped transcript reads were then processed in R (version 4.0.3). Reads with raw counts lower than ten were removed from the data set. The library sizes of all the samples were normalised by the size factor defined within the DESeq2 package (version 3.12), following the pipeline described in (Love et al., 2014). The log2 fold changes of individual genes per treatment level by comparing treatment conditions versus control levels (negative control mentioned earlier) were further calculated by the DESeq2 package in the Danube River case study. The normalised transcriptomics data in the Chaobai case

study and the log2 fold change patterns in the Danube case study were used in the downstream analysis.

3.3.6 Metabolome profiling and pre-processing

Each treatment sample had six biological replicates for each ion mode, and each ion mode had six negative controls (borehole media with 0.08 % methanol), six extraction blanks, and twenty-six quality control samples as recommended by (Kirwan et al., 2014). The extraction blank sample reflects potential contamination of experimental processes, from extraction to loading, which contains the extraction solvents. The quality control reflects the stability of the system performance, which is a pooled sample of all the treatment samples with equal volumes (2 µl of re-suspended polar aliquots per sample).

As described in section 3.3.4, there were two tubes of polar aliquots per sample. One tube was suspended in 30 µl methanol: H_2O with 0.25 % formic acid (volume ratio 4:1) for mass-to-charge (m/z) detection under the positive ion mode (polarpos), and the other tube was suspended in 30 µl methanol: 25 mM aqueous ammonium acetate (volume ratio 4:1) for m/z detection under the negative ion mode (polarneg). Then, 15 µl of each re-suspended sample was loaded onto two 384-well plates. The metabolomics profiling was performed by Direct Infusion Mass Spectrometry (DIMS), which consists of the LTQ-Orbitrap Elite mass spectrometer (MS; Thermo Fisher Scientific, Germany) being attached with a chip-based direct infusion nano-electrospray ionisation assembly (nESI; Triversa, Advion Biosciences, U.S.A.). The DIMS for polar metabolites was conducted under both positive and negative ion modes separately. The DIMS settings followed the description in (Kirwan et al., 2014). And the SIM-stitching approach for obtaining metabolites ranging from 50 mass-to-charge

(m/z) and 620 m/z was applied to both ion modes, as described in (Southam et al., 2017). Each sample was scanned four times as internal technical replicates. After collecting the mass spectra data, 72 and 70 treatment samples were successfully profiled for positive and negative ion mode, respectively.

Since the total number of polarpos and polarneg samples were different, merging the two data sets into one may be problematic. For all remaining processes of this study, including the data analysis and integration, these two modalities from the metabolome are treated as separate data sets.

All the mass spectra were processed by the DIMSpy pipeline implemented on the Galaxy (Ralf and Zhou, 2020), following the methods described by (Taylor et al., 2010). In short, m/z peaks were filtered by their detection among technical replicates (those shared by three out of four technical replicates were retained), and peak signals against extraction blanks (those with signal-to-noise ratio above 3.0 were also retained), and occurrences among all the samples (greater than 50 %). Missing values were imputed by the k-nearest neighbours (KNN) algorithm (Malarvizhi and Thanamani, 2012). All the intensity values were transformed under probabilistic quotient normalisation (PQN; Dieterle et al., 2006) and generalised log transformation (glog; Parsons et al., 2007) to normalise the variance of peaks. The resulting data matrix was used for downstream analysis.

The m/z peak annotation was achieved by a customised algorithm named BEAMSpy (v1.1.0, https://github.com/computational-metabolomics/beamspy), which performs *in silico* prediction of the molecular formula and putative annotation of the compound identity (Taylor et al., 2010). Briefly, the BEAMSpy predicts the molecular formula by the mass-to-charge (m/z) value of the polarpos/polarneg peak with an acceptable error

range of 5 ppm, and then maps the m/z value of the peak against reference compounds within the KEGG Compound database (Kanehisa et al., 2007) with consideration of the parent compound and their positive adducts ($[M + H]^+$, $[M + K]^+$, $[M + Na]^+$) as well as negative adducts ($[M - H]^-$, $[M + Cl]^-$, $[M + HAc - H]^-$, [M + K - 2H]-, $[M + Na - 2H]^-$). The putatively annotated peak list was then used for functional interpretation and pathway analysis.

3.3.7 Module identification

The co-responsive networks were built from omics data of each case study based on the adjacency matrix (Singh and Sharma, 2012). As such, the features under investigation can either be genes or metabolites.

For feature i and feature j, the adjacency matrix s_{ij} was calculated as follows:

$$s_{ij} = || \, \mathrm{cor}(x_i, x_j) \, ||$$

where x_i and x_j were the (normalised) expression levels of feature i and feature j, and s_{ij} represented the absolute value of the Pearson's correlation coefficient of feature i and feature j. Due to the noisy nature of the omics data (Ma and Zhang, 2019), soft thresholding is utilised to retain the stronger correlation and suppress the weaker correlation caused by noise (Langfelder and Horvath, 2008; Zhang and Horvath, 2005). The soft thresholding was applied to generate a weighted adjacency a_{ij} matrix by calculating the following function:

$$a_{ij} = s_{ij}^{\beta}$$

where β was a value to power the s_{ij}.

The goal of soft thresholding is to generate a resulting network that is believed to have a scale-free topology (Barabási and Oltvai, 2004), which is characterised by the degree distribution of the network following a power-law module:

$$P(k) = ak^{-\gamma}$$

Where k represents the number of connections of a node (or known as degree), $P(k)$ represents the fraction of nodes with k degree in the network, a is a constant variable.

After log transformation, the $logP(k)$ depends linearly on $logk$ as following:

$$logP(k) = (-\gamma)loga + (-\gamma)logk$$

So that it can be diagnostic of selecting a proper power β. By visualising the impact of β, ranging from 1 to 30, the mean degree and the goodness-of-fit (R^2) of the linear model between $logk$ and $logP(k)$ of the resulting network were plotted against different β values. The power β is selected when the mean degree is closer to 5 (Fortunato, 2010).

The modules within the network were identified by the multi-level modularity optimisation algorithm (Blondel et al., 2008), and the quality of modularity was determined by the following algorithm:

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\sigma(c_i, c_j)$$

where m equals the total number of links, $\mathrm{A_{ij}}$ is the weight of the edge between gene i and gene j, $\mathrm{k_i}$ equals the sum of weights of the edges linked to gene i, and $\mathrm{c_i}$ is the community (module) the gene i is assigned to.

3.3.8 Sparse Canonical Correlation Analysis and sparsity permutation

Data integration consists of identifying correlated features among four data types: chemical distribution among samples, the transcriptomics data, and two modalities of metabolomic data (polarpos and polarneg). The correlations between the individual chemical distribution (caffeine or carbamazepine), transcriptomic features, and

metabolomic features were identified with sparse Canonical Correlation Analysis (sCCA) algorithm (Tenenhaus et al., 2014; Tenenhaus and Tenenhaus, 2011). The sCCA is applied to identify the subset of transcriptomic and metabolomic features linearly correlated to the chemical distribution pattern when projected to a common latent space. With a certain sparsity level (between 0 and 1), subsets of biomolecular features were selected from the transcriptomics data block and the two metabolomics data blocks. I assumed that at a small sparsity level, only a few dozens of features were included in the sCCA model, and those features were most likely to belong to the same module(s); by increasing the sparsity level, more features were included, as well as less correlated modules, so that based on the guilty by association (Piovesan et al., 2015), all the other features within those modules were most likely associated with the chemical distribution pattern.

To clarify the relationship between selected features from sCCA (responsive features) and module assignment, I performed sCCA with multiple sparsity levels (sparsity permutation) and recorded the selected feature sets corresponding to all the sparsity levels. For sparsity s, the module enrichment score was calculated as:

$$\text{Module Enrichment Score}\ (\mathrm{m}) = \left(\frac{\mathrm{G}_{\mathrm{m},s}}{\mathrm{s}}\right)$$

where $\mathrm{G}_{\mathrm{m,s}}$ is the G-statistics of likelihood ratio of sCCA selected features assigned to module m at the sparsity level s. Then I selected the modules based on their P-values (P-value < 0.01), then ranked the modules by their module enrichment scores so that the highest ranked modules were assumed to be most closely associated with the concentration distribution of the chemical compound.

3.3.9 Cross-species extrapolation and pathway overrepresentation analysis

The orthologous relationships between *Daphnia magna* and *Drosophila melanogaster* showed that 8018 *Drosophila melanogaster* genes have orthologs in 10228 *Daphnia magna* genes, belonging to 5301 ortholog groups (OGs) that defined at the *Arthropoda* taxonomic level as the most recent common ancestor, based on the orthoDB database (version 10.1; Kriventseva et al., 2019). Among them, 4042 *Daphnia* genes were annotated with both OGs and corresponding *Drosophila melanogaster*'s pathways information. Previous investigation revealed that all 137 *Drosophila melanogaster* pathways recorded in the KEGG Pathway database (version 96.0; Kanehisa et al., 2007) have unique OGs pattern that can be used to distinguish different pathways. The OGs-pathway associations were thereby established based on OGs and pathways of *Drosophila melanogaster*, as summarised in Table 3.1. These 137 pathways consist of general metabolic pathways (64 %), genetic information processing (16 %), environmental information processing (9 %), cellular processes (7 %) and organismal systems (4 %).

To perform a statistical overrepresentation test for the enrichment of pathways by exposure-responsive *Daphnia* genes, chi-square was normally used to estimate the significance levels for the difference between the observed frequency and the expected frequency (Zhou *et al.* 2017). The null hypothesis of the chi-square test is that the genes are normally distributed across all the known pathways. However, the frequency of ortholog groups may not follow a uniform distribution but instead follow a multinomial (categorical) distribution, resulting in inapplicability of a normal chi-square test. Moreover, it would be nearly impossible to establish the null distribution for ortholog groups as only 62 % of the *Daphnia magna* genes had OGs information.

Table 3. 1 Summary of the 137 pathways in the KEGG pathway database.

Class	Number of pathways [a]	Number of OGs [b]
Cellular processes	**9**	**390**
Cell growth and death	2	46
Transport and catabolism	7	344
Environmental information processing	**13**	**390**
Membrane transport	1	8
Signal transduction	10	349
Signalling molecules and interaction	2	33
Genetic information processing	**22**	**910**
Folding, sorting and degradation	7	271
Replication and repair	7	142
Transcription	3	139
Translation	5	358
Metabolism	**88**	**2050**
Amino acid metabolism	13	200
Biosynthesis of other secondary metabolites	1	2
Carbohydrate metabolism	14	268
Energy metabolism	3	106
Global and overview maps	6	863
Glycan biosynthesis and metabolism	13	133
Lipid metabolism	12	160
Metabolism of cofactors and vitamins	12	97
Metabolism of other amino acids	7	60
Metabolism of terpenoids and polyketides	2	29
Nucleotide metabolism	2	78
Xenobiotics biodegradation and metabolism	3	54
Organismal systems	**5**	**96**
Ageing	1	25
Development and regeneration	1	20
Environmental adaptation	1	8
Immune system	1	27
Sensory system	1	16
Sum	**137**	**3836**

a. The number of pathways recorded in the KEGG Pathway database.
b. The number of ortholog groups (OGs) assigned to the pathways.

A permutation chi-square test was used instead (Beh and Lombardo 2014) to relax the requirement of a uniform distribution and help generate a robust estimation of significance (P-value) directly from resampling detected *Daphnia magna* genes annotated within ortholog groups. Thus, a permutation chi-square test was performed on each *Daphnia magna* re-annotated gene set for each pathway (with their corresponding OGs pattern) 100,000 times. The P-values of the permutation chi-square tests were further corrected followed the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) with a false discovery rate at 0.05.

For the metabolomic co-responsive modules, the pathway overrepresentation test was performed by a chi-square test (McHugh, 2013) only with the putatively annotated peaks against the KEGG Compound database (v96.0), which linked to KEGG Pathway database. The *P-value*s of the chi-square tests were also corrected following the Benjamini-Hochberg procedure with a false discovery rate at 0.05.

3.4 Results

3.4.1 Chemical distribution pattern in surface water samples

As shown in Figure 3.1, the caffeine was detected within the chemical mixtures of waters collected from 29 of the 30 sampled sites from the Chaobai River, ranging from 0 ng/L (B05) to 64.69 ng/L (M06). Carbamazepine was detected in 25 of 30 sampled sites of the Chaobai River, mostly downriver, ranging from 0 ng/L (B01, B03, B04, B05, C04) to 35.23 ng/L (M11) and averaging at 7.23 ng/L across all sites. In the Danube River, carbamazepine was detected in all 12 sites, ranging from 12 ng/L (D04) to 37 ng/L (D06), with an average concentration level of 27.17 ng/L.

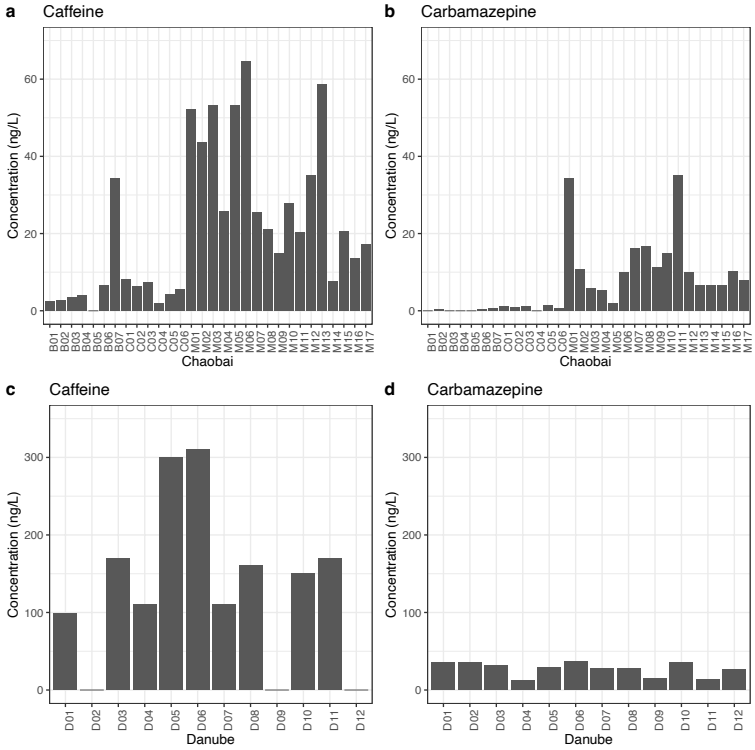

Figure 3. 1 Concentration of caffeine and carbamazepine in surface water samples. Sample sites distributed across the Chaobai River Basin and the Danube River Basin. Plots (a) and (b) show the distribution patterns of these two chemicals within the Chaobai River. Plots (c) and (d) show the same chemical distribution patterns within the Danube River.

3.4.2 Chaobai River case study

Transcriptomic co-responsive network

In the Chaobai River case study, a total of 95 samples were included in the transcriptomic assays. Each sample had 12 million mapped transcript reads on average. Genes with average raw counts under ten were removed, leaving a total of 14705 genes. These 14705 genes were used to construct the co-responsive network. The soft thresholding algorithm was utilised to suppress the weaker correlations among transcriptomic genes (noise). The power values were set to 23, the mean degree of the resulting network is approaching 5, and the linear regression model is

generated by log(Fraction of nodes) and log(degree), with an R^2 value of 0.91 (Figure S3.1b), indicating that the resulting network is scale-free. Based on the multi-level modularity optimisation algorithm, a total of 25 modules (with more than 5 features) are identified from 5699 genes. The largest module (CB_0) consists of 1068 genes, and the smallest module (CB_24) consists of only 21 genes (detailed information listed in Table S3.1).

Sparse CCA analysis between chemical and transcriptomic features

The sCCA was applied to discover a subset of transcriptomic features that were linearly correlated with caffeine or carbamazepine concentrations in the mixture. As revealed in Figure S3.2, the subset of transcriptomic features can explain a portion of the total variance in the whole transcriptomic data, and simultaneously correlated with the chemical concentration values. The subset of transcriptomic features can explain 56.7 % of the total variance of the transcriptomic profiles and correlate with caffeine with a correlation coefficient of 0.525. At the same time, another subset of transcriptomic features can account for 10.2 % of the total variance of the transcriptomic profiles and correlate with carbamazepine with a correlation coefficient of 0.617.

Chemical-associated co-responsive modules

The co-responsive modules that are associated with caffeine are ranked based on their module enrichment scores and annotated with their corresponding levels of P-value, as shown in Figure S3.3a. At a P-value threshold of under 0.01, eight co-responsive modules are regarded as caffeine-associated modules, namely CB_0, CB_1, CB_2, CB_3, CB_4, CB_7, CB_15, and CB_20. Similarly, sixteen co-responsive modules are regarded as carbamazepine-associated modules, namely CB_0, CB_1, CB_2, CB_3,

CB_4, CB_5, CB_6, CB_7, CB_8, CB_9, CB_10, CB_11, CB_14, CB_19, CB_21, and CB_24, as shown in Figure S3.3b.

Pathway overrepresentation analysis

The statistical overrepresentation tests of *Daphnia magna* genes within functional pathways were performed by permutation chi-square test. The *Daphnia magna* genes that are mapped to orthologous *D.melanogaster* genes and the KEGG pathway database are summarised in Table S3.1. A total of 116 pathways are identified to be significantly enriched in at least one co-responsive module. The adjusted P-values of overrepresentation tests on the KEGG pathways in the 18 co-responsive modules associated with caffeine and/or carbamazepine are listed in Appendix 1. The adjusted P-values of overrepresentation tests on 20 selected pathways are plotted in Figure 3.2, which represent pathways of xenobiotic metabolisms, apoptosis and general metabolic pathways.

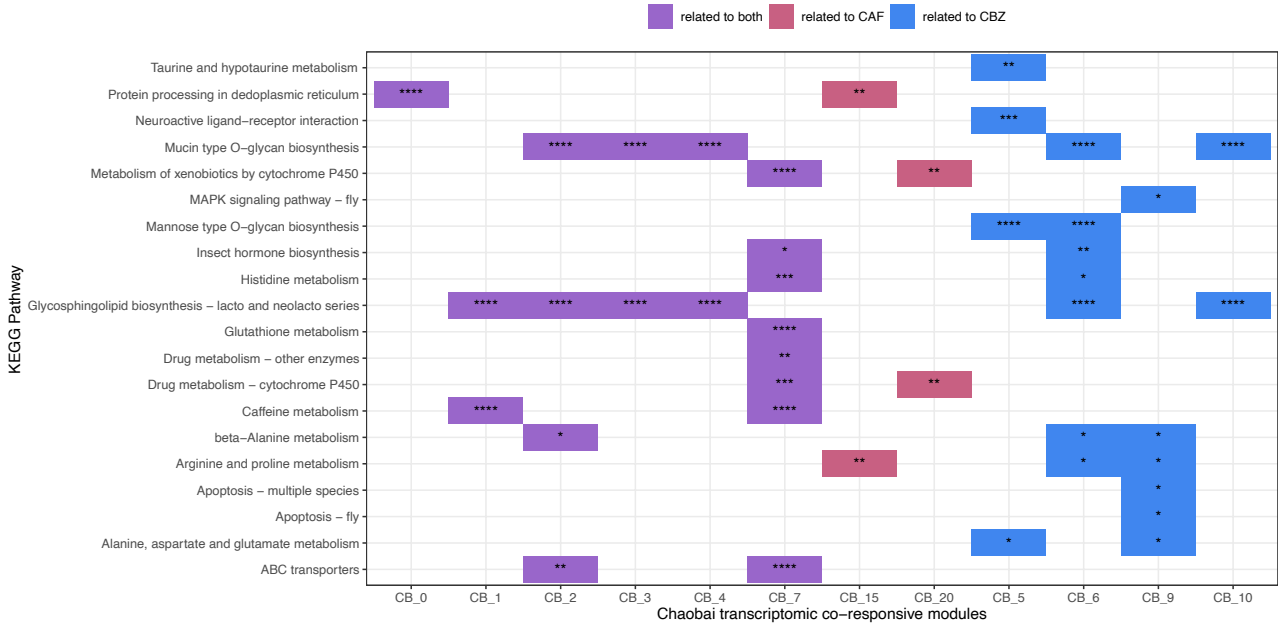

Figure 3. 2 Chaobai case study: overrepresentation tests of selected KEGG pathways by permutation chi-square test. Selected pathways and modules are coloured based on (1) modules associated with both caffeine (CAF) and carbamazepine (CBZ) concentrations in

the mixtures (in purple), (2) modules associated with CAF (in red), and (3) modules associated with CBZ (in blue). The adjusted P-values are labelled as follows: P < 0.05, *; P < 0.01, **; P < 0.001, ***; P < 0.0001, ****.

Modules that are associated with both caffeine and carbamazepine

As shown in Figure 3.2 and Appendix 1, there are six modules that fall within both caffeine- and carbamazepine-associated module lists, namely CB_0, CB_1, CB_2, CB_3, CB_4, and CB_7. Modules CB_0, CB_1, CB_2, CB_3 and CB_4 report extremely significant P-values (P-value lower than 0.00001) at enriching pathways related to carbohydrate metabolism, amino acid metabolism, lipid metabolism, glycan biosynthesis and metabolism, mRNA translation, DNA transcription, lysosome, peroxisome, and protein processing, suggested that these five modules may be responsible for general metabolism of cellular components (carbohydrate, amino acids, lipids and proteins) in response of external chemical exposure.

At gene level, modules associated with caffeine and carbamazepine consist of 31 CYP, 14 GST, and 44 ABC genes. Notably, two genes for xanthine dehydrogenase/oxidase (XDH) are found within module CB_7. Among the significantly enriched pathways in CB_7, there are various pathways known to be functionally related to xenobiotic metabolisms, such as pathways mediated by cytochrome P450, ABC transporter, drug metabolism, and glutathione metabolism. The caffeine metabolism is also found to be significantly enriched by genes within module CB_7.

Modules that are associated with caffeine only

Module CB_15 and module CB_20 are caffeine-specific co-responsive modules. Based on the overrepresentation tests, three pathways are significantly enriched by genes within module CB_15, namely arginine and proline metabolism, various types

of N-glycan biosynthesis, and protein processing in the endoplasmic reticulum. Except for two pathways mediated by cytochrome P450, seven other pathways are significantly enriched by genes within module CB_20. These pathways are tyrosine metabolism, pyruvate metabolism, glycolysis/gluconeogenesis, fatty acid degradation, retinol metabolism, phototransduction and endocytosis.

Modules that are associated with carbamazepine only

There are ten modules associated with carbamazepine and not to caffeine, namely modules CB_5, CB_6, CB_8, CB_9, CB_10, CB_11, CB_14, CB_19, CB_21 and CB_24. Most of the ten modules are associated with the biosynthesis of glycan, carbohydrate, amino acids, cofactors, and lipids. For example, genes from module CB_6 are shown to significantly enrich glycan biosynthesis pathways, histidine metabolism, arginine-proline metabolism, beta-alanine metabolism, and insect hormone biosynthesis; CB_9 is significantly enriching seventeen metabolic pathways related to carbohydrate, amino acid, cofactors, glycan and lipid; CB_10 may be composed of genes that function in glycan metabolism. The ten carbamazepine-associated modules are also associated with pathways in neuroactive ligand-receptor interaction (CB_5), the apoptosis pathways and MAPK signalling pathway (CB_9), autophagy (CB_11, CB_19, CB_24), Toll and Imd signalling pathway (CB_14, CB_21), and signal transduction (CB_11, CB_24).

At gene level, based on ortholog group functional annotation, carbamazepine-related features consist of 34 CYP, 16 GST, 46 ABC, 2 Glutathione peroxidase (GPX), 22 Sulfotransferase (SF), and 6 Superoxide dismutase (SOD) *Daphnia* genes. There are also neuroactive ligand-receptors, such as neurotransmitter-gated ion-channel ligand-

binding domain coding genes (13 genes), sodium channel proteins (2 genes) and sodium neurotransmitter symporters (11 genes).

Generally, these ten carbamazepine-specific modules suggest the potential impact of carbamazepine exposure as apoptosis, glycan biosynthesis variation, neuroactive receptor binding, accompanied with effects on general metabolism.

3.4.3 Danube River case study

Transcriptomic co-responsive network

In the Danube River case study, a total of 96 samples were included in the transcriptomic sequencing, which generated 6 million mapped reads per sample with an average mapping rate of 83 %. With the power value of 18, the mean degree of the weighted co-responsive network is close to 5, and the goodness-of-fit (R^2) value is 0.92 (Figure S3.4). Subsequently, based on the multi-level modularity optimisation algorithm, a total of 36 modules are identified from 4465 genes. The largest module (DA_0) was with 803 genes, and the smallest module (DA_35) was with 20 genes. The descriptions of 36 co-responsive modules are listed in Table S3.2.

Metabolomic co-responsive network

For the polar metabolite profiles, after data pre-processing by DIMSpy, 1285 peaks were detected and selected among 72 treatment samples under the detection of positive ion mode (polarpos shorts for polar-positive metabolite); 2331 peaks were detected and selected among 70 treatment samples under the detection of negative ion mode (polarneg shorts for polar-negative metabolite). These two metabolomic modalities were used for constructing the co-responsive networks separately.

For both polarpos (Figure S3.5) and polarneg (Figure S3.6), when the power is set at 16, the resulting weighted correlation network is scale-free, with an R^2 value of 0.96. Based on the multi-level modularity optimisation algorithm, a total of 56 modules are identified from 975 peaks of polarpos data set. The largest module (DAp_0) consists of 89 peaks and the smallest module (DAp_55) consists of 5 peaks (Table 5.3). Using the same method, a total of 70 modules are identified from 1958 peaks of polarneg data set. The largest module (DAn_0, 208) consists of peaks 40 times more the peaks included in the smallest module (DAn_65, 5). The total number of peaks, the number of peaks with chemical formulae and the number of peaks with KEGG putative annotation are summarised for each module in Table S3.3 and Table S3.4.

Sparse CCA analysis between chemical, transcriptomic and metabolomics features

The sCCA was applied to discover a subset of transcriptomic features and two subsets of metabolomic features (polarpos and polarneg) that were linearly correlated with caffeine or carbamazepine concentrations in the mixture. As revealed in Figure S3.7 (a, c, e), the subset of transcriptomic features can explain 7.87 % of the total variance and correlate with caffeine with a correlation coefficient of -0.610. In the same sCCA model, the subset of polarpos features can explain 10.63 % of the total variance in polarpos dataset and correlate with caffeine (0.56). The subset of polarneg features selected in this model can explain 7.4 % of the total variance in the polarneg dataset and correlated with caffeine with a coefficient of -0.570.

The plots b, d, and f in Figure S3.7 shows the correlation relationships between subsets of transcriptomic or metabolomic features selected by sCCA model and the concentration levels of carbamazepine. To be specific, the subset of transcriptomic

features account for 11.7 % of the total variance and correlate with carbamazepine with a coefficient of -0.550. The polarpos subset account for 19.0 % of the total variance and correlate to carbamazepine with a coefficient of -0.42. In the same sCCA model, the subset of polarneg features explain 22.6 % of the total variance and correlate with carbamazepine with a coefficient of 0.470.

Chemical-associated co-responsive modules

The co-responsive modules that are associated with caffeine are ranked based on their module enrichment scores, as shown in Figure S3.8. With consideration of a P-value threshold at 0.01, nine transcriptomic modules, (DA_1, DA _5, DA _11, DA _0, DA _13, DA _10, DA _2, DA _17 and DA_14), nine polarpos modules (DAp_45, DAp_55, DAp_18, DAp_2, DAp_4, DAp_10, DAp_12, DAp_0, DAp_14), and forteen polarneg modules (DAn_39, DAn_42, DAn_59, DAn_13, DAn_8, DAn_18, DAn_5, DAn_2, DAn_0, DAn_27, DAn_6, DAn_1, DAn_9, DAn_33) are regarded as caffeine-associated modules.

Similarly, the co-responsive modules that are carbamazepine-associated are ranked by their module enrichment scores, as shown in Figure S3.9. The carbamazepine-related co-responsive modules are eleven transcriptomic modules (DA_1, DA_2, DA_0, DA_5, DA_8,DA_4, DA_30, DA_12, DA_7, DA_18, DA_22), nine polarpos modules (DAp_38, DAp_16, DAp_18, DAp_0, DAp_9, DAp_10, DAp_4, DAp_7, DAp_5), and fifteen polarneg modules (DAn_55, DAn_33, DAn_41, DAn_57, DAn_1, DAn_32, DAn_2, DAn_66, DAn_3, DAn_8, DAn_4, DAn_6, DAn_12, DAn_27, DAn_35), based on P-value thresholding at 0.01.

Pathway overrepresentation analysis

For transcriptomic co-responsive modules, the statistical overrepresentation tests of *Daphnia magna* genes within functional pathways were performed by permutation chi-square test. The *Daphnia magna* genes that can be mapped to orthologous *D.melanogaster* genes and KEGG Pathway database are summarised in Table S3.2. A total of 119 pathways are identified to be significantly enriched in at least one transcriptomic co-responsive module. The adjusted P-values of overrepresentation tests on the 119 KEGG pathways are listed in Appendix 2. The adjusted P-values of overrepresentation tests on the same 20 pathways are plotted in Figure 3.3a.

For metabolomic co-responsive modules, the statistical overrepresentation tests of polar metabolite peaks (polarpos and polarneg) within functional pathways were performed by chi-square test. The polar metabolite peaks that have molecular formula and putative annotation based on KEGG Compound database are summarised in Table S3.3 and S3.4. Jointly, a total of 49 pathways are identified to be significantly enriched in at least one metabolomic co-responsive modules (polarpos and polarneg). The adjusted P-values of overrepresentation tests on the 49 pathways are listed in Appendix 3. The adjusted P-values of overrepresentation tests on the same 20 pathways mentioned earlier are plotted in Figure 3.3b.

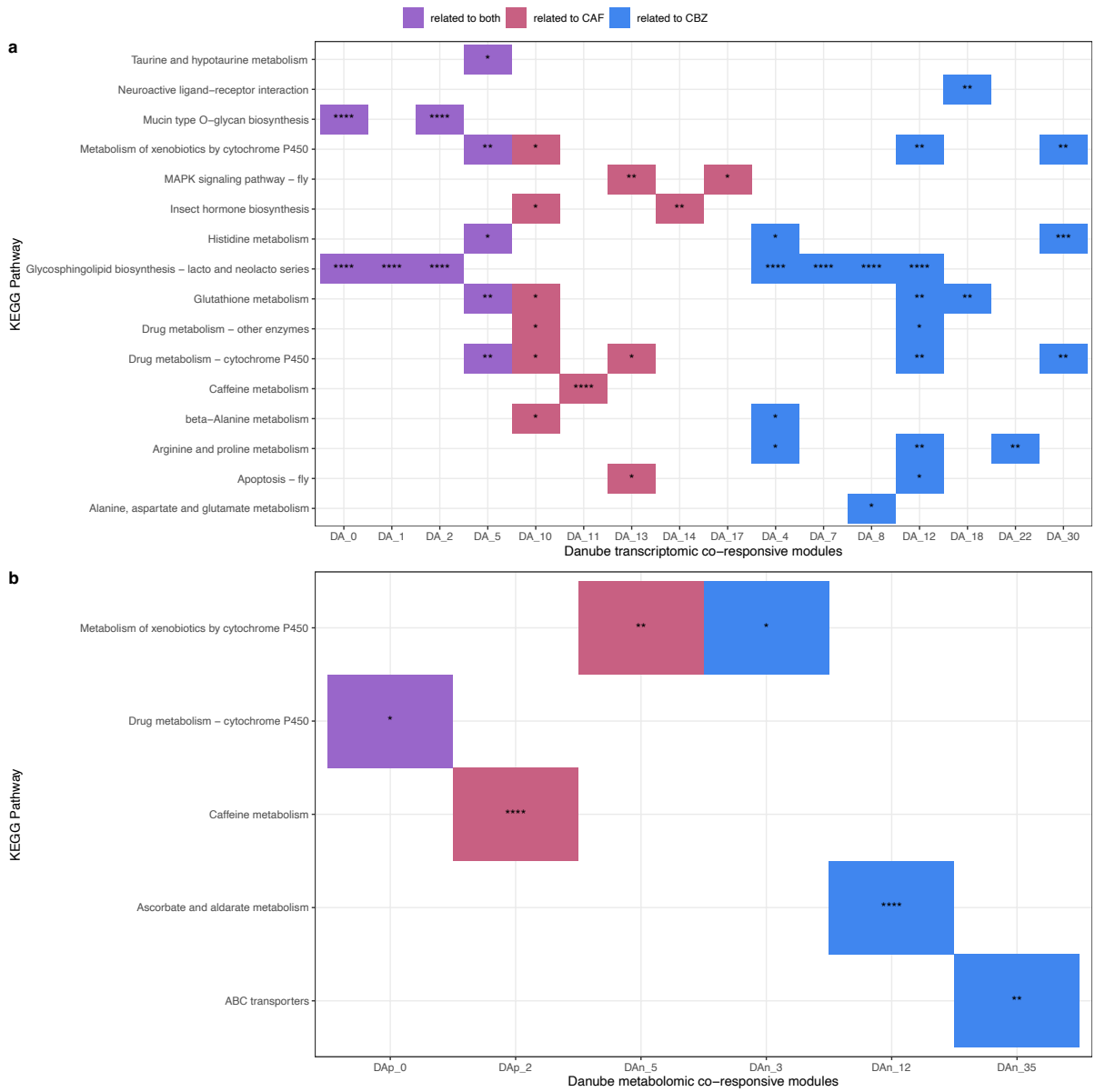

Figure 3. 3 Danube case study: overrepresentation tests of selected KEGG pathways by (a) permutation chi-square test and (b) chi-square test. The modules are selected from (a) transcriptomic and (b) metabolomic co-responsive modules. Selected pathways and modules are coloured based on (1) module associated with both caffeine (CAF) and carbamazepine (CBZ) concentrations in the mxitures (in purple), (2) module associated with CAF (in red), and (3) module associated with CBZ (in blue). The adjusted P-values are labelled as follows: $P < 0.05$, *; $P < 0.01$, **; $P < 0.001$, ***; $P < 0.0001$, ****.

Modules that are associated with both caffeine and carbamazepine

There are four transcriptomic modules associated with both caffeine and carbamazepine, namely DA_0, DA_1, DA_2 and DA_5. Modules DA_0, DA_1 and DA_2 report extremely significant P-values (P-value lower than 0.00001) at enriching pathways related to glycan biosynthesis and steroid biosynthesis. Genes in the module DA_5 are significantly enriched in carbohydrate metabolism, amino acid metabolism, lipid metabolism, glycan biosynthesis and metabolism, cofactors and vitamins metabolism, and xenobiotic biodegradation by cytochrome P450. At gene level, these four modules consist of 34 CYP, 10 GST and 45 ABC genes.

There are nine metabolic modules associated with both chemicals, namely DAp_0, DAp_4, DAp_10, DAn_0, DAn_1, DAn_2, DAn_6, DAn_8, and DAn_27. Metabolites in these nine modules are significantly enriched in drug metabolism by cytochrome P450 (DAp_0), carbohydrate metabolism (DAn_0, DAn_6), Isoflavonoid biosynthesis (DAp_10, DAn_0, DAn_2, DAn_8), flavonoid biosynthesis (DAn_2, DAn_8), and type II polyketide products biosynthesis (DAp_10, DAn_8).At peak level, three peaks are putatively annotated as metabolites involved in xenobiotic metabolism (DAn_1 and DAn_8) and drug metabolism (DAp_0), listed in Table S3.5.

Modules that are associated with caffeine only

There are five transcriptomic modules associated with caffeine only, namely DA_10, DA_11, DA_13, DA_14 and DA_17. The caffeine metabolism pathway is reported to be significantly enriched only in DA_11. The insect hormone biosynthesis pathway is significantly enriched in DA_10 and DA_14. The module DA_10 is reported to be significantly enriched in a few pathways related to xenobiotic metabolism, glutathione metabolism, transcription, and protein folding. Genes in the module DA_13 may be

related to drug metabolism and apoptosis-related pathways. At gene level, the same two genes coding for xanthine dehydrogenase/oxidase (XDH) are found in module DA_11.

For metabolomic profiles, there are four caffeine-specific metabolomic modules, namely DAp_2, DAn_5, DAn_13 and DAn_18. Among the polar positive peak set of DAp_2, there are three peaks putatively annotated as metabolites involved in the caffeine metabolism (Table S3.5), which substantiate the significant enrichment of caffeine metabolism pathway in module DAp_2. However, further confirmation of the chemical structure of the isomers will be needed to verify the structure of the caffeine metabolites in *Daphnia magna*. There are four peaks in DAn_5, one peak in DAn_13, and two peaks in DAn_18, which are tentatively annotated as metabolites involved in xenobiotic metabolism (Table 5.5); although the P-values of overrepresentation tests on xenobiotic metabolism with metabolites from modules DAn_13 and DAn_18 are not significant. On the contrary, flavonoid biosynthesis is reported to be significantly enriched in both DAn_13 and DAn_18.

Modules that only related to carbamazepine

Seven transcriptomic modules are only associated with carbamazepine, namely DA_4, DA_7, DA_8, DA_12, DA_18, DA_22, and DA_30. Modules DA_4, DA_7, DA_12, DA_22 and DA_30 are reported to be significantly enriched in pathways related to carbohydrate metabolism, amino acid metabolism, and lipid metabolism. The glycosphingolipid biosynthesis is significantly enriched in modules DA_4, DA_7, DA_8 and DA_12. The xenobiotic metabolism pathways are significantly enriched in modules DA_12 and DA_30. And glutathione metabolism is significantly enriched in modules DA_12 and DA_18. Notably, apoptosis pathway is only reported to be enriching in

module DA_12; and the neuroactive ligand-receptor interaction is only reported to be significantly enriched in module DA_18.

For metabolomics profiles, there are six metabolic modules particularly associated with carbamazepine, namely DAp_16, DAp_38, DAn_3, DAn_12, DAn_35, DAn_66. Module DAn_3 is significantly enriched in xenobiotic biodegradation pathway, and the DAn_35 is reported with significant enrichment in ABC transporters. Notably, one peak in the polarneg data set was putatively annotated as carbamazepine-o-quinone, which is involved in the carbamazepine-associated metabolomic module (DAn_17).

3.5 Discussion

3.5.1 Chemical associated co-responsive features integrated from transcriptomic and metabolomic

In this study, two cases are included to investigate the chemical component (caffeine and carbamazepine) associated effect with newly developed methods. Integration of transcriptomic and metabolomic features is achieved by multi-block correlation modelling, which highlights the linear dependent relationships between a set of biological feature readouts and the concentration level of a specific chemical compound.

Xenobiotic metabolism is a well-studied chemical detoxification pathway, which includes three phases: monooxygenases (cytochrome P450s, CYPs; Guéguen et al., 2006), conjugation (i.e., glutathione, glutathione S-transferase and sulfotransferase; Townsend and Tew, 2003), and xenobiotic transport (i.e., multidrug resistance-associated proteins; Xu et al., 2005). In both case studies, modules associated with both caffeine and carbamazepine consist of CYP, GST, and ABC genes, all of which

play significant roles in xenobiotic detoxification and excretion *Daphnia magna* (Campos et al., 2014; Lee et al., 2019). These modules are characterised by xenobiotic processes mediated by ABC transporter, cytochrome P450 and glutathione metabolism, suggesting that the xenobiotic metabolic pathways mediated by ABC, CYPs and GST can be successfully captured by this new method, and the resulting features are align with prior knowledge on general chemical metabolism (Campos et al., 2014; Lee et al., 2019).

Caffeine metabolism in *Daphnia magna*

In both studies, transcriptomic co-responsive modules include two genes annotated as xanthine dehydrogenase/oxidase (XDH) (CB_7 in Chaobai case, DA_11 in Danube case), which are mediators for the biotransformation of dimethylxanthine into dimethyluric acid (Begas et al., 2007). In the Danube case, metabolomic co-responsive modules include features putatively annotated as metabolites involved in the caffeine metabolism, which further substantiate that the method can effectively identify caffeine metabolic associated features in both transcriptomic and metabolomic profiles. Based on prior knowledge, there are a few potential routes of caffeine metabolism (Figure 3.4): (1) caffeine can be 1-demethylation as theobromine by CYP then transformed into 3,7-dimethyluric acid by xanthine oxidase; (2) Caffeine can be 3-demethylation as paraxanthine by CYP then transformed into 1,7-dimethyluric acid by xanthine oxidase or CYP; (3) Caffeine can be transformed to Paraxanthine then 1- or 7-Methylxanthine by CYP and subsequently to 1- or 7-Methyluric acid by xanthine oxidase; (4) Caffeine might be also catalysed C-8 oxidation by caffeine dehydrogenase (cdh). The first three potential routes are the same with the caffeine metabolism in human (Nehlig, 2018) and the fourth route is mediated by co-exist microbes (Summers et al., 2015; Yu et al.,

2008). Structure confirmation of the m/z 167.055911 peak (putatively annotated as 1-/3-/7-Methylxanthine) will be needed for further clarification of which metabolic process (route) is the major caffeine metabolic pathways in *Daphnia magna*.

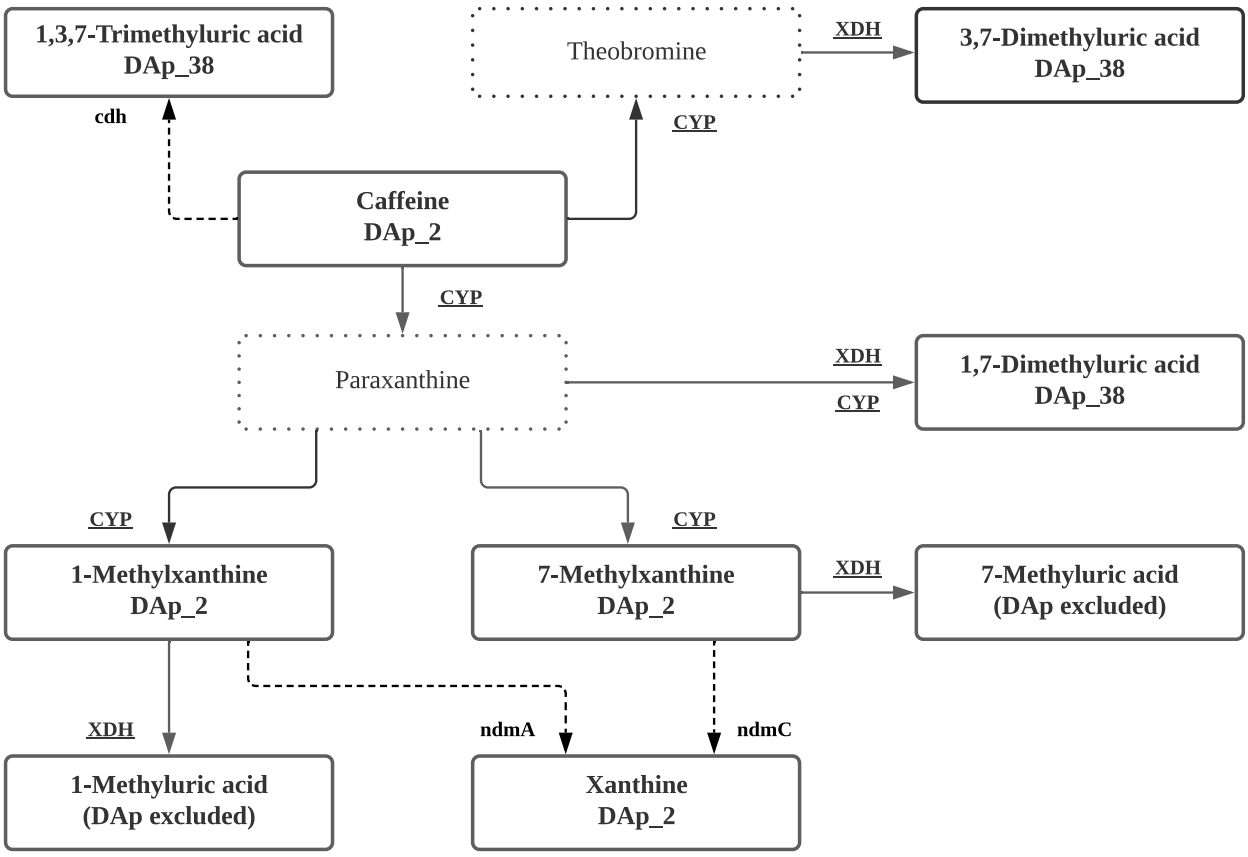

Figure 3. 4 Caffeine metabolism in *Daphnia magna*. The hypothetical metabolic process of caffeine is constructed based on KEGG (map00232), HMDB (SMP0000028) and the biomolecular features of caffeine-associated transcriptomic/metabolomic modules. The underlined genes are the genes detected in the caffeine-associated transcriptomic modules; the bold text within black-solid boxes are the putatively annotated metabolites in the caffeine-associated metabolomic modules. The dashed lines characterise potential microbe-mediated reactions. Abbreviation: XDH, xanthine dehydrogenase/oxidase; CYP, cytochrome P450; cdh, caffeine dehydrogenase; ndmA, methylxanthine N1-demethylase; ndmC, methylxanthine N7-demethylase.

Apart from caffeine biodegradation pathway, pathways significantly enriched in these caffeine-specific modules suggest a potential impact of caffeine exposure on endoplasmic reticulum, endocytosis process, and by disrupting a critical pathway like pyruvate metabolism that may further link to gluconeogenesis. This finding suggests that the modes of action of caffeine may be concentration-dependent, as caffeine exposure at lower levels (below 65 ng/L) might lead to endocytosis inhibition (Gonzalez *et al.* 1990) via suppressing amyloid-beta protein precursor (Li *et al.* 2015). The pathways reported within these caffeine-associated modules imply that the impact of caffeine exposure may be stress-induced apoptosis (Saiki *et al.* 2011) and endocrine disturbance represented by variation in insect hormone synthesis (Coelho *et al.* 2015).

Carbamazepine metabolism in *Daphnia magn*a

In both cases, transcriptomic co-responsive modules consist of drug metabolism-related features. Based on ortholog group functional annotation, carbamazepine-related features consist of CYP, GST, ABC, Glutathione peroxidase (GPX), Sulfotransferase (SF), and Superoxide dismutase (SOD) genes. The biochemical effects of carbamazepine are already known to induce oxidative stress in *Daphnia magna*, represented by significant suppression of SOD, catalase and glutathione reductase (Nkoom et al., 2019). Similar inhibition was also reported in the mussel *Dreissena polymorpha* under carbamazepine exposure for seven days (Contardo-Jara *et al.* 2011), the brachyuran crab *Carcinus maenas* under 50 µg/L carbamazepine for 28 days (Aguirre-Martínez *et al.* 2013), the clam *Ruditapes philippinarum* under 9 µg/L carbamazepine for 28 days (Almeida *et al.* 2015), and the clam *Corbicula flumniea* under 10 and 50 µg/L of carbamazepine for 21 days (Aguirre-Martínez *et al.* 2015). Biotransformation related enzymes like GSTs and CYPs have significantly increased

expression after carbamazepine exposure (Pires *et al.* 2016), as CYPs are involved in hydroxylation and bioactivation of the carbamazepine (Pearce *et al.* 2002, 2008; Aguirre-Martínez *et al.* 2015, 2016) and GSTs catalyse the oxidation of carbamazepine (Vernouillet *et al.* 2010). As shown in Figure 3.5, for carbamazepine, the potential metabolic routes are that carbamazepine can be transformed to 2,3-dihydroxycarbamazepine via carbamazepine-2,3-epoxide and 2-hydroxycarbamazepine or 3-hydroxycarbamazepine, which are similar to the minor metabolic pathways reported in human (Kitteringham et al. 1996; Thorn et al. 2011). However, there is no other evidence to support the potential role of carbamazepine-o-quinone detected in the Danube case, except for one study reported the carbamazepine-o-quinone as one of the metabolites detected in the sea anemones (Vitale et al. 2020). Future biochemical analysis will be needed to confirm the role of this metabolite and the metabolic processes proposed in this work.

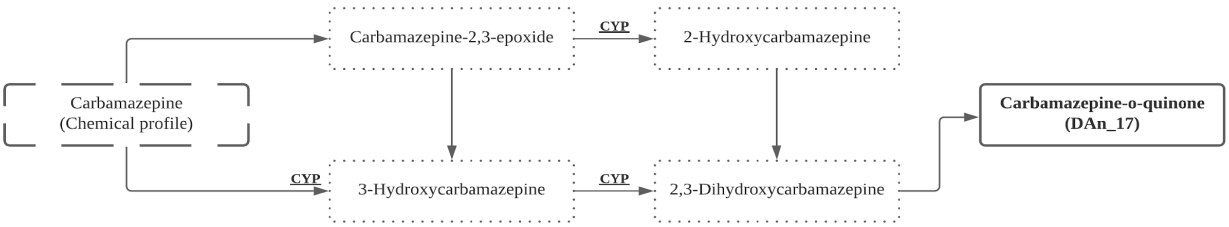

Figure 3. 5 Carbamazepine metabolism in *Daphnia magna*. The hypothetical metabolic process is constructed based on KEGG (map00982), HMDB (SMP0000634) and the features of carbamazepine-associated transcriptomic/metabolomic modules. The underlined genes are the genes in the carbamazepine-associated transcriptomic modules; the bold text within black-solid box is the putatively annotatted metabolite in the carbamazepine-associated metabolomic module. Abbreviation: CYP, cytochrome P450.

3.5.2 Transcriptome-based case comparison

The robustness of this data-driven method in identifying the co-responsive features associated with caffeine and metabolism is assessed by comparing the genes of transcriptomic co-responsive modules independently identified in the Chaobai and Danube case. The shared number of *Daphnia* genes between the selected co-responsive modules are shown in Figure S3.10. Pathway analysis of those shared *Daphnia magna* genes are summarised in Appendix 4 and plotted in Figure 3.6.

For the modules that are associated with both chemicals, up to 356 *Daphnia* genes are common between CB_4 and DA_2, which are significantly enriched in pathways like mucin type O-glycan biosynthesis and glycosphingolipid biosynthesis. There are 340 *Daphnia* gene shared by CB_3 and DA_1, which are also significantly enriched in glycosphingolipid biosynthesis. A total of 191 *Daphnia* genes are common between modules CB_2 and DA_5, which are significantly enriching glutathione metabolism, drug metabolism and xenobiotics metabolism by cytochrome P450, neuroactive ligand-receptor interaction, carbohydrate metabolism, glycan biosynthesis, and lipid metabolism. For the modules that are associated with to carbamazepine not to caffeine, the 31 *Daphnia* genes shared by modules CB_6 and DA_7 share are significantly enriched in drug metabolism pathways, xenobiotic metabolism by cytochrome P450 and glutathione metabolism. As listed above, the significantly enriched pathways of those intersect gene sets are similar to the significantly enriched pathways of those two modules being compared, suggested that the data-driven approach to identify molecular features associated with chemical component in the the mixture can identify the functional genes that are core to xenobiotic metabolism and other chemical exposure-related pathways.

However, there are no genes shared between any caffeine-associated modules (CB_15, CB_20; DA_10, DA_11, DA_13, DA_14, DA_17) from the two case studies. The discrepancy in the genes among these caffeine-specific modules might be due to major differences in the concentration levels of caffeine detected in the two rivers. As lower levels of caffeine exposure (below 65 ng/L in Chaobai River case) are found to be associated with endocytosis effect and higher levels of caffeine exposure (100-310 ng/L in Danube River case) are associated with apoptotic activation, these results suggest that the modes of action of caffeine in the mixtures may be concentration dependent and the adverse outcomes are determined by the concentration levels (Saiki *et al.* 2011). It's worthwhile to perform validation on the responsive patterns of caffeine at different concentration levels within natural chemical mixtures in order to substantiate the mode of action of caffeine in the environmental chemical mixture.

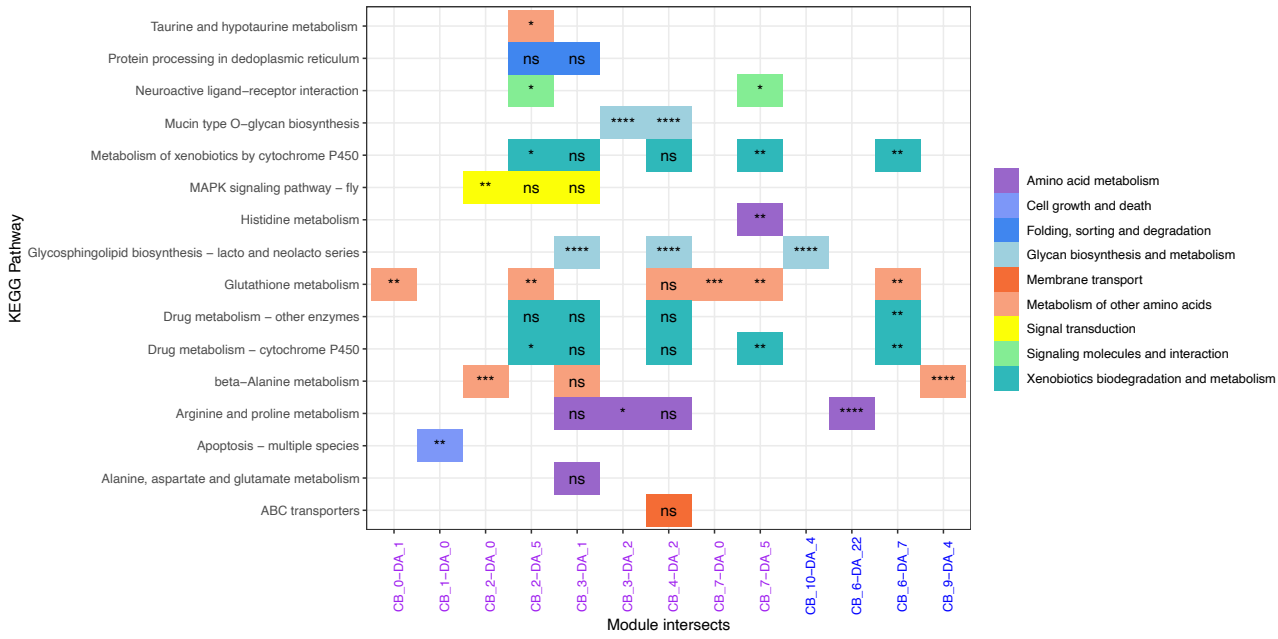

Figure 3. 6 Case comparison: pathway overrepresentation tests by permutation chi-square test. The module pairs are selected from the transcriptomic co-responsive modules of Chaobai (CB) and Danube (DA) case. Selected module pairs are coloured based on (1) modules that are

both associated with caffeine (CAF) and carbamazepine (CBZ) concentrations in the mixtures (in purple), and (2) modules that are both associated with CBZ (in blue). The adjusted P-value are annotated as following: P > 0.05, ns; P < 0.05, *; P < 0.01, **; P < 0.001, ***; P < 0.0001, ****.

3.5.3 Comparison of transcriptomic and metabolomic co-responsive modules in the Danube River case study

In the Danube River case study, the biological responses of the organic extracts from the Danube River were depicted by both transcriptome and metabolome. The identified pathways included in the selected co-responsive modules in the Danube case are compared between the two omics approaches to evaluate the insights provided by two types of omics assays.

In general, the transcriptomic assays consist of more diverse information than the metabolomic assays. There are 119 pathways in 16 transcriptomic modules that are associated with caffeine and/or carbamazepine; but for metabolomic co-responsive modules, there are only 67 pathways in 37 selected modules that are associated with caffeine and/or carbamazepine. Those 16 selected transcriptomic modules mainly consist of genes involved in pathways of carbohydrate metabolism (17 %), amino acid metabolism (13 %), signal transduction (11 %), lipid metabolism (11 %), glycan biosynthesis and metabolism (8 %), xenobiotic biodegradation and metabolism (5 %), metabolism of cofactors and vitamins (5 %), transport and catabolism (5 %), and translation (4 %). While the 37 metabolomic modules mainly comprise of metabolites that participate in pathways of biosynthesis of secondary metabolites (30 %), carbohydrate metabolism (18 %), xenobiotic biodegradation and metabolism (6 %), and membrane transport (6 %).

For chemical associated co-responsive modules, the Venn diagrams (Figure S3.11) reveal that there is consistency between pathways identified in the transcriptomes and those in the metabolomes. For example, there are two pathways associated with both chemicals that are identified by both omics assays, which are drug metabolism via cytochrome P450, and starch and sucrose metabolism. Two pathways that are associated with caffeine only in both omics assays, which were caffeine metabolism, and metabolism of xenobiotics by cytochrome P450. There are 12 pathways that are commonly found in carbamazepine-associated co-responsive modules of two omics, which are seven carbohydrate metabolic pathways, two nucleotides metabolism pathways, xenobiotic metabolism, amino acid metabolism and signal transduction.

The chemical-associated co-responsive modules identified by the two omics assays also consist of pathways that are complementary to each other. For example, for co-responsive modules that are associated with caffeine only, the transcriptomic co-responsive modules consist of genes from the pathways of aging, amino acid metabolism, transcription, translation, and signal transduction, while the metabolomic co-responsive modules consist of metabolites mostly participating in the metabolism and/or catabolism.

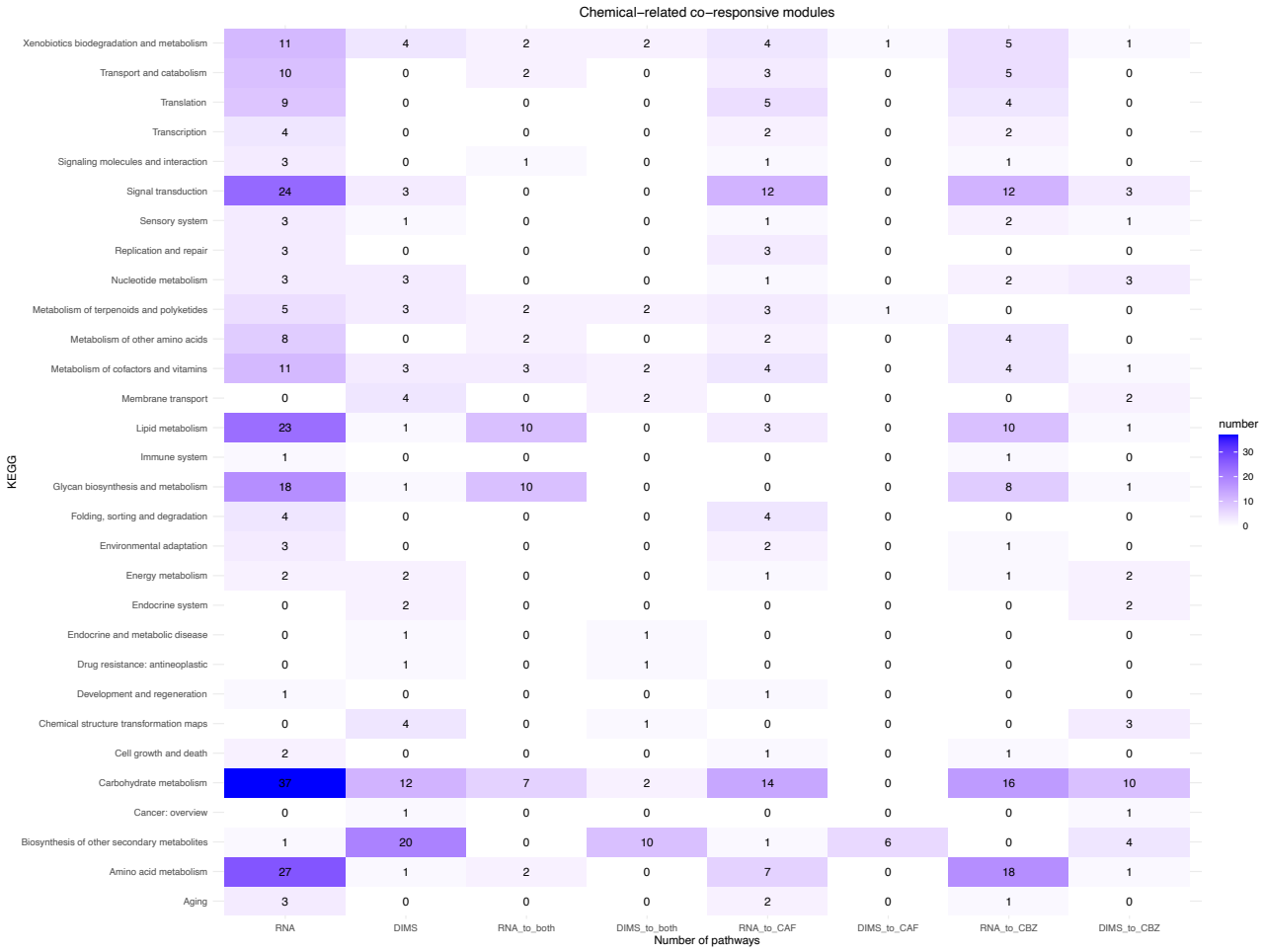

Figure 3. 7 Danube case study: summary of the numbers of pathways identified in the chemical-associated co-responsive modules in transcriptome and metabolome. Abbreviation: RNA for transcriptome, DIMS for metabolome, CAF for caffeine, CBZ for carbamazepine, both refers to caffeine and carbamazepine.

3.6 Conclusion

In this work, the co-responsive features of two well-studied chemical compounds, caffeine and carbamazepine, were investigated in two case studies. Omics-based bioassays, transcriptomic and metabolomic, were applied to generate systematic biological responses. Compared with prior knowledge, the method purposed here, which combines the multi-block correlation modelling and network analysis, can

effectively identify biological features that are reasonably related to caffeine or carbamazepine metabolism. The highly consistency in the transcriptomic co-responsive features generated by two independent case studies indicate that the data-drive approach proposed in this work can identify not only the general metabolic pathways but also potential concentration-dependent adverse effects, as constituents of the chemical mode of action. The integration of transcriptome and metabolome assists in constructing the hypothetical metabolic processes of caffeine and carbamazepine in *Daphnia magna*, but also provides a more comprehensive profiles that joint the benefits of complementary views of distinctive types of omics assays. Most importantly, it substantiated the practicality of the Precision Environmental Health framework in identifying chemical component molecular mode of actions. With the knowledge and evidence of chemical component related effect, it may be able to identify the harmful chemical component within the environmental chemical mixture, which is the premise of establishing a practical and efficient environmental monitoring and regulation to manage the pollutants that may lead to health threats.

3.7 Supplementary

Table S3. 1 Chaobai case study: transcriptomic co-responsive module gene lists mapping summary.

Module ID	Number of genes [a]	Genes with orthologs [b]	Genes with orthologs and pathways [c]
CB_0	1068	881 (83%)	514 (48%)
CB_1	939	590 (63%)	204 (22%)
CB_2	763	412 (54%)	165 (22%)
CB_3	673	327 (49%)	78 (12%)
CB_4	622	243 (39%)	55 (9 %)
CB_5	305	213 (70%)	43 (14 %)
CB_6	253	129 (51%)	45 (18%)
CB_7	189	121 (64%)	55 (29%)
CB_8	147	130 (88%)	97 (66%)
CB_9	142	74 (52%)	20 (14%)
CB_10	103	80 (78%)	21 (20%)
CB_11	75	16 (21%)	7 (9%)
CB_12	52	31 (60%)	13 (25%)
CB_13	42	19 (45%)	8 (19%)
CB_14	41	17 (42%)	3 (7%)
CB_15	40	14 (35%)	5 (13%)
CB_16	35	22 (63%)	5 (14%)
CB_17	33	4 (12%)	2 (6%)
CB_18	28	19 (68%)	9 (32%)
CB_19	28	2 (7%)	1 (4%)
CB_20	26	22 (85%)	12 (46%)
CB_21	26	6 (23%)	3 (12%)
CB_22	25	16 (64%)	11 (44%)
CB_23	23	13 (57%)	4 (17%)
CB_24	21	8 (38%)	2 (10%)

a. The total number of *Daphnia magna* genes in each transcriptomic co-responsive module.

b. The total number of *Daphnia* magna genes shared orthologous relationships with *Drosophila melanogaster* genes.

c. The total number of *Daphnia magna* genes with *Drosophila melanogaster* orthologs and pathway information recorded in the KEGG database.

Table S3. 2 Danube case study: transcriptomic co-responsive module gene lists mapping summary.

Module ID	Number of genes [a]	Genes with orthologs [b]	Genes with orthologs and pathways [c]
DA_0	803	396 (49%)	144 (18%)
DA_1	553	293 (53%)	64 (12%)
DA_2	508	198 (39%)	55 (11%)
DA_3	470	323 (69%)	166 (35%)
DA_4	269	164 (61%)	66 (25%)
DA_5	235	154 (66%)	63 (27%)
DA_6	207	170 (82%)	120 (58%)
DA_7	189	83 (44%)	21 (11%)
DA_8	182	116 (64%)	38 (21%)
DA_9	114	75 (66%)	26 (23%)
DA_10	93	55 (59%)	23 (25%)
DA_11	79	43 (54%)	20 (25%)
DA_12	66	33 (50%)	9 (14%)
DA_13	59	36 (61%)	14 (24%)
DA_14	49	28 (57%)	11 (22%)
DA_15	45	24 (53%)	12 (27%)
DA_16	42	26 (62%)	13 (31%)
DA_17	41	22 (54%)	5 (12%)
DA_18	39	20 (51%)	7 (18%)
DA_19	33	22 (67%)	5 (15%)
DA_20	32	17 (53%)	5 (16%)
DA_21	30	17 (57%)	7 (23%)
DA_22	28	20 (71%)	8 (29%)
DA_23	28	18 (64%)	6 (21%)
DA_24	27	11 (41%)	2 (7%)
DA_25	25	19 (76%)	5 (20%)
DA_26	24	13 (54%)	3 (13%)
DA_27	23	13 (57%)	5 (22%)
DA_28	23	17 (74%)	10 (44%)
DA_29	23	14 (61%)	5 (22%)
DA_30	22	18 (82%)	11 (50%)
DA_31	22	15 (68%)	6 (27%)
DA_32	21	3 (14%)	0
DA_33	21	12 (57%)	4 (19%)
DA_34	20	0	0
DA_35	20	9 (45%)	3 (15%)

a. The total number of *Daphnia magna* genes in each transcriptomic co-responsive module.

b. The total number of *Daphnia magna* genes shared orthologous relationships with *Drosophila melanogaster* genes.

c. The total number of *Daphnia magna* genes with *Drosophila melanogaster* orthologs and pathway information recorded in the KEGG database

Table S3. 3 Danube case study: metabolomic (polar positive) co-responsive module peak lists mapping summary.

Module	Number [a]	Formulae [b]	KEGG [c]	Module	Number	Formulae	KEGG
DAp_0	89	27	23	DAp_28	10	2	2
DAp_1	67	44	37	DAp_29	10	6	5
DAp_2	62	35	28	DAp_30	10	7	5
DAp_3	41	10	10	DAp_31	10	4	3
DAp_4	40	24	17	DAp_32	10	4	4
DAp_5	34	7	5	DAp_33	10	1	1
DAp_6	32	19	16	DAp_34	9	3	2
DAp_7	32	5	4	DAp_35	9	6	4
DAp_8	31	21	17	DAp_36	9	5	4
DAp_9	29	13	9	DAp_37	9	1	1
DAp_10	28	21	15	DAp_38	8	4	2
DAp_11	26	2	2	DAp_39	8	4	3
DAp_12	26	12	10	DAp_40	8	2	2
DAp_13	25	6	5	DAp_41	8	5	2
DAp_14	22	10	8	DAp_42	8	1	1
DAp_15	21	16	14	DAp_43	7	2	2
DAp_16	20	0	1	DAp_44	7	0	0
DAp_17	15	3	2	DAp_45	7	8	6
DAp_18	15	7	6	DAp_46	6	6	3
DAp_19	15	3	2	DAp_47	6	1	1
DAp_20	14	0	0	DAp_48	6	5	3
DAp_21	14	12	10	DAp_49	5	0	0
DAp_22	12	2	2	DAp_50	5	0	0
DAp_23	12	0	0	DAp_51	5	1	1
DAp_24	12	9	8	DAp_52	5	5	4
DAp_25	11	4	3	DAp_53	5	2	2
DAp_26	10	2	2	DAp_54	5	5	5
DAp_27	10	0	0	DAp_55	5	0	0

a. The total number of peaks in polar positive ion datasets.

b. The number of peaks that could be assigned empirical formulae.

c. The number of peaks with putative annotation based on KEGG database.

Table S3. 4 Danube case study: metabolomic (polar negative) co-responsive module peak lists mapping summary.

Module	Number [a]	Formulae [b]	KEGG [c]	Module	Number	Formulae	KEGG
DAn_0	208	75	72	DAn_35	12	8	8
DAn_1	161	93	70	DAn_36	11	7	7
DAn_2	145	49	39	DAn_37	11	10	10
DAn_3	118	86	60	DAn_38	11	5	3
DAn_4	71	24	23	DAn_39	11	9	6
DAn_5	66	51	38	DAn_40	11	12	9
DAn_6	59	47	31	DAn_41	10	4	4
DAn_7	57	31	24	DAn_42	10	8	6
DAn_8	53	92	45	DAn_43	10	10	6
DAn_9	50	11	12	DAn_44	10	3	3
DAn_10	46	31	24	DAn_45	9	3	2
DAn_11	45	17	12	DAn_46	9	16	9
DAn_12	44	41	27	DAn_47	9	6	6
DAn_13	43	38	25	DAn_48	9	0	0
DAn_14	37	25	17	DAn_49	9	12	7
DAn_15	36	4	5	DAn_50	9	0	1
DAn_16	34	6	6	DAn_51	9	8	4
DAn_17	34	24	20	DAn_52	9	6	3
DAn_18	32	39	21	DAn_53	8	4	3
DAn_19	31	3	3	DAn_54	8	6	4
DAn_20	29	23	19	DAn_55	8	0	1
DAn_21	28	23	17	DAn_56	8	3	3
DAn_22	25	12	9	DAn_57	7	3	2
DAn_23	24	7	11	DAn_58	7	11	5
DAn_24	23	17	10	DAn_59	6	7	4
DAn_25	22	12	12	DAn_60	6	1	2
DAn_26	21	9	8	DAn_61	6	5	4
DAn_27	20	18	13	DAn_62	6	1	1
DAn_28	19	17	13	DAn_63	6	3	2
DAn_29	18	12	10	DAn_64	5	2	2
DAn_30	16	19	12	DAn_65	5	2	3
DAn_31	16	16	10	DAn_66	5	6	3
DAn_32	15	8	3	DAn_67	5	8	5
DAn_33	14	11	11	DAn_68	5	4	2
DAn_34	13	23	10	DAn_69	5	5	3

a. The total number of peaks in polar positive ion datasets.

b. The number of peaks that could be assigned empirical formulae.

c. The number of peaks with putative annotation based on KEGG database.

Table S3. 5 Danube case study: metabolomic peaks in the metabolomic co-responsive modules that are associated with both caffeine and carbamazepine.

Module	m/z	KEGG ID	Putative annotation	Pathway
DAp_0	314.981944	C07643	4-hydroxycyclophosphamide	Drug metabolism – cytochrome P450
		C07645	Aldophosphamide	
		C16553	4-hydroxyifosfamidem	
		C16556	Aldoifosfamide	
DAn_5	378.026687	C16619	6-Thioguanosine monophosphate	
	392.042435	C16620	6-Methylthioguanosine monophosphate	
DAn_17	303.019088	C16606	Carbamazepine-o-quinone	
DAn_1	321.067854	C14852	Benzo[a]pyrene-7,8-diol	Metabolism of xenobiotics by cytochrome P450
DAn_8	371.074978	C19590	6-[2,3-dihydroxy-1-(hydroxymethyl)propyl]-1,2-dihydro-7-hydroxy-9-methoxy-cyclopenta[c][1]benzopyran-3,4-dione	
DAn_5	400.058901	C14862	2-S-glutathionyl acetate	
	386.042701	C14871	s-(formylmethyl)glutathione	
	369.076449	C14874	glutathione episulfonium ion	
	307.090129	C19489	1a,11b-Dihydro-4,9-dimethylbenz[a]anthra[3,4-b]oxirene	
		C19561	7-Hydroxymethyl-12-methylbenz[a]anthracene	
		C19604	7,12-Dimethylbenz[a]anthracene 5,6-oxide	
DAn_13	258.066151	C19563	4-[(Hydroxymethyl)nitrosoamino]-1-(3-pyridinyl)-1-butanone	
		C19566	4-Hydroxy-4-(methylnitrosoamino)-1-(3-pyridinyl)-1-butanone	
		C19602	4-(Methylnitrosamino)-1-(1-oxido-3-pyridinyl)-1-butanone	
DAn_18	252.099133	C19564	4-(Nitrosoamino)-1-(3-pyridinyl)-1-butanone	
	254.114784	C19581	alpha-[3-(Nitrosoamino)propyl]-3-pyridinemethanol	
DAp_2	153.040183	C00385	Xanthine	Caffeine metabolism
	195.087058	C07481	Caffeine	
	167.055911	C16353	7- Methylxanthine	
		C16357	3- Methylxanthine	
		C16358	1- Methylxanthine	

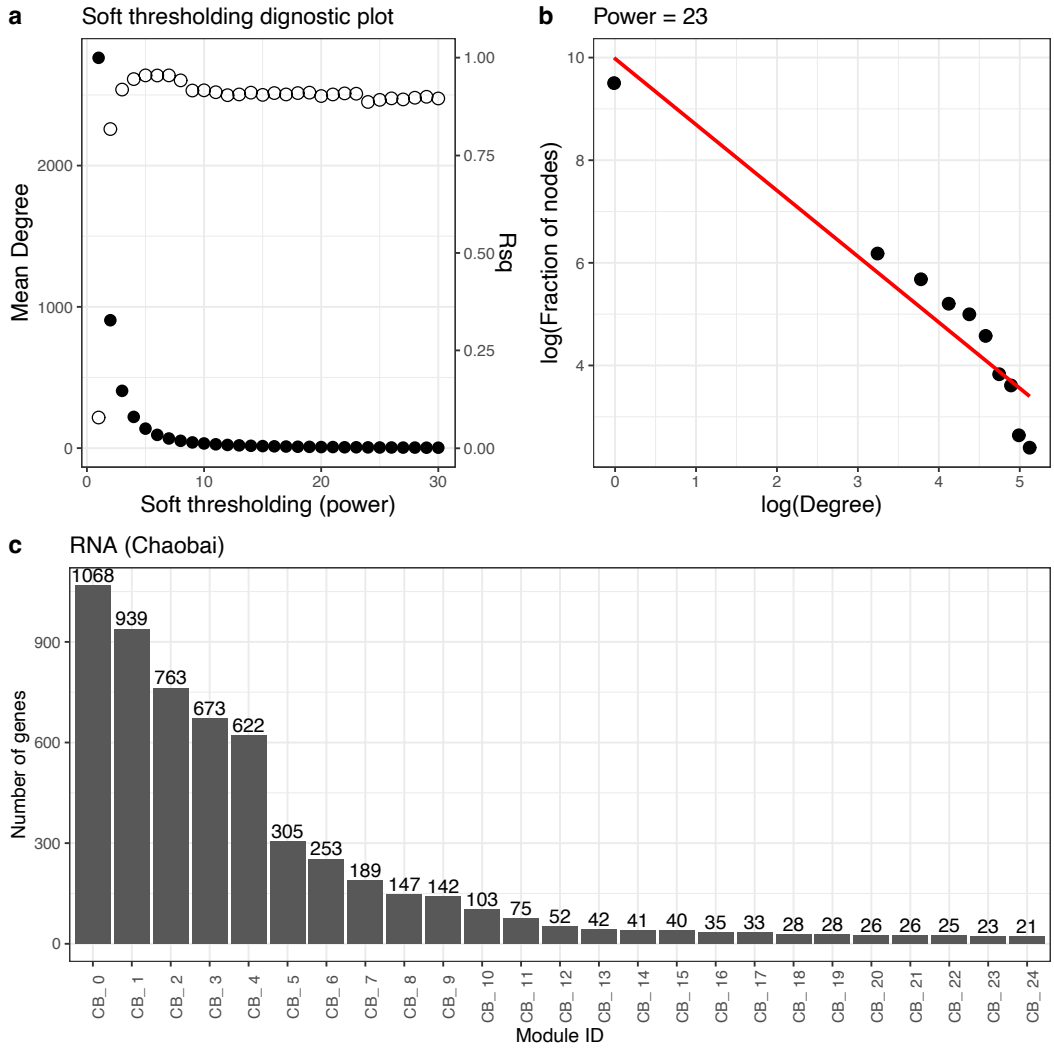

Figure S3. 1 Chaobai case study: transcriptomic co-responsive network and modules. In plot (a), the power value for soft thresholding is on the x-axis, the solid dots (Mean degree, y-axis on the left) represent the mean degree value of the network generated by the weighted adjacency matrix corresponding to a power value, and the hollow dots (Rsq, y-axis on the right) represent the goodness-of-fit (R^2) value of the linear regression model built upon the log-transformed degree and log-transformed fraction of nodes. The power value is selected when the mean degree of the resulting network is closer to 5. Plot (b) shows the linear regression model with the log-transformed degree on the x-axis and the log-transformed fraction of nodes on the y-axis, with a power of 23. Plot (c) shows the number of genes of 25 modules in the transcriptomic co-responsive network.

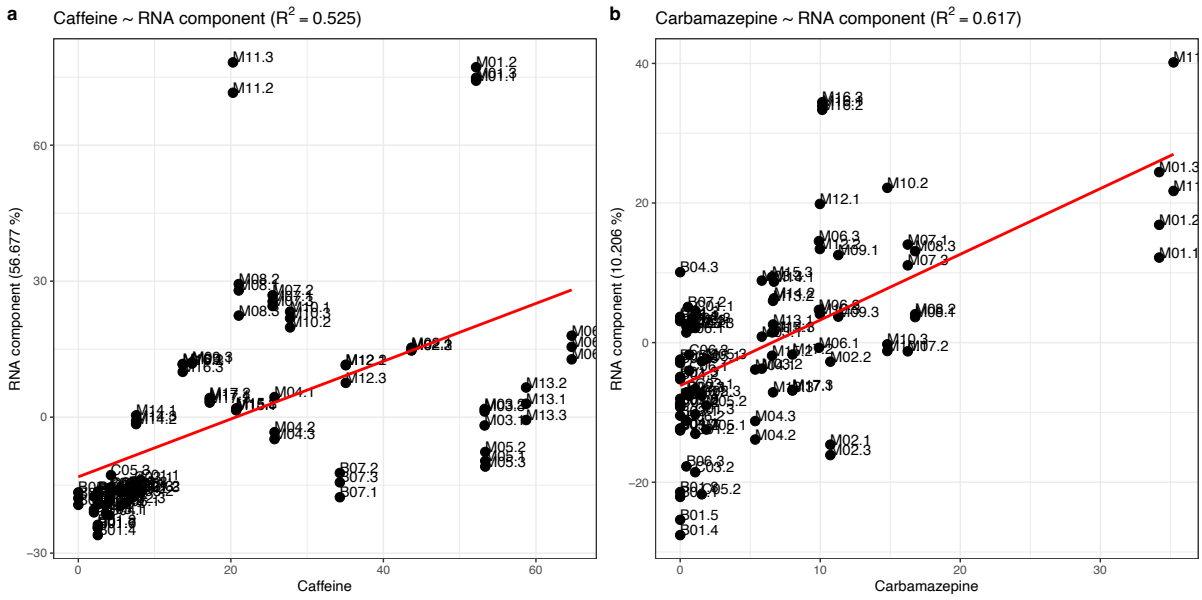

Figure S3. 2 Chaobai case study: sCCA analysis of relationship (a) between transcriptomic component and caffeine, and (b) between transcriptomic component and carbamazepine. The R^2 value is calculated based on the Pearson correlation coefficient between caffeine/carbamazepine concentration values and the subset of transcriptomic features.

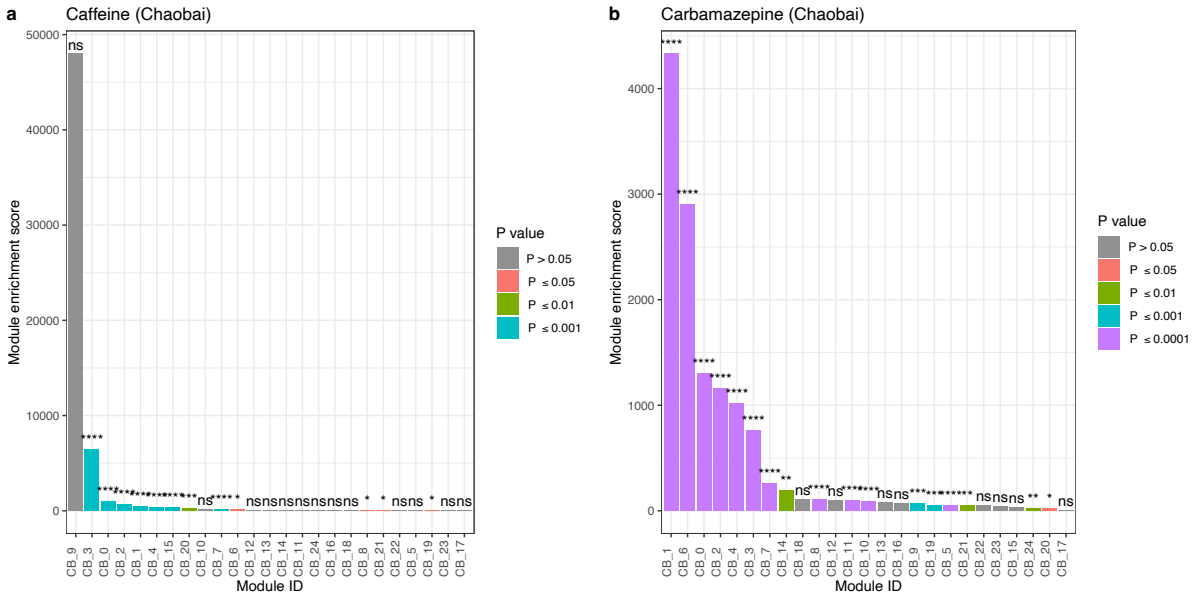

Figure S3. 3 Chaobai case study: transcriptomics co-responsive modules are ranked by their module enrichment scores corresponding to their association with (a) caffeine and (b) carbamazepine concentrations in mixtures. The modules on the x-axis are ordered by the value of the module enrichment score. The P-values of G statistics are stratified into five groups and annotated as follows: P > 0.05, ns (not significant); P < 0.05, *; P < 0.01, **; P < 0.001, ***; P < 0.0001, ****.

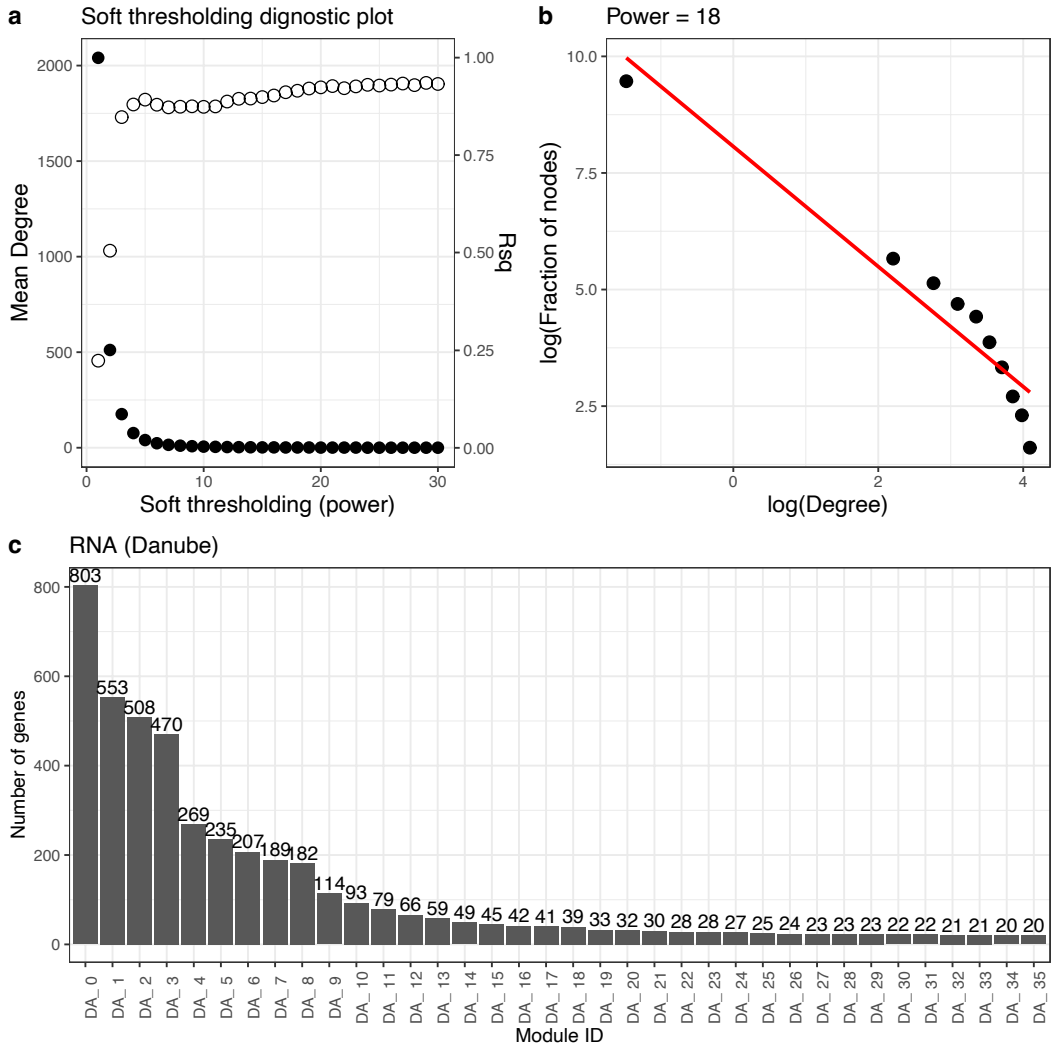

Figure S3. 4 Danube case study: transcriptomic co-responsive network and modules. In plot (a), the power value for soft thresholding is on the x-axis, the solid dots (Mean degree, y-axis on the left) represent the mean degree value of the network generated by the weighted adjacency matrix corresponding to a power value, and the hollow dots (Rsq, y-axis on the right) represent the goodness-of-fit (R^2) value of the linear regression model built upon the log-transformed degree and log-transformed fraction of nodes. Plot (b) shows the linear regression model with the log-transformed degree on the x-axis and the log-transformed fraction of nodes on the y-axis, with a power of 18. Plot (c) shows the number of genes of 36 modules in the transcriptomic co-responsive network.

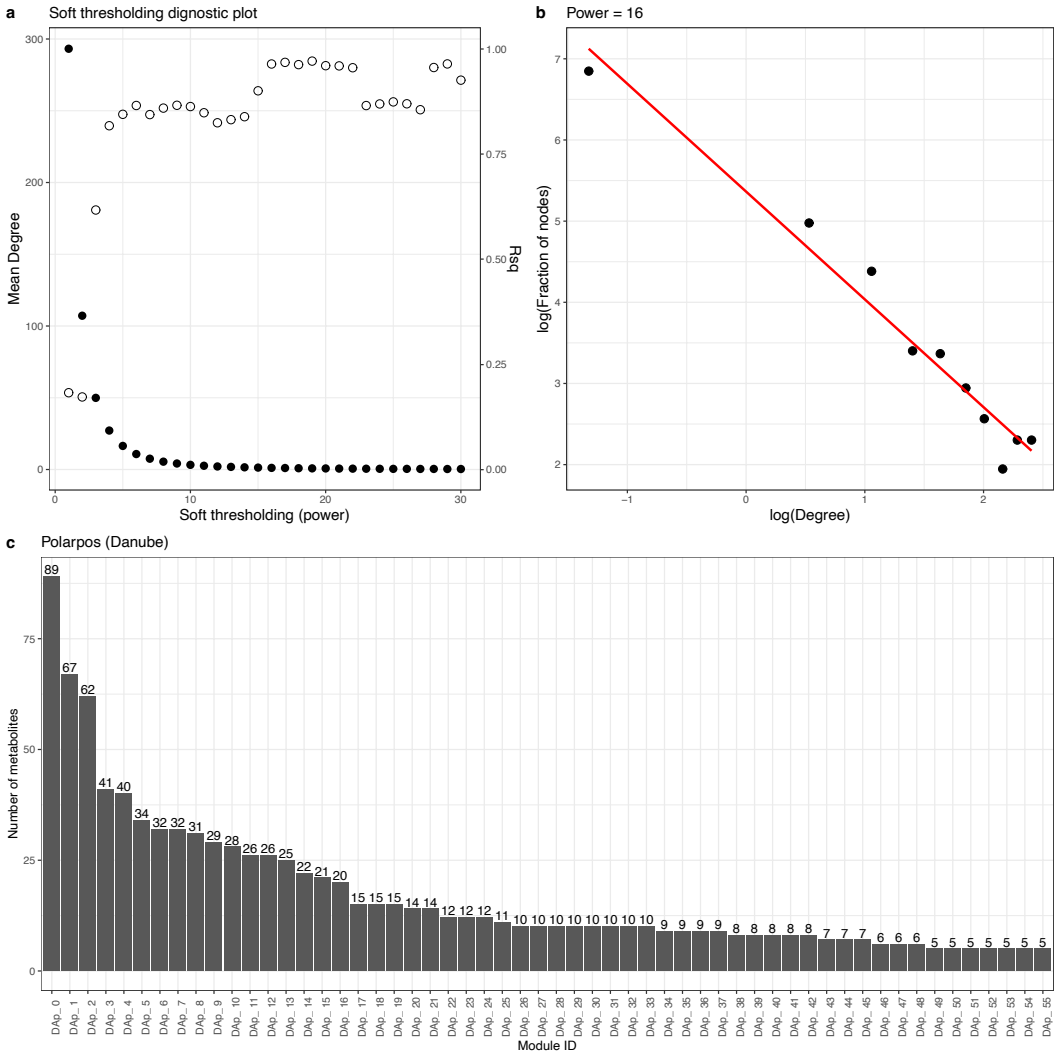

Figure S3. 5 Danube case study: metabolomic (polar positive) peaks co-responsive network and modules. In plot (a), the power value for soft thresholdoing is on the x-axis, the solid dots (Mean degree, y-axis on the left) represent the mean degree value of the network generated by the weighted adjacency matrix corresponding to a power value, and the hollow dots (Rsq, y-axis on the right) represent the goodness-of-fit (R^2) value of the linear regression model built upon the log-transformed degree and the log-transformed fraction of nodes. Plot (b) shows the linear regression model with the log-transformed degree on the x-axis and the log-transformed fraction of nodes on the y-axis, with a power of 16. Plot (c) shows the number of metabolites (peaks) of 56 modules in the metabolomic co-responsive network.

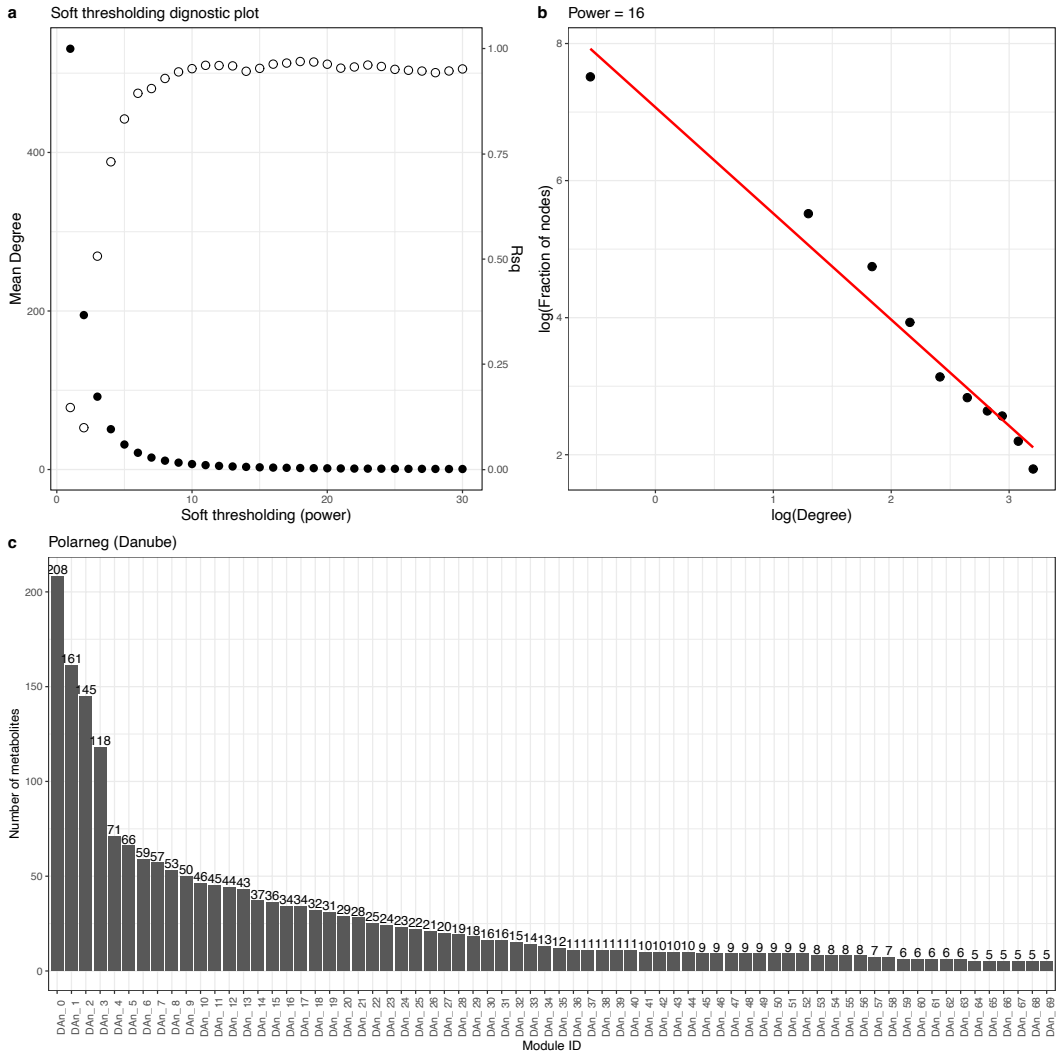

Figure S3. 6 Danube case: metabolomic (polar negative) peaks co-responsive network. In plot (a), the power value for soft thresholdoing is on the x-axis, the solid dots (Mean degree, y-axis on the left) represent the mean degree value of the network generated by the weighted adjacency matrix corresponding to a power value, and the hollow dots (Rsq, y-axis on the right) represent the goodness-of-fit (R^2) value of the linear regression model built upon the log-transformed degree and the log-transformed fraction of nodes. Plot (b) shows the linear regression model with the log-transformed degree on the x-axis and the log-transformed fraction of nodes on the y-axis, with a power of 16. Plot (c) shows the the number of metabolites (peaks) of 70 modules in the metabolomic co-responsive network.

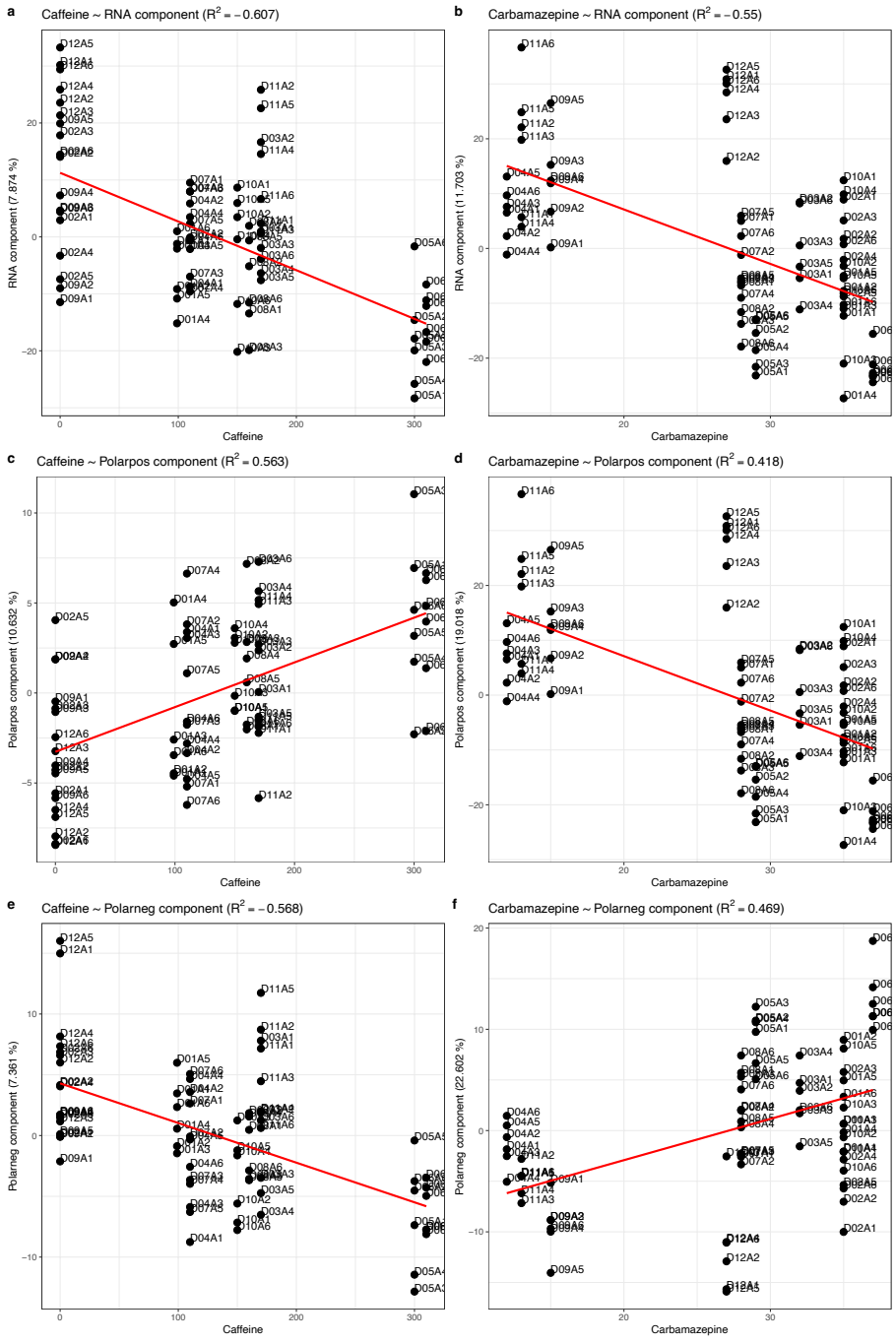

Figure S3. 7 Danube case study: sCCA analysis of relationship between omics features and two chemical compounds. Plots on the left panel show the correlations between concentration of caffeine and (a) transcriptomic selected features, (c) polarpos selected features, and (e) polarneg selected features. Plots on the right panel show the correlations between concentration of carbamazepine and (b) transcriptomic selected features, (d) polarpos selected features, and (f) polarneg selected features.

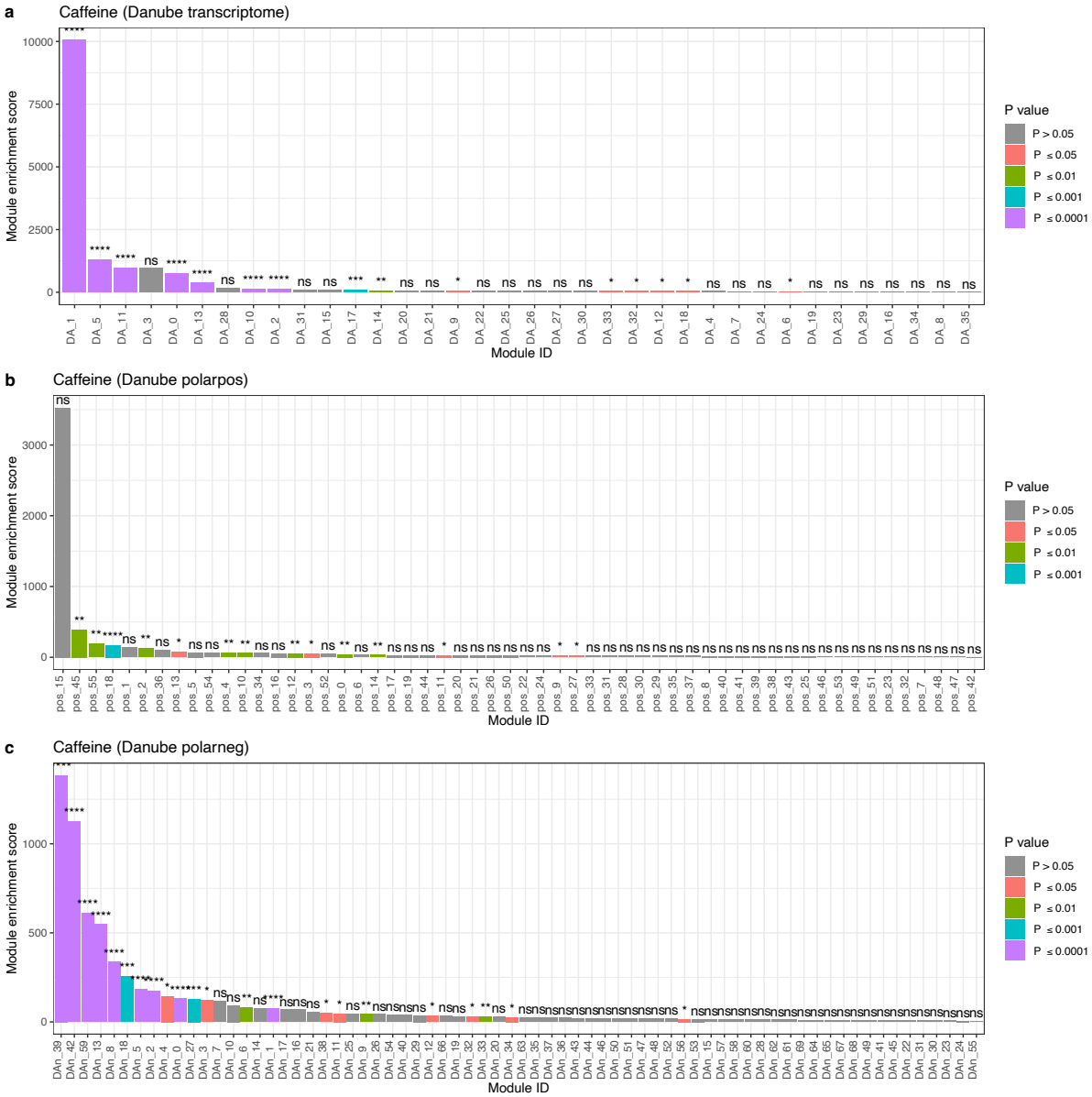

Figure S3. 8 Danube case study: caffeine-associated co-responsive modules are ranked by their module enrichment scores. The co-responsive modules are from (a) transcriptomic, (b) polarpos and (c) polarneg co-responsive networks. The modules on the x-axis are ordered by the value of module enrichment score. The P-values of G statistics are stratified into five groups and annotated as follows: P > 0.05, ns (not significant); P < 0.05, *; P < 0.01, **; P < 0.001, ***; P < 0.0001, ****.

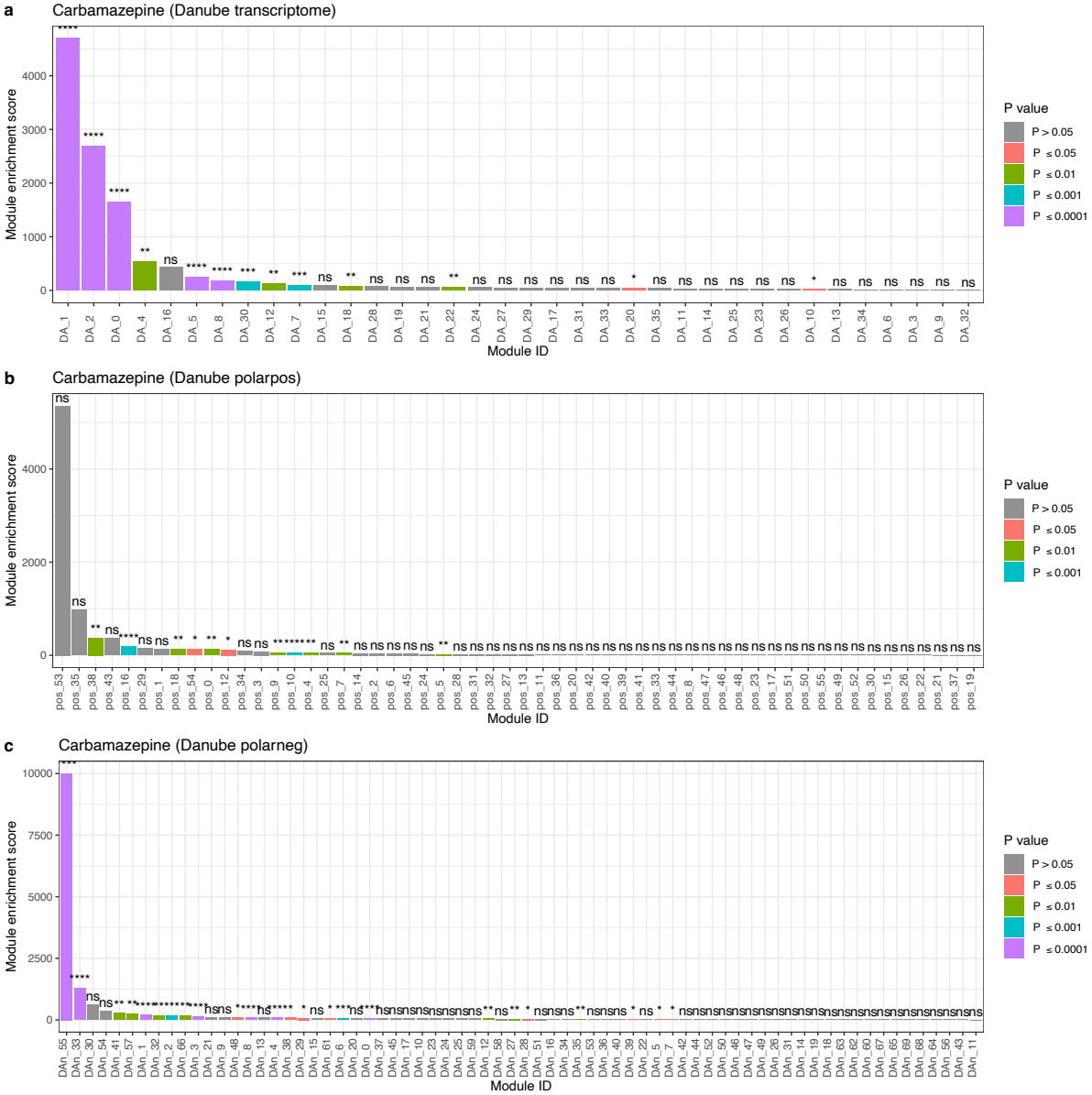

Figure S3. 9 Danube case study: carbamazepine-associated co-responsive modules are ranked by their module enrichment scores. The co-responsive modules are from (a) transcriptomic, (b) polarpos and (c) polarneg co-responsive networks. The modules on the x-axis are ordered by the value of module enrichment score. The P-values of G statistics are stratified into five groups and annotated as follows: P > 0.05, ns (not significant); P < 0.05, *; P < 0.01, **; P < 0.001, ***; P < 0.0001, ****.

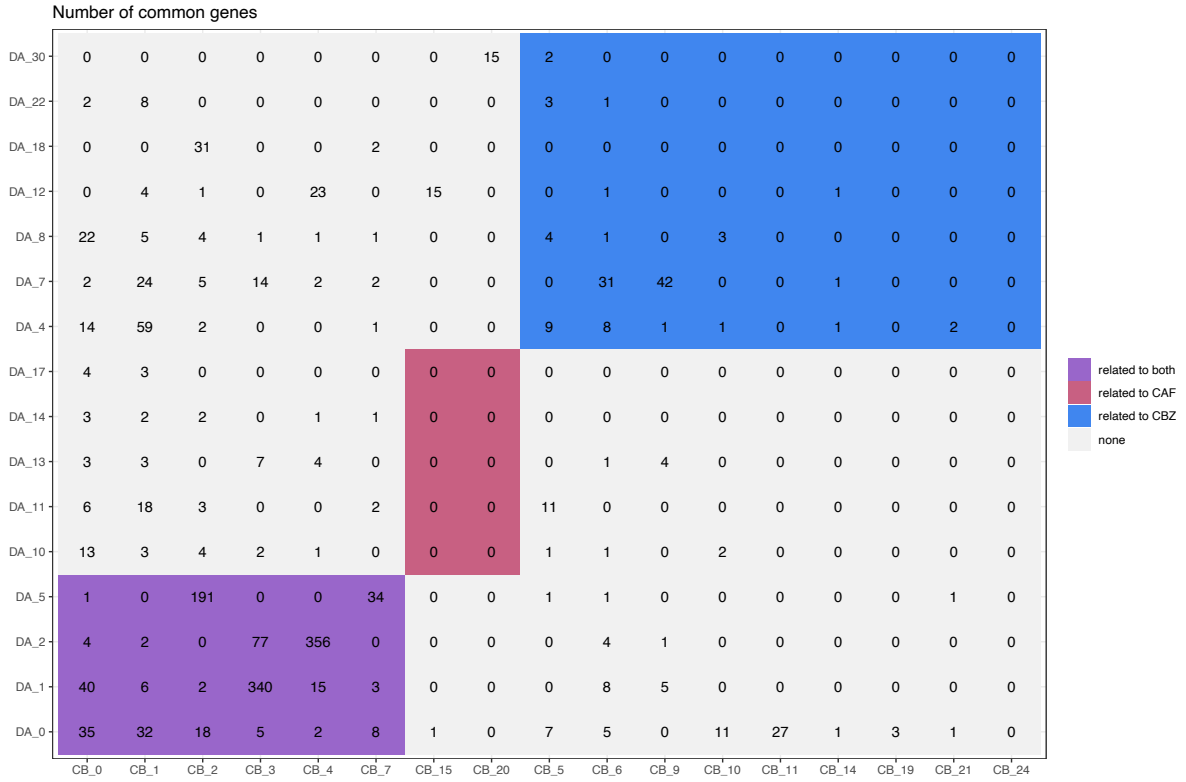

Figure S3. 10 Case comparison: number of common genes between Chaobai transcriptomic modules and Danube transcriptomic modules. Modules are coloured based on (1) modules that are associated with both caffeine (CAF) and carbamazepine (CBZ) concentrations in the mixture (in purple), (2) module that are associated with CAF only (in red), and (3) modules that are associated with CBZ only (in blue).

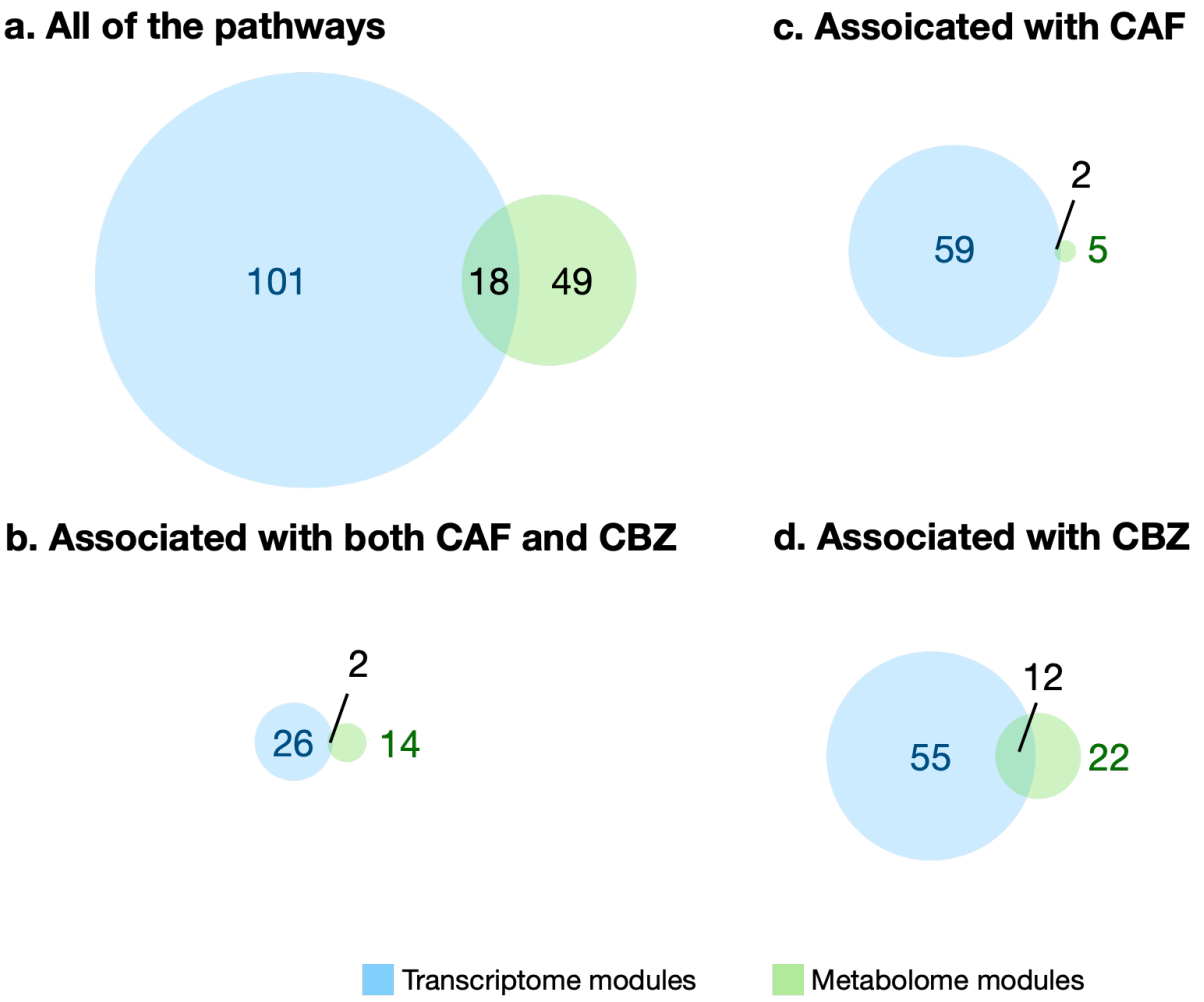

Figure S3. 11 Danube case study: summary of the numbers of pathways identified commonly in transcriptomic and metabolomic co-responsive modules. The venn diagrams show (a) the number of pathways in 16 transcriptomic modules and 37 metabolomic modules, (b) the number of pathways in the modules that are associated with both caffeine and carbamazepine, (c) the number of pathways in the modules that are associated with caffeine only, and (d) the number of pathways in the modules that are associated with carbamazepine only.

3.8 Reference

Altenburger, R., Ait-Aissa, S., Antczak, P., Backhaus, T., Barceló, D., Seiler, T.-B., Brion, F., Busch, W., Chipman, K., de Alda, M.L., de Aragão Umbuzeiro, G., Escher, B.I., Falciani, F., Faust, M., Focks, A., Hilscherova, K., Hollender, J., Hollert, H., Jäger, F., Jahnke, A., Kortenkamp, A., Krauss, M., Lemkine, G.F., Munthe, J., Neumann, S., Schymanski, E.L., Scrimshaw, M., Segner, H., Slobodnik, J., Smedes, F., Kughathas, S., Teodorovic, I., Tindall, A.J., Tollefsen, K.E., Walz, K.-H., Williams, T.D., Van den Brink, P.J., van Gils, J., Vrana, B., Zhang, X., Brack, W., 2015. Future water quality monitoring — Adapting tools to deal with mixtures of pollutants in water resource management. Sci. Total Environ. 512–513, 540–551. https://doi.org/10.1016/j.scitotenv.2014.12.057

Altenburger, R., Brack, W., Burgess, R.M., Busch, W., Escher, B.I., Focks, A., Mark Hewitt, L., Jacobsen, B.N., de Alda, M.L., Ait-Aissa, S., Backhaus, T., Ginebreda, A., Hilscherová, K., Hollender, J., Hollert, H., Neale, P.A., Schulze, T., Schymanski, E.L., Teodorovic, I., Tindall, A.J., de Aragão Umbuzeiro, G., Vrana, B., Zonja, B., Krauss, M., 2019. Future water quality monitoring: improving the balance between exposure and toxicity assessments of real-world pollutant mixtures. Environ. Sci. Eur. 31, 12. https://doi.org/10.1186/s12302-019-0193-1

Andreozzi, R., 2002. Carbamazepine in water: persistence in the environment, ozonation treatment and preliminary assessment on algal toxicity. Water Res. 36, 2869–2877. https://doi.org/10.1016/S0043-1354(01)00500-0

Barabási, A.-L., Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5, 101–113. https://doi.org/10.1038/nrg1272

Barbosa, M.O., Moreira, N.F.F., Ribeiro, A.R., Pereira, M.F.R., Silva, A.M.T., 2016. Occurrence and removal of organic micropollutants: An overview of the watch list of EU Decision 2015/495. Water Res. 94, 257–279. https://doi.org/10.1016/j.watres.2016.02.047

Bean, T.G., Rattner, B.A., Lazarus, R.S., Day, D.D., Burket, S.R., Brooks, B.W., Haddad, S.P., Bowerman, W.W., 2018. Pharmaceuticals in water, fish and osprey nestlings in Delaware River and Bay. Environ. Pollut. 232, 533–545. https://doi.org/10.1016/j.envpol.2017.09.083

Begas, E., Kouvaras, E., Tsakalof, A., Papakosta, S., Asprodini, E.K., 2007. In vivo evaluation of CYP1A2, CYP2A6, NAT-2 and xanthine oxidase activities in a Greek population sample by the RP-HPLC monitoring of caffeine metabolic ratios. Biomed. Chromatogr. 21, 190–200. https://doi.org/10.1002/bmc.736

Ben, W., Zhu, B., Yuan, X., Zhang, Y., Yang, M., Qiang, Z., 2018. Occurrence, removal and risk of organic micropollutants in wastewater treatment plants across China: Comparison of wastewater treatment processes. Water Res. 130, 38–46. https://doi.org/10.1016/j.watres.2017.11.057

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Beutler, A.S., Li, S., Nicol, R., Walsh, M.J., 2005. Carbamazepine is an inhibitor of histone deacetylases. Life Sci. 76, 3107–3115. https://doi.org/10.1016/j.lfs.2005.01.003

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Campos, B., Altenburger, R., Gómez, C., Lacorte, S., Piña, B., Barata, C., Luckenbach, T., 2014. First evidence for toxic defense based on the multixenobiotic resistance (MXR) mechanism in Daphnia magna. Aquat. Toxicol. 148, 139–151. https://doi.org/10.1016/j.aquatox.2014.01.001

Cappelletti, S., Daria, P., Sani, G., Aromatario, M., 2015. Caffeine: Cognitive and Physical Performance Enhancer or Psychoactive Drug? Curr. Neuropharmacol. 13, 71–88. https://doi.org/10.2174/1570159X13666141210215655

Carelli-Alinovi, C., Ficarra, S., Russo, A.M., Giunta, E., Barreca, D., Galtieri, A., Misiti, F., Tellone, E., 2016. Involvement of acetylcholinesterase and protein kinase C in the protective effect of caffeine against β-amyloid-induced alterations in red blood cells. Biochimie 121, 52–59. https://doi.org/10.1016/j.biochi.2015.11.022

Carter, S.L., Brechbuhler, C.M., Griffin, M., Bond, A.T., 2004. Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics 20, 2242–2250. https://doi.org/10.1093/bioinformatics/bth234

Cunningham, V.L., Perino, C., D'Aco, V.J., Hartmann, A., Bechter, R., 2010. Human health risk assessment of carbamazepine in surface waters of North America and Europe. Regul. Toxicol. Pharmacol. 56, 343–351. https://doi.org/10.1016/j.yrtph.2009.10.006

Davidson, E.H., Erwin, D.H., 2006. Gene Regulatory Networks and the Evolution of Animal Body Plans. Science 311, 796–800. https://doi.org/10.1126/science.1113832

Dieterle, F., Ross, A., Schlotterbeck, G., Senn, H., 2006. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in ^{1}H NMR Metabonomics. Anal. Chem. 78, 4281–4290. https://doi.org/10.1021/ac051632c

Fortunato, S., 2010. Community detection in graphs. Phys. Rep. 486, 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

Guéguen, Y., Souidi, M., Baudelin, C., Dudoignon, N., Grison, S., Dublineau, I., Marquette, C., Voisin, P., Gourmelon, P., Aigueperse, J., 2006. Short-term hepatic effects of depleted uranium on xenobiotic and bile acid metabolizing cytochrome P450 enzymes in the rat. Arch. Toxicol. 80, 187–195. https://doi.org/10.1007/s00204-005-0027-3

Gunnarsson, L., Jauhiainen, A., Kristiansson, E., Nerman, O., Larsson, D.G.J., 2008. Evolutionary Conservation of Human Drug Targets in Organisms used for Environmental Risk Assessments. Environ. Sci. Technol. 42, 5807–5813. https://doi.org/10.1021/es8005173

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P., 2019. The reactome pathway

knowledgebase. Nucleic Acids Res. gkz1031. https://doi.org/10.1093/nar/gkz1031
Josyula, N., Andersen, M.E., Kaminski, N.E., Dere, E., Zacharewski, T.R., Bhattacharya, S., 2020. Gene co-regulation and co-expression in the aryl hydrocarbon receptor-mediated transcriptional regulatory network in the mouse liver. Arch. Toxicol. 94, 113–126. https://doi.org/10.1007/s00204-019-02620-5
Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., Yamanishi, Y., 2007. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 36, D480–D484. https://doi.org/10.1093/nar/gkm882
Kato, Y., Kobayashi, K., Oda, S., Colbourn, J.K., Tatarazako, N., Watanabe, H., Iguchi, T., 2008. Molecular cloning and sexually dimorphic expression of DM-domain genes in Daphnia magna. Genomics 91, 94–101. https://doi.org/10.1016/j.ygeno.2007.09.002
Khatri, P., Sirota, M., Butte, A.J., 2012. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput. Biol. 8, e1002375. https://doi.org/10.1371/journal.pcbi.1002375
Kim, J., Park, S.-M., Cho, K.-H., 2013. Discovery of a kernel for controlling biomolecular regulatory networks. Sci. Rep. 3, 2223. https://doi.org/10.1038/srep02223
Kirwan, J.A., Weber, R.J.M., Broadhurst, D.I., Viant, M.R., 2014. Direct infusion mass spectrometry metabolomics dataset: a benchmark for data processing and quality control. Sci. Data 1, 140012. https://doi.org/10.1038/sdata.2014.12
Kolahdouzan, M., Hamadeh, M.J., 2017. The neuroprotective effects of caffeine in neurodegenerative diseases. CNS Neurosci. Ther. 23, 272–290. https://doi.org/10.1111/cns.12684
Koonin, E.V., 2005. Orthologs, Paralogs, and Evolutionary Genomics. Annu. Rev. Genet. 39, 309–338. https://doi.org/10.1146/annurev.genet.39.073003.114725
Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F.A., Zdobnov, E.M., 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 47, D807–D811. https://doi.org/10.1093/nar/gky1053
Kustatscher, G., Grabowski, P., Schrader, T.A., Passmore, J.B., Schrader, M., Rappsilber, J., 2019. Co-regulation map of the human proteome enables identification of protein functions. Nat. Biotechnol. 37, 1361–1371. https://doi.org/10.1038/s41587-019-0298-5
Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559. https://doi.org/10.1186/1471-2105-9-559
Langfelder, P., Horvath, S., 2007. Eigengene networks for studying the relationships between co-expression modules. BMC Syst. Biol. 1, 54. https://doi.org/10.1186/1752-0509-1-54
Lee, B.-Y., Choi, B.-S., Kim, M.-S., Park, J.C., Jeong, C.-B., Han, J., Lee, J.-S., 2019. The genome of the freshwater water flea Daphnia magna: A potential use for freshwater molecular ecotoxicology. Aquat. Toxicol. 210, 69–84. https://doi.org/10.1016/j.aquatox.2019.02.009

Lee, S.-A., Chan, C., Tsai, C.-H., Lai, J.-M., Wang, F.-S., Kao, C.-Y., Huang, C.-Y.F., 2008. Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. BMC Bioinformatics 9, S11. https://doi.org/10.1186/1471-2105-9-S12-S11

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550. https://doi.org/10.1186/s13059-014-0550-8

Ma, T., Zhang, A., 2019. Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE). BMC Genomics 20, 944. https://doi.org/10.1186/s12864-019-6285-x

Malarvizhi, R., Thanamani, D.A.S., n.d. K-Nearest Neighbor in Missing Data Imputation 3.

McHugh, M.L., 2013. The Chi-square test of independence. Biochem. Medica 143–149. https://doi.org/10.11613/BM.2013.018

Mutiyar, P.K., Gupta, S.K., Mittal, A.K., 2018. Fate of pharmaceutical active compounds (PhACs) from River Yamuna, India: An ecotoxicological risk assessment approach. Ecotoxicol. Environ. Saf. 150, 297–304. https://doi.org/10.1016/j.ecoenv.2017.12.041

Neale, P.A., Ait-Aissa, S., Brack, W., Creusot, N., Denison, M.S., Deutschmann, B., Hilscherová, K., Hollert, H., Krauss, M., Novák, J., Schulze, T., Seiler, T.-B., Serra, H., Shao, Y., Escher, B.I., 2015. Linking in Vitro Effects and Detected Organic Micropollutants in Surface Water Using Mixture-Toxicity Modeling. Environ. Sci. Technol. 49, 14614–14624. https://doi.org/10.1021/acs.est.5b04083

Nehlig, A., 2018. Interindividual Differences in Caffeine Metabolism and Factors Driving Caffeine Consumption. Pharmacol. Rev. 70, 384–411. https://doi.org/10.1124/pr.117.014407

Nkoom, M., Lu, G., Liu, J., Yang, H., Dong, H., 2019. Bioconcentration of the antiepileptic drug carbamazepine and its physiological and biochemical effects on Daphnia magna. Ecotoxicol. Environ. Saf. 172, 11–18. https://doi.org/10.1016/j.ecoenv.2019.01.061

Oropesa, A.L., Floro, A.M., Palma, P., 2016. Assessment of the effects of the carbamazepine on the endogenous endocrine system of Daphnia magna. Environ. Sci. Pollut. Res. 23, 17311–17321. https://doi.org/10.1007/s11356-016-6907-7

Orsini, L., Brown, J.B., Shams Solari, O., Li, D., He, S., Podicheti, R., Stoiber, M.H., Spanier, K.I., Gilbert, D., Jansen, M., Rusch, D.B., Pfrender, M.E., Colbourne, J.K., Frilander, M.J., Kvist, J., Decaestecker, E., De Schamphelaere, K.A.C., De Meester, L., 2018. Early transcriptional response pathways in *Daphnia magna* are coordinated in networks of crustacean-specific genes. Mol. Ecol. 27, 886–897. https://doi.org/10.1111/mec.14261

Orsini, L., Gilbert, D., Podicheti, R., Jansen, M., Brown, J.B., Solari, O.S., Spanier, K.I., Colbourne, J.K., Rusch, D.B., Decaestecker, E., Asselman, J., De Schamphelaere, K.A.C., Ebert, D., Haag, C.R., Kvist, J., Laforsch, C., Petrusek, A., Beckerman, A.P., Little, T.J., Chaturvedi, A., Pfrender, M.E., De Meester, L., Frilander, M.J., 2016. Daphnia magna transcriptome by RNA-Seq across 12 environmental stressors. Sci. Data 3, 160030. https://doi.org/10.1038/sdata.2016.30

Parsons, H.M., Ludwig, C., Günther, U.L., Viant, M.R., 2007. Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. BMC Bioinformatics 8, 234. https://doi.org/10.1186/1471-2105-8-234

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., Kingsford, C., 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14, 417–419. https://doi.org/10.1038/nmeth.4197

Piovesan, D., Giollo, M., Ferrari, C., Tosatto, S.C.E., 2015. Protein function prediction using guilty by association from interaction networks. Amino Acids 47, 2583–2592. https://doi.org/10.1007/s00726-015-2049-3

Posthuma, L., Altenburger, R., Backhaus, T., Kortenkamp, A., Müller, C., Focks, A., de Zwart, D., Brack, W., 2019. Improved component-based methods for mixture risk assessment are key to characterize complex chemical pollution in surface waters. Environ. Sci. Eur. 31, 70. https://doi.org/10.1186/s12302-019-0246-5

Ramirez, A.J., Brain, R.A., Usenko, S., Mottaleb, M.A., O'Donnell, J.G., Stahl, L.L., Wathen, J.B., Snyder, B.D., Pitt, J.L., Perez-Hurtado, P., Dobbins, L.L., Brooks, B.W., Chambliss, C.K., 2009. OCCURRENCE OF PHARMACEUTICALS AND PERSONAL CARE PRODUCTS IN FISH: RESULTS OF A NATIONAL PILOT STUDY IN THE UNITED STATES. Environ. Toxicol. Chem. 28, 2587. https://doi.org/10.1897/08-561.1

Rivetti, C., Campos, B., Faria, M., De Castro Català, N., Malik, A., Muñoz, I., Tauler, R., Soares, A.M.V.M., Osorio, V., Pérez, S., Gorga, M., Petrovic, M., Mastroianni, N., de Alda, M.L., Masiá, A., Campo, J., Picó, Y., Guasc, H., Barceló, D., Barata, C., 2015. Transcriptomic, biochemical and individual markers in transplanted Daphnia magna to characterize impacts in the field. Sci. Total Environ. 503–504, 200–212. https://doi.org/10.1016/j.scitotenv.2014.06.057

Sakurai, N., Ara, T., Ogata, Y., Sano, R., Ohno, T., Sugiyama, K., Hiruta, A., Yamazaki, K., Yano, K., Aoki, K., Aharoni, A., Hamada, K., Yokoyama, K., Kawamura, S., Otsuka, H., Tokimatsu, T., Kanehisa, M., Suzuki, H., Saito, K., Shibata, D., 2011. KaPPA-View4: a metabolic pathway database for representation and analysis of correlation networks of gene co-expression and metabolite co-accumulation and omics data. Nucleic Acids Res. 39, D677–D684. https://doi.org/10.1093/nar/gkq989

Schulze, T., Ahel, M., Ahlheim, J., Aït-Aïssa, S., Brion, F., Di Paolo, C., Froment, J., Hidasi, A.O., Hollender, J., Hollert, H., Hu, M., Kloß, A., Koprivica, S., Krauss, M., Muz, M., Oswald, P., Petre, M., Schollée, J.E., Seiler, T.-B., Shao, Y., Slobodnik, J., Sonavane, M., Suter, M.J.-F., Tollefsen, K.E., Tousova, Z., Walz, K.-H., Brack, W., 2017. Assessment of a novel device for onsite integrative large-volume solid phase extraction of water samples to enable a comprehensive chemical and effect-based analysis. Sci. Total Environ. 581–582, 350–358. https://doi.org/10.1016/j.scitotenv.2016.12.140

Seetharam, A., Stuart, G.W., 2013. A study on the distribution of 37 well conserved families of C2H2 zinc finger genes in eukaryotes 7.

Silva, C.P., Lima, D.L.D., Schneider, R.J., Otero, M., Esteves, V.I., 2014. Evaluation of the anthropogenic input of caffeine in surface waters of the north and center of Portugal by ELISA. Sci. Total Environ. 479–480, 227–232. https://doi.org/10.1016/j.scitotenv.2014.01.120

Singh, H., Sharma, R., 2012. Role of Adjacency Matrix & Adjacency List in Graph Theory. Int. J. Comput. Technol. 3, 179–183. https://doi.org/10.24297/ijct.v3i1c.2775

Sousa, J.C.G., Ribeiro, A.R., Barbosa, M.O., Pereira, M.F.R., Silva, A.M.T., 2018. A review on environmental monitoring of water organic pollutants identified by EU guidelines. J. Hazard. Mater. 344, 146–162. https://doi.org/10.1016/j.jhazmat.2017.09.058

Southam, A.D., Weber, R.J.M., Engel, J., Jones, M.R., Viant, M.R., 2017. A complete workflow for high-resolution spectral-stitching nanoelectrospray direct-infusion mass-spectrometry-based metabolomics and lipidomics. Nat. Protoc. 12, 310–328. https://doi.org/10.1038/nprot.2016.156

Stuart, J.M., 2003. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science 302, 249–255. https://doi.org/10.1126/science.1087447

Su, D., Ben, W., Strobel, B.W., Qiang, Z., 2020. Occurrence, source estimation and risk assessment of pharmaceuticals in the Chaobai River characterized by adjacent land use. Sci. Total Environ. 712, 134525. https://doi.org/10.1016/j.scitotenv.2019.134525

Summers, R.M., Mohanty, S.K., Gopishetty, S., Subramanian, M., 2015. Genetic characterization of caffeine degradation by bacteria and its potential applications. Microb. Biotechnol. 8, 369–378. https://doi.org/10.1111/1751-7915.12262

Sun, S., Chen, Y., Lin, Y., An, D., 2018. Occurrence, spatial distribution, and seasonal variation of emerging trace organic pollutants in source water for Shanghai, China. Sci. Total Environ. 639, 1–7. https://doi.org/10.1016/j.scitotenv.2018.05.089

Taylor, N.S., Weber, R.J.M., White, T.A., Viant, M.R., 2010. Discriminating between Different Acute Chemical Toxicities via Changes in the Daphnid Metabolome. Toxicol. Sci. 118, 307–317. https://doi.org/10.1093/toxsci/kfq247

Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K.-A., Grill, J., Frouin, V., 2014. Variable selection for generalized canonical correlation analysis. Biostatistics 15, 569–583. https://doi.org/10.1093/biostatistics/kxu001

Tenenhaus, A., Tenenhaus, M., 2011. Regularized Generalized Canonical Correlation Analysis. Psychometrika 76, 257–284. https://doi.org/10.1007/s11336-011-9206-8

Thorn, C.F., Aklillu, E., McDonagh, E.M., Klein, T.E., Altman, R.B., 2012. PharmGKB summary: caffeine pathway. Pharmacogenet. Genomics 22, 389–395. https://doi.org/10.1097/FPC.0b013e3283505d5e

Thorn, C.F., Leckband, S.G., Kelsoe, J., Steven Leeder, J., Müller, D.J., Klein, T.E., Altman, R.B., 2011. PharmGKB summary: carbamazepine pathway. Pharmacogenet. Genomics 21, 906–910. https://doi.org/10.1097/FPC.0b013e328348c6f2

Tilden, A.R., McCoole, M.D., Harmon, S.M., Baer, K.N., Christie, A.E., 2011. Genomic identification of a putative circadian system in the cladoceran crustacean Daphnia pulex. Comp. Biochem. Physiol. Part D Genomics Proteomics 6, 282–309. https://doi.org/10.1016/j.cbd.2011.06.002

Tolou-Ghamari, Z., Zare, M., Habibabadi, J.M., Najafi, M.R., 2013. A quick review of carbamazepine pharmacokinetics in epilepsy from 1953 to 2012. J. Res. Med. Sci. 5.

Townsend, D.M., Tew, K.D., 2003. The role of glutathione-S-transferase in anti-cancer drug resistance. Oncogene 22, 7369–7375. https://doi.org/10.1038/sj.onc.1206940

Xu, C., Li, C.Y.-T., Kong, A.-N.T., 2005. Induction of phase I, II and III drug metabolism/transport by xenobiotics. Arch. Pharm. Res. 28, 249–268. https://doi.org/10.1007/BF02977789

Yang, Y.-Y., Zhao, J.-L., Liu, Y.-S., Liu, W.-R., Zhang, Q.-Q., Yao, L., Hu, L.-X., Zhang, J.-N., Jiang, Y.-X., Ying, G.-G., 2018. Pharmaceuticals and personal care products (PPCPs) and artificial sweeteners (ASs) in surface and ground waters and their application as indication of wastewater contamination. Sci. Total Environ. 616–617, 816–823. https://doi.org/10.1016/j.scitotenv.2017.10.241

Yu, C.L., Kale, Y., Gopishetty, S., Louie, T.M., Subramanian, M., 2008. A Novel Caffeine Dehydrogenase in *Pseudomonas* sp. Strain CBB1 Oxidizes Caffeine to Trimethyluric Acid. J. Bacteriol. 190, 772–776. https://doi.org/10.1128/JB.01390-07

Zdobnov, E.M., Kuznetsov, D., Tegenfeldt, F., Manni, M., Berkeley, M., Kriventseva, E.V., 2021. OrthoDB in 2020: evolutionary and functional annotations of orthologs. Nucleic Acids Res. 49, D389–D393. https://doi.org/10.1093/nar/gkaa1009

Zhang, B., Horvath, S., 2005. A General Framework for Weighted Gene Co-Expression Network Analysis. Stat. Appl. Genet. Mol. Biol. 4. https://doi.org/10.2202/1544-6115.1128

3.9 Appendix 1

Pathway analysis of 18 transcriptomic modules in the Chaobai case study.

Statistical over-representation tests were performed on 137 *Drosophila melanogaster* pathways from KEGG database with permutation chi-squared test. The resulting P values were adjusted by FDR at 0.05. Only the pathways with adjusted P values lower than 0.05 are listed in this table. Modules are selected and labelled as (1) “Both” means module associated with both caffeine and carbamazepine concentrations in the mixture, (2) “CAF” means modules associated with caffeine only, and (3) “CBZ” means modules associated with carbamazepine only.

	Chemical association	Both	Both	Both	Both	Both	CBZ	CBZ	Both	CBZ	CBZ	CBZ	CBZ	CBZ	CAF	CBZ	CAF	CBZ	CBZ
Pathway	Description	CB_0	CB_1	CB_2	CB_3	CB_4	CB_5	CB_6	CB_7	CB_8	CB_9	CB_10	CB_11	CB_14	CB_15	CB_19	CB_20	CB_21	CB_24
dme04214	Apoptosis - fly										0.031								
dme04215	Apoptosis - multiple species										0.037								
dme04137	Mitophagy - animal																		
dme04140	Autophagy - animal												0.020			0.000			0.005
dme04142	Lysosome			0.000	0.038														0.004
dme04144	Endocytosis												0.031				0.011		0.007
dme04145	Phagosome	0.000							0.020										0.002
dme04146	Peroxisome			0.001	0.000														

dme02010	ABC transporters			0.010					0.000										
dme04013	MAPK signaling pathway - fly										0.048								
dme04068	FoxO signaling pathway		0.013										0.011						
dme04070	Phosphatidylinositol signaling system												0.006						
dme04150	mTOR signaling pathway												0.017						
dme04310	Wnt signaling pathway												0.015						
dme04330	Notch signaling pathway																		
dme04341	Hedgehog signaling pathway - fly																		0.001
dme04350	TGF-beta signaling pathway		0.033																0.001
dme04391	Hippo signaling pathway - fly							0.010					0.013						
dme04392	Hippo signaling pathway - multiple species		0.047					0.031											
dme04080	Neuroactive ligand-receptor interaction						0.000												
dme04512	ECM-receptor interaction			0.024								0.032							
dme03018	RNA degradation																		
dme03050	Proteasome	0.000								0.019									
dme03060	Protein export	0.000																	
dme04120	Ubiquitin mediated proteolysis									0.021			0.013						0.003
dme04130	SNARE interactions in vesicular transport																		

dme04141	Protein processing in dedoplasmic reticulum	0.000												0.009					
dme03030	DNA replication	0.007																	
dme03410	Base excision repair														0.000				
dme03440	Homologous recombination												0.002						
dme03460	Fanconi anemia pathway																		
dme03022	basal transcription factors																		
dme03040	Spliceosome	0.000	0.010									0.001		0.012					
dme00970	Aminoacyl-tRNA biosynthesis	0.004																	
dme03008	Ribosome biogenesis in eukaryotes	0.000										0.025	0.008						
dme03010	Ribosome	0.000	0.001							0.000			0.003						
dme03013	RNA transport	0.000								0.015									
dme00220	Arginine biosynthesis										0.017								
dme00250	Alanine, aspartate and glutamate metabolism						0.042				0.011								
dme00260	Glycine, serine and threonine metabolism												0.004						
dme00270	Cysteine and methionine metabolism	0.014							0.000				0.003						
dme00280	Valine, leucine and isoleucine degradation	0.000																	
dme00310	Lysine degradation	0.002																	
dme00330	Arginine and proline metabolism							0.047			0.011				0.001				

dme00340	Histidine metabolism							0.034	0.000											
dme00350	Tyrosine metabolism		0.007														0.001			
dme00360	Phenylalanine metabolism		0.041						0.008											
dme00380	Tryptophan metabolism	0.002							0.003											
dme00400	Phenylalanine, tyrosine and tryptophan biosynthesis	0.000																		
dme00232	Caffeine metabolism		0.000						0.000											
dme00010	Glycolysis / Gluconeogenesis	0.000							0.012								0.002			
dme00020	Citrate cycle (TCA cycle)	0.000			0.035															
dme00030	Pentose phosphate pathway	0.000							0.046		0.027									
dme00040	Pentose and glucuronate interconversions			0.041					0.000		0.047									
dme00051	Fructose and mannose metabolism	0.006							0.044											
dme00053	Ascorbate and aldarate metabolism			0.024					0.000		0.040									
dme00500	Starch and sucrose metabolism	0.014		0.034																
dme00520	Amino sugar and nucleotide sugar metabolism				0.000															
dme00562	Inositol phosphate metabolism										0.049									
dme00620	Pyruvate metabolism	0.000							0.008								0.002			
dme00630	Glyoxylate and dicarboxylate metabolism	0.000																		
dme00640	Propanoate metabolism	0.000		0.047		0.041					0.046									

dme00650	Butanoate metabolism										0.029									
dme00190	Oxidative phosphorylation	0.000	0.005							0.002										
dme00910	Nitrogen metabolism							0.045			0.014									
dme00920	Sulfur metabolism								0.022		0.015									
dme00510	N-Glycan biosynthesis								0.033		0.017									
dme00511	Other glycan degradation			0.000							0.030									
dme00512	Mucin type O-glycan biosynthesis			0.000	0.000	0.000		0.000				0.000								
dme00513	Various types of N-glycan biosynthesis	0.009													0.003					
dme00514	Other types of O-glycan biosynthesis							0.001												
dme00515	Mannose type O-glycan biosynthesis						0.000	0.000												
dme00531	Glycosaminoglycan degradation			0.003																
dme00532	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate							0.032												
dme00601	Glycosphingolipid biosynthesis - lacto and neolacto series		0.000	0.000	0.000	0.000		0.000				0.000								
dme00603	Glycosphingolipid biosynthesis - globo and isoglobo series			0.000								0.040								
dme00604	Glycosphingolipid biosynthesis - ganglio series			0.000																
dme00061	Fatty acid biosynthesis	0.032																		
dme00062	Fatty acid elongation				0.006				0.006											
dme00071	Fatty acid degradation	0.043		0.024													0.002			

dme00072	Synthesis and degradation of ketone bodies										0.014									
dme00100	Steroid biosynthesis		0.000	0.000	0.000	0.000														
dme00561	Glycerolipid metabolism				0.038				0.000											
dme00564	Glycerophospholipid metabolism				0.002															
dme00565	Ether lipid metabolism				0.000															
dme00590	Arachidonic acid metabolism								0.012											
dme00592	alpha-Linolenic acid metabolism			0.020	0.011															
dme00600	Sphingolipid metabolism			0.000																
dme01040	Biosynthesis of unsaturated fatty acids			0.000	0.001				0.006											
dme00130	Ubiquinone and other terpenoid-quinone biosynthesis	0.031								0.042										
dme00670	One carbon pool by folate							0.046												
dme00730	Thiamine metabolism								0.020		0.014									
dme00750	Vitamin B6 metabolism			0.022																
dme00760	Nicotinate and nicotinamide metabolism				0.033						0.022									
dme00770	Pantothenate and CoA biosynthesis							0.006	0.020											
dme00790	Folate biosynthesis								0.010		0.034									
dme00830	Retinol metabolism			0.000					0.001								0.001			
dme00860	Porphyrin and chlorophyll metabolism	0.000							0.003											

dme00410	beta-Alanine metabolism			0.020				0.025			0.031								
dme00430	Taurine and hypotaurine metabolism						0.003												
dme00440	Phosphonate and phosphinate metabolism								0.006										
dme00450	Selenocompound metabolism										0.014								
dme00471	D-Glutamine and D-glutamate metabolism		0.000																
dme00480	Glutathione metabolism								0.000										
dme00900	Terpenoid backbone biosynthesis	0.000			0.043														
dme00981	Insect hormone biosynthesis							0.010	0.030										
dme00230	Purine metabolism																		
dme00240	Pyrimidine metabolism	0.007																	
dme00980	Metabolism of xenobiotics by cytochrome P450								0.000								0.001		
dme00982	Drug metabolism - cytochrome P450								0.000								0.001		
dme00983	Drug metabolism - other enzymes								0.005										
dme04213	Longevity regulating pathway - multiple species												0.008						
dme04320	Dorso-ventral axis formation										0.047								
dme04711	Circadian rhythm - fly								0.030										
dme04624	Toll and Imd signaling pathway													0.009				0.000	
dme04745	Phototransduction - fly												0.001				0.000		

3.10 Appendix 2

Pathway analysis of 16 transcriptomic modules in the Danube case study in Chapter 5.

Statistical over-representation tests were performed on 137 *Drosophila melanogaster* pathways from KEGG database with permutation chi-squared test. The resulting P values were adjusted by FDR at 0.05. Only the pathways with adjusted P values lower than 0.05 were listed in this table. Modules are selected and labelled as (1) “Both” means module associated with both caffeine and carbamazepine concentrations in the mixture, (2) “CAF” means modules associated with caffeine only, and (3) “CBZ” means modules associated with carbamazepine only.

	Chemical association	Both	Both	Both	CBZ	Both	CBZ	CBZ	CAF	CAF	CBZ	CAF	CAF	CAF	CBZ	CBZ	CBZ
Pathway	Description	DA_0	DA_1	DA_2	DA_4	DA_5	DA_7	DA_8	DA_10	DA_11	DA_12	DA_13	DA_14	DA_17	DA_18	DA_22	DA_30
dme00980	Metabolism of xenobiotics by cytochrome P450					0.004			0.046		0.002						0.001
dme00982	Drug metabolism - cytochrome P450					0.005			0.048		0.002	0.021					0.001
dme00983	Drug metabolism - other enzymes								0.032		0.030						
dme00220	Arginine biosynthesis				0.016												
dme00250	Alanine, aspartate and glutamate metabolism							0.048									
dme00260	Glycine, serine and threonine metabolism									0.030		0.026					
dme00270	Cysteine and methionine metabolism												0.012				
dme00280	Valine, leucine and isoleucine degradation				0.022		0.029										

dme00290	Valine, leucine and isoleucine biosynthesis						0.000										
dme00310	Lysine degradation								0.040			0.018				0.007	
dme00330	Arginine and proline metabolism				0.023						0.009					0.006	
dme00340	Histidine metabolism				0.049	0.016											0.000
dme00350	Tyrosine metabolism						0.043					0.011					0.000
dme00360	Phenylalanine metabolism				0.031	0.011											0.000
dme00380	Tryptophan metabolism								0.038							0.006	0.001
dme00400	Phenylalanine, tyrosine and tryptophan biosynthesis				0.000												
dme00232	Caffeine metabolism									0.000							
dme00010	Glycolysis / Gluconeogenesis						0.032					0.007	0.019			0.012	0.002
dme00020	Citrate cycle (TCA cycle)				0.001		0.017			0.021						0.007	
dme00030	Pentose phosphate pathway		0.000									0.009					
dme00040	Pentose and glucuronate interconversions					0.020										0.006	
dme00051	Fructose and mannose metabolism										0.005	0.009					
dme00052	Galactose metabolism					0.016	0.049					0.002					
dme00053	Ascorbate and aldarate metabolism		0.012			0.016										0.006	
dme00500	Starch and sucrose metabolism					0.016	0.049					0.002					
dme00520	Amino sugar and nucleotide sugar metabolism										0.014	0.000			0.006		

dme00562	Inositol phosphate metabolism																
dme00620	Pyruvate metabolism						0.024			0.029			0.016			0.009	0.001
dme00630	Glyoxylate and dicarboxylate metabolism								0.034								
dme00640	Propanoate metabolism								0.038				0.010				
dme00650	Butanoate metabolism					0.048			0.023								
dme00190	Oxidative phosphorylation																
dme00910	Nitrogen metabolism						0.017			0.006							
dme00920	Sulfur metabolism																
dme00510	N-Glycan biosynthesis							0.029									
dme00511	Other glycan degradation					0.049											
dme00512	Mucin type O-glycan biosynthesis	0.000		0.000													
dme00513	Various types of N-glycan biosynthesis							0.036			0.021						
dme00514	Other types of O-glycan biosynthesis																
dme00515	Mannose type O-glycan biosynthesis																
dme00531	Glycosaminoglycan degradation		0.019			0.029											
dme00534	Glycosaminoglycan biosynthesis - heparan sulfate / heparin																
dme00563	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis																
dme00601	Glycosphingolipid biosynthesis - lacto and neolacto series	0.000	0.000	0.000	0.000		0.000	0.000			0.000						

dme00603	Glycosphingolipid biosynthesis - globo and isoglobo series										0.004						
dme00604	Glycosphingolipid biosynthesis - ganglio series		0.000			0.000											
dme00061	Fatty acid biosynthesis						0.001										
dme00062	Fatty acid elongation		0.038			0.045			0.024								
dme00071	Fatty acid degradation								0.048								0.001
dme00100	Steroid biosynthesis	0.000	0.000	0.000		0.000		0.000									
dme00561	Glycerolipid metabolism						0.017									0.007	
dme00564	Glycerophospholipid metabolism						0.012										
dme00565	Ether lipid metabolism						0.034										
dme00590	Arachidonic acid metabolism					0.016	0.049				0.007						
dme00592	alpha-Linolenic acid metabolism						0.019										
dme00600	Sphingolipid metabolism					0.004								0.007			
dme01040	Biosynthesis of unsaturated fatty acids		0.038			0.048											
dme00730	Thiamine metabolism					0.022											
dme00750	Vitamin B6 metabolism												0.001				
dme00760	Nicotinate and nicotinamide metabolism		0.032				0.029									0.003	
dme00770	Pantothenate and CoA biosynthesis																
dme00780	Biotin metabolism				0.016												

dme00790	Folate biosynthesis									0.016							
dme00830	Retinol metabolism					0.002											0.001
dme00860	Porphyrin and chlorophyll metabolism								0.038	0.020							
dme00410	beta-Alanine metabolism				0.046				0.025								
dme00430	Taurine and hypotaurine metabolism					0.016											
dme00471	D-Glutamine and D-glutamate metabolism				0.000												
dme00480	Glutathione metabolism					0.002			0.020		0.003				0.007		
dme00900	Terpenoid backbone biosynthesis		0.044	0.042									0.000				
dme00981	Insect hormone biosynthesis								0.015				0.004				
dme00230	Purine metabolism											0.018				0.037	
dme00240	Pyrimidine metabolism															0.008	
dme04213	Longevity regulating pathway - multiple species												0.023	0.015		0.014	
dme04320	Dorso-ventral axis formation								0.040								
dme04711	Circadian rhythm - fly				0.017					0.008			0.004				
dme04624	Toll and Imd signaling pathway														0.002		
dme04745	Phototransduction - fly									0.040						0.013	0.000
dme04214	Apoptosis - fly										0.021	0.041					
dme04215	Apoptosis - multiple species																

dme04136	Autophagy - other							0.048	0.035								
dme04137	Mitophagy - animal																
dme04140	Autophagy - animal															0.039	
dme04142	Lysosome		0.032			0.001										0.036	
dme04144	Endocytosis															0.008	0.009
dme04145	Phagosome												0.027				
dme04146	Peroxisome													0.027			
dme02010	ABC transporters																
dme04013	MAPK signaling pathway - fly											0.004		0.028			
dme04068	FoxO signaling pathway				0.023								0.006	0.022		0.021	
dme04070	Phosphatidylinositol signaling system															0.012	
dme04150	mTOR signaling pathway												0.009			0.033	
dme04310	Wnt signaling pathway				0.046						0.038		0.044			0.027	
dme04330	Notch signaling pathway									0.016							
dme04341	Hedgehog signaling pathway - fly				0.023								0.009				
dme04350	TGF-beta signaling pathway				0.011												
dme04391	Hippo signaling pathway - fly									0.017	0.034		0.038			0.023	
dme04392	Hippo signaling pathway - multiple species						0.045			0.016			0.008				

dme04080	Neuroactive ligand-receptor interaction														0.009		
dme04512	ECM-receptor interaction					0.028				0.000							
dme03018	RNA degradation																
dme03050	Proteasome								0.039								
dme03060	Protein export																
dme04120	Ubiquitin mediated proteolysis								0.014								
dme04130	SNARE interactions in vesicular transport									0.011			0.005				
dme04141	Protein processing in dedoplasmic reticulum																
dme03030	DNA replication													0.007			
dme03410	Base excision repair																
dme03420	Nucleotide excision repair								0.047								
dme03440	Homologous recombination																
dme03450	Non-homologous end-joining											0.002					
dme03020	RNA polymerase							0.048	0.036								
dme03022	basal transcription factors								0.010								
dme03040	Spliceosome															0.030	
dme00970	Aminoacyl-tRNA biosynthesis												0.013			0.008	
dme03008	Ribosome biogenesis in eukaryotes											0.042					

dme03010	Ribosome																
dme03013	RNA transport													0.038		0.035	
dme03015	mRNA surveillance pathway				0.033					0.040			0.022			0.014	

3.11 Appendix 3

Pathway analysis of 19 metabolomic modules in the Danube case study in Chapter 5.

Statistical over-representation tests were performed on 137 *Drosophila melanogaster* pathways from KEGG database with permutation chi-squared test. The resulting P values were adjusted by FDR at 0.05. Only the pathways with adjusted P values lower than 0.05 are listed in this table. Modules are selected and labelled as (1) “Both” means module associated with both caffeine and carbamazepine concentrations in the mixture, (2) “CAF” means modules associated with caffeine only, and (3) “CBZ” means modules associated with carbamazepine only.

	Chemical association	Both	CAF	Both	Both	CBZ	CBZ	Both	Both	Both	CBZ	CAF	Both	Both	CBZ	CAF	CAF	Both	CBZ	CBZ
Pathway	Description	DAp_0	DAp_2	DAp_4	DAp_10	DAp_16	DAp_38	DAn_0	DAn_1	DAn_2	DAn_3	DAn_5	DAn_6	DAn_8	DAn_12	DAn_13	DAn_18	DAn_27	DAn_35	DAn_66
map00525	Acarbose and validamycin biosynthesis					0.023														
map02010	ABC transporters																		0.004	
map02060	Phosphotransferase system (PTS)							0.001					0.000		0.000					
map04022	cGMP-PKG signaling pathway														0.045					
map04070	Phosphatidylinositol signaling system														0.030					
map04152	AMPK signaling pathway														0.001					
map05204	Chemical carcinogenesis										0.023									
map01523	Antifolate resistance				0.000															

map04931	Insulin resistance												0.020							
map00400	Phenylalanine, tyrosine and tryptophan biosynthesis														0.045					
map00232	Caffeine metabolism		0.000																	
map00254	Aflatoxin biosynthesis													0.000						
map00261	Monobactam biosynthesis			0.001					0.012											
map00521	Streptomycin biosynthesis														0.019					
map00901	Indole alkaloid biosynthesis																0.002			
map00941	Flavonoid biosynthesis									0.000				0.000		0.000	0.002			
map00943	Isoflavonoid biosynthesis				0.016			0.002		0.000				0.000		0.001				
map00950	Isoquinoline alkaloid biosynthesis										0.002							0.000		
map00966	Glucosinolate biosynthesis																0.000			
map00010	Glycolysis / Gluconeogenesis														0.000					
map00030	Pentose phosphate pathway														0.000					
map00040	Pentose and glucuronate interconversions						0.000								0.000					
map00051	Fructose and mannose metabolism														0.000					
map00052	Galactose metabolism														0.000					
map00053	Ascorbate and aldarate metabolism														0.000					
map00500	Starch and sucrose metabolism							0.000							0.000					

map00520	Amino sugar and nucleotide sugar metabolism												0.000		0.000				
map00562	Inositol phosphate metabolism														0.000				
map01060	Biosynthesis of plant secondary metabolites														0.005				
map01061	Biosynthesis of phenylpropanoids													0.000					
map01065	Biosynthesis of alkaloids derived from histidine and purine														0.000				
map01070	Biosynthesis of plant hormones														0.006				
map00680	Methane metabolism														0.000				
map00710	Carbon fixation in photosynthetic organisms														0.000				
map00541	O-Antigen nucleotide sugar biosynthesis New!														0.002				
map00121	Secondary bile acid biosynthesis																		0.000
map00670	One carbon pool by folate				0.000														
map00740	Riboflavin metabolism				0.036														
map00750	Vitamin B6 metabolism														0.030				
map00253	Tetracycline biosynthesis											0.020							
map01057	Biosynthesis of type II polyketide products				0.000									0.000					
map00230	Purine metabolism														0.000			0.002	
map00240	Pyrimidine metabolism														0.005				

map00624	Polycyclic aromatic hydrocarbon degradation									0.046										
map00980	Metabolism of xenobiotics by cytochrome P450										0.046	0.005								
map00982	Drug metabolism - cytochrome P450	0.034																		
map04918	Thyroid hormone synthesis														0.014					
map04922	Glucagon signaling pathway														0.002					
map04740	Olfactory transduction														0.030					

3.12 Appendix 4

Pathway analysis of shared genes between chemical-associated transcriptomic modules of Chaobai case and Danube case.

Statistical over-representation tests were performed on 137 *Drosophila melanogaster* pathways from KEGG database with permutation chi-squared test. The resulting P values were adjusted by FDR at 0.05. Only the pathways with adjusted P values lower than 0.05 are listed in this table. Modules are selected and labelled as (1) “Both” means module associated with both caffeine and carbamazepine concentrations in the mixture, and (2) “CBZ” means modules associated with carbamazepine only.

	Chemical association	Both	Both	Both	Both	Both	Both	Both	Both	Both	Both	Both	Both	Both	Both	CBZ	CBZ	CBZ	CBZ	CBZ
Pathway	Description	CB_0-DA_0	CB_0-DA_1	CB_0-DA_2	CB_1-DA_0	CB_1-DA_1	CB_2-DA_0	CB_2-DA_5	CB_3-DA_0	CB_3-DA_1	CB_3-DA_2	CB_4-DA_1	CB_4-DA_2	CB_7-DA_0	CB_7-DA_5	CB_5-DA_22	CB_6-DA_22	CB_6-DA_7	CB_9-DA_4	CB_10-DA_4
dme00980	Metabolism of xenobiotics by cytochrome P450							0.0162							0.0056			0.0013		
dme00982	Drug metabolism - cytochrome P450							0.0162							0.0059			0.0013		
dme00983	Drug metabolism - other enzymes																	0.0041		
dme00260	Glycine, serine and threonine metabolism											0.0017								
dme00270	Cysteine and methionine metabolism	0.0142	0.0043																	
dme00280	Valine, leucine and isoleucine degradation										0.0206								0.0000	

dme00330	Arginine and proline metabolism										0.0121						0.0000			
dme00340	Histidine metabolism														0.0012					
dme00350	Tyrosine metabolism														0.0036					
dme00360	Phenylalanine metabolism														0.0006					
dme00380	Tryptophan metabolism														0.0044					
dme00010	Glycolysis / Gluconeogenesis				0.0100															
dme00020	Citrate cycle (TCA cycle)										0.0122									
dme00030	Pentose phosphate pathway		0.0000											0.0001						
dme00040	Pentose and glucuronate interconversions							0.0124		0.0310										
dme00052	Galactose metabolism				0.0042			0.0102												
dme00053	Ascorbate and aldarate metabolism							0.0097		0.0235										
dme00500	Starch and sucrose metabolism		0.0030		0.0042			0.0103												
dme00520	Amino sugar and nucleotide sugar metabolism			0.0002							0.0176									
dme00562	Inositol phosphate metabolism				0.0052														0.0000	
dme00640	Propanoate metabolism						0.0007												0.0000	
dme00650	Butanoate metabolism										0.0070									
dme00190	Oxidative phosphorylation	0.0000																		

dme00910	Nitrogen metabolism		0.0009								0.0034		0.0450					0.0003		
dme00920	Sulfur metabolism				0.0017															
dme00511	Other glycan degradation							0.0344												
dme00512	Mucin type O-glycan biosynthesis										0.0000		0.0000							
dme00514	Other types of O-glycan biosynthesis										0.0135									
dme00531	Glycosaminoglycan degradation				0.0021			0.0210		0.0397										
dme00601	Glycosphingolipid biosynthesis - lacto and neolacto series									0.0000			0.0000							0.0000
dme00603	Glycosphingolipid biosynthesis - globo and isoglobo series																			0.0000
dme00604	Glycosphingolipid biosynthesis - ganglio series							0.0000												
dme00062	Fatty acid elongation		0.0018							0.0131					0.0026					
dme00071	Fatty acid degradation						0.0008													
dme00072	Synthesis and degradation of ketone bodies										0.0034									
dme00100	Steroid biosynthesis							0.0000		0.0000			0.0000							
dme00561	Glycerolipid metabolism													0.0002						
dme00564	Glycerophospholipid metabolism													0.0005						
dme00590	Arachidonic acid metabolism							0.0469							0.0040					

dme00592	alpha-Linolenic acid metabolism						0.0002						0.0450							
dme00600	Sphingolipid metabolism							0.0021												
dme01040	Biosynthesis of unsaturated fatty acids		0.0020				0.0005			0.0149					0.0028					
dme00670	One carbon pool by folate								0.0000											
dme00730	Thiamine metabolism							0.0162												
dme00760	Nicotinate and nicotinamide metabolism								0.0000									0.0004		
dme00770	Pantothenate and CoA biosynthesis									0.0313										
dme00790	Folate biosynthesis				0.0035			0.0408	0.0000											
dme00830	Retinol metabolism							0.0011												
dme00410	beta-Alanine metabolism						0.0004												0.0000	
dme00430	Taurine and hypotaurine metabolism							0.0114												
dme00450	Selenocompound metabolism		0.0009																	
dme00480	Glutathione metabolism		0.0080					0.0089						0.0004	0.0014			0.0022		
dme00900	Terpenoid backbone biosynthesis		0.0000								0.0003									
dme00230	Purine metabolism								0.0003									0.0068		
dme00240	Pyrimidine metabolism								0.0001									0.0013		
dme04213	Longevity regulating pathway - multiple species	0.0271				0.0000					0.0270					0.0000				

dme04624	Toll and Imd signaling pathway				0.0208										0.0191					
dme04745	Phototransduction - fly															0.0001				
dme04215	Apoptosis - multiple species				0.0038															
dme04136	Autophagy - other				0.0042															
dme04140	Autophagy - animal				0.0325	0.0000										0.0004				
dme04142	Lysosome		0.0215	0.0007	0.0310			0.0033		0.0331					0.0287					
dme04144	Endocytosis						0.0062							0.0021		0.0003				
dme04145	Phagosome						0.0019				0.0319									
dme04146	Peroxisome						0.0028				0.0462	0.0045								
dme04013	MAPK signaling pathway - fly						0.0030													
dme04068	FoxO signaling pathway															0.0001				
dme04070	Phosphatidylinositol signaling system															0.0001				
dme04150	mTOR signaling pathway				0.0282	0.0000										0.0002				
dme04310	Wnt signaling pathway	0.0107					0.0031									0.0001				
dme04330	Notch signaling pathway	0.0010																		
dme04350	TGF-beta signaling pathway					0.0000														
dme04391	Hippo signaling pathway - fly															0.0002		0.0044		
dme04392	Hippo signaling pathway - multiple species																	0.0008		

dme04080	Neuroactive ligand-receptor interaction							0.0344							0.0107					
dme04512	ECM-receptor interaction							0.0200												
dme03050	Proteasome				0.0053															
dme04120	Ubiquitin mediated proteolysis				0.0029															
dme03030	DNA replication	0.0021																		
dme03420	Nucleotide excision repair	0.0143																		
dme03430	Mismatch repair	0.0063																		
dme03440	Homologous recombination	0.0098																		
dme03460	Fanconi anemia pathway	0.0108																		
dme03020	RNA polymerase				0.0044															
dme03040	Spliceosome	0.0118																		
dme00970	Aminoacyl-tRNA biosynthesis													0.0002						
dme03010	Ribosome	0.0038																		
dme03013	RNA transport	0.0124																		
dme03015	mRNA surveillance pathway	0.0258																		

4 Chaobai case study

4.1 Abstract

Environmental chemical pollution severely threatens the health of humans and the environment. Yet, the problem is so complex that no solutions are available for protection without research that better understands the toxicity of chemicals as real-world mixtures. The first step is to identify samples from the environment that may be hazardous, by assessing the toxicity of the sampled environment as a whole, while detecting its chemical constituents. Here the relative toxicity of waters sampled from the Chaobai River (China) are assessed based on gene expression of the model test species *Daphnia magna*. Two-steps hierarchical clustering was applied to cluster the transcriptomic data into multiple co-responsive gene clusters then group sampled waters within gene clusters. To characterise and classify the biological effect of exposure to the chemical mixture at environmental levels, the functional roles of gene clusters were determined by an ortholog-based pathway overrepresentation analysis. Results show that expression-based clustering analysis of five gene clusters revealed that the environmental chemical mixture of a single site (M16) induced relatively higher expression levels in stress response and cellular homeostasis, and these differences are significantly related to Dibenz[a,h]anthracene, Erythromycin and Trimethoprim. These results demonstrated the feasibility of classifying the biological effect of exposure to environmental chemical mixtures based on gene expression at environmental relevant levels.

4.2 Introduction

Environmental chemical pollution is a global and persistent problem that threatens the health of living organisms and is a primary cause of biodiversity declines in natural ecosystems (Amoatey and Baawain, 2019; Landrigan et al., 2018; Vermeulen et al., 2020). Chemical safety legislation is designed to minimise the adverse impacts of substances on humans and the environment by regulating those specific chemicals that pose a threat to life (European Chemical Agency (ECHA) and European Food Safety Authority (EFSA) with the technical support of the Joint Research Centre (JRC) et al., 2018; Fantke et al., 2020). For example, the surface water quality criteria in China (GB3838-2002) focus on monitoring pre-selected chemical substances in the environment (Su et al., 2017). Yet pollution control requires the development of a holistic effect assessment of chemical mixture in the environment, especially for those rivers receiving multiple sources of pollutants. One of the river, the Chaobai River (Beijing, China), receives both treated and untreated reclaimed water generated from the wastewater treatment plants, industrial outlets, and agricultural runoffs (He et al., 2018). These pollutants, just to name a few, include large amounts of nutrients, metals, pharmaceuticals, polycyclic aromatic hydrocarbons (PAHs), which enter the river daily. The reclaimed water that originates from the treated wastewater contains a relatively high amount of nitrogen, phosphorus, salts, metals, un-removed organic compounds and pathogens compared to freshwater, which may lead to excessive loads of nutrient or eutrophication (Yu et al., 2020), increases in soil salinity (Chen et al., 2013b), and potential contamination in groundwater (Chen et al., 2013a). Varying in time, the effluents of the wastewater treatment plant (WWTP) may also contain a variety of pharmaceuticals and a high amount of antibiotic (Iwane et al., 2001) that are not

completely removed by the sewage treatment process (Eggen et al., 2014; Falås et al., 2016), which may pose a potentially harmful effect on aquatic species. In the populated area, the incomplete combustion of biomass and fossil fuel also contribute to the emission of PAHs (Qian et al., 2017), which may induce genotoxicity in non-targeted aquatic species (Yu, 2002). Multiple-source pollutants from domestic, agricultural, and industrial activities could also pose threats to ecosystem stability, as the chemical pollutants may determine the structure of zooplankton communities (Xiong et al., 2017). For these reasons and concerns for human health and the environment in this region, the Chaobai River system is an ideal model to study the potential biological effects of river waters with multiple-source pollutants.

To evaluate the joint effect of pollutants in the natural river, omics-based bioassays that interrogate the global effects of chemicals on biomolecular pathways linked to health can be applied to capture the biological signatures of chemical toxicity at the molecular level. For example, gene expression profiles measured by the transcriptome are able to uncover adversity-related functional pathways under the joint chemical mixture exposure conditions that result in observed adversity (Watanabe et al., 2008). Transcriptome profiling has successfully characterised the biological responses to chemicals and environmental mixtures in zebrafish embryos (Wang et al., 2018), Atlantic eels (Baillon et al., 2015), oysters (Lüchmann et al., 2015), and waterfleas (Orsini et al., 2016). By interpreting the transcriptome, enriched biomolecular pathways, including metabolic pathways, which are responsive to chemical exposure, can be further characterised. For example, environmental chemicals may induce alternation in inter-correlated biological processes, such as xenobiotic metabolism and stress response. Xenobiotic metabolism is responsible for detoxification and

biotransformation of exogenous substances, which is represented by biomarkers that are diagnostic of these pathways, such as cytochrome P450 (CYP), ATP-binding cassette transporter (ABC), glutathione S-transferase (GST), and glutathione peroxidase (GPX) (Hassan et al., 2015)(Hassan *et al.* 2015). In transcriptomes, pronounced expression in these biomarkers might be interpreted as activation of xenobiotic defence (Liu et al., 2017). Exposure to environmental chemicals might also trigger oxidative stress responses; enhanced expression of glutathione reflects activated antioxidant defence (Regoli and Giuliani, 2014). Bioactivity of glutathione, GST and glutathione reductase play important roles in neutralising reactive oxygen species and avoiding further damage caused by exogenous compounds (Oliveira et al., 2015), which are considered as the biomarkers of oxidative stress response in *Daphnia magna* (Barata et al., 2005). These are but a few examples of detecting toxicity based on the observed changes in the expression of a defined set of genes, whose functions are sufficiently well understood to permit an assessment of exposure-related adversity to substances in the environment. But transcriptomes are rich in data, representing global changes in gene expression that can be harnessed for a more systemic and pathway-based understanding of molecular toxicology.

To classify the biological effects of environmental chemicals based on toxicity pathways, unsupervised learning methods like clustering may be used to identify groups of co-variant genes (gene-based clustering) or groups of homogeneous samples (sample-based clustering). Genes that share similar expression patterns are often assumed to be under the control of shared regulatory pathways, and therefore functionally related and biologically relevant (Gasch et al., 2000). Hierarchical clustering algorithms like DIANA and the Hierarchical Ordered Partitioning and

Collapsing Hybrid (HOPACH, Pollard, 2005) methods generate a hierarchical clustering tree, which identifies gene clusters within the transcriptome. The reasons that the HOPACH algorithm might outcompete the DIANA algorithm is that the DIANA requires manual selection of a height value for cutting the tree to determine the number of clusters in a hierarchical clustering tree, which can be problematic (Slonim, 2002); while the HOPACH automatically finds the optimal number of gene clusters from their expression patterns at each level of the clustering tree based on the Median Split Silhouette criterium, resulting in a robust clustering pattern. Further sample-based clustering analysis may characterise the grouping patterns of each gene cluster so that the structure of the gene expression data can be revealed in greater details. As each gene cluster may consist of genes of similar function, the grouping patterns of the gene clusters may assist in distinguishing the general differences in the overall transcriptomic profiles with respect to their biological roles, so that the functions potentially perturbed by the environmental chemicals may be revealed at a systematic level. Biological interpretation of the gene clusters may require comprehensive information of gene function and pathway. A cross-species extrapolation can be employed to annotate genes of the poorly defined species by referred to well-studied species based on their orthologs. As the functions of unknown genes can be putatively annotated by the corresponding OGs' function, the gene-pathway association can be thereby transformed into an OGs-pathway association. If the OGs composition of every pathway is unique, the OGs-pathway association can be used to (1) distinguish different pathways and (2) applied as the reference data, similar to gene sets serving as background knowledge in the pathway overrepresentation analysis.

The Chaobai River is selected as the natural river system, I combined targeted chemical analysis of river water with non-targeted transcriptomics measurements of exposure-related effects on *Daphnia magna* using sampled surface river samples from 30 sites along the Chaobai River Basin. The biological effects of these natural surface waters were assessed without pre-concentration or extraction of chemicals. The specific objectives of this study are to (1) identify gene clusters within the transcriptome, and (2) functionally annotated the gene clusters via pathway overrepresentation analysis, and finally (3) characterise the joint effect of the environmental chemical mixture through the grouping pattern of gene clusters.

4.3 Methods

4.3.1 Site description and sampling regime

The Chaobai river is the 2nd largest river in Beijing, having a total length of 458 kilometres (km). It is a vital drinking water resource in the Hebei-Beijing-Tianjin region that covers 13,846 km^2 (Wang et al., 2009) and sustains a human population of 100.8 million (Sun et al., 2019). The Chaobai River starts at the confluence of two headwaters called the Bai River and the Chao River, which both originate from the mountainous area in Hebei province, converge at Miyun of Beijing and join as the Chaobai river. The Chaobai River flows through the populated urban area in Beijing and the agricultural region in Tianjin, then flows into the Pacific Ocean at Bohai gulf. The whole river basin has several water sources. One of the water sources, the reclaimed water from Wenyu River, contributes to 38 billion m^3 of water into the Chaobai River per year (Huang *et al*., 2010), assisting in the government-initiated restoration of the eco-environment and at refilling the groundwater. Another source is the effluents of wastewater treatment

plants along the river that accounts for 9.2 billion m^3 of water into the Chaobai River annually (Su et al., 2020). The river also receives industrial outlets in the urban area and agricultural runoffs in the farmland. Several reaches of the river have severe eutrophication problem (Figure S4.1), which is another long-lasting problem for citizens and the river ecology. The flow velocity is nearly zero in the mainstream of the Chaobai River (He *et al*. 2017). There are over 20 dams across the whole basin to regulate the flows in the upstream and distribute surface waters to the irrigation-intensive area. Along the whole river basin, thirty sites were selected for field sampling, including seven sites in the Bai River, six sites in the Chao River and seventeen sites in the Chaobai River (Figure S4.2). The GPS locations of sampling sites are summarised in Table S3.1.

All surface water samples were collected in the middle of the river (if there is a bridge) or 1 – 2 meter offshore and stored in Duran amber glass bottles in September 2017. Water samples for the needs of chemical analysis were collected simultaneously: 500 ml for salts and heavy metals measurement, 4L for PAHs detection, and 3L for organic micropollutants extraction and measurement (especially pharmaceuticals). The pH and total dissolved solids (TDS) were measured on-site, while the total organic carbon (TOC) was measured in the laboratory within two weeks. All water samples were transported at room temperature (19 - 23 °C), filtered with 0.7 μm glass fibre membrane filters (GF/F, Whatman, U.S.A) and stored at four °C before chemical analysis. And 1 L of the water samples were collected for the exposure experiment.

4.3.2 Chemical analysis

For generating the chemical profiles of the chemical mixtures in the water samples collected from the Chaobai River, chemical analysis was performed by targeted

analysis. The targeted analysis consisted of quantitative measurements for (1) salts and metals that are included in the regulatory monitoring under the Environmental quality standard of surface water (GB 3838-2002); (2) 16 polycyclic aromatic hydrocarbons (PAHs) that are priority compounds for regulation by the US Environmental Protection Agency (Keith L.H. 2015) but not included in the current surface water quality standards in China (GB 3838-2002); and (3) 22 organic micropollutants, including caffeine and 21 widely-occurred pharmaceuticals in China (Su *et al.* 2020), that are also not included in the current standards (GB 3838-2002). These three groups of chemicals can reveal the basic status of nutrient and salts in the river waters and potential pollutants from the urban runoffs or WWTP outlets. The methods applied for targeted analysis of these three groups of chemical substances are listed below.

Salts and metals

Water samples were filtered with 0.7 μm glass fibre membrane filters (GF/F, Whatman, U.S.A) and stored at 4 ℃ before measurement. The total nitrogen (TN), ammonia (NH4), nitrate (NO3), and total phosphorus were determined using the methods described in Liao *et al.* (2019). The rest were determined by methods described in Xiong *et al*. (2017). The detailed information of salts and metals were listed in Table S4.2.

PAHs

For measurements of prioritised polycyclic aromatic hydrocarbons (PAHs), filtered water samples were extracted with C18 cartridges (500mg, 6ml, Supelco) and HLB cartridges (500mg, 6ml, Waters), then eluted with the organic solvent. The internal standards (100 ng/L) were injected into each sample before instrumental analysis.

Targeted analysis of 16 prioritised PAHs was conducted on the Agilent 7890A gas chromatography (GC) equipped with a 5795C mass spectrometry (MS) detector with electrospray ionisation (EI) sources in the selective ion monitoring mode. The instrument setting followed the description in Qiao *et al.* (2014, 2020). The limits of quantifications for 16 PAHs are listed in Table S4.3.

Organic micropollutants

As for organic micropollutants, water samples were filtered with 0.7 μm glass fibre membrane filters (GF/F, Whatman, U.S.A) and spiked with 100 ng internal standards (Sulfamethazine-$^{13}C_6$, Ofloxacin-D_3, Caffeine-$^{13}C_3$). The organic substances were extracted by SPE with Oasis HLB cartridges (500mg, 6ml, Waters, U.S.A.), eluted with methanol, dried under nitrogen at room temperature, and dissolved in 40% methanol solvent (v:v). Targeted analysis of 22 organic micropollutants was conducted on the Agilent 1290 ultra-performance liquid chromatography (UPLC) system equipped with the Agilent 6420 Triple Quad mass spectrometer (MS). The extraction method and instrumental settings were described in Ben *et al.* (2018) and Su *et al.* (2020), respectively. Limits of quantification (LOQ) for 22 organic micropollutants were listed in Table S4.4.

4.3.3 Sample preparation and bioassay

A total of 30 water samples were collected from the Chaobai River, filtered with 0.7 μm glass fibre membrane (GF/F, Whatman, USA) and stored at 20 ℃ in the dark in preparation of the bioassay with *Daphnia* neonates. Each water sample had three biological replicates except for B01 with eight replicates, resulting in a total of 95 samples in the Chaobai case study.

For the bioassay, water samples from case one and treatment media from case two were transferred to glass vials before the bioassay. Exposure tests were conducted using *Daphnia magna* (Bham2 strain) neonates according to OECD guideline 202 (OECD 202). Neonates hatched within 24 hours were exposed to filtered Chaobai River waters or spike-in borehole waters (spike with organic extracts from Danube River) for 48 hours without feeding. After the 48 hours of exposure, the number of immobilised neonates were recorded, and all the exposed neonates were collected, flash-frozen within liquid nitrogen and stored at -80 ℃ before RNA extraction.

4.3.4 Total RNA extraction and transcriptome sequencing

For the Chaobai River case study, the frozen pooled neonates for each biological replicate contained 5 exposed neonates. Frozen pooled neonates were homogenised in GenoGrinder (SPEX SamplePrep, U.S.A.) for 45 seconds at the speed of 1750 rpm. Total RNA extraction was performed using the Agencourt RNAdvance Tissue Total RNA kit (Beckman Coulter, U.S.A.), as the total RNA was captured onto magnetic beads, washed twice for removing unwanted salts, and eluted in 100 µl RNAse-free H_2O, following the manufacturer's instructions. The concentration of total RNA concentrations was quantified by Nanodrop 8000 Spectrophotometer (Labtech Ltd., U.K.). The quality of extracted total RNA, both integrity and purity, was measured on TapeStation 2200 (Agilent Technologies, U.S.A.). A cDNA library was generated for each sample from 150 ng of RNA using NEBNext Ultra II Directional RNA Library Prep Kit for Illumina, following the manufacturer's instructions. All of the sample libraries were then normalised to the same molecular weight and pooled together using the adapter indices supplied by the manufacturer. Transcriptome sequencing (RNA-seq) was performed on the Hiseq4000 (Illumina, U.S.A.) at BGI.

4.3.5 Sequence pre-processing

Reads from the two case studies were processed separately. Raw reads were firstly trimmed in Trimmomatic (version 0.32; Bolger *et al.* 2014) to remove sequencing adapter and obtain sequences with phred scores of more than 30. FastQC (version 0.11.9; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) were used to screen the overall sequence quality. Transcript reads were mapped to an established transcriptome reference database (Orsini *et al.* 2016) using Salmon (Version 0.8.2; Patro *et al.* 2017). The quasi-mapping function of Salmon was used with GC and positional bias correction. For each sample, both paired ends from two lanes were run together. Then the mapped transcript reads were processed in R (version 4.0.3). Low count reads (reads with raw count lower than 10) and outlier samples were identified and removed from the data set. The read counts were normalised by the size factor defined in the DESeq2 package (version 1.30.0; Love *et al.* 2014) in this study..

4.3.6 Similarity analysis of transcriptomic and chemical profiles

Normalised gene counts and targeted chemical concentrations in the Chaobai River case study were used for similarity analysis. Principal component analysis (PCA; Konishi T. 2015) was used to reveal the overall similarity based on the first two principal components, which explained a considerable proportion of the overall variance. Hierarchical clustering analysis (HCA; Eisen *et al.* 1998) was conducted based on the Euclidean distance with the ward.D2 clustering method. Pearson correlation coefficient was calculated in pairwise treatment levels to reveal the co-variation (another perspective of similarity) of any two treatment levels.

4.3.7 Gene cluster identification

The highly variable genes were selected by scran package with the normalised gene counts of both case studies (version 1.18.7; Lun *et al.* 2016). Selected genes were clustered by Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH) algorithm in the hopach package (version 2.52.0; Pollard and Van Der Laan 2003) on R. Cosine distance was chosen to capture the similarity between any two genes, as suggested in Eisen *et al.* (1998). Sample bootstrapping was performed to confirm the variability of the composition of gene clusters. The median value of genes belonging to the gene cluster was used as the reference value. For each pseudo-replication, samples were randomly selected to generate a new cluster pattern based on the gene cluster reference values. The gene cluster assignments were recorded. The frequencies of genes assigned to each cluster were summarised from the records of 10,000 repeats. Gene clusters were further used for clustering the samples based on Euclidean distance measurement with the ward.D2 clustering method.

4.3.8 Pathway analysis of co-responsive modules

As described in Chapter 3, a cross-species KEGG pathway overrepresentation test was performed for *Daphnia magna* gene set pathway analysis. The *Daphnia magna* genes in the transcriptomic co-responsive modules were re-annotated by their corresponding ortholog group IDs, based on the orthologous relationships between *Daphnia magna* and *Drosophila melanogaster* from the OrthoDB database (v10.1; Kriventseva *et al.* 2019). A permutation chi-square test was performed over 100,000 iterations to generate a robust P-value estimation directly from resampling detected *Daphnia* genes annotated with ortholog groups. The P-values of the permutation chi-

square tests were further corrected following the Benjamini-Hochberg procedure with a false discovery rate at 0.05 (Benjamini and Hochberg 1995).

4.3.9 Correlation analysis of eigengenes and chemical components

The eigengene of each gene cluster is the first principal component of the gene cluster matrix. Pearson correlation coefficients were calculated between each eigengene and individual chemical, in order to identify close associations between chemical component and gene clusters.

4.4 Results

4.4.1 Chemical analysis

(1) Distribution of PAHs in Chaobai River – US EPA prioritised PAHs were measured using surface water samples collected from 30 sites along the Chaobai River Basin (Figure S4.3). The indeno[1,2,3-cd]pyrene (IncdP) were under detection levels among all the sites. The naphthalene (Nap), acenaphthylene (Acy) and benzo[k]fluoranthene (BkF) detected the highest concentration levels at site B07, which is near the urban wastewater outlet. Dibenz[a,h]anthracene (DBA) detected 16.72 ng/L at site M16, which is also close to the municipal wastewater treatment plant (WWTP) outlet. The highest levels of acenaphthene (Ace), pyrene (Pyr), chrysene (Chry), benz[a]anthrancene (BaA) were detected at site M06; and the highest levels of fluorene (Fluo), phenanthrene (Phe), fluoranthene (Flua), benzo[a]pyrene (BaP) and benzo[g,h,i]perylene (BghiP) were detected at site M08; these two sites are located in the urban area close to megacity Beijing (population over 21.54 million).

(2) Distribution of organic micropollutants in Chaobai River – A total of 21 organic substances were measured using surface water samples collected from 30 sites along

the Chaobai River Basin (Figure S4.4). The tetracycline (TCN) was under detection levels among all 30 sites, and the atenolol (ATN), chlortetracycline (CTC), norfloxacin (NOR), oxytetracycline (OTC), propranolol (PROP), and sulfamerazine (SMR) could be only detected in one site along the river. Among 21 organic substances, caffeine (CAF), carbamazepine (CBZ) and erythromycin (ERY) were detected in more than 50% of all sampling sites (29, 25 and 27 sites, respectively) along the Chaobai River. The CAF could be detected in 29 sites ranging from 2.1 to 64.7 ng/L, observing its highest level at site M06. The CBZ could be detected in 25 sites ranging from 0.4 to 35.2 ng/L, with the highest concentration level at site M11 and the second highest (34.2 ng/L) at site M01. The ERY could be detected in 27 sites and reached the highest level (593.7 ng/L) at site M16 and the second-highest level (311.2 ng/L) at site M17, which are the lowest two sites in the Chaobai River and close to the municipal WWTP outlet.

(3) Similarity analysis of targeted chemicals – The principal component analysis (PCA) plot reveals the general similarity of measured chemicals among 30 water samples in the Chaobai River case. It is obvious that site B07, C03 and M06 are different from the rest. Table S4.5 summarise the relative contribution of each chemical factor to the first two components. The variance of the first component is largely contributed by organic chemicals (BF, CLA, ROX, and SMX) and salts (PO4 and TP); while the variance of the second component is largely contributed by heavy metals (Cr, Cu, Fe, Ni, and Zn) and organic micropollutants (CIP, ENR, and LOM). Thus, the differences between M06 and the rest 29 water samples is closely related to the concentration differences of BF, CLA, ROX, SMX, PO4 and TP; while the difference between C03 and the rest is contributed by Cr, Cu, Fe, Ni, Zn, CIP, ENR and LOM.

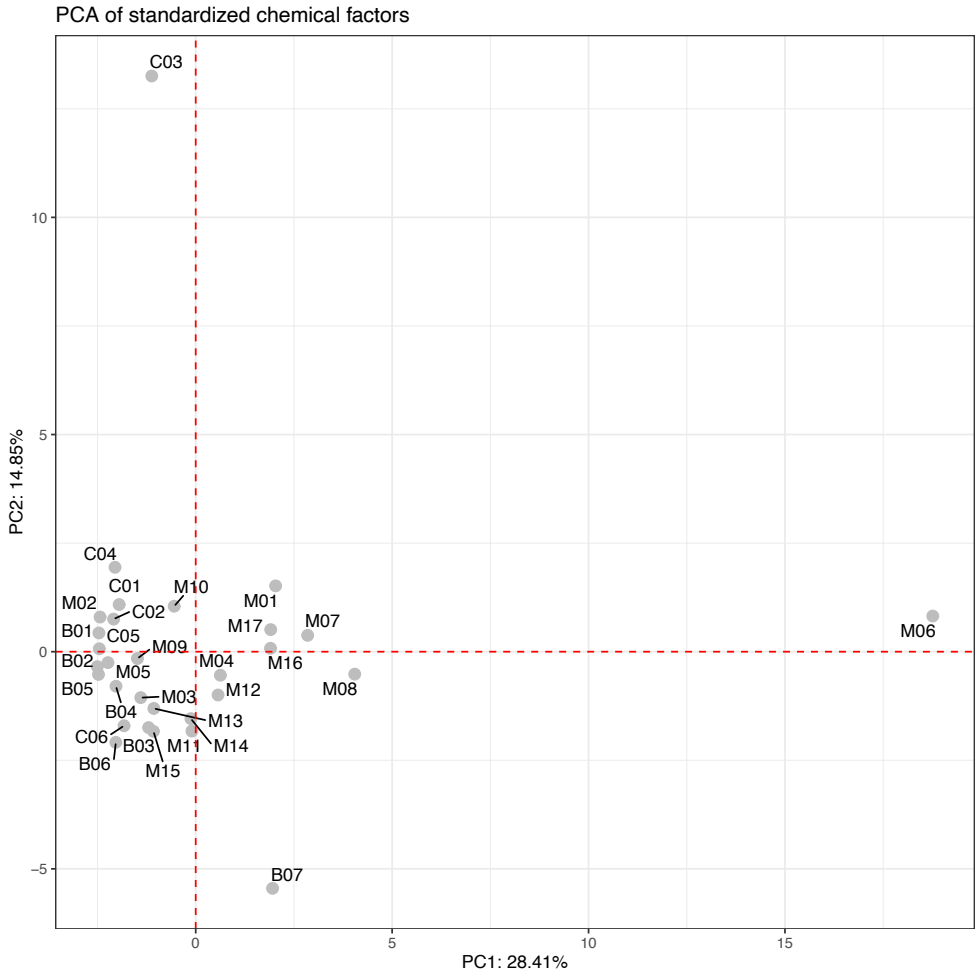

Figure 4. 1 PCA plot of targeted chemicals in water samples of the Chaobai River. The sampling site names are described in Table S4.1 and Figure S4.2.

4.4.2 Immobility rate of 48 hours exposure

The immobility rate of exposed *Daphnia* is plotted in Figure S4.5. After 48 hours of exposure, only the water sample from site C04 observed a 20% immobility rate (1 out of 5 neonates immobilised, the same among three biological replicates). Water samples from sites C05, M02, M03, M05, and M10 also recorded a 6.7% of immobility rate (1 out of 5 neonates immobilised in 1 biological replicate). The rest 24 water samples did not induce any immobilised neonates after 48 hours of exposure.

4.4.3 RNAseq data pre-processing

Although the transcriptome can be interpreted at the level of alternative splice variants of genes, for this study, all the successfully mapped transcriptome data were summarised at the gene level (Figure S4.6). With an average mapping rate of 98.13%, RNA sequencing of each sample produced 12 million reads, on average. Genes with raw read counts under 10 were also removed from the downstream analysis, leaving a final total of 14,705 genes for my investigations.

4.4.4 Similarity analysis of the transcriptomics profiles

To demonstrate the overall similarity of the transcriptomics profiles (Figure 4.2), principal component analysis (PCA) was performed using the mean values of the 14,705 genes of individual river treatments. Figure 4.2a shows both the first principal component (explaining 91.3 % of the total variance) and the second component (2.2 % of the total variance) could clearly distinguish M16 samples from the others. The hierarchical clustering dendrogram plot in Figure 4.2b was generated by ward.D2 clustering method based on Euclidean distance. Similar to the PCA plot, M16 formed a unique branch apart from the other clustered samples, based on gene expression similarities. A pairwise Pearson correlation coefficient matrix shown in Figure 4.2c revealed that most of the sites in the Chaobai River induced highly similar transcriptomic profiles (coefficient larger than 0.8) in the exposed daphnids except for M16 and M17.

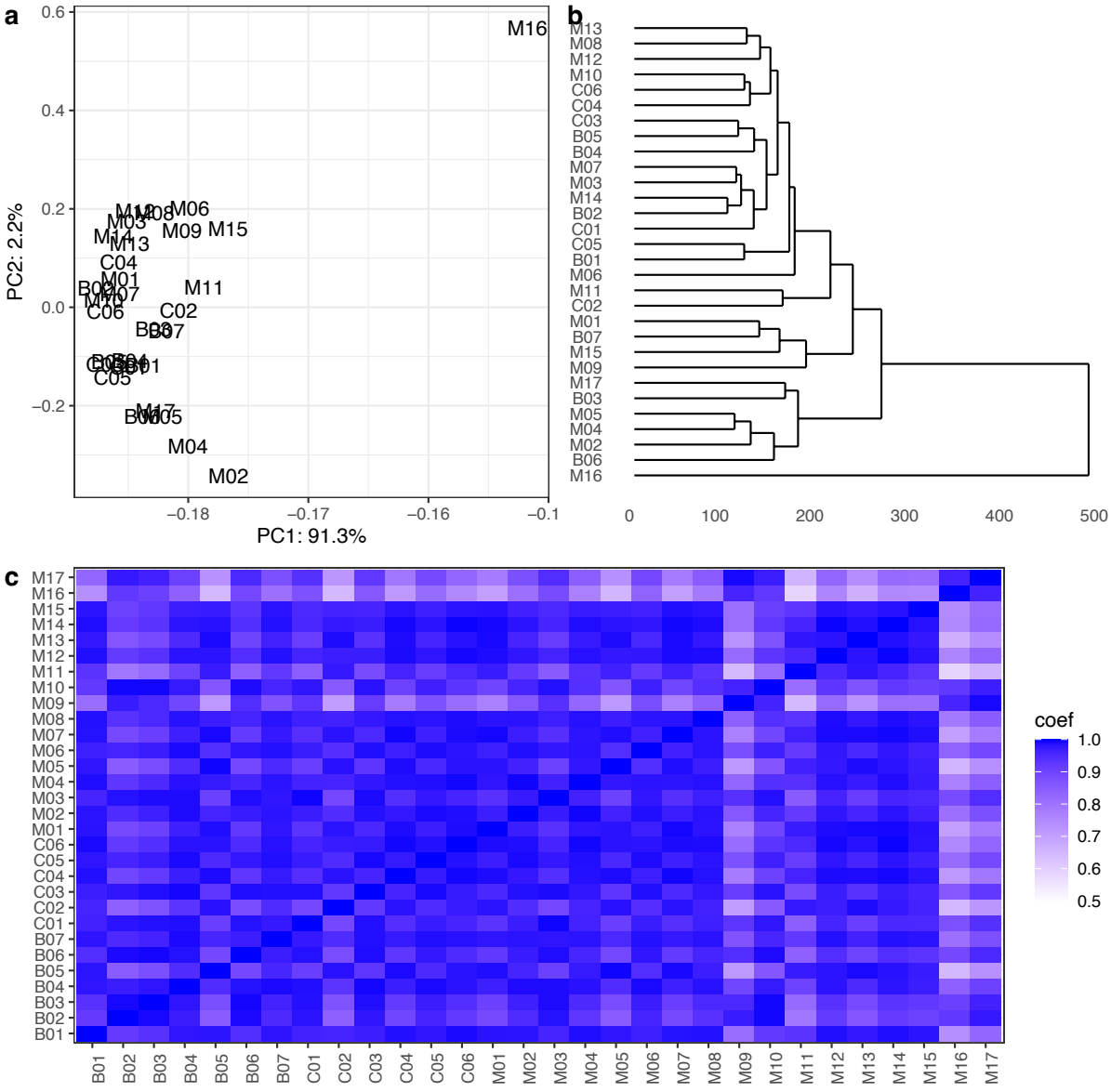

Figure 4. 2 Similarity analysis of transcriptomic profiles in the Chaobai case. (a) Score plot of Principal Component Analysis, (b) Dendrogram of Hierarchical Clustering Analysis, and (c) Heatmap of pairwise Pearson correlation coefficient (coef). Sample site locations are shown in Figure S4.2.

4.4.5 Gene cluster identification

A total of 2796 genes were identified as highly variable genes with the scran package. A total of 27 gene clusters were generated by the HOPACH algorithm, including 13 smaller clusters (each cluster consists of 1 to 4 genes) and 14 larger clusters (each cluster consists of 25 to 756 genes). Among them, cluster C8000 was the largest consisting of 756 genes (27 %), followed by cluster C6200 (622 genes, 22 %) and cluster C3000 (325 genes, 12 %). The detailed information of 14 larger gene clusters is listed in Table S4.6.

The robustness of the gene clustering pattern was evaluated by sample bootstrapping, and the reappearance frequencies of gene clusters are plotted in Figure S4.7. Most of the 14 larger gene clusters were relatively stable. The average reappearance frequencies were all above 50 % except for cluster C4401; cluster C1200, C5200, and C9000 had average reappearance frequencies levels greater than 87 %. In brief, the gene clustering pattern is considered to be robust, as most of the gene clusters consisted of stable members verified in the bootstrapping analysis.

4.4.6 Functional analysis of gene cluster

The statistical overrepresentation tests for gene sets forming clusters were performed by permutation chi-squared tests. This analysis was conducted to investigate whether the membership of similarly expressed genes in each cluster is reflective of their functions within known KEGG pathways. The adjusted *P* values of all KEGG pathways are listed in Appendix 1. The adjusted *P* values of pathways related to xenobiotic metabolism are plotted in Figure 4.3.

Figure 4.3 shows that xenobiotic biodegradation and metabolism pathways were significantly enriched in clusters C2351, C3000, C7000 and C8000. The ABC

transporter related to transmembrane transportation was significantly enriched in C7000 and C5200. The antioxidant defence system, which includes glutathione and ascorbate, was also over-represented in C2351, C3000 and C7000. The functional roles of genes within C2351, C3000, C5200, C7000 and C8000 might be similar as they consist of functional genes related to xenobiotic detoxification processes.

In addition to the pathway enrichment results for the six clusters in Figure 10, autophagy-related pathways were significantly enriched in C1120, C2100 and C5100. Metabolic pathways related to the metabolism of carbohydrate, lipid, glycan, and protein processing were also reported as significantly enriched in C2100 (Appendix 1), suggesting that C1120, C2100 and C5100 might be related to turnover of cellular components and maintaining cellular homeostasis.

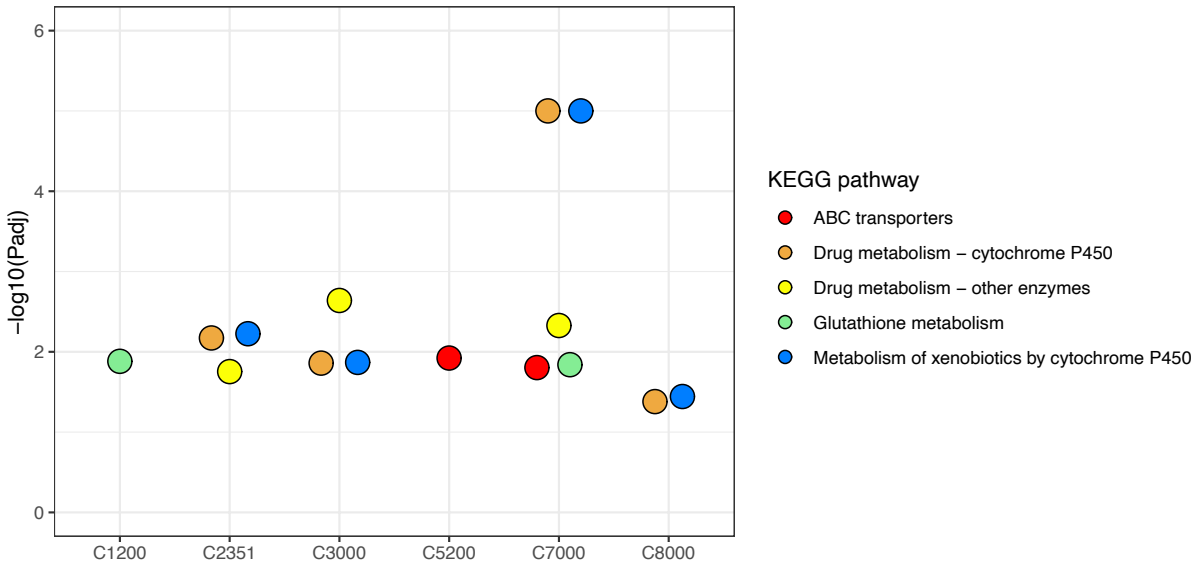

Figure 4. 3 Overrepresentation analysis of xenobiotic metabolism-related pathways among the Chaobai River gene clusters. The significance levels of selected pathways are shown in the plot with their negative logarithms (-log10) of the adjusted P values.

4.4.7 Clustering pattern of xenobiotic-related gene clusters

The HCA plots of 14 gene clusters and a combined set of 13 other gene clusters are shown in Figure S4.8. Among them, gene clusters C2100, C2351, C3000, C4401, and C7000 revealed distinctive expression pattern in M16 compared to the other sites, which reflects the gene-centric pattern shown in Figure 4.2a. As mentioned earlier, gene clusters C2351, C3000 and C7000 are signalling the xenobiotic metabolic functions, and C2100 is associated with cellular homeostasis. Thus, the clustering pattern of these five gene clusters suggested that the chemical mixture within the sampled waters from site M16 induced a distinctive transcriptomics profile which might be related to higher expression levels (compared against site B01) in xenobiotic degradation pathways and potential cellular damage that require faster turnover of cellular construct molecules.

4.4.8 Correlation analysis between eigengenes of 14 gene clusters and chemical factors

The eigengene is the first principal component of the gene cluster matrix. As genes in each gene cluster share similar variation pattern across all the samples, Based on the Pearson correlation between eigengenes and chemical factors. It is clear that cluster C4401 associated with most of the chemical factors, including salts (NO2, TN, PO4, TP), Mn, PAHs (Ace, Pyr, BaA, Chry) and organic chemicals (ATE, AZN, BF, CLA, ERY, MET, ROX, SDZ, SMX, TMP). Cluster C2351, C3000, and C7000 are significantly correlated with DBA, ERY and TMP. Cluster C2100 is significantly associated with PAHs (Fluo, Phe, Flua, DBA) and CTC.

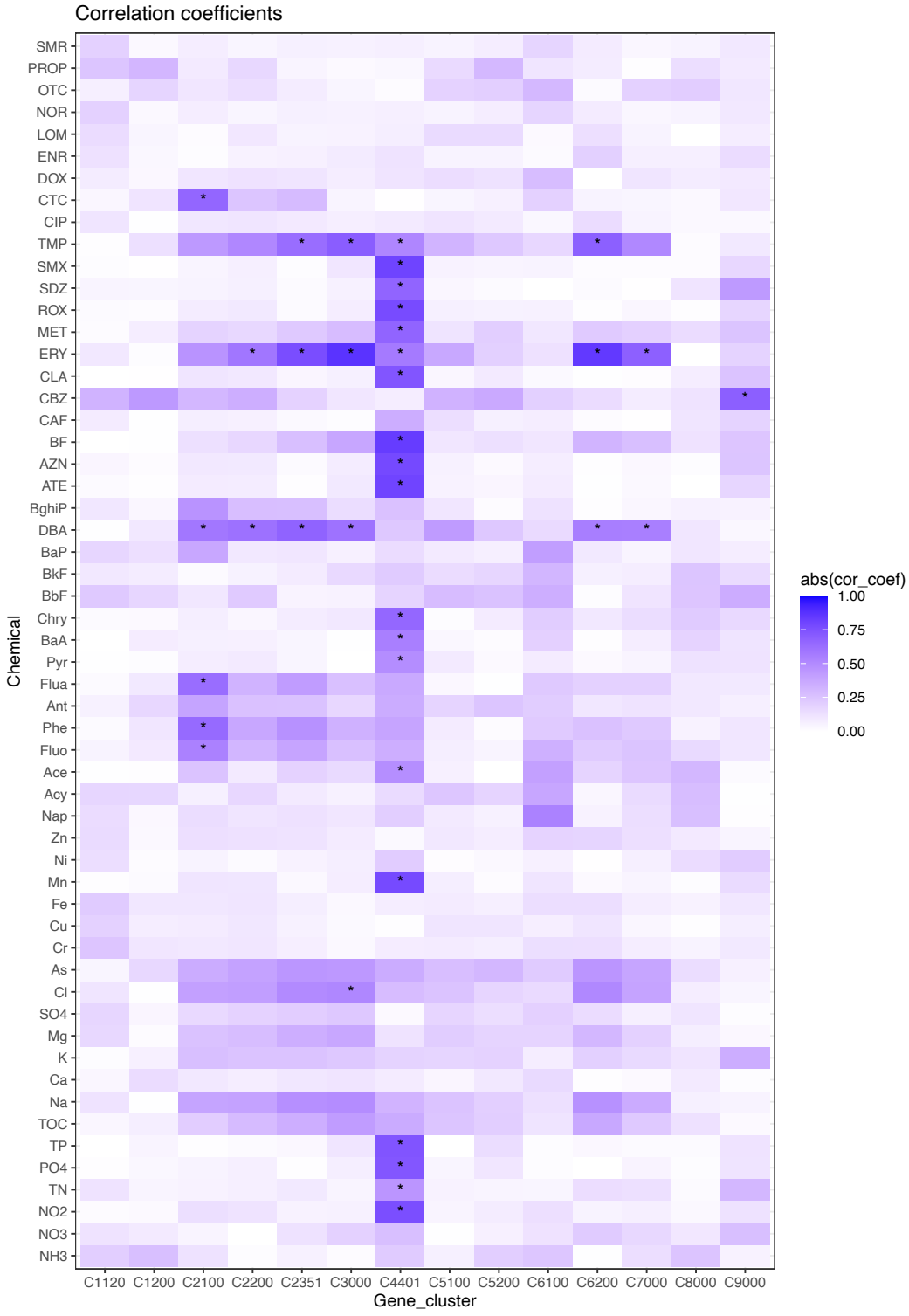

Figure 4. 4 Correlation analysis between eigengenes of 14 gene clusters and chemical factors. The color code corresponds to the absolute value of Pearson correlation coefficient values. The asterisks highlight the significant correlation with P values lower than 0.05.

4.5 Discussion

In the Chaobai River case, 65 chemicals and transcriptomic profiles of exposed daphnids are measured after 48 hours exposure. However, the general similarity analysis of chemical factors and transcriptomic profiles reveal different grouping pattern. One of the reasons is that the targeted chemical profiles is limited to tens of chemical component, which may not be adequate to reveal the overall differences of the water samples. Another possibility is that the concentration levels of most of the chemicals are not exceed their corresponding EC50 levels (refer to Table S4.3 and S4.4), and they are not efficient to induce observable variations in immobility rate (Figure S4.5), suggesting they may not be responsible for the differences observed in the transcriptomes.

PCA plot of transcriptomic profiles reveals that sampled water from site M16 induced obvious variations in the xenobiotic metabolism pathway and potentially trigger oxidative stress responses in the exposed *Daphnia* neonates after 48 hours of exposure. Specifically, the expression pattern of clusters C2351, C3000 and C7000 in Daphnia's response to water sampled at M16 revealed obviously higher transcripts levels in genes participating in xenobiotic biodegradation, compared against expression profiles of site B01. Correlation analysis further suggested that the differences between M16 and the rest might be related to DBA, ERY and TMP, which is partially suggested by the PCA analysis of the targeted chemical profiles (Table S4.5).

Targeted chemicals measurements of site M16 reveal that relatively high levels of phenanthrene (PHE, 110.12 ng/L), dibenz[a,h]anthrancene (DBA, 16.72 ng/L), erythromycin (ERY, 593.68 ng/L), trimethoprim (TMP, 132.18 ng/L) were detected at

this sampling site, compared against site B01. Previous studies revealed similar expression patterns of exposed organisms to either individual chemicals or chemical mixtures containing those four chemical compounds. The PHE induced alternations in xenobiotic related genes in aquatic amphipods by the up-regulations of CYP and GST but no significant changes in the expression of ABC, PHE also activated stress response genes such as heat shock protein in the amphipod *Lepeophtheirus salmonis* (Shatilina *et al.* 2020), which is also a malacostraca crustacean. Although there are only a few past studies investigating gene expression changes from exposure to DBA (Labib *et al.* 2016) or part of the mixture treatment (Labib *et al.* 2017; Liu *et al.* 2019; Sun *et al.* 2021), past studies suggested that DBA shared similar mode of action with PAH mixtures (Sun *et al.* 2021) and may trigger signalling pathways (such as p53, apoptosis, cell cycle, AhR, circadian rhythm) and xenobiotic metabolic pathways (Labib *et al.* 2017). Recent studies revealed that the mechanism of action of ERY includes growth inhibition of algae by down-regulating DNA replication, interrupting carbon assimilation and chlorophyll-a biosynthesis, and reducing detoxification activities in xenobiotic metabolism (Machado and Soares 2019; Guo et al. 2021). TMP may suppress reproduction in *Daphnia magna* at the concentration level of 13 mg/L (Dalla Bona *et al.* 2015), which is 100,000 higher than the concentration levels observed at site M16 (132 ng/L) and hardly comparable to the environmental chemicals' exposure scenario described in the Chaobai case. Thus, the activation of xenobiotic related genes and potentially oxidative stress responses might be related to the relatively high concentrations of PHE, DBA, and ERY. Further investigation of the organic substances by non-targeted screening assay will be needed to identify the chemical components that trigger oxidative stress in the exposed *Daphnia*.

4.6 Conclusion

In this Chaobai River case study, transcriptomic profile is used to characterise the effects of environmental chemical mixtures from a natural river in China. Genes clusters support the differences between site M16 and the rest are closely related to xenobiotic metabolism and stress response. And these clusters are significantly correlated to organic chemicals like DBA, ERY and TMP.

4.7 Supplementary

Tables and Figures

Table S4. 1 Description of the sampling sites along the Chaobai River Basin. Chemical attributes at each site were measured on-site from 19th to 22nd September in 2017.

Site	Longitude(E)	Latitude(N)	Reach	pH	TDS [a] (mg/L)	TOC [b] (mg/L)
B01	116.624	40.730	Bai river	8.19	312.8	1.26
B02	116.775	40.654	Bai river	8.32	190.4	1.21
B03	116.782	40.632	Bai river	8.49	148.2	1.06
B04	116.796	40.613	Bai river	8.28	312.7	1.63
B05	116.802	40.565	Bai river	8.02	278.0	2.03
B06	116.847	40.408	Bai river	8.23	138.7	1.90
B07	116.836	40.370	Bai river	8.27	93.4	2.53
C01	117.162	40.694	Chao river	8.13	447.0	1.28
C02	117.127	40.680	Chao river	8.24	316.3	1.24
C03	117.177	40.650	Chao river	7.95	420.9	1.66
C04	116.996	40.438	Chao river	8.00	478.6	1.74
C05	116.911	40.391	Chao river	7.97	488.9	2.14
C06	116.837	40.348	Chao river	8.26	206.9	5.87
M01	116.817	40.348	Chaobai river	8.09	718.2	6.37
M02	116.678	40.136	Chaobai river	8.12	453.8	3.46
M03	116.732	40.102	Chaobai river	8.29	337.5	6.08
M04	116.765	40.046	Chaobai river	8.33	462.3	8.56
M05	116.763	39.971	Chaobai river	8.14	199.6	4.33
M06	116.781	39.908	Chaobai river	7.34	485.7	4.79
M07	116.842	39.857	Chaobai river	7.93	652.0	3.75
M08	116.973	39.785	Chaobai river	7.98	629.9	3.48
M09	117.129	39.738	Chaobai river	8.38	327.3	2.64
M10	117.209	39.710	Chaobai river	7.92	569.9	3.90
M11	117.290	39.680	Chaobai river	8.10	272.8	2.78
M12	117.388	39.610	Chaobai river	7.89	441.2	3.58
M13	117.479	39.470	Chaobai river	8.28	342.5	2.95
M14	117.504	39.390	Chaobai river	8.11	418.4	3.75
M15	117.578	39.278	Chaobai river	8.19	252.3	5.75
M16	117.661	39.155	Chaobai river	7.60	916.3	7.90
M17	117.734	39.110	Chaobai river	8.58	927.4	3.65

a. TDS, total dissolved solids.

b. TOC, total organic carbon.

Table S4. 2 Inorganic chemicals in surface water samples from the Chaobai River Basin.

Parameter	Unit	Occurrence [a]	Range [b]	LOQ [c]
Ammonia (NH3)	mg/L	22	0-1.43 (M10)	0.02
Nitrate (NO3)	mg/L	30	0.07-7.74 (M01)	0.004
Nitrite (NO2)	mg/L	14	0-0.69 (M06)	0.004
Total N (TN)	mg/L	29	0-16.14 (M06)	0.01
Phosphate (PO4)	mg/L	8	0-1.43 (M06)	0.01
Total P (TP)	mg/L	10	0-1.44 (M06)	0.01
Sodium (Na)	mg/L	30	3.35-170.44 (M17)	0.005
Calcium (Ca)	mg/L	30	16.16-44.10 (C01)	0.011
Potassium (K)	mg/L	30	1.11-28.36 (M01)	0.020
Magnesium (Mg)	mg/L	30	4.93-29.33 (M17)	0.013
Sulphate (SO4)	mg/L	30	17.78-125.41 (M17)	0.75
Chloride (Cl)	mg/L	30	5.08-247.07 (M17)	0.15
Arsenic (As)	μg/L	30	0.69-7.01 (M17)	0.01
Chromium (Cr)	μg/L	30	2.48-97.30 (C03)	0.01
Copper (Cu)	μg/L	30	0.07-44.90 (C03)	0.01
Iron (Fe)	μg/L	30	26.3-414.0 (C03)	0.01
Manganese (Mn)	μg/L	30	0.04-119.00 (M06)	0.01
Nickel (Ni)	μg/L	30	0.33-10.40 (C03)	0.01
Zinc (Zn)	μg/L	30	4.48-112.00 (C03)	0.01

a. The number of sites with detectable and quantifiable measurements.

b. Zero measurement means under the limit of quantification (LOQ).

c. LOQ, limit of quantification.

Table S4. 3 PAHs in surface water samples from the Chaobai River Basin.

Parameter	Occ [a]	Range [b] (ng/L)	LOQ (ng/L)	EC50 [c] (mg/L)	Reference
Naphthalene (Nap)	28	0-34.55 (B07)	11.46	7.92	Ha 2019
Acenaphthylene (Acy)	28	0-4.25 (B07)	1.24	/	/
Acenaphthene (Ace)	28	0-6.07 (M06)	1.54	1.275	Munoz 1993
Fluorene (Fluo)	28	0-47.32 (M08)	4.52	1.34	Ha 2019
Phenanthrene (Phe)	28	0-246.47 (M08)	15.30	0.46	Ha 2019
Anthracene (Ant)	28	0-22.69 (M07)	1.67	0.095	Munoz 1993
Fluoranthene (Flua)	28	0-73.8 (M08)	0.70	0.106	Clement 2000
Pyrene (Pyr)	28	0-42.67 (M06)	0.65	0.058	Ha 2019
Benz[a]anthrancene (BaA)	28	0-81.30 (M06)	2.04	0.0015	Lampi 2006
Chrysene (Chry)	28	0-160.22 (M06)	3.93	/	/
Benzo[b]fluoranthene (BbF)	24	0-26.20 (M12)	2.22	0.12	Lampi 2006
Benzo[k]fluoranthene (BkF)	16	0-15.83 (B07)	0.04	/	/
Benzo[a]pyrene (BaP)	26	0-34.07 (M08)	0.62	0.0016	Lampi 2006
Indeno[1,2,3-cd] pyrene (IncdP)	0	/	1.10	/	/
Dibenz[a,h]anthracene (DBA)	28	0-16.72 (M16)	0.19	0.0016	Lampi 2006
Benzo[g,h,i]perylene (BghiP)	28	0-2.75 (M08)	1.26	0.0010	Lampi 2006

a. Occ, occurrence, the number of sites with detectable and quantifiable measurements.

b. Zero measurement means under the limit of quantification (LOQ).

c. EC50, concentration that induces 50 % immobility rate of *Daphnia* after 48 hours exposure.

Table S4. 4 Organic micropollutants in surface water samples from the Chaobai River Basin.

Parameter	CAS	Occ [a]	Range [b] (ng/L)	LOQ	EC50[c] (mg/L)	Reference
Atenolol (ATE)	29122-68-7	1	0-5.84 (M06)	1.82	/	/
Azithromycin (AZN)	83905-01-5	3	0-4.99 (M06)	0.34	3.23	Kuzmanovic 2014
Bezafibrate (BF)	41859-67-0	13	0-10.86 (M06)	0.41	240	Duarte 2019
Caffeine (CAF)	58-08-2	29	0-64.69 (M06)	1.40	1079	Lomba 2020
Carbamazepine (CBZ)	298-46-4	25	0-35.23 (M11)	0.36	9.53	Tongur 2020
Clarithromycin (CLA)	81103-11-9	3	0-4.60 (M06)	0.74	>2	Baumann 2016
Erythromycin (ERY)	114-07-8	27	0-593.68 (M16)	0.86	8.617	Kuzmanovic 2014
Metoprolol (MET)	37350-58-6	10	0-52.73 (M06)	1.08	133	Moermond 2016
Roxithromycin (ROX)	80214-83-1	5	0-29.48 (M06)	0.63	74.3	Choi 2008
Sulfadiazine (SDZ)	68-35-9	4	0-22.62 (M06)	1.31	97.28	Liu 2016
Sulfamethoxazole (SMX)	57-68-1	12	0-260.20 (M06)	1.37	189.2	Kim 2007
Trimethoprim (TMP)	738-70-5	16	0-132.18 (M16)	0.69	167	Choi 2008
Ciprofloxacin (CIP)	85721-33-1	3	0-7.42 (C03)	0.79	>100	Załęska-Radziwiłł 2011
Chlortetracycline (CTC)	64-72-2	1	0-1.58 (M08)	0.99	>400	Kim 2010
Doxycycline (DOX)	564-25-0	5	0-18.69 (M17)	1.03	156	Fernandez 2004
Enrofloxacin (ENR)	93106-60-6	6	0-6.62 (C03)	0.55	45.8	Kim 2010
Lomefloxacin (LOM)	98079-51-7	2	0-5.09 (C03)	0.52	166	Luo 2018
Norfloxacin (NOR)	70458-96-7	1	0-3.87 (C06)	1.12	/	/
Oxytetracycline (OTC)	79-57-2	1	0-2.09 (B07)	0.93	>400	Zounková 2011
Propranolol (PROP)	526-66-6	1	0-4.99 (M12)	0.66	2.19	Nielsen 2018
Sulfamerazine (SMR)	127-79-7	1	0-4.96 (C06)	1.62	205	Jung 2008
Tetracycline (TET)	60-54-8	0	-	1.37	8.16	Havelkova 2016

a. The number of sites with detectable and quantifiable measurements.

b. Zero measurement means under the limit of quantification (LOQ).

c. EC50, concentration that induces 50 % immobility rate.

Table S4. 5 Relative contribution of chemical factors to first two components in Chaobai case.

	Standardized			Standardized	
	PC1 (28.41%)	PC2 (14.85%)		PC1 (28.41%)	PC2 (14.85%)
NH3	0.37	0.01	As	0.95	0.02
NO3	0.00	2.62	Cr	0.09	**9.36**
NO2	4.21	0.06	Cu	0.00	**9.57**
TN	3.11	2.00	Fe	0.08	**9.63**
PO4	**5.42**	0.03	Mn	4.69	0.46
TP	**5.44**	0.02	Ni	0.98	**7.72**
TOC	0.69	0.06	Zn	0.01	**9.36**
Na	0.93	0.06	ATE	4.96	0.03
Ca	0.03	3.34	AZN	4.96	0.05
K	0.92	0.08	BF	**5.44**	0.03
Mg	0.05	1.25	CAF	1.53	0.15
SO4	0.00	2.33	CBZ	0.58	0.01
Cl	0.67	0.11	CLA	**5.17**	0.08
Nap	0.00	0.31	ERY	0.52	0.01
Acy	1.14	2.60	MET	4.62	0.10
Ace	3.53	1.87	ROX	**5.18**	0.05
Fluo	2.03	1.25	SDZ	4.60	0.11
Phe	1.99	0.66	SMX	**5.29**	0.07
Ant	2.39	0.01	TMP	1.02	0.00
Flua	2.05	0.56	CIP	0.03	**8.51**
Pyr	3.37	1.10	CTC	0.23	0.01
BaA	3.52	0.58	DOX	0.01	0.08
Chry	4.66	0.78	ENR	0.00	**8.43**
BbF	0.04	0.95	LOM	0.00	**8.89**
BkF	0.00	0.63	NOR	0.05	0.15
BaP	1.05	1.50	OTC	0.05	1.53
DBA	0.04	0.58	PROP	0.00	0.05
BghiP	1.25	0.00	SMR	0.05	0.15

a. Abbreviations are described in Table S4.2, S4.3, S4.4.

Table S4. 6 Summary of 14 gene clusters in the Chaobai case.

Module ID	Number of genes [a]	Genes with orthologs [b]	Genes with orthologs and pathways [c]
C1120	25	6	2
C1200	131	23	4
C2100	58	22	5
C2200	84	26	6
C2351	73	35	4
C3000	325	162	29
C4401	75	21	4
C5100	62	18	8
C5200	194	63	14
C6100	148	39	17
C6200	622	243	79
C7000	129	75	25
C8000	756	365	132
C9000	86	18	9

a. The total number of *Daphnia* genes in the gene cluster.

b. The number of *Daphnia* genes with orthologs in *Drosophila melanogaster* at the *Arthropoda* level in the gene cluster, the corresponding proportion listed in the parentheses.

c. The number of *Daphnia* genes with *Drosophila melanogaster* ortholog and KEGG pathway information.

Figure S4. 1 The eutrophication area of the Chaobai River. The image was taken on 20th September 2017 from a sampling site (M08) contributing to this study.

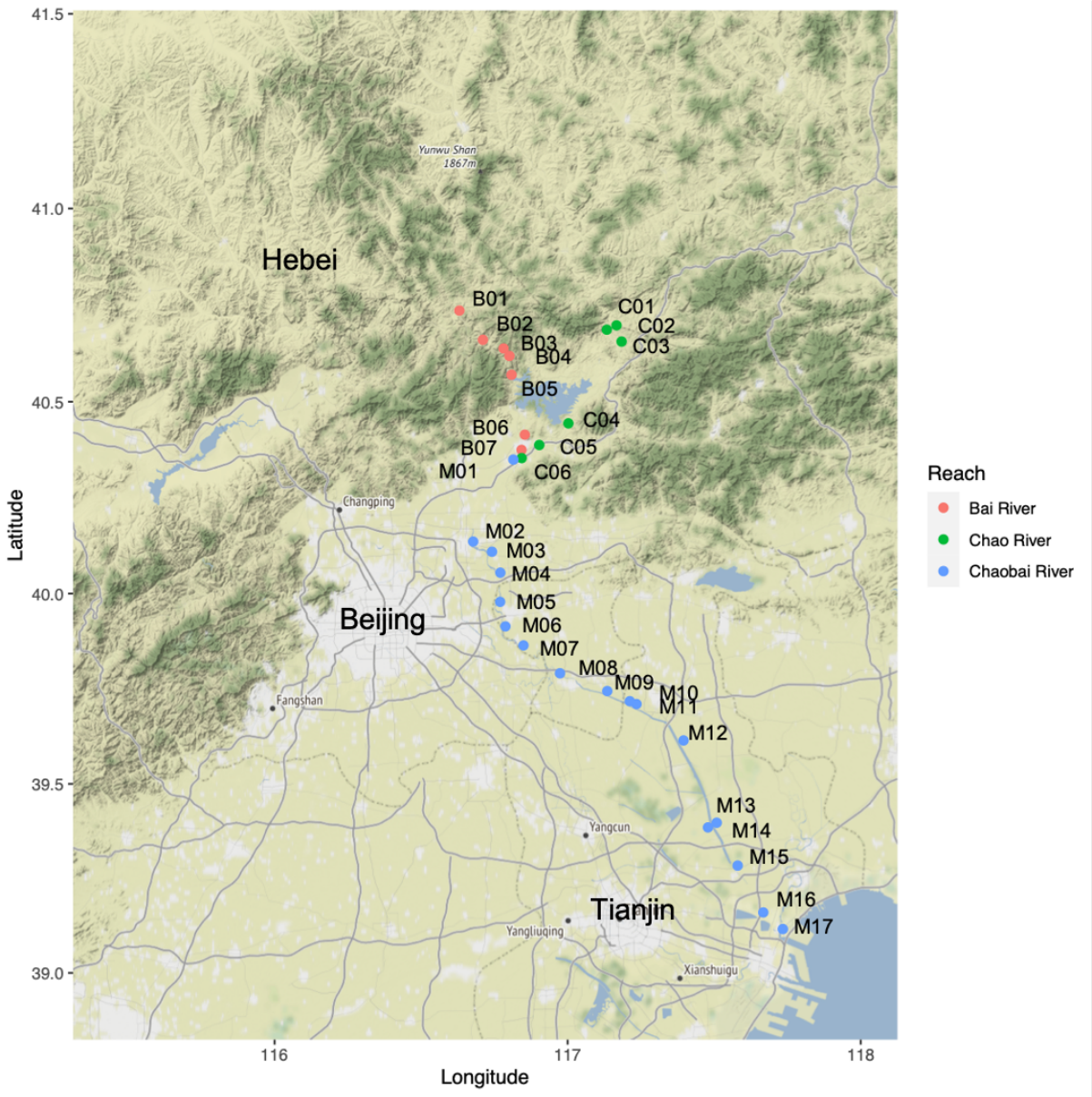

Figure S4. 2 The sampling sites of the Chao River, the Bai River and the Chaobai River. The headwaters originate from the Yunwu Mountains northeast of Beijing (sites B01-B06 and sites C01-C05), of which the region covered are regarded as a mountainous area. Recycled urban wastewaters (treated and untreated) enter the river between sites B07/C06 and M11, of which the region covered are regarded as urban area. There are two major WWTP outlets near B07-C06-M01 and M16. Agricultural runoff enters the river between sites M12 and M17, of which the region covered are labelled as agricultural area.

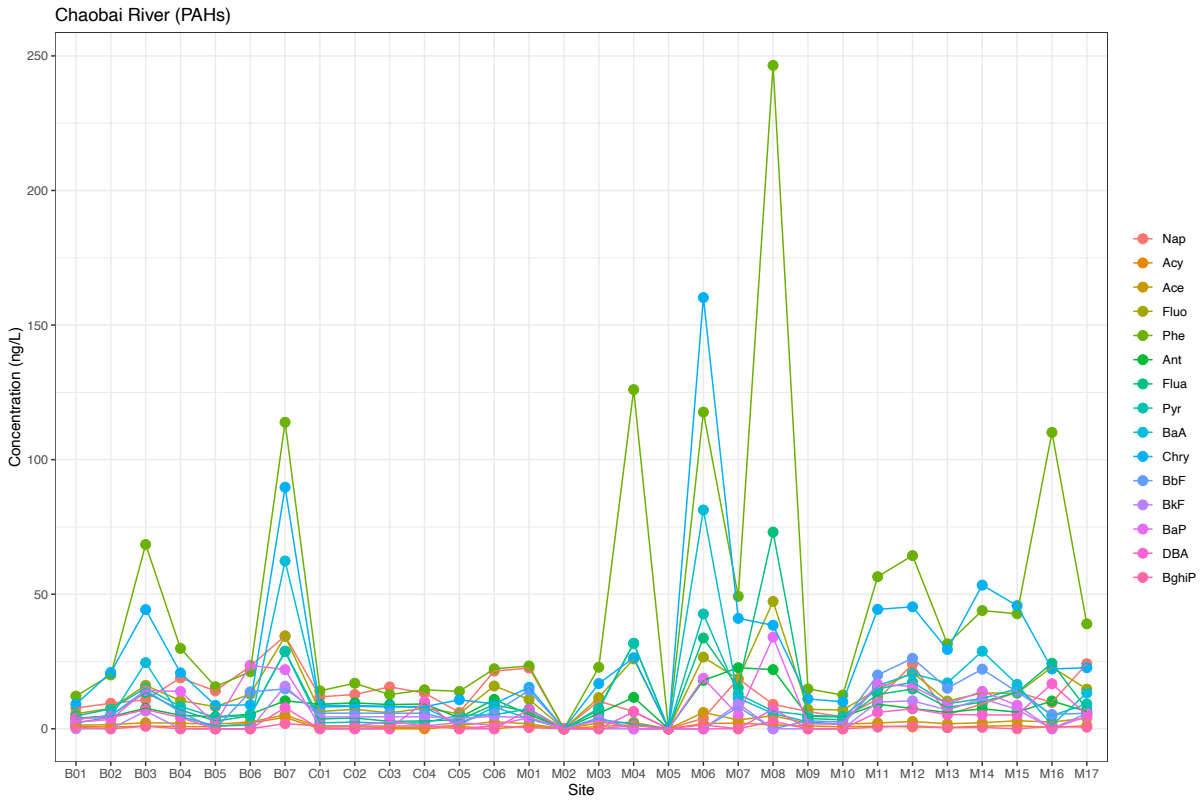

Figure S4. 3 Distribution of PAHs in the Chaobai River. The abbreviations of the PAHs are referred to in Table S4.3. The sites are described in Figure S4.2.

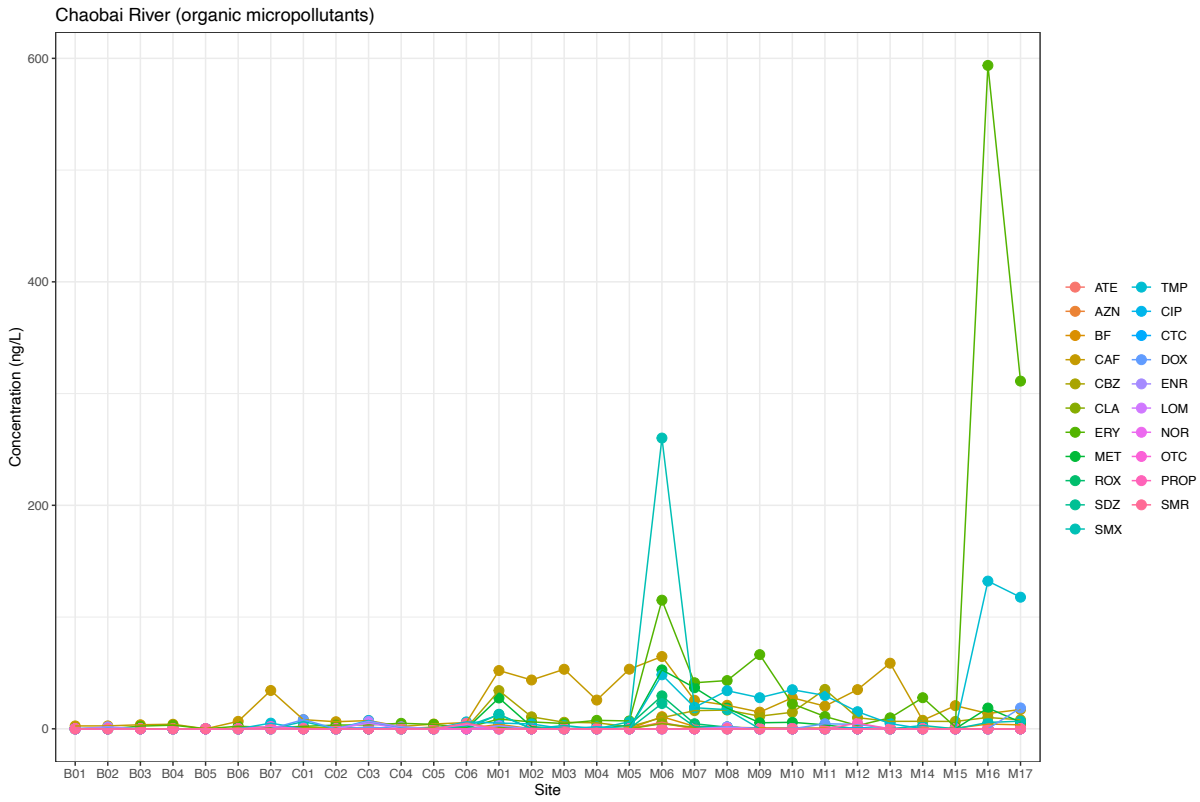

Figure S4. 4 Distribution of organic micropollutants in the Chaobai River. The abbreviations of the organic micropollutants are referred to in Table S4.4. The sites are described in Figure S4.2.

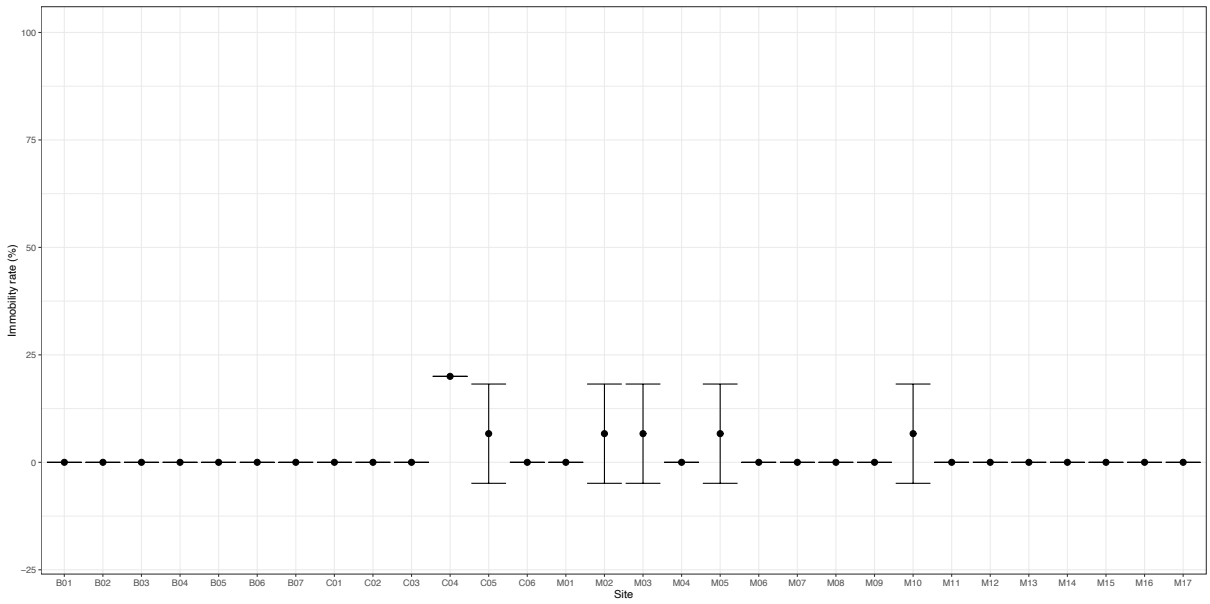

Figure S4. 5 Immobility rate of *Daphnia magna* after 48 hours exposure to filtered surface waters from the Chaobai River. The sites are described in Table S4.1 and Figure S4.2.

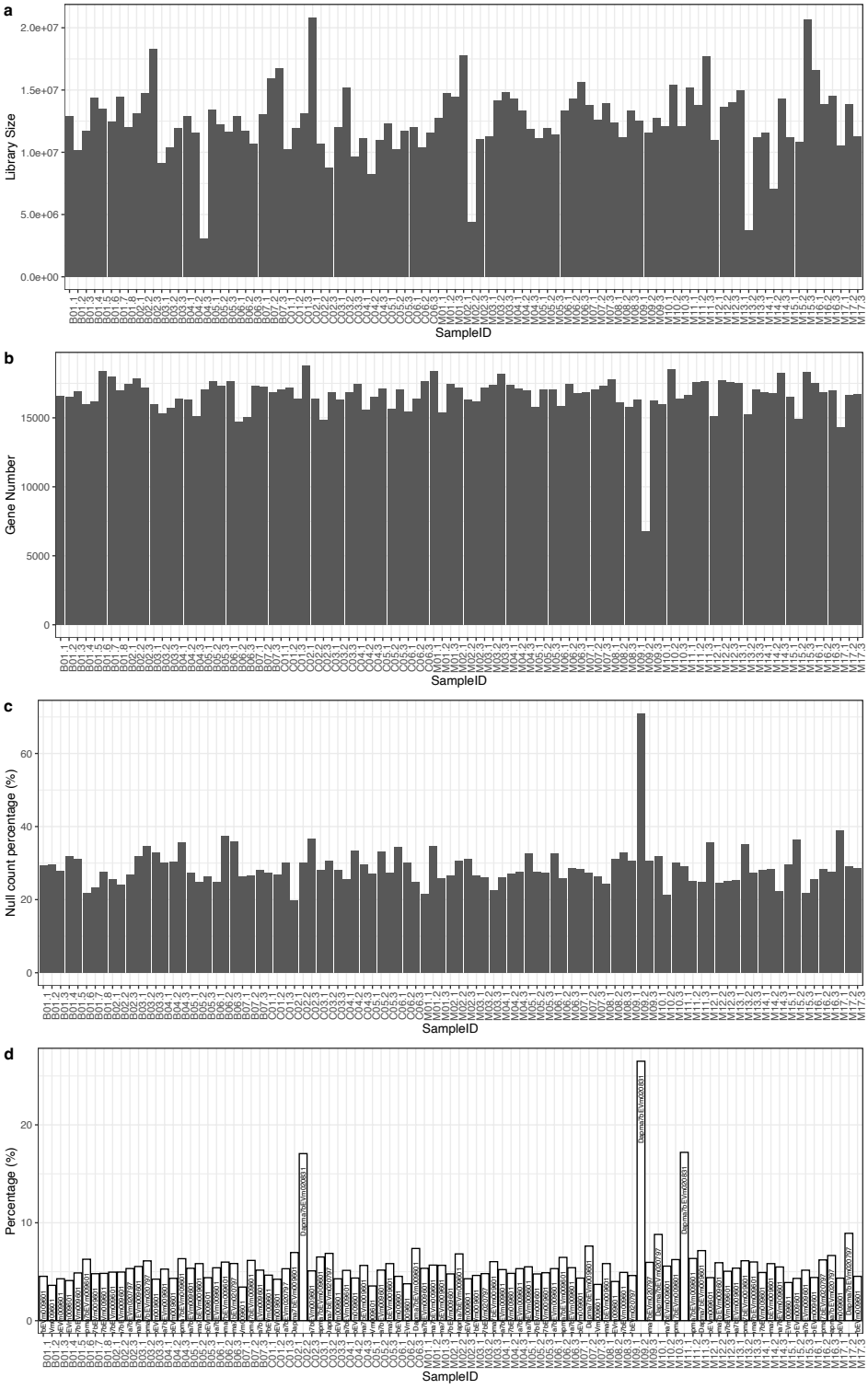

Figure S4. 6 Overview of Chaobai transcriptome data sets. (a) the total number of read counts in each sample; (b) the total number of genes in each sample; (c) the percentage of genes with null count in each sample; (d) over-representative genes in transcriptome profiles.

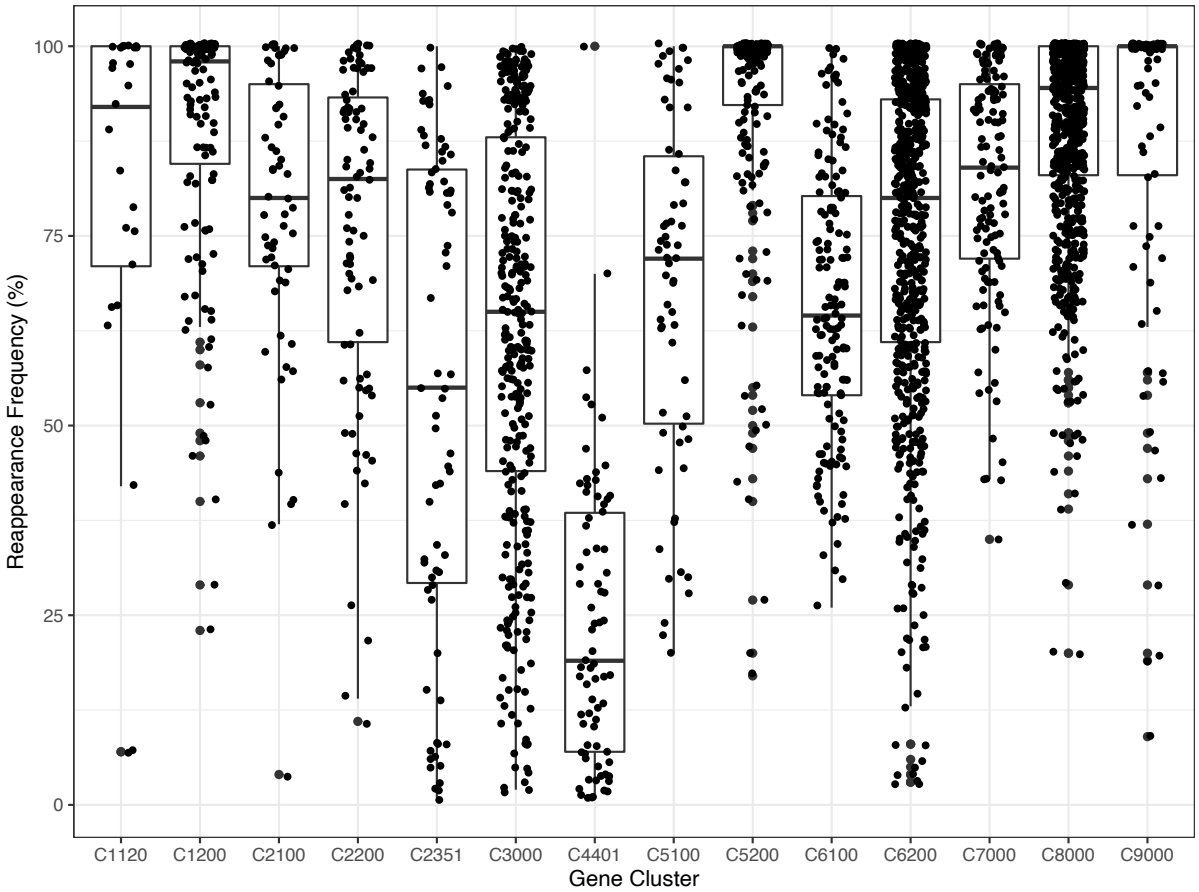

Figure S4. 7 Robustness of transcriptomic gene clusters in the Chaobai case. After 10,000 times bootstrap resampling of the genes. The frequencies of individual genes assigning to the same gene cluster are plotted as a boxplot representing the first-second-third quartiles of the frequencies, with scatter points representing individual genes' reappearance frequencies.

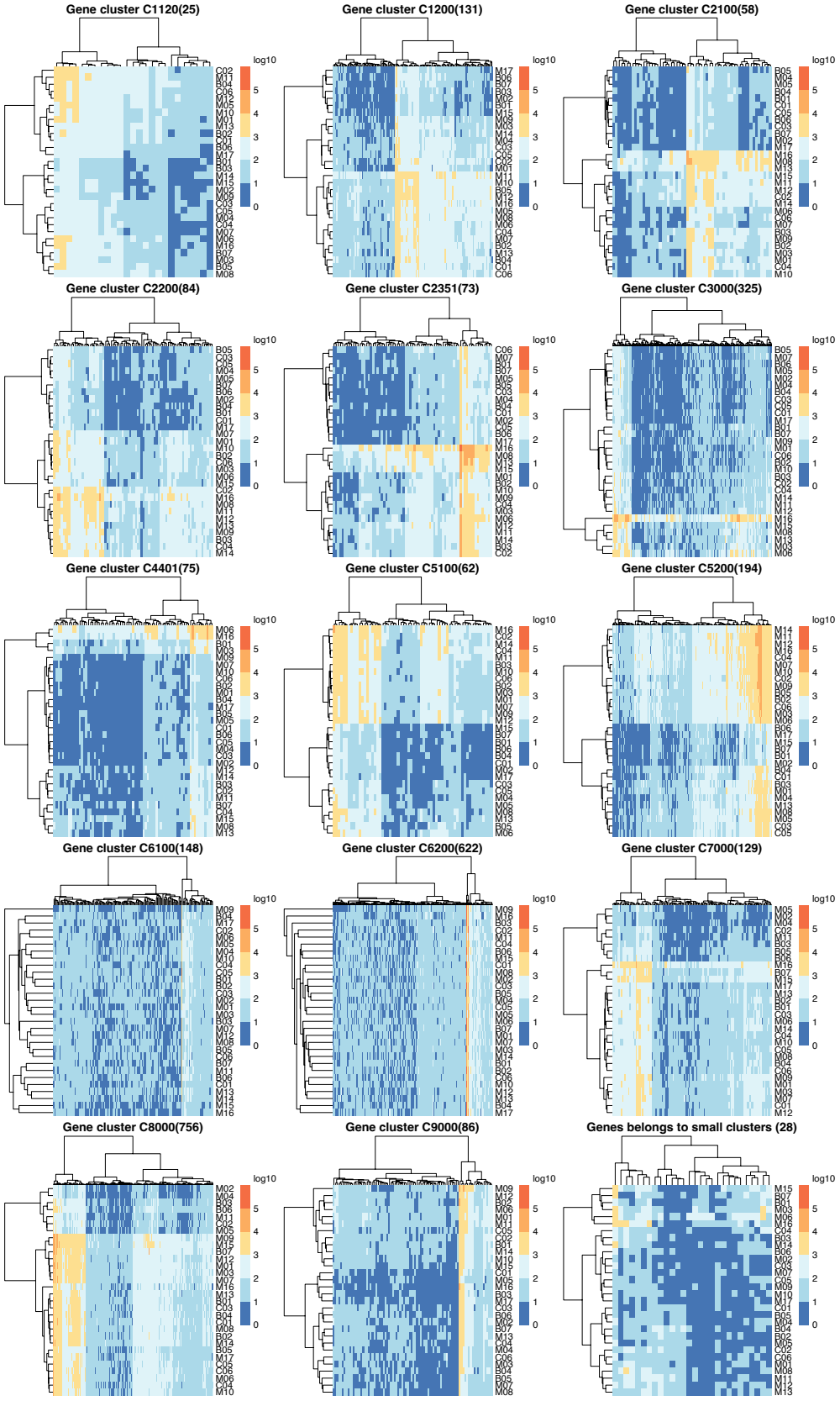

Figure S4. 8 Hierarchical clustering of gene expression in selected Chaobai gene clusters. All the gene counts were log10 transformed.

4.8 Reference

Amoatey, P., Baawain, M.S., 2019. Effects of pollution on freshwater aquatic organisms. Water Environment Research 91, 1272–1287. https://doi.org/10.1002/wer.1221

Baillon, L., Pierron, F., Coudret, R., Normendeau, E., Caron, A., Peluhet, L., Labadie, P., Budzinski, H., Durrieu, G., Sarraco, J., Elie, P., Couture, P., Baudrimont, M., Bernatchez, L., 2015. Transcriptome profile analysis reveals specific signatures of pollutants in Atlantic eels. Ecotoxicology 24, 71–84. https://doi.org/10.1007/s10646-014-1356-x

Barata, C., Varo, I., Navarro, J.C., Arun, S., Porte, C., 2005. Antioxidant enzyme activities and lipid peroxidation in the freshwater cladoceran Daphnia magna exposed to redox cycling compounds. Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology 140, 175–186. https://doi.org/10.1016/j.cca.2005.01.013

Chen, W., Lu, S., Jiao, W., Wang, M., Chang, A.C., 2013a. Reclaimed water: A safe irrigation water source? Environmental Development 8, 74–83. https://doi.org/10.1016/j.envdev.2013.04.003

Chen, W., Lu, S., Pan, N., Jiao, W., 2013b. Impacts of long-term reclaimed water irrigation on soil salinity accumulation in urban green land in Beijing: Soil Salinity Accumulation with Reclaimed Water Irrigation. Water Resour. Res. 49, 7401–7410. https://doi.org/10.1002/wrcr.20550

Eggen, R.I.L., Hollender, J., Joss, A., Schärer, M., Stamm, C., 2014. Reducing the Discharge of Micropollutants in the Aquatic Environment: The Benefits of Upgrading Wastewater Treatment Plants. Environ. Sci. Technol. 48, 7683–7689. https://doi.org/10.1021/es500907n

European Chemical Agency (ECHA) and European Food Safety Authority (EFSA) with the technical support of the Joint Research Centre (JRC), Andersson, N., Arena, M., Auteri, D., Barmaz, S., Grignard, E., Kienzler, A., Lepper, P., Lostia, A.M., Munn, S., Parra Morte, J.M., Pellizzato, F., Tarazona, J., Terron, A., Van der Linden, S., 2018. Guidance for the identification of endocrine disruptors in the context of Regulations (EU) No 528/2012 and (EC) No 1107/2009. EFS2 16. https://doi.org/10.2903/j.efsa.2018.5311

Falås, P., Wick, A., Castronovo, S., Habermacher, J., Ternes, T.A., Joss, A., 2016. Tracing the limits of organic micropollutant removal in biological wastewater treatment. Water Research 95, 240–249. https://doi.org/10.1016/j.watres.2016.03.009

Fantke, P., Aurisano, N., Provoost, J., Karamertzanis, P.G., Hauschild, M., 2020. Toward effective use of REACH data for science and policy. Environment International 135, 105336. https://doi.org/10.1016/j.envint.2019.105336

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., Brown, P.O., 2000. Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes□D. Molecular Biology of the Cell 11, 17.

Hassan, I., Jabir, N.R., Ahmad, S., Shah, A., Tabrez, S., 2015. Certain Phase I and II Enzymes as Toxicity Biomarker: An Overview. Water Air Soil Pollut 226, 153. https://doi.org/10.1007/s11270-015-2429-z

He, B., He, J., Wang, J., Li, J., Wang, F., 2018. Characteristics of GHG flux from water-air interface along a reclaimed water intake area of the Chaobai River in Shunyi, Beijing. Atmospheric Environment 172, 102–108. https://doi.org/10.1016/j.atmosenv.2017.10.060

Iwane, T., Urase, T., Yamamoto, K., 2001. Possible impact of treated wastewater discharge on incidence of antibiotic resistant bacteria in river water. Water Science and Technology 43, 91–99. https://doi.org/10.2166/wst.2001.0077

Landrigan, P.J., Fuller, R., Acosta, N.J.R., Adeyi, O., Arnold, R., Basu, N. (Nil), Baldé, A.B., Bertollini, R., Bose-O'Reilly, S., Boufford, J.I., Breysse, P.N., Chiles, T., Mahidol, C., Coll-Seck, A.M., Cropper, M.L., Fobil, J., Fuster, V., Greenstone, M., Haines, A., Hanrahan, D., Hunter, D., Khare, M., Krupnick, A., Lanphear, B., Lohani, B., Martin, K., Mathiasen, K.V., McTeer, M.A., Murray, C.J.L., Ndahimananjara, J.D., Perera, F., Potočnik, J., Preker, A.S., Ramesh, J., Rockström, J., Salinas, C., Samson, L.D., Sandilya, K., Sly, P.D., Smith, K.R., Steiner, A., Stewart, R.B., Suk, W.A., van Schayck, O.C.P., Yadama, G.N., Yumkella, K., Zhong, M., 2018. The Lancet Commission on pollution and health. The Lancet 391, 462–512. https://doi.org/10.1016/S0140-6736(17)32345-0

Liu, Y., Wang, L., Pan, B., Wang, C., Bao, S., Nie, X., 2017. Toxic effects of diclofenac on life history parameters and the expression of detoxification-related genes in Daphnia magna. Aquatic Toxicology 183, 104–113. https://doi.org/10.1016/j.aquatox.2016.12.020

Lüchmann, K.H., Clark, M.S., Bainy, A.C.D., Gilbert, J.A., Craft, J.A., Chipman, J.K., Thorne, M.A.S., Mattos, J.J., Siebert, M.N., Schroeder, D.C., 2015. Key metabolic pathways involved in xenobiotic biotransformation and stress responses revealed by transcriptomics of the mangrove oyster Crassostrea brasiliana. Aquatic Toxicology 166, 10–20. https://doi.org/10.1016/j.aquatox.2015.06.012

Oliveira, B.R.R., Deslandes, A.C., Santos, T.M., 2015. Differences in exercise intensity seems to influence the affective responses in self-selected and imposed exercise: a meta-analysis. Front. Psychol. 6. https://doi.org/10.3389/fpsyg.2015.01105

Orsini, L., Gilbert, D., Podicheti, R., Jansen, M., Brown, J.B., Solari, O.S., Spanier, K.I., Colbourne, J.K., Rusch, D.B., Decaestecker, E., Asselman, J., De Schamphelaere, K.A.C., Ebert, D., Haag, C.R., Kvist, J., Laforsch, C., Petrusek, A., Beckerman, A.P., Little, T.J., Chaturvedi, A., Pfrender, M.E., De Meester, L., Frilander, M.J., 2016. Daphnia magna transcriptome by RNA-Seq across 12 environmental stressors. Sci Data 3, 160030. https://doi.org/10.1038/sdata.2016.30

Pollard, K.S., n.d. Cluster Analysis of Genomic Data with Applications in R 27.

Qian, X., Liang, B., Liu, Xuan, Liu, Xinhui, Wang, J., Liu, F., Cui, B., 2017. Distribution, sources, and ecological risk assessment of polycyclic aromatic hydrocarbons in surface sediments from the Haihe River, a typical polluted urban river in Northern China. Environ Sci Pollut Res 24, 17153–17165. https://doi.org/10.1007/s11356-017-9378-6

Regoli, F., Giuliani, M.E., 2014. Oxidative pathways of chemical toxicity and oxidative stress biomarkers in marine organisms. Marine Environmental Research 93, 106–117. https://doi.org/10.1016/j.marenvres.2013.07.006

Slonim, D.K., 2002. From patterns to pathways: gene expression data analysis comes of age. Nat Genet 32, 502–508. https://doi.org/10.1038/ng1033

Su, D., Ben, W., Strobel, B.W., Qiang, Z., 2020. Occurrence, source estimation and risk assessment of pharmaceuticals in the Chaobai River characterized by adjacent land use. Science of The Total Environment 712, 134525. https://doi.org/10.1016/j.scitotenv.2019.134525

Su, J., Ji, D., Lin, M., Chen, Y., Sun, Y., Huo, S., Zhu, J., Xi, B., 2017. Developing surface water quality standards in China. Resources, Conservation and Recycling 117, 294–303. https://doi.org/10.1016/j.resconrec.2016.08.003

Sun, J., Zhou, Q., Hu, X., 2019. Integrating multi-omics and regular analyses identifies the molecular responses of zebrafish brains to graphene oxide: Perspectives in environmental criteria. Ecotoxicology and Environmental Safety 180, 269–279. https://doi.org/10.1016/j.ecoenv.2019.05.011

Vermeulen, R., Schymanski, E.L., Barabási, A.-L., Miller, G.W., 2020. The exposome and health: Where chemistry meets biology. Science 367, 392–396. https://doi.org/10.1126/science.aay3164

Wang, G., Xia, J., Chen, J., 2009. Quantification of effects of climate variations and human activities on runoff by a monthly water balance model: A case study of the Chaobai River basin in northern China: CLIMATE VARIATIONS AND HUMAN ACTIVITIES. Water Resour. Res. 45. https://doi.org/10.1029/2007WR006768

Wang, P., Xia, P., Yang, J., Wang, Z., Peng, Y., Shi, W., Villeneuve, D.L., Yu, H., Zhang, X., 2018. A Reduced Transcriptome Approach to Assess Environmental Toxicants Using Zebrafish Embryo Test. Environ. Sci. Technol. 52, 821–830. https://doi.org/10.1021/acs.est.7b04073

Watanabe, H., Kobayashi, K., Kato, Y., Oda, S., Abe, R., Tatarazako, N., Iguchi, T., 2008. Transcriptome profiling in crustaceans as a tool for ecotoxicogenomics: Daphnia magna DNA microarray. Cell Biol Toxicol 24, 641–647. https://doi.org/10.1007/s10565-008-9108-4

Xiong, W., Ni, P., Chen, Y., Gao, Y., Shan, B., Zhan, A., 2017. Zooplankton community structure along a pollution gradient at fine geographical scales in river ecosystems: The importance of species sorting over dispersal. Mol Ecol 26, 4351–4360. https://doi.org/10.1111/mec.14199

Yu, H., 2002. ENVIRONMENTAL CARCINOGENIC POLYCYCLIC AROMATIC HYDROCARBONS: PHOTOCHEMISTRY AND PHOTOTOXICITY. Journal of Environmental Science and Health, Part C 20, 149–183. https://doi.org/10.1081/GNC-120016203

Yu, Y., Song, X., Zhang, Y., Zheng, F., 2020. Assessment of Water Quality Using Chemometrics and Multivariate Statistics: A Case Study in Chaobai River Replenished by Reclaimed Water, North China. Water 12, 2551. https://doi.org/10.3390/w12092551

4.9 Appendix 1

Pathway analysis of 14 gene clusters in the Chaobai case study.

The *Daphnia magna* genes from 14 gene clusters were re-labelled by their corresponding ortholog groups shared with *Drosophila melanogaster* at *Arthropoda*. Statistical over-representation tests were performed with 137 *Drosophila melanogaster* pathways from KEGG database by permutation chi-squared test. The resulting *P* values are adjusted by FDR at 0.05. Only the pathways with adjusted *P* values lower than 0.05 are listed in this table.

Pathway	Description	C1120	C1200	C2100	C2200	C2351	C3000	C4401	C5100	C5200	C6100	C6200	C7000	C8000	C9000
dme00980	Metabolism of xenobiotics by cytochrome P450					0.006	0.014						0.000	0.036	
dme00982	Drug metabolism - cytochrome P450					0.007	0.014						0.000	0.042	
dme00983	Drug metabolism - other enzymes					0.018	0.002						0.005		
dme00480	Glutathione metabolism		0.013										0.014		
dme00260	Glycine, serine and threonine metabolism								0.005	0.032	0.027				0.006
dme00270	Cysteine and methionine metabolism										0.020				0.005
dme00330	Arginine and proline metabolism					0.005	0.048								
dme00340	Histidine metabolism												0.006		
dme00350	Tyrosine metabolism												0.022		
dme00360	Phenylalanine metabolism												0.003		
dme00380	Tryptophan metabolism						0.048						0.031		0.004

dme00010	Glycolysis / Gluconeogenesis			0.008											
dme00040	Pentose and glucuronate interconversions			0.004		0.005	0.010						0.000	0.028	
dme00051	Fructose and mannose metabolism												0.016		
dme00052	Galactose metabolism												0.026		
dme00053	Ascorbate and aldarate metabolism			0.003		0.004	0.009						0.000	0.016	
dme00500	Starch and sucrose metabolism													0.022	
dme00520	Amino sugar and nucleotide sugar metabolism									0.033					
dme00620	Pyruvate metabolism			0.007									0.047		
dme00630	Glyoxylate and dicarboxylate metabolism														0.003
dme00650	Butanoate metabolism						0.028		0.002						
dme00910	Nitrogen metabolism						0.014		0.001						
dme00511	Other glycan degradation												0.017	0.000	
dme00512	Mucin type O-glycan biosynthesis		0.000	0.000							0.000			0.000	
dme00513	Various types of N-glycan biosynthesis										0.042				
dme00514	Other types of O-glycan biosynthesis		0.007	0.005							0.004				
dme00531	Glycosaminoglycan degradation						0.018							0.022	
dme00534	Glycosaminoglycan biosynthesis - heparan sulfate / heparin	0.000								0.010					
dme00601	Glycosphingolipid biosynthesis - lacto and neolacto series											0.000		0.000	

dme00603	Glycosphingolipid biosynthesis - globo and isoglobo series													0.000	
dme00604	Glycosphingolipid biosynthesis - ganglio series													0.000	
dme00061	Fatty acid biosynthesis										0.008				
dme00062	Fatty acid elongation						0.027						0.017		
dme00071	Fatty acid degradation										0.022				
dme00072	Synthesis and degradation of ketone bodies								0.001						
dme00100	Steroid biosynthesis			0.000					0.000	0.000				0.000	
dme00561	Glycerolipid metabolism			0.005									0.007		
dme00564	Glycerophospholipid metabolism				0.011					0.046					
dme00565	Ether lipid metabolism				0.003					0.013					
dme00590	Arachidonic acid metabolism				0.005					0.019			0.026		
dme00592	alpha-Linolenic acid metabolism				0.002					0.007					
dme00600	Sphingolipid metabolism						0.048							0.007	
dme01040	Biosynthesis of unsaturated fatty acids						0.028						0.017		
dme00730	Thiamine metabolism												0.008		
dme00790	Folate biosynthesis												0.003		
dme00830	Retinol metabolism					0.004	0.009						0.000	0.000	
dme00860	Porphyrin and chlorophyll metabolism					0.005	0.011						0.000	0.030	

dme00440	Phosphonate and phosphinate metabolism													0.030	
dme00450	Selenocompound metabolism										0.006				
dme00900	Terpenoid backbone biosynthesis						0.027								
dme00240	Pyrimidine metabolism						0.014								
dme02010	ABC transporters									0.012			0.016		
dme04068	FoxO signaling pathway														0.002
dme04070	Phosphatidylinositol signaling system														0.008
dme04150	mTOR signaling pathway														0.021
dme04310	Wnt signaling pathway										0.015				0.018
dme04330	Notch signaling pathway										0.013				
dme04341	Hedgehog signaling pathway - fly										0.016				
dme04391	Hippo signaling pathway - fly														0.016
dme03018	RNA degradation										0.030				
dme04130	SNARE interactions in vesicular transport										0.008				
dme04141	Protein processing in dedoplasmic reticulum			0.030											0.029
dme03030	DNA replication						0.011								
dme03440	Homologous recombination				0.004						0.013				
dme03460	Fanconi anemia pathway				0.005						0.016	0.010			

dme03020	RNA polymerase										0.016				0.004
dme00970	Aminoacyl-tRNA biosynthesis										0.021				0.000
dme03008	Ribosome biogenesis in eukaryotes														0.011
dme03010	Ribosome														0.027
dme04213	Longevity regulating pathway - multiple species				0.012										0.001
dme04624	Toll and Imd signaling pathway		0.025		0.020	0.020	0.048	0.001	0.001				0.028		
dme04745	Phototransduction - fly										0.036				0.009
dme04140	Autophagy - animal	0.002		0.026	0.033				0.021						0.026
dme04142	Lysosome	0.002		0.025	0.030		0.010		0.002				0.043	0.000	
dme04144	Endocytosis														0.038
dme04145	Phagosome	0.001			0.014				0.009						
dme04146	Peroxisome		0.025	0.016	0.020	0.020	0.049		0.013						0.016

5 Danube Case Study

5.1 Abstract

Organic micropollutants in the natural rivers have great impact on the health of humans and nontargeted species. To identify harmful chemical component in the environment by assessing the biological effects of the sampled environment is a complicate problem. Here the relative toxicity of organic chemical components in the waters sampled from the Danube River (Europe) were assessed based on gene expression of exposed model test species, *Daphnia magna*. Unsupervised method like clustering was applied to cluster the transcriptomic data into multiple co-responsive gene clusters then group sampled waters within gene clusters. The functional roles of gene clusters were determined by an ortholog-based cross-species extrapolation method with pathway overrepresentation analysis. In this case study, similarity analysis of chemical profiles and transcriptomic profiles reveal similar grouping pattern, as expression-based clustering analysis of gene clusters revealed that distinctive transcriptomic profiles of two sites (D11 and D12) reveal down-regulation of xenobiotic biodegradation and antioxidative response pathways.

These results demonstrated the feasibility of classifying the biological effect of exposure to environmental chemical mixtures and assessing the joint effects of specific compounds based on gene expression with a whole-mixture approach.

5.2 Introduction

Organic micropollutants raise considerable concerns as complex environmental pollutants pose health threats on both human and wildlife (Brunsch, 2021). Statistics report from EUROSTAT in 2020 revealed that more than 50 % of the total chemical production in European countries are environmentally harmful compounds (EUROSTAT, 2020). Previous environmental monitoring showed that organic micropollutants detected in the parts of the Danube River induce estrogenic activation, xenobiotic metabolism (pregnane X receptor), ligan-dependent transcription factor activation (aryl hydrocarbon receptor), inflammation (NF-κB-bla) and oxidative stress (Neale et al., 2015), where various *in vivo* test systems are used to detect toxicological effects from exposure to the water. Minor genotoxicity was detected in the lower reach of the Danube River water samples via umuC testing with prior S9 activation (Kittinger et al., 2016). The DNA damage was also detected in mussels from the middle reach of the Danube River, which were thought to be affected by untreated wastewater effluents associated with the distribution of polycyclic aromatic hydrocarbons, dioxin and emerging pollutants (Oxazepam, Chloridazon-desphenyl) (Kolarević et al., 2016). Seventeen pesticides and contaminants derived from wastewater were detected in invertebrates (gammarids) collected from the Danube River (Inostroza et al., 2016). The measured concentration of emerging pollutants like perfluorooctanesulfonic acid (PFOS, new priority substance of the WFD, 2013/39/EU, EU, 2013) and diclofenac (pharmaceutical, new priority substance of the WFD) exceeded the environmental quality standard threshold in the upper (JDS20, the upper reach of D01 in Figure S5.1) and lower (JDS58, between D09 and D10 in Figure S5.1) reach, respectively (Loos et al., 2017). Thus, organic micropollutants from multiple sources, especially wastewater

treatment plants, are believed to pose a health hazard on target and nontarget aquatic organisms.

To evaluate the joint effect of an environmental chemical mixture in a holistic way, omics-based bioassays can be applied to capture the biological signatures of chemical effects at the molecular level and interrogate the global effects of chemicals on biomolecular pathways linked to health. Environmental chemicals may induce alternation in inter-correlated biological processes, such as xenobiotic metabolism and stress response. Xenobiotic metabolism is responsible for detoxification and biotransformation of exogenous substances, which is represented by biomarkers that are diagnostic of these pathways, such as cytochrome P450 (CYP), ATP-binding cassette transporter (ABC), glutathione S-transferase (GST), and glutathione peroxidase (GPX) (Hassan et al., 2015). In transcriptomes, pronounced expression in these biomarkers might be interpreted as activation of xenobiotic defence (Campos et al., 2014). Exposure to environmental chemicals might also trigger oxidative stress responses; enhanced expression of glutathione reflects activated antioxidant defence (Regoli and Giuliani, 2014). Bioactivity of glutathione, GST and glutathione reductase play important roles in neutralising reactive oxygen species and avoiding further damage caused by exogenous compounds (Oliveira et al., 2015), which are considered as the biomarkers of oxidative stress response in *Daphnia magna* (Barata et al., 2005).

To classify the biological effects of environmental chemicals based on toxicity pathways, unsupervised learning methods like clustering may be used to identify groups of co-variant genes (gene-based clustering) or groups of homogeneous samples (sample-based clustering). Genes that share similar expression patterns are

often assumed to be under the control of shared regulatory pathways, and therefore functionally related and biologically relevant (Gasch et al., 2000). Hierarchical clustering algorithms like DIANA and the Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH, Pollard, 2005) methods generate a hierarchical clustering tree, which identifies gene clusters within the transcriptome. The reasons that the HOPACH algorithm might outcompete the DIANA algorithm is that the DIANA requires manual selection of a height value for cutting the tree to determine the number of clusters in a hierarchical clustering tree, which can be problematic (Slonim, 2002); while the HOPACH automatically finds the optimal number of gene clusters from their expression patterns at each level of the clustering tree based on the Median Split Silhouette criterium, resulting in a robust clustering pattern. Further sample-based clustering analysis may characterise the grouping patterns of each gene cluster so that the structure of the gene expression data can be revealed in greater details.

The co-responsive clusters establish the basis of structuralising the omics data into multiple co-varying sets that links the mathematical modelling (pairwise correlation of biological features) with functional relationships (co-regulation of biological features) (Josyula et al., 2020; Kustatscher et al., 2019). Chemical components in a mixture may affect the biological systems in an independent and additive way, as the molecular features that are associated with one specific chemical component in single substance-based exposure testing may be observed in the mixture exposure testing. If the chemical substance is bioactive and the associated features reveal concentration-dependent response, the linear combination of responses of these associated features may be also correlated to the concentration levels of the chemical components in the mixture. Such chemical-associated features can be identified via multi-block

correlation analysis, which identifies the linear combination of biomolecular features that are correlated with the chemical components of interest within the chemical mixtures (Tenenhaus et al., 2014; Tenenhaus and Tenenhaus, 2011). The multi-block correlation analysis exemplified by the Canonical Correlation Analysis (CCA; Jun et al., 2018) studies the inter-connection between multiple data sources. The sparse version of Regularized Generalized Canonical Correlation Analysis (RGCCA/SGCCA) is particularly appropriate in this case as it may find the subset of chemical features that are linearly correlated with the individual gene cluster. In that case, those identified chemical features may be associated the gene cluster, which provides the insight of association between biological responses and chemical factors.

As each gene cluster may consist of genes of similar function, the grouping patterns of the gene clusters may assist in distinguishing the general differences in the overall transcriptomic profiles with respect to their biological roles, so that the functions potentially perturbed by the environmental chemicals may be revealed at a systematic level. Biological interpretation of the gene clusters may require comprehensive information of gene function and pathway. A cross-species extrapolation can be employed to annotate genes of the poorly defined species by referred to well-studied species based on their orthologs. As the functions of unknown genes can be putatively annotated by the corresponding OGs' function, the gene-pathway association can be thereby transformed into an OGs-pathway association. If the OGs composition of every pathway is unique, the OGs-pathway association can be used to (1) distinguish different pathways and (2) applied as the reference data, similar to gene sets serving as background knowledge in the pathway overrepresentation analysis.

The Danube River is selected for this case study. I combined targeted chemical analysis of river water with non-targeted transcriptomics and metabolomics measurements of exposure-related effects on *Daphnia magna* using sampled surface river samples from 12 sites along the Danube River Basin. The biological effects of these organic extracts were assessed at environmental relevant levels. The specific objectives of this investigation are to (1) identify gene clusters within the transcriptome, and (2) functionally annotated the gene clusters via pathway overrepresentation analysis, and finally (3) identify associations between chemical profiles and gene clusters with sparse CCA modelling.

5.3 Methods

5.3.1 Site description

The Danube River is 2850 km long, which ranks the second in Europe. The Danube River Basin is more than 800,000 km^2 and sustains over 80 million people along the river (Liška *et al.* 2015). It is an essential source of drinking water in 10 countries in Central and Eastern Europe. This transborder river originates from the Black Forest in Germany, flows through or touches the borders of Bavaria, Austria, Slovakia, Hungary, Croatia, Serbia, Romania, Bulgaria, Moldova and Ukraine, then empties into the Black Sea via the Danube Delta in Romanis. The river water stems from the glaciers and precipitation of the Alps and the Carpathian Mountains with additional freshwater from its tributaries (like the Sava River, the Tisza River and the Drava River), and effluents from municipal wastewater treatment plants (WWTPs) along the river. As the Danube River flows through populated areas, organic micropollutants like pharmaceuticals,

pesticides, personal care products and industrial compounds are discharged from wastewater treatment plants, urban runoffs and agricultural outlets.

5.3.2 Sampling regime

The environmental samples used in the Danube River case study are from the Joint Danube Survey 3 (JDS 3), which was the most significant river research expedition ever conducted in 2013 (Liška *et al.* 2015). Water sampling was completed in the JDS 3 from 13th August to 25th September in 2013. The stretch of the Danube River investigated in this study was 2581 km long, which flowed across eight countries of Central and Eastern Europe. A total of 68 sites were sampled during this expedition from the upper, middle and lower reaches of the river, mostly in the mainstream and major tributaries; only 12 of them were selected in this case study. The sampling map of 12 selected sites are revealed in Figure S5.1, and the detailed information is listed in Table S5.1. Over 500 L of surface water were pumped into the stainless-steel tank filled with sorbents for neutral, anionic, and cationic ions for each site, according to the description in Schuzle *et al.* (2017). The SPE was performed on-site with the large volume solid phase extraction (LVSPE) device by the JDS3 team (Schulze et al., 2017). The elutes were dried under nitrogen and stored at - 20 °C. The dried extracts were then shipped to the University of Birmingham, maintained in methanol and stored - 20 °C.

5.3.3 Chemical analysis

The chemical profiles in the Danube case only account for the profiles of organic chemical mixtures in the water samples. Targeted chemical analysis was performed on the same water samples for the polar organic substances, including pesticides, pharmaceuticals and their transformed products.

Non-targeted screening analysis of organic substances was performed by Ultimate 3000 LC system (Thermo Scientific) with a quadrupole-Orbitrap MS (QExactive Plus, Thermo Scientific), according to Hashmi *et al.* (2019). Valid peak lists were generated by MZmine 2.21 (Pluskal et al., 2010). All the non-targeted screening data were provided by collaborators in the SOLUTIONs project (Escher *et al.* 2014), thanks to Dr Tobias Schulze from Helmholtz Centre for Environmental Research (UFZ). Among these peaks, 91 organic micropollutants were quantified refer to internal standards (semi-quantification). The detailed information of these 91 semi-quantified organic substances is listed in Table S5.2.

5.3.4 Daphnia strain and culture conditions

The *Daphnia magna* (Bham2 strain) isolate was once again used in this study, permitting consistency with the Chaobai Case study for comparative analyses. This isolate has been maintained for more than ten years by the Environmental Genomic Groups at the University of Birmingham. The culture medium was prepared by filtering borehole water through charcoal and maintaining it at 20 ± 1℃ overnight before use. The borehole water was collected near the channel flowing through the Edgbaston campus of the University of Birmingham (52°27'20.08"N, 1°55'43.81"W). The stock culture of *Daphnia magna* (Bham2) was incubated in the borehole waters at 20 ± 1℃ with a photoperiod of 14h: 10h (light: dark). The main food source was *Chlorella vulgaris* Beijerinck (1890) strain 211/11B, which was fed 0.5 mg per 10 daphnids per day. Neonates of *Daphnia magna,* which hatched within 12-24 h from the 2nd to 3rd broods of the same 900 matured *Daphnia* population, were collected for exposure testing.

5.3.5 Organic chemical mixture exposure of neonates

The organic chemical mixtures of each river water samples were used for my exposure experiments. Organic extracts from 12 selected sites were maintained in methanol and stored at - 40 ℃. Based on previous results listed in Table S5.1, extracts from all 12 sites except for D02 (EC50 at 83x) had an EC50 value over REF 100x; no immobility incidences were recorded at the REF 1x (original environmental level). The relative enrichment factor (REF, Escher *et al.* 2014) refers to the ratio between the volume of extracted river water and the volume of re-suspended solvent; for example, REF 100 means enriching the concentration level 100 times the environmental level. All the organic extracts were diluted in borehole media to generate a stock solution of REF 1x (reproducing the environmental level). And the amount of methanol solvent within the exposure solutions was normalised to 0.08 % (v: v). The borehole media with 0.08 % methanol (v: v) was referred as the reference (negative control) level in this case, as methanol being the carrier of the organic chemical mixtures.

For each treatment group, fifteen neonates were transferred to a 20 ml glass vial with 15 ml exposure solutions (spike-in borehole media); and each treatment group had six biological replicates. The control treatment (borehole media with 0.08% methanol) had twenty-four biological replicates. After 48h exposure, all the exposed neonates were flash-frozen with liquid nitrogen and stored at -80 ℃ before transcriptome and metabolome extraction. A total of 96 samples were prepared for downstream multi-omics profiling.

5.3.6 Total RNA extraction and transcriptome sequencing

For the Danube River case study, the frozen pooled neonates for each biological replicate contained 15 exposed neonates. Frozen pooled neonates were homogenised

in GenoGrinder (SPEX SamplePrep, U.S.A.) for 90 seconds at the speed of 1750 rpm. Total RNA extraction was performed using the Agencourt RNAdvance Tissue Total RNA kit (Beckman Coulter, U.S.A.), as the total RNA was captured onto magnetic beads, washed twice for removing unwanted salts, and eluted in 100 µl RNAse-free H_2O, following the manufacturer's instructions. The concentration of total RNA concentrations was quantified by Nanodrop 8000 Spectrophotometer (Labtech Ltd., U.K.). The quality of extracted total RNA, both integrity and purity, was measured on TapeStation 2200 (Agilent Technologies, U.S.A.). A cDNA library was generated for each sample from 150 ng of RNA using NEBNext Ultra II Directional RNA Library Prep Kit for Illumina, following the manufacturer's instructions. All of the sample libraries were then normalised to the same molecular weight and pooled together using the adapter indices supplied by the manufacturer. Transcriptome sequencing (RNA-seq) was performed on the DNBseq at BGI.

5.3.7 Sequence pre-processing

Reads from the two case studies were processed separately. Raw reads were firstly trimmed in Trimmomatic (version 0.32; Bolger *et al.* 2014) to remove sequencing adapter and obtain sequences with phred scores of more than 30. FastQC (version 0.11.9; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) were used to screen the overall sequence quality. Transcript reads were mapped to an established transcriptome reference database (Orsini *et al.* 2016) using Salmon (Version 0.8.2; Patro *et al.* 2017). The quasi-mapping function of Salmon was used with GC and positional bias correction. For each sample, both paired ends from two lanes were run together. Then the mapped transcript reads were processed in R (version 4.0.3). Low count reads (reads with raw count lower than 10) and outlier samples were identified

and removed from the data set. The read counts were normalised by the size factor defined in the DESeq2 package (version 1.30.0; Love *et al.* 2014) in this case. The log2 fold changes of individual genes per treatment level were further calculated by the DESeq2 package against control level (negative control).

5.3.8 Similarity analysis of transcriptomic and chemical profiles

Log2 fold changes of transcripts and concentration levels of targeted chemicals were used for similarity analysis. Principal component analysis (PCA; Konishi T. 2015) was used to reveal the overall similarity based on the first two principal components, which explained a considerable proportion of the overall variance. Hierarchical clustering analysis (HCA; Eisen *et al.* 1998) was conducted based on the Euclidean distance with the ward.D2 clustering method. Pearson correlation coefficient was calculated in pairwise treatment levels to reveal the co-variation (another perspective of similarity) of any two treatment levels.

5.3.9 Gene cluster identification

The highly variable genes were selected by scran package with the normalised gene counts of both case studies (version 1.18.7; Lun *et al.* 2016). Selected genes were clustered by Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH) algorithm in the hopach package (version 2.52.0; Pollard and Van Der Laan 2003) on R. Cosine distance was chosen to capture the similarity between any two genes, as suggested in Eisen *et al.* (1998). Sample bootstrapping was performed to confirm the variability of the composition of gene clusters. The median value of genes belonging to the gene cluster was used as the reference value. For each pseudo-replication, samples were randomly selected to generate a new cluster pattern based on the gene cluster reference values. The gene cluster assignments were recorded. The

frequencies of genes assigned to each cluster were summarised from the records of 10,000 repeats. Gene clusters were further used for clustering the samples based on Euclidean distance measurement with the ward.D2 clustering method.

5.3.10 Pathway analysis of co-responsive modules

A cross-species KEGG pathway overrepresentation test was performed for *Daphnia magna* gene set pathway analysis. The *Daphnia magna* genes in the transcriptomic co-responsive modules were re-annotated by their corresponding ortholog group IDs, based on the orthologous relationships between *Daphnia magna* and *Drosophila melanogaster* from the OrthoDB database (v10.1; Kriventseva *et al.* 2019). A permutation chi-square test was performed over 100,000 iterations to generate a robust P-value estimation directly from resampling detected *Daphnia* genes annotated with ortholog groups. The P-values of the permutation chi-square tests were further corrected following the Benjamini-Hochberg procedure with a false discovery rate at 0.05 (Benjamini and Hochberg 1995).

5.3.11 Correlation analysis of eigengenes and chemical components

The eigengene of each gene cluster is the first principal component of the gene cluster matrix. Pearson correlation coefficients were calculated between each eigengene and individual chemical, in order to identify close associations between chemical component and gene clusters.

5.3.12 Sparse CCA modelling of chemical components and gene clusters

The correlations between the individual chemical and transcriptomic features were identified with sparse Canonical Correlation Analysis (sCCA) algorithm (Tenenhaus et al., 2014; Tenenhaus and Tenenhaus, 2011). The sCCA is applied to identify the subset of transcriptomic features linearly correlated to the chemical distribution pattern

when projected to a common latent space. With a certain sparsity level (between 0 and 1), subsets of chemical features were selected. I assumed that only a few chemical features (with similar mode of action) were associated with individual gene cluster (functional unit). The loading values of selected chemical components were plotted so that chemical components that are consistently selected are proposed to be associated with the gene clusters and pathways within.

5.4 Results

5.4.1 Chemical analysis

The principal component analysis (PCA) plot reveals the general similarity of measured chemicals among 12 water samples in the Danube River case. It is obvious that site D11 and D12 are different from the rest. Table S5.3 summarise the relative contribution of each chemical factor to the first two components. Of the first component, 46 % of the total variance is contributed by 14 organic chemicals (ACF, BP4, CBZ, CDZ, DHC, FAA, HEX, MBT, MEC, MLC, MLCE, PNZ, PSA, SUC), which may account for the differences between D11 and the rest. Of the second component, 44 % of the total variance is contributed by the other 9 organic micropollutants (BEN, CAR, CPP, DIA, MFM, OXA, SIM, SMX, TBH)., which may account for the differences between D12 and the rest.

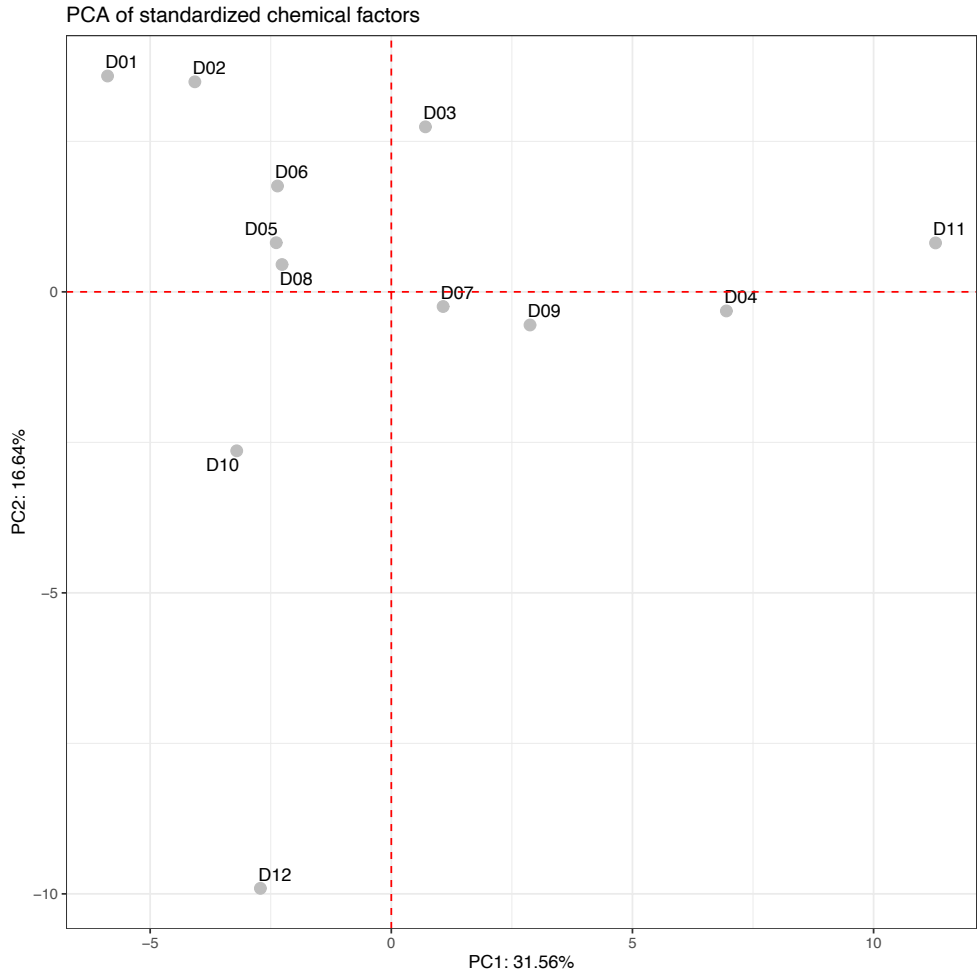

Figure 5. 1 PCA plot of targeted chemicals in water samples of the Danube River. The sampling site names are described in Table S5.1 and Figure S5.1.

5.4.2 Immobility rate of 48 hours exposure

None of the tested water samples observed immobilised neonates after 48 hours of exposure.

5.4.3 Similarity analysis of the transcriptomics variations

The mapped transcriptomics data that were obtained from the Danube River experiments are summarised at the gene level (Figure S5.2). With an average mapping rate of 83 %, each sample retrieved 6.61 million reads, on average. By removing low count genes and outlier samples, a total of 13,670 genes were selected for the

downstream analysis. To evaluate the similarity of the transcriptomic variations induced by organic extracts from the 12 sites, principal component analysis was performed based on the log2 fold change profiles of 12 site treatments (Figure 5.2a). The PCA plot showed that both the first (63.8 % of total variance) and the second (14.3 %) principal components could clearly distinguish D11 and D12 from the rest of the ten sites, suggesting the organic mixtures collected from the lowest reaches of the Danube River induce different transcriptomics variation in the exposed daphnids compared to the upper reaches of the river. The HCA plot (Figure 5.2b) also reveals a similar pattern as the D11 and D12 sampling sites formed a unique branch apart from the others. The pairwise Pearson correlation coefficient matrix (Figure 5.2c) revealed relatively lower similarity among the upstream sites (average at 0.56) but higher similarity between two downstream sites, D11 and D12 (0.83).

5.4.4 Identify gene clusters of the highly variable genes

Again, by using the scran package, 1451 genes were identified as highly variable genes. The gene clusters of these highly variable genes were identified by the HOPACH algorithm, produced a total of 99 clusters based on the observed similarities in the log2 fold change pattern of selected genes, including 85 smaller clusters and 14 larger clusters (containing more than 20 genes). Among 14 larger clusters, cluster D4000 was the largest consisting of 150 genes, followed by cluster D3100 (136 genes) and cluster D7100 (134 genes). The detailed information of 14 genes clusters is listed in Table S5.4.

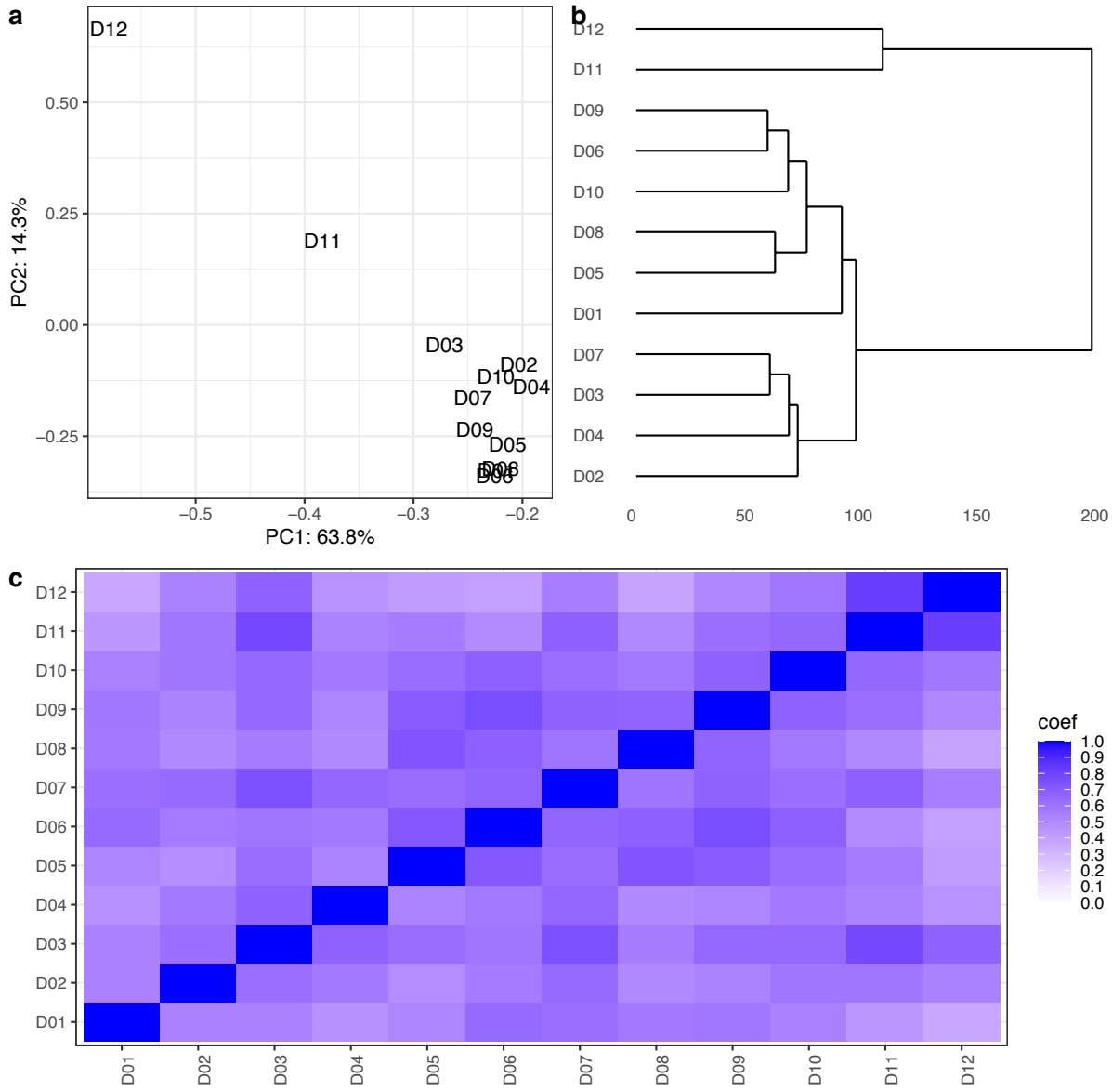

Figure 5. 2 Similarity analysis of log2 fold change patterns in transcriptomic profiles of the Danube case. (a) Score plot of Principal Component Analysis, (b) Dendrogram of Hierarchical Clustering Analysis, and (c) Heatmap of pairwise Pearson correlation coefficients (coef). Sample site locations are shown in Figure S5.1.

The robustness of the gene clustering pattern was also evaluated by sample bootstrapping for 10,000 iterations, and the gene reappearance frequencies are plotted in Figure S5.3. Among those 14 larger clusters, most of the gene clusters had a large

portion of genes with relatively higher reappearance frequencies. For example, cluster D2200, D5000 and D1440 had relatively higher average reappearance frequency levels, at 87 %, 82 % and 80 %, respectively. By contrast, clusters D1270, D7100, and D7200 only had the average reappearance frequencies of 34 %, 26 %, and 23 %, respectively.

5.4.5 Functional analysis of gene clusters

The adjusted *P* values of overrepresentation tests on all KEGG pathways are listed in Appendix 1. The adjusted *P* values of pathways related to xenobiotic metabolism are summarised and plotted in Figure 5.3.

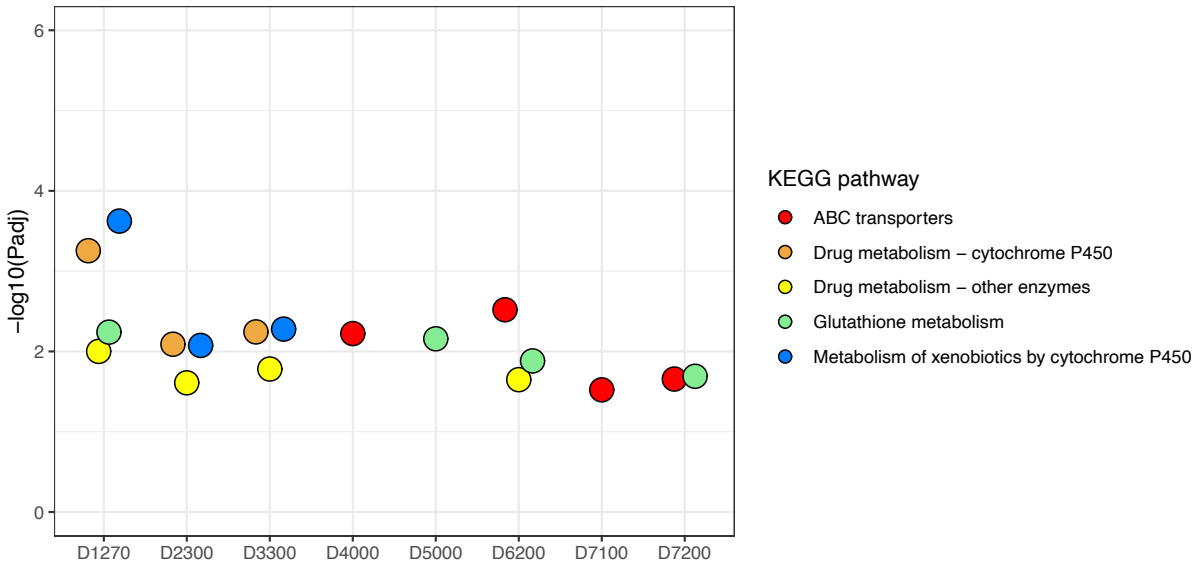

Figure 5. 3 Overrepresentation analysis of xenobiotic metabolism-related pathways among the selected Danube River gene clusters. The significance levels of selected pathways are shown in the plot with their negative logarithms (-log10) of the adjusted P values

Figure 5.3 shows that the xenobiotic biodegradation and metabolism pathways were significantly enriched in clusters D1270, D2300, and D3300. The drug metabolism

pathway was reported to be significantly enriched by genes from D6200. The ABC transporter related to transmembrane transportation was significantly enriched by genes within clusters D4000, D6200, D7100 and D7200. Genes that function in the antioxidant defence system represented by glutathione metabolism were overrepresented in clusters D1270, D4000, D5000, and D7200. Thus, the functional role of genes within clusters D1270, D2300, D3300, D4000, D5000, D6200, D7100 and D7200 might be closely related to xenobiotic detoxification and oxidative stress response.

In addition to the pathway enrichment results for the six clusters in Figure 5.3, autophagy-related pathways were significantly enriched in clusters D2100, D5000 and D6200. Pathways related to glycan biosynthesis and amino acid metabolism were also found to be significantly enriched by genes within cluster D2100 (Appendix 2). Fatty acid degradation and serine metabolism were also found to be significantly enriched by genes within cluster D5000. Jointly the results suggested that clusters D2100, D5000 and D6200 might be related to the turnover of cellular substances like amino acids and fatty acids.

5.4.6 Clustering pattern of xenobiotic-related gene clusters

The HCA plots of 14 gene clusters and a combined set of 13 other gene clusters are shown in Figure 5.4. Extracts from sampled waters from the D12 site drastically induced down-regulation (negative log2 fold change values) in gene clusters D2100, D2200, D2300, D3100, D3200, D3300 and D5000, while D11 shared a similar regulation pattern with D12 in gene clusters D2100, D2200, D2300, D4000 and D6200. Down-regulation in D2100, D2300, D3300, D4000, D5000 and D6200 suggested a reduced expression level of xenobiotic biodegradation pathways and potent inhibition

of antioxidant defence pathways, the distinctive transcriptomics profiles induced by organic extracts from D11 and D12 might be associated with lower levels of xenobiotic metabolism compared to the other sites. While comparing the expressional patterns of D11 and D12, cluster D1270 reveal different patterns.

5.4.7 Correlation analysis between eigengenes of 14 gene clusters and chemical factors

The eigengene is the first principal component of the gene cluster matrix. As genes in each gene cluster share similar variation pattern across all the samples, As revealed in Figure 5.5. it is obvious that cluster D1270 associated with five organic pollutants, including CAR, CHL, CPP, DIA, and SIM. Cluster D2100 and D2200 are significantly correlated with DPP. While D2300, D4000 and D6200 are not significantly correlated with any of these organic pollutants.

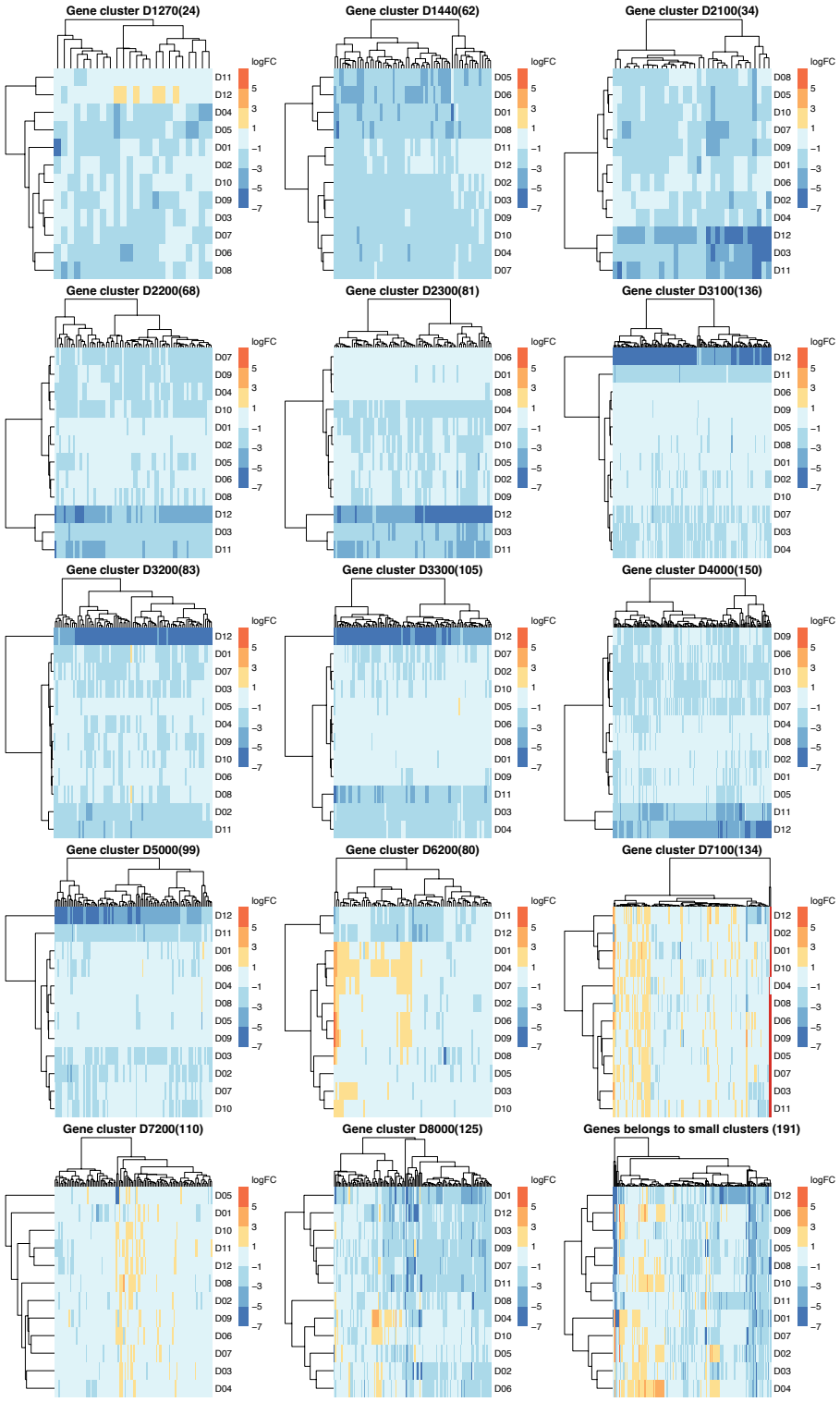

Figure 5. 4 Hierarchical clusterings of transcriptomic profiles of selected Danube gene clusters. All the genes were represented by their log2 fold change (logFC) values

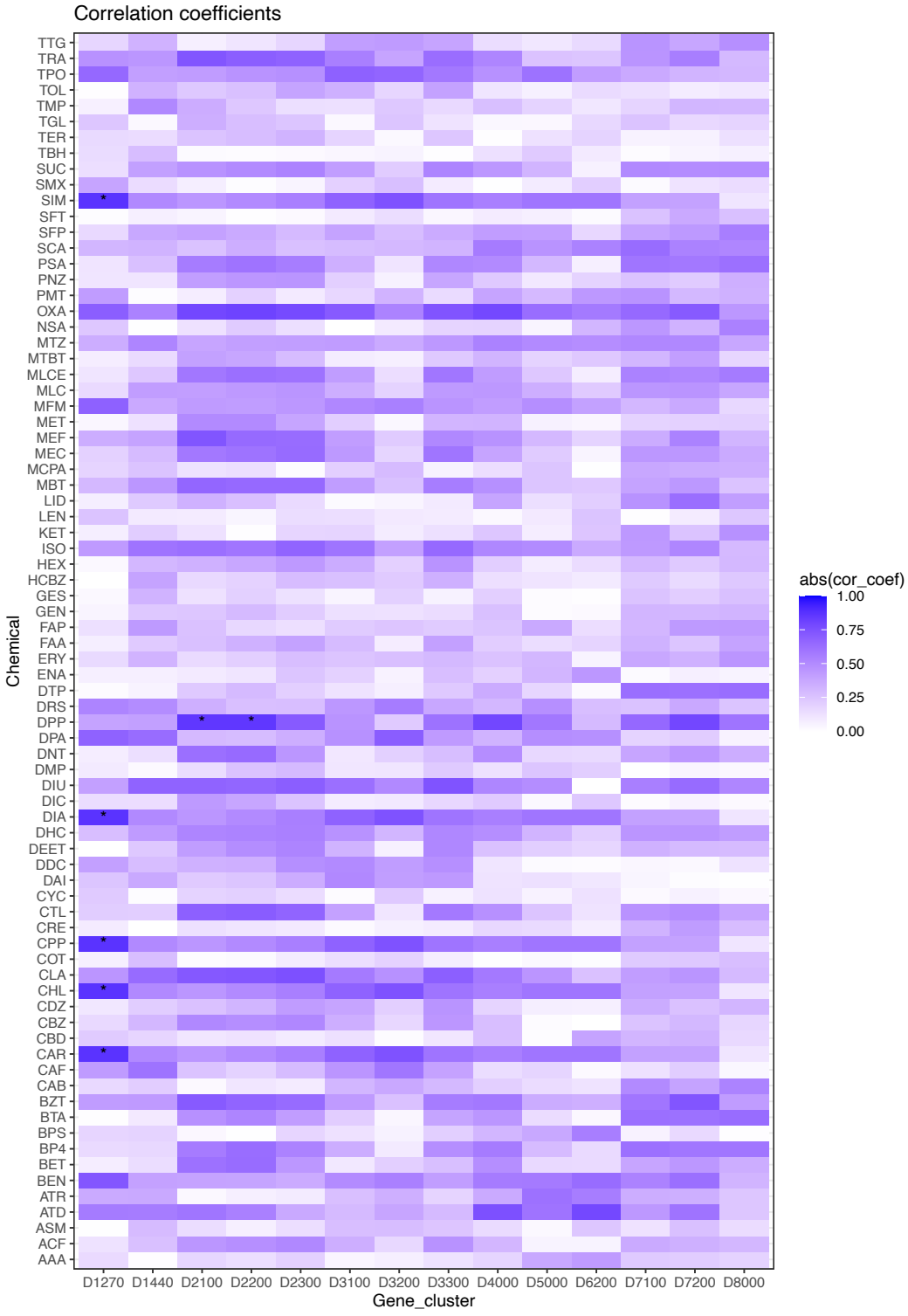

Figure 5. 5 Correlation analysis between eigengenes of 14 gene clusters and chemical factors. The color code corresponds to the absolute value of Pearson correlation coefficient values. The asterisks highlight the significant correlation with P values lower than 0.05.

5.5 Discussion

In this study, organic micropollutant mixtures extracted from 12 water samples collected from the Danube River were used as exposure and transcriptomic profiles of exposed waterflea (*Daphnia magna*) were measured to characterise the biological effects. Similarity analysis of chemical profiles and transcriptomic profiles reveal similar grouping pattern, as two sites in the downstream (D11 and D12) reveal dissimilar chemical compositions and transcriptomic profiles (log2 fold changes) against the rest sites.

To be specific, the organic extracts from sampled waters of D11 and D12 induced down-regulate xenobiotic metabolic activities and potent inhibition of antioxidant defence compared to the other sites, represented by the negative log2 fold change (down-regulation compared against the control level) in six gene clusters significantly enriched in xenobiotic metabolism. Chemical measurements revealed that relatively higher levels of Sulfamethoxazole (SMX, 25 ng/L), Atrazine (ATZ, 11 ng/L), Terbutylazine (TER, 14 ng/L), Bentazone (BEN, 35 ng/L), and Metazachlor (MTZ, 26 ng/L) were detected in D11; while Carbamazepine (CBZ, 12 ng/L), 1H-Benzotriazole (BZT, 82 ng/L), 5-Methyl-1H-Benzotriazole (MBZ, 87 ng/L), Acesulfame (ACF, 410 ng/L), Atrazine (ATZ, 6 ng/L), Cotinine (COT, 23 ng/L), n-Acetyl-4-aminoantipyrine (AAA, 80 ng/L) were detected in D12. The composition of detectable substances among these 16 organic pollutants was highly dissimilar, even though the similarity of the overall transcriptomic profiles of D11 and D12 was relatively high (Pearson correlation coefficient at 0.83, in Figure 4.11c), suggested that the dissimilarities in the chemical compositions (among the 16 detected organic substances) might not induce observable differences in overall transcriptomic expression patterns. Based on the

EC50 values listed in Table 3.7 (Chapter 3), none of these organic substances exceeds their EC50 levels, which may partially explain the zero-immobility rate observed after 48 hours of exposure.

Multivariate association analysis between chemical features and gene clusters are revealed by sparse CCA modelling. To be specific, D1270 is significantly correlated to carbaryl (carbamate insecticide), chlorophene (biocide and preservatives in cosmetics), chlorpropham (herbicide), diazinon (insecticide), and Simazine (herbicide). While D2100 and D2200 is significantly correlated with diphenylphosphate (an additive for paints and coating). Previous studies revealed that pharmaceuticals like CBZ are known endocrine disruptors at 10 to 200 µg/L (Oropesa *et al.* 2016). My results agreed with a previous study that the expression levels of features related to glutathione metabolism are reduced after 48-hour exposure to CBZ (Nkoom *et al.* 2019). The activity of glutathione reductase, which relies on the glutathione as substrate, might also be reduced as *Daphnia* was overwhelmed by the CBZ-related stress (Li *et al.* 2009). However, a previous study reported that ATZ at environmentally relevant levels (4 to 8 µg/L) triggered up-regulation of glutathione metabolism and xenobiotic metabolism after 96-hour exposure (Schmidt *et al.* 2017). The disagreement between my observation and theirs might be related to the differences in concentration levels of ATZ. The EC_{50} level of SMX was at 205.2 mg/L for 48 hours of exposure (Jung *et al.* 2008), and chronic exposure of SMX at 10 µg/L (accompanied with four other antibiotics) may affect the sex ratio of the first brood (Flaherty and Dodson 2005). However, in my studies, the highest concentration level was far below the EC_{50} levels, which might not induce a detrimental effect on the sex ratio. Thus, it is still unclear which organic substances may be the cause of the downregulation in the xenobiotic

metabolism compared against the control level. Further investigation of the organic substances accounted effects will be needed to identify the combinations of organic chemical components that may trigger lower levels of xenobiotic activities in the exposed *Daphnia*.

5.6 Conclusion

In this Danube River case study, transcriptomic profile is used to characterise the effects of environmental chemical mixtures from a natural river in China. Genes clusters support the differences between site D11 and D12 and the rest are closely related to xenobiotic metabolism and stress response. And these clusters are significantly correlated to different concentration levels of biocides. Further verification study will be needed to confirm the relative contribution of biocides, which may help define acceptable environmental exposure levels.

5.7 Supplementary

Tables and Figures

Table S5. 1 Description of 12 selected sites along the Danube River Basin. Site ID is the water sample labels used in this current study. JDS ID is the water sample labels used by the Joint Danube Survey 3 expedition.

Site ID	JDS ID	Sampling site	Reach	REF [a] EC50 [b]
D01	JDS27	Hercegszanto	Upper	> 200
D02	JDS32	Novi Sad upstream	Middle	83
D03	JDS36	Tisa downstream/ Sava upstream (Belegis)	Middle	> 200
D04	JDS37	Sava	Middle	139
D05	JDS39	Pancevo downstream	Middle	> 200
D06	JDS44	Irongate reservoir (Golubac/Koronin)	Lower	101
D07	JDS53	Zimnicea downstream/Svishtov	Lower	> 200
D08	JDS55	Jantra downstream	Lower	> 200
D09	JDS57	Ruse downstream/Giurgiu	Lower	> 200
D10	JDS60	Chiciu/Silistra	Lower	> 200
D11	JDS64	Prut	Lower	> 200
D12	JDS67	Sulina	Lower	167

a. REF stands for relative enrichment factor, and it is calculated by the relative enrichment factor as the ratio between the extracted water volume and the elution volume (Escher *et al.* 2013).

b. EC50 stands for effective REF-based concentration level that induces 50% immobility rate in *Daphnia magna* exposure experiments.

Table S5. 2 Nontargeted screening analysis of organic substances in surface water samples from the Danube River Basin. EC50 values are recorded in ECOTOX knowledge database.

Chem ID	ChemName	CAS	EC50 (mg/L)	Chem ID	ChemName	CAS	EC50 (mg/L)
DHC	10,11-Dihydro-10-hydroxycarbamazepine	29331-92-8	/	ENA	Enalapril	75847-73-3	0
DDC	10,11-Dihydro-10,11-dihydroxy-carbamazepine	58955-93-4	/	ERY	Erythromycin	114-07-8	207.83
BZT	1H-Benzotriazole	95-14-7	15.8	GEN	Genistein	446-72-0	0
CAB	2-(2-(Chlorophenyl)amino)benzaldehyde	71758-44-6	0	GES	Gestoden	60282-87-3	0
MTBT	2-(Methylthio)benzothiazole	615-22-5	0	HEX	Hexa(methoxymethyl)melamine	68002-20-0	0
BTA	2-Benzothiazolesulfonic acid	941-57-1	0	ISO	Isoproturon	34123-59-6	1
HCBZ	2-Hydroxycarbamazepine	68011-66-5	0	KET	Ketoprofen	22071-15-4	0
NSA	2-Naphthalenesulfonic acid	120-18-3	0	LEN	Lenacil	2164-08-1	0
DPA	2,4-Dichlorphenoxyacetic acid	94-75-7	0	LID	Lidocaine	137-58-6	0
DTP	2,4-Dinitrophenol	51-28-5	4.39	LOR	Lorazepam	846-49-1	0
FAP	4-Formyl-antipyrine	950-81-2	0	MCPA	MCPA	94-74-6	180
MBT	5-methyl-1H-benzotriazole	136-85-6	51.6	MEC	Mecoprop	93-65-2	0
ACF	Acesulfame	33665-90-6	0	MEF	Mefenamic acid	61-68-7	0
AMP	Acetamiprid	135410-20-7	50	MFM	Metformin	657-24-9	64
ASM	Acetyl-Sulfamethoxazole	21312-10-7	0	MLC	Metolachlor	51218-45-2	4.25
ATR	Atrazine	1912-24-9	4.6	MLCE	Metolachlor ESA	171118-09-5	0
BEN	Bentazone	25057-89-0	0	MET	Metoprolol	37350-58-6	0
BP3	Benzophenone-3	131-57-7	0	AAA	N-Acetyl-4-aminoantipyrine	83-15-8	0
BP4	Benzophenone-4	4065-45-6	0	FAA	N-Formyl-4-aminoantipyrine	1672-58-8	0
BTZ	Benzothiazole	95-16-9	0	OXA	Oxazepam	604-75-1	0
BF	Bezafibrate	41859-67-0	75.79	NIT	p-Nitrophenol	100-02-7	4.7
BPS	Bisphenol S	80-09-1	0	TOL	p-Toluenesulfonamide	70-55-3	0
CAF	Caffeine	58-08-2	177.8	PFA	Perfluoroheptanoic acid	375-85-9	1019
CBZ	Carbamazepine	298-46-4	77.7	PNZ	Phenazone	60-80-0	0.117
CAR	Carbaryl	63-25-2	0.00225	PSA	Phenylbenzimidazole sulfonic acid	27503-81-7	0
CBD	Carbendazim	10605-21-7	0.0876	PMT	Prometryn	7287-19-6	9.7
BET	Cetirizine	83881-51-0	0	PPZ	Propyphenazone	479-92-5	0
CDZ	Chloridazon	1698-60-8	0	ROX	Roxithromycin	80214-83-1	0
CHL	Chlorophene	120-32-1	0.59	SCA	Salicylic acid	69-72-7	870
CTL	Chlorotoluron	15545-48-9	0	SIM	Simazine	122-34-9	1
CPP	Chlorpropham	101-21-3	3.7	SUC	Sucralose	56038-13-2	0
CHO	Cholic acid	81-25-4	0	SFT	Sulfamethazine	57-68-1	31.4
CLA	Clarithromycin	81103-11-9	8.16	SMX	Sulfamethoxazole	723-46-6	96.7
CFA	Clofibric acid	882-09-7	72	SFP	Sulfapyridine	144-83-2	0

COT	Cotinine	486-56-6	0	TBN	Tebuconazole	107534-96-3	2.88
CRE	creatinine	60-27-5	0	TER	Terbuthylazine	5915-41-3	21.2
CYC	Cyclamate	100-88-9	0	TBH	Terbuthylazine-2-hydroxy	66753-07-9	0
DAI	Daidzein	486-66-8	0	TTG	Tetraglyme	143-24-8	0
DEET	DEET	134-62-3	75	TRA	Tramadol	27203-92-5	0
DNT	Denatonium	3734-33-6	0	TBP	Tri(butoxyethyl)phosphate	78-51-3	0
DRS	Desethylatrazine	6190-65-4	35.6	TCS	Triclosan	3380-34-5	0.0998
DIA	Diazinon	333-41-5	0.00052	TTC	Triethylcitrate	77-93-0	0
DIC	Diclofenac	15307-86-5	67	TGL	Triglyme	112-49-2	0
DMP	Dimethylaminophenazone	58-15-1	0	TMP	Trimethoprim	738-70-5	92
DPP	Diphenylphosphate	838-85-7	0	TPO	Triphenylphosphine oxide	791-28-6	0
DIU	Diuron	330-54-1	7.2				

Table S5. 3 Relative contribution of chemical factors to first two principal components in Danube case.

	Standardized			Standardized	
	PC1 (31.56%)	PC2 (16.64%)		PC1 (31.56%)	PC2 (16.64%)
AAA	0.82	0.00	FAA	**3.24**	0.80
ACF	**3.41**	0.08	FAP	0.12	0.03
ASM	0.55	1.86	GEN	0.47	1.26
ATD	0.05	1.41	GES	0.27	0.72
ATR	0.05	0.03	HCBZ	2.47	1.30
BEN	0.22	**5.28**	HEX	**3.44**	0.28
BET	0.88	1.62	ISO	2.56	0.28
BP4	**3.29**	0.24	KET	0.13	0.24
BPS	0.79	0.01	LEN	0.12	1.19
BTA	2.79	0.05	LID	0.35	0.24
BZT	1.50	2.44	MBT	**3.35**	0.37
CAB	0.09	0.04	MCPA	0.98	0.04
CAF	0.00	0.85	MEC	**3.41**	0.25
CAR	0.12	**5.78**	MEF	1.09	2.58
CBD	0.39	2.43	MET	2.38	0.01
CBZ	**3.10**	0.29	MFM	0.38	**5.69**
CDZ	**3.20**	0.01	MLC	**3.12**	0.20
CHL	0.12	5.78	MLCE	**3.45**	0.13
CLA	2.29	0.52	MTBT	0.69	0.03
COT	0.08	0.39	MTZ	2.08	0.04
CPP	0.12	**5.78**	NSA	1.55	0.73
CRE	0.09	0.72	OXA	1.43	**3.85**
CTL	2.99	0.79	PMT	0.23	1.86
CYC	1.50	0.80	PNZ	**3.12**	0.24
DAI	0.67	0.04	PSA	**3.33**	0.03
DDC	0.38	2.18	SCA	1.37	0.01
DEET	2.32	0.21	SFP	1.37	0.01
DHC	**3.17**	0.11	SFT	0.17	0.41
DIA	0.12	**5.78**	SIM	0.12	**5.78**
DIC	0.65	0.22	SMX	1.07	**3.29**
DIU	0.89	0.83	SUC	**3.64**	0.07
DMP	0.01	0.44	TBH	0.53	**3.19**
DNT	0.89	1.62	TER	0.01	2.83
DPA	0.31	1.44	TGL	1.00	0.79
DPP	0.91	2.26	TMP	0.70	0.25
DRS	0.45	2.19	TOL	2.96	0.69
DTP	0.33	0.37	TPO	0.01	2.33
ENA	1.03	0.13	TRA	1.98	2.91
ERY	0.08	0.01	TTG	0.64	0.01

Table S5. 4 Summary of 14 gene clusters in the Danube case.

Module ID	Number of genes [a]	Genes with orthologs [b]	Genes with orthologs and pathways [c]
D1270	24	17	9
D1440	62	35	14
D2100	34	17	3
D2200	68	15	3
D2300	81	38	5
D3100	136	66	11
D3200	83	30	2
D3300	105	53	9
D4000	150	47	14
D5000	99	17	4
D6200	80	26	10
D7100	134	66	28
D7200	110	61	19
D8000	125	28	9

a. The total number of *Daphnia* genes in the gene cluster.

b. The number of *Daphnia* genes with orthologs in *Drosophila melanogaster* at the *Arthropoda* level in the gene cluster.

c. The number of *Daphnia* genes with *Drosophila melanogaster* ortholog and KEGG pathway information.

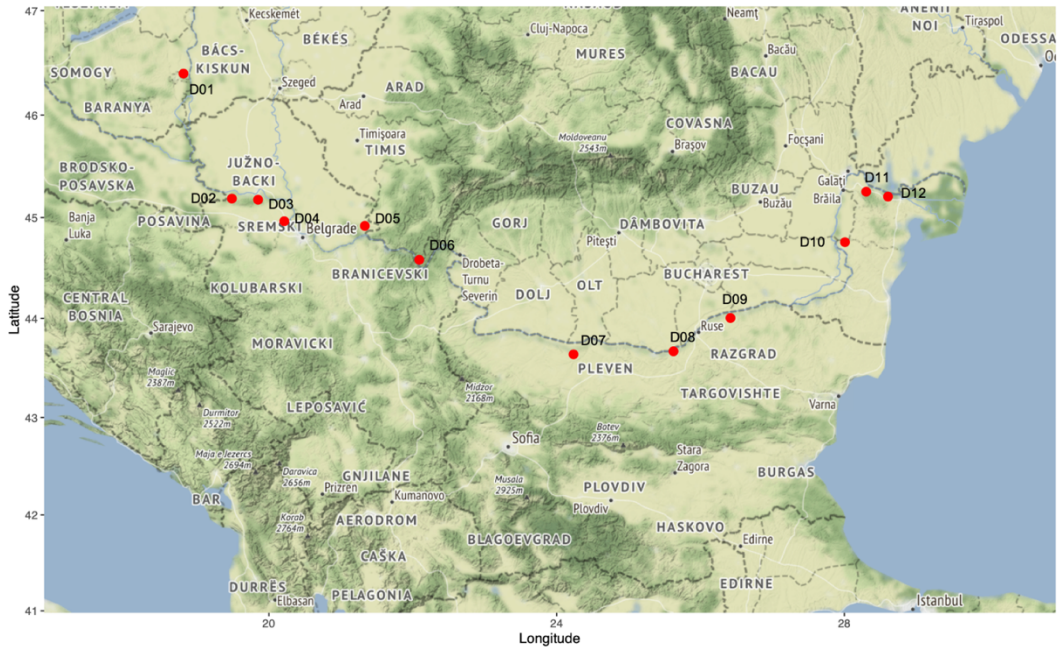

Figure S5. 1 The sampling sites of the Danube River from which water samples were used in this present study. Sites D01 to D06 locate in the middle stretch, while sites D07 to D12 locate in the lower stretch. Among these sites, the first five sites (D01 to D05) located in Croatia, D06 on the border between Serbia Monten and Romania, four sites (D07 to D10) on the border between Romania and Bulgaria, the last two sites (D11 and D12) the border between Romania and Ukraine.

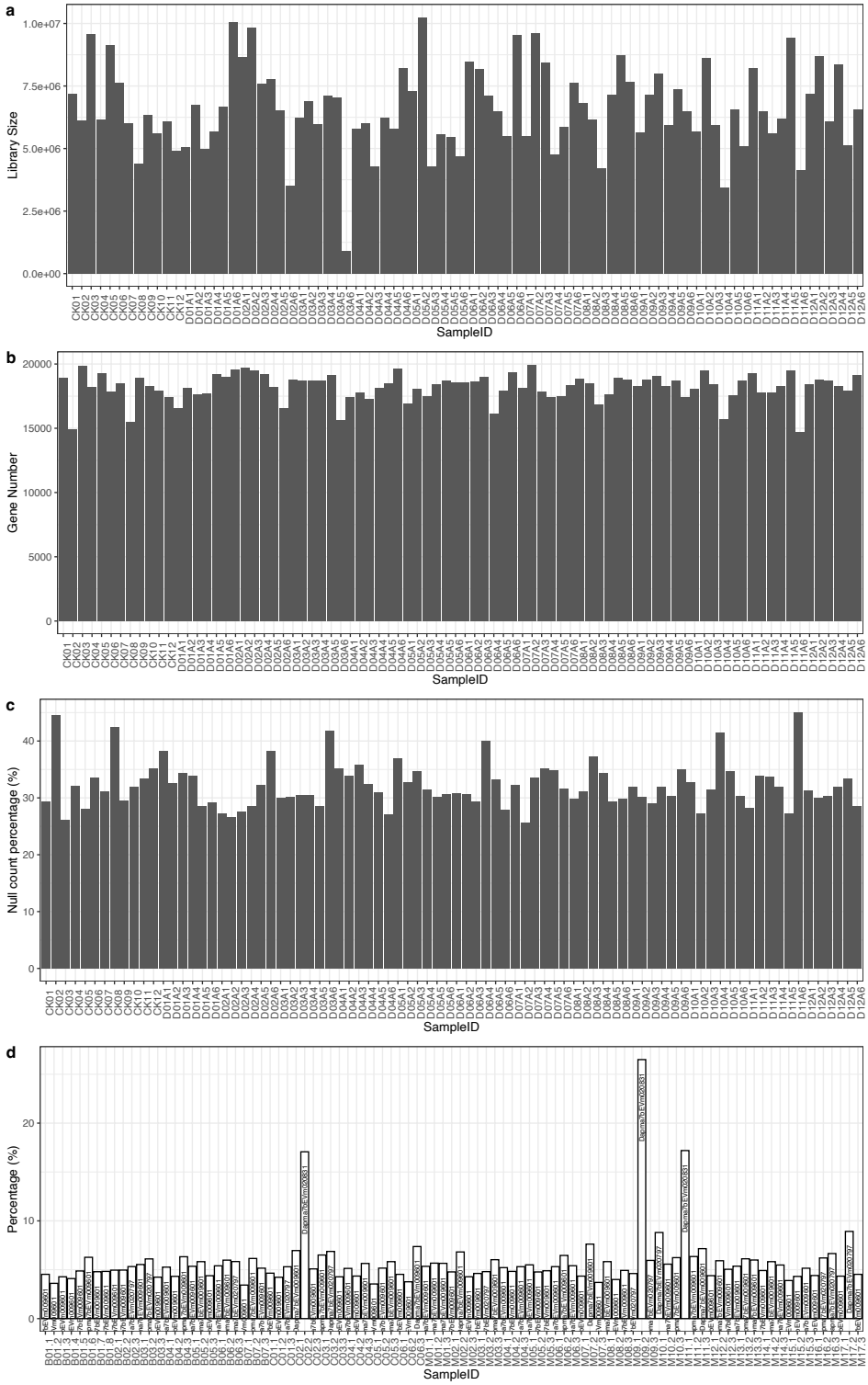

Figure S5. 2 Overview of Danube transcriptome data sets. (a) the total number of read counts in each sample; (b) the total number of genes in each sample; (c) the percentage of genes with null count in each sample; (d) over-representative genes in transcriptome profiles.

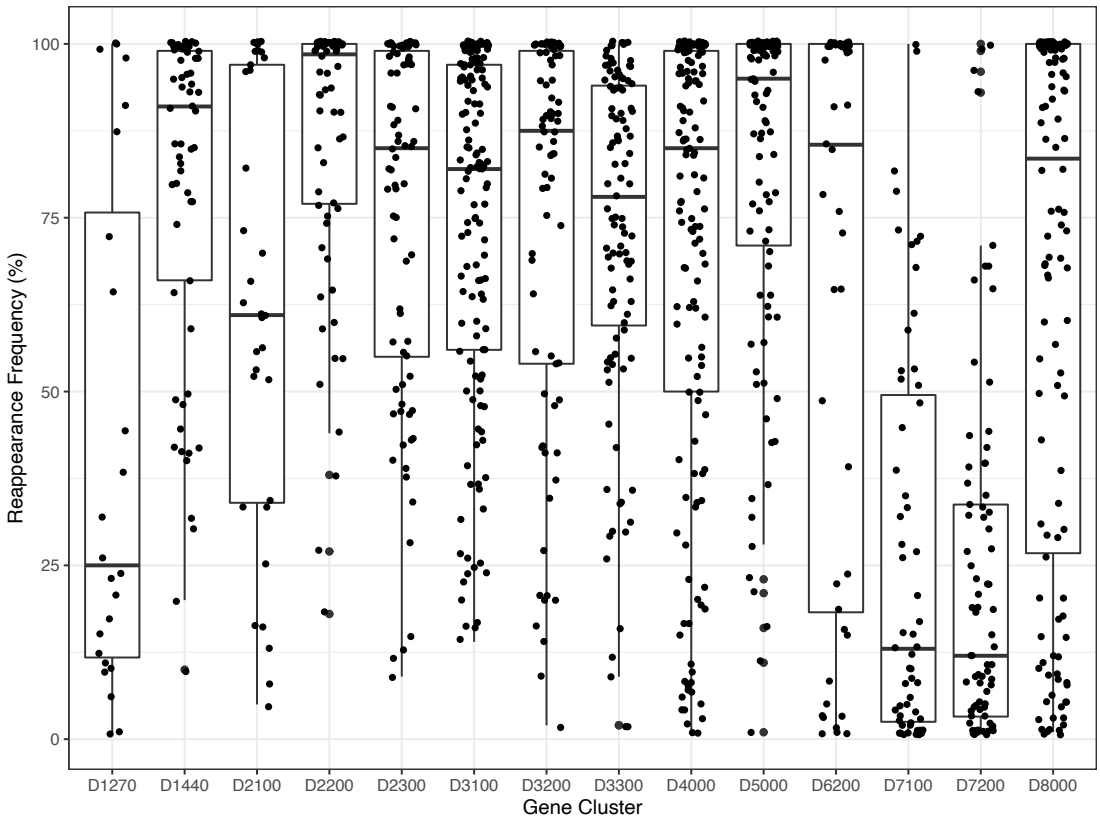

Figure S5. 3 Robustness of Danube gene clusters. After 10,000 times bootstrap resampling of the genes. The frequencies of individual genes assigning to the same gene cluster are plotted as a boxplot representing the first-second-third quartiles of the frequencies, with scatter points representing individual genes' reappearance frequencies.

5.8 Reference

Barata, C., Varo, I., Navarro, J.C., Arun, S., Porte, C., 2005. Antioxidant enzyme activities and lipid peroxidation in the freshwater cladoceran Daphnia magna exposed to redox cycling compounds. Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology 140, 175–186. https://doi.org/10.1016/j.cca.2005.01.013

Brunsch, A.F., 2021. Managing organic micropollutants in rivers : From monitoring to mitigation. Wageningen University. https://doi.org/10.18174/549910

Campos, B., Altenburger, R., Gómez, C., Lacorte, S., Piña, B., Barata, C., Luckenbach, T., 2014. First evidence for toxic defense based on the multixenobiotic resistance (MXR) mechanism in Daphnia magna. Aquatic Toxicology 148, 139–151. https://doi.org/10.1016/j.aquatox.2014.01.001

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., Brown, P.O., 2000. Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes□D. Molecular Biology of the Cell 11, 17.

Hassan, I., Jabir, N.R., Ahmad, S., Shah, A., Tabrez, S., 2015. Certain Phase I and II Enzymes as Toxicity Biomarker: An Overview. Water Air Soil Pollut 226, 153. https://doi.org/10.1007/s11270-015-2429-z

Inostroza, P.A., Wicht, A.-J., Huber, T., Nagy, C., Brack, W., Krauss, M., 2016. Body burden of pesticides and wastewater-derived pollutants on freshwater invertebrates: Method development and application in the Danube River. Environmental Pollution 214, 77–85. https://doi.org/10.1016/j.envpol.2016.03.064

Josyula, N., Andersen, M.E., Kaminski, N.E., Dere, E., Zacharewski, T.R., Bhattacharya, S., 2020. Gene co-regulation and co-expression in the aryl hydrocarbon receptor-mediated transcriptional regulatory network in the mouse liver. Arch Toxicol 94, 113–126. https://doi.org/10.1007/s00204-019-02620-5

Kittinger, C., Lipp, M., Baumert, R., Folli, B., Koraimann, G., Toplitsch, D., Liebmann, A., Grisold, A.J., Farnleitner, A.H., Kirschner, A., Zarfel, G., 2016. Antibiotic Resistance Patterns of Pseudomonas spp. Isolated from the River Danube. Front. Microbiol. 7. https://doi.org/10.3389/fmicb.2016.00586

Kolarević, S., Kračun-Kolarević, M., Kostić, J., Slobodnik, J., Liška, I., Gačić, Z., Paunović, M., Knežević-Vukčević, J., Vuković-Gačić, B., 2016. Assessment of the genotoxic potential along the Danube River by application of the comet assay on haemocytes of freshwater mussels: The Joint Danube Survey 3. Science of The Total Environment 540, 377–385. https://doi.org/10.1016/j.scitotenv.2015.06.061

Kustatscher, G., Grabowski, P., Schrader, T.A., Passmore, J.B., Schrader, M., Rappsilber, J., 2019. Co-regulation map of the human proteome enables identification of protein functions. Nat Biotechnol 37, 1361–1371. https://doi.org/10.1038/s41587-019-0298-5

Loos, R., Tavazzi, S., Mariani, G., Suurkuusk, G., Paracchini, B., Umlauf, G., 2017. Analysis of emerging organic contaminants in water, fish and suspended particulate matter (SPM) in the Joint Danube Survey using solid-phase extraction followed by UHPLC-MS-MS and GC–MS analysis. Science of The

Total Environment 607–608, 1201–1212. https://doi.org/10.1016/j.scitotenv.2017.07.039
Neale, P.A., Ait-Aissa, S., Brack, W., Creusot, N., Denison, M.S., Deutschmann, B., Hilscherová, K., Hollert, H., Krauss, M., Novák, J., Schulze, T., Seiler, T.-B., Serra, H., Shao, Y., Escher, B.I., 2015. Linking in Vitro Effects and Detected Organic Micropollutants in Surface Water Using Mixture-Toxicity Modeling. Environ. Sci. Technol. 49, 14614–14624. https://doi.org/10.1021/acs.est.5b04083
Oliveira, B.R.R., Deslandes, A.C., Santos, T.M., 2015. Differences in exercise intensity seems to influence the affective responses in self-selected and imposed exercise: a meta-analysis. Front. Psychol. 6. https://doi.org/10.3389/fpsyg.2015.01105
Pluskal, T., Castillo, S., Villar-Briones, A., Orešič, M., 2010. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11, 395. https://doi.org/10.1186/1471-2105-11-395
Pollard, K.S., n.d. Cluster Analysis of Genomic Data with Applications in R 27.
Regoli, F., Giuliani, M.E., 2014. Oxidative pathways of chemical toxicity and oxidative stress biomarkers in marine organisms. Marine Environmental Research 93, 106–117. https://doi.org/10.1016/j.marenvres.2013.07.006
Schulze, T., Ahel, M., Ahlheim, J., Aït-Aïssa, S., Brion, F., Di Paolo, C., Froment, J., Hidasi, A.O., Hollender, J., Hollert, H., Hu, M., Kloß, A., Koprivica, S., Krauss, M., Muz, M., Oswald, P., Petre, M., Schollée, J.E., Seiler, T.-B., Shao, Y., Slobodnik, J., Sonavane, M., Suter, M.J.-F., Tollefsen, K.E., Tousova, Z., Walz, K.-H., Brack, W., 2017. Assessment of a novel device for onsite integrative large-volume solid phase extraction of water samples to enable a comprehensive chemical and effect-based analysis. Science of The Total Environment 581–582, 350–358. https://doi.org/10.1016/j.scitotenv.2016.12.140
Slonim, D.K., 2002. From patterns to pathways: gene expression data analysis comes of age. Nat Genet 32, 502–508. https://doi.org/10.1038/ng1033
Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K.-A., Grill, J., Frouin, V., 2014. Variable selection for generalized canonical correlation analysis. Biostatistics 15, 569–583. https://doi.org/10.1093/biostatistics/kxu001
Tenenhaus, A., Tenenhaus, M., 2011. Regularized Generalized Canonical Correlation Analysis. Psychometrika 76, 257–284. https://doi.org/10.1007/s11336-011-9206-8

5.9 Appendix 1

Pathway analysis of 14 gene clusters in the Danube case study.

The *Daphnia magna* genes from 14 gene clusters were re-labelled by their corresponding ortholog groups shared with *Drosophila melanogaster* at *Arthropoda*. Statistical over-representation tests were performed with 137 *Drosophila melanogaster* pathways from KEGG database by permutation chi-squared test. The resulting *P* values are adjusted by FDR at 0.05. Only the pathways with adjusted *P* values lower than 0.05 are listed in this table.

Pathway	Description	D1270	D1440	D2100	D2200	D2300	D3100	D3200	D3300	D4000	D5000	D6200	D7100	D7200	D8000
dme00980	Metabolism of xenobiotics by cytochrome P450	0.0002				0.0084			0.0053						
dme00982	Drug metabolism - cytochrome P450	0.0005				0.0081			0.0057						
dme00983	Drug metabolism - other enzymes	0.0100				0.0246			0.0166			0.0224			
dme00480	Glutathione metabolism	0.0057									0.0070	0.0131		0.0204	
dme00260	Glycine, serine and threonine metabolism									0.0172	0.0052	0.0099			
dme00270	Cysteine and methionine metabolism											0.0073			
dme00280	Valine, leucine and isoleucine degradation													0.0171	
dme00330	Arginine and proline metabolism			0.0011										0.0392	
dme00380	Tryptophan metabolism													0.0389	
dme00010	Glycolysis / Gluconeogenesis					0.0126						0.0115			0.0195
dme00040	Pentose and glucuronate interconversions	0.0026				0.0005			0.0047						

dme00052	Galactose metabolism											0.0052			0.0089
dme00053	Ascorbate and aldarate metabolism	0.0025				0.0006			0.0041						
dme00500	Starch and sucrose metabolism		0.0034									0.0051			
dme00520	Amino sugar and nucleotide sugar metabolism		0.0066							0.0164					
dme00620	Pyruvate metabolism					0.0109									
dme00630	Glyoxylate and dicarboxylate metabolism	0.0020												0.0348	
dme00640	Propanoate metabolism													0.0389	
dme00650	Butanoate metabolism				0.0010									0.0239	
dme00910	Nitrogen metabolism						0.0012			0.0035		0.0018			
dme01100	Metabolic pathways	0.0370										0.0234			
dme01200	Carbon metabolism	0.0134												0.0434	
dme01212	Fatty acid metabolism								0.0106						0.0228
dme01230	Biosynthesis of amino acids											0.0105			
dme01240	Biosynthesis of cofactors New!	0.0199	0.0276			0.0071			0.0041						
dme00510	N-Glycan biosynthesis		0.0053												
dme00511	Other glycan degradation													0.0256	
dme00512	Mucin type O-glycan biosynthesis			0.0000									0.0000		
dme00513	Various types of N-glycan biosynthesis												0.0299	0.0032	0.0263

dme00514	Other types of O-glycan biosynthesis			0.0012											
dme00531	Glycosaminoglycan degradation						0.0017							0.0162	
dme00534	Glycosaminoglycan biosynthesis - heparan sulfate / heparin				0.0005										
dme00601	Glycosphingolipid biosynthesis - lacto and neolacto series											0.0000	0.0000		
dme00603	Glycosphingolipid biosynthesis - globo and isoglobo series												0.0244	0.0194	
dme00604	Glycosphingolipid biosynthesis - ganglio series													0.0000	
dme00062	Fatty acid elongation								0.0026						
dme00071	Fatty acid degradation										0.0039				
dme00072	Synthesis and degradation of ketone bodies				0.0005									0.0121	
dme00100	Steroid biosynthesis								0.0000	0.0000					0.0000
dme00561	Glycerolipid metabolism					0.0075									
dme00564	Glycerophospholipid metabolism				0.0040										
dme00565	Ether lipid metabolism				0.0012										
dme00590	Arachidonic acid metabolism	0.0023			0.0016										
dme00592	alpha-Linolenic acid metabolism				0.0005										
dme00600	Sphingolipid metabolism						0.0040								
dme01040	Biosynthesis of unsaturated fatty acids								0.0027						0.0067
dme00130	Ubiquinone and other terpenoid-quinone biosynthesis								0.0009						

dme00670	One carbon pool by folate												0.0156		
dme00730	Thiamine metabolism		0.0010												
dme00790	Folate biosynthesis		0.0028												
dme00830	Retinol metabolism	0.0025				0.0059			0.0041						
dme00860	Porphyrin and chlorophyll metabolism	0.0026				0.0070			0.0047						
dme00410	beta-Alanine metabolism													0.0258	
dme00450	Selenocompound metabolism											0.0019			
dme00900	Terpenoid backbone biosynthesis					0.0039									
dme00230	Purine metabolism	0.0172										0.0056			
dme00240	Pyrimidine metabolism											0.0067			
dme04213	Longevity regulating pathway - multiple species	0.0062													
dme04320	Dorso-ventral axis formation														0.0106
dme04624	Toll and Imd signaling pathway	0.0115	0.0170				0.0169	0.0040	0.0189	0.0082					
dme04745	Phototransduction - fly	0.0059													
dme04140	Autophagy - animal		0.0258	0.0074		0.0425					0.0206	0.0391			
dme04142	Lysosome		0.0256	0.0074			0.0004		0.0276	0.0110	0.0194	0.0372			
dme04145	Phagosome			0.0033						0.0298	0.0088	0.0175			
dme04146	Peroxisome	0.0110							0.0182	0.0414	0.0129			0.0124	

dme02010	ABC transporters									0.0060		0.0030	0.0301	0.0221	
dme04013	MAPK signaling pathway - fly														0.0414
dme04068	FoxO signaling pathway											0.0203			
dme04310	Wnt signaling pathway														0.0444
dme04080	Neuroactive ligand-receptor interaction	0.0062	0.0097												
dme03018	RNA degradation											0.0113			
dme04141	Protein processing in dedoplasmic reticulum		0.0288												
dme03430	Mismatch repair													0.0221	
dme03440	Homologous recombination											0.0051	0.0413		0.0083
dme03460	Fanconi anemia pathway												0.0112	0.0360	0.0092
dme03040	Spliceosome														0.0485
dme03013	RNA transport														0.0102

6 Conclusion

Aquatic environmental pollution is a complex and long-lasting problem. The potential health effects of complex chemical mixtures are alarming thereby requiring novel approaches that can reduce the complexity of mixture toxicology by taking advantage of data-driven methodologies at identifying potential pollution hotspots in the environment and identifying potentially harmful chemical substance for monitoring and regulation. These were the primary goals of this thesis

The **Precision Environmental Health framework** that I proposed in Chapter 2 is a viable solution to solve the puzzle of identifying environmental mixture effects for effective pollution control. Within this framework, non-targeted investigations of chemical mixtures and biomolecular signatures of environmental exposures are its two basic components (Figure 2.1), to assemble a holistic assessment of the environmental chemical mixture that may reflect the realistic environmental exposure scenarios. The multilevel approach to reveal the mode of action of mixture effect via molecular key events (Figure 2.2) and further the modes of action of chemical components of this mixture provides the required evidence for classifying the mixture and identifying harmful components.

Conventional approaches, such as individual chemical toxicity testing with targeted adversity, are unable at establishing toxicity testing results that reflects realistic environmental exposure scenarios, and the component-based toxicity testing are ruled by reductionism that includes manually constructing artificial mixtures that are insufficient for studying the complex environmental chemical mixture with limited

information of the chemical composition and component identities. The work I presented in my thesis is focused on developing a practical approach that reflects the nature of the environmental chemical mixture—by treating it as an entity. **The whole-mixture approach** is the key component that ensures the environmental chemical mixtures under study mirrors the chemical mixtures detected in the environment, and the concentration levels of the chemical mixtures remain the same as the original environmental levels. Two kinds of the environmental chemical mixture were investigated, as in the Chaobai River case, the whole mixture referred to the filtered unconcentrated river waters, while in the Danube River case, the whole-mixture referred to the organic mixtures extracted from the river waters (re-diluted to their original environmental concentration levels). Theoretically, the whole-mixtures I used in these two case studies consisted of most of, or at least a considerable amount of, the chemical components that can be captured from the real-world environment. Although the stability of chemical components within the mixtures during the pre-processing and exposure testing still need further validation, this approach is under considerate and comparable processes of treating and maintaining the water samples, which is feasible to apply in the short-term exposure testing (i.e., 48 hours).

The omics-based approaches are also highlighted in three chapters as such non-targeted assays may provide a systematic perspective of the biological responses that are free of prior expectation (of adversity). It is feasible to apply a single omics approach to characterise the subtle differences in effects of environmental chemical mixtures, for example Chaobai case study in Chapter 4 and Danube case study in Chapter 5. And the strength of applying the multi-omics approach at characterising the modes of action of the chemical components in the mixture, not only the directly related

metabolic pathways but also the associated downstream adverse effect, is emphasised in Chapter 3. The **integration of multi-omics** may provide an interconnected and complementary view of the same biological processes from different types of biomolecular readouts (different omics) that further achieve the mechanistic understanding of the biological responses.

Data analysis that establishes **co-varying** characteristics of the biological features (gene clusters in Chapter 4 and 5) and **co-responsive** characteristics of the biological network (gene modules in Chapter 3) stratify the omics profiles into a limited number of functional groups, so that the biological roles of the functional groups can be interpreted, and the variation patterns of the functional subgroups can be visualised and understood. The data-driven approach that relies on **pairwise correlation analysis** (Pearson correlation between eigengene and chemical factors in Chapter 4 and 5) and **multi-view learning** (sCCA in Chapter 5) can model the correlations between several data sources and identify co-responsive biomolecular features that are associated with chemical data. Notably, in the Chaobai case, Dibenz[a,h]anthracene, Erythromycin and Trimethoprim are associated with distictive expressional patterns in xenobiotic activities. In the Danube case, gene clusters that define the differences between D11/D12 and the rest account for mild external stress levels while one specific gene cluster significantly associate with five biocides (carbaryl, chlorophene, chlorpropham, diazinon, and Simazine. Further investigation on these selected gene clusters may be the basis for biomarkers of specific subsets of chemical components, which are assumed to be the driving factors in the environmental chemical mixture. Regulation of environmental chemical mixture is always challenging. A promising way is to reveal potential toxicity driver among thousands of chemical

components. One character of the toxicity driver is that the concentration levels of that chemical components are high enough to be bioactive and capable of inducing (contributing) to the toxicity observed under the whole mixture exposure. This can be verified by artificial chemical mixture exposure v.s. single chemical compound exposure, so that the relative contribution of identified toxicity driving component can be revealed. Another character of the toxicity driver is critical to the adverse outcome, that is to say, by removing the toxicity driver, no adverse outcome or comparable adversity is observed. Once the toxicity driver of environmental chemical mixture is revealed and verified, the regulator can further monitor the small subset of toxicity driver and set regulation on the acceptable concentration levels in the environment.

Biological interpretation of the gene sets of a premier model species (like *Daphnia magna*) was accomplished by making reference to gene and pathway homologies to functional and exquisitely well annotated genes in the other well-annotated genetic and biomedical model species (like *Drosophila melanogaster*). Such **cross-species extrapolation** was here achieved by referring to evolutionarily conserved ortholog groups that are shared by both species despite their deep evolutionary divergence. A hypothesis of the pathway construction is purposed as the ortholog groups are the essential elements that define a pathway and reveal its functionality. The premise of this hypothesis is that the ortholog group patterns are unique in each pathway (defined in the knowledge-based pathway databases, like KEGG or Reactome) and conserved across multiple species. Investigation on the KEGG pathway database provided preliminary reference data for extrapolating general (potentially conserved) pathways in *Daphnia magna* from *Drosophila melanogaster*. The permutation chi-square test is subsequently applied to fulfil the need for a **statistical overrepresentation test**. This

cross-species pathway extrapolation method is novel and important in that (1) it facilitates pathway analysis of gene sets from the responsive genome of any test species for (eco)toxicology by making reference to the superior annotated genomes of biomedical based on the orthology of genes and pathways that are shared by common ancestry; (2) it assists at identifying chemical-responsive pathways that are conserved across multiple species (even human); (3) it implies chemical-induced adverse outcomes via conserved functional pathways that are shared across multiple species thereby indicating potential exposure hazards to the health of animals beyond the target test species. Future verification of this cross-species pathway analysis will be needed as a sound reference database like that can assist at establishing environmental effect assessment based on evolutionarily conserved pathways that are shared by multiple species in the environment, including those that are also necessary for human health.

The framework of Precision Environmental Health is not only theoretically sound but also practically feasible. Two approaches applied on two independent case studies (Chapter 4 and 5) present the **proof-of-concept** that the framework can be applied on the case-based investigations of (1) characterising and classifying the effect of environmental chemical mixtures by the omics-based assays of the test system and (2) identifying the modes of action of the chemical components in the mixture as a basis for identifying harmful component that may contribute to adverse effect in the mixture.

Future verification work remains challenging in applying this framework in regular environmental monitoring and chemical prioritisation process. The major challenge is the development of a **meta study** that covers a large number of natural water samples

and a more comparable bioassay. Such a study will verify that the responsive signatures generated by a single case study can generate robust signatures that may be diagnostic of the harmful chemical components (case comparison in Chapter 5). With a substantial amount of chemical components detected in the natural environments investigated by the omics assays, the modes of action of detected (or even identified) chemical components will further provide the basis of molecular classification of the chemical effect, which may assist better identification of the harmful chemical substances.