OMICS FOR AQUATIC ECOTOXICOLOGY: CONTROL OF EXTRANEOUS VARIABILITY TO ENHANCE THE ANALYSIS OF ENVIRONMENTAL EFFECTS

DENINA B.D. SIMMONS,*† JONATHAN P. BENSKIN,‡ JOHN R. COSGROVE,†‡ BERNARD P. DUNCKER,§ DREW R. EKMAN,∥

CHRISTOPHER J. MARTYNIUK,# and JAMES P. SHERRY†

Emerging Methods Section, Aquatic Contaminants Research Division, Water Science & Technology Directorate, Environment Canada, Ontario, Canada

Axys Analytical Services, Sidney, British Columbia, Canada

Department of Biology, University of Waterloo, Ontario, Canada

Ecosystems Research Division, National Exposure Research Laboratory, Office of Research and Development, US Environmental Protection Agency, Athens, Georgia, USA

#Center for Environmental and Human Toxicology & Department of Physiological Sciences, University of Florida, Gainesville, Florida, USA

(Submitted 8 October 2014; Returned for Revision 9 January 2015; Accepted 30 March 2015)

Abstract: There are multiple sources of biological and technical variation in a typical ecotoxicology study that may not be revealed by traditional endpoints but that become apparent in an omics dataset. As researchers increasingly apply omics technologies to environmental studies, it will be necessary to understand and control the main source(s) of variability to facilitate meaningful interpretation of such data. For instance, can variability in omics studies be addressed by changing the approach to study design and data analysis? Are there statistical methods that can be employed to correctly interpret omics data and make use of unattributed, inherent variability? The present study presents a review of experimental design and statistical considerations applicable to the use of omics methods in systems toxicology studies. In addition to highlighting potential sources that contribute to experimental variability, this review suggests strategies with which to reduce and/or control such variability so as to improve reliability, reproducibility, and ultimately the application of omics data for systems toxicology. Environ Toxicol Chem 2015;34:1693–1704. © 2015 SETAC

Keywords: Variability Systems toxicology Metabolomics Proteomics Transcriptomics

INTRODUCTION

Systems biology aims to integrate global responses within an organism from genotype to phenotype and is informally referred to as the integrated study of transcriptomics, proteomics, metabolomics, and other biological -omics subdisciplines [1,2]. The term systems toxicology has been used to describe the integration of systems biology approaches with traditional toxicology. Although the concept is still developing within the field of aquatic ecotoxicology, the intention is to provide an integrated perspective among transcriptomic, proteomic, metabolomic, and whole-organism (or even population-level) responses to specific physiological changes, which may have resulted from an environmental exposure [3]. To accomplish this, modern approaches to systems toxicology make use of highly innovative omics technologies and bioinformatics software. Specifically, next-generation RNA sequencing and microarrays quantify responses in the transcriptome of organisms in response to environmental pollutants. Nuclear magnetic resonance as well as separation techniques (e.g., chromatography, electrophoresis) paired with mass spectrometry can be used to detect changes in the metabolome, lipidome, and proteome. The major advantage of using omics methods in a systems approach is that one can assess hundreds to thousands of molecular responses simultaneously within an organism, facilitating a more holistic understanding of the organism’s physiological status. A disadvantage, however, is the significant amount of data that require novel statistical analyses and investigative methods compared with traditional endpoints.

Bioinformatics methods, such as gene set enrichment analysis or pathway analysis software, are used to address these issues by integrating biological function and interaction networks within larger datasets. The combination of high-throughput data generation and bioinformatics has transformed toxicology from its hypothesis-driven experimental phase to its emergence as a science that is both powerfully deductive and, in its opened-ended variants, an impressively inductive science.

However, a significant challenge is that the omics data generated from systems toxicology studies include natural and experimental variability from multiple endpoints. That variability must be characterized and attributed, to correctly interpret biological responses to environmental stressors [4]. There are many potential sources of biological variation in a conventional ecotoxicology study. For instance, seasonal and temporal variation may affect reproductive and metabolic endpoints [5]. Complex mixtures, such as effluent discharges, and differences in abiotic factors between sampling locations also add variability [6,7]. However, ubiquitous sources of variation (such as genetic variation among individuals, life history, and trophic interactions), which are often apparent when measuring conventional endpoints, can become increasingly problematic and can confound omics datasets [8,9]. Thus the experimental design and the strategy for the statistical analyses become critically important in systems toxicology studies. The interpretation of results from an omics study can easily overwhelm the investigator and confound the ability to draw clear conclusions. As researchers, we need to challenge ourselves to understand variability when working with samples of ecological relevance. For instance, can we better address variability in omics data by changing our approach to study design and data analysis? How can we improve omic methods to reduce variability? Are there statistical methods we can employ...
to correctly interpret omics data by reducing the effect of technical variability? The present review addresses these questions by highlighting the issues surrounding variation and also provides recommendations for addressing variability in aquatic systems toxicology.

**BIOLOGICAL VARIATION**

Biological variation includes the measured differences between cells, tissues, organisms, or groups of organisms and results from the combination of genetic, phenotypic, and environmental differences represented by each individual.

**Laboratory and field studies**

The use of model laboratory organisms can address some of the challenges associated with biological variation because their genomes are well characterized and they can be manipulated and tested within a controlled laboratory environment. However, studies that systematically quantify genetic variability are lacking. Martyniuk and Houlanah [10] recently addressed variability in the fathead minnow transcriptome, a widely used toxicological laboratory model organism, using >230 microarray datasets. Their study found that the mean coefficient of variation (CV) across fathead minnow gene probes was 0.56. However, a limitation of their study was that it addressed transcriptome variability using multiple tissues. That variability was comparable to other reproductive endpoints in female fathead minnows that included steroids (17β-estradiol, CV = 0.66; testosterone, CV = 0.74) and liver somatic index (CV = 0.33) [5]. It was concluded that transcripts can show a wide range of variability among individuals and that specific functionally related groups of genes can show lower (or higher) variability as a whole (i.e., variability appears not to be random in the transcriptome and is related, in part, to biological process). More recently, a meta-analysis of approximately 600 microarrays in the fathead minnow and zebrafish ovary revealed that the biological variability of the transcriptome in the ovary can be lower within tissue (CV = ~0.05) [11].

A recent study [12] of the plasma proteome among individual field-captured white sucker (Catostomus commersonii; a cyprinid fish species widely used in environmental effects monitoring programs) determined that the mean CV for protein expression levels across all detected plasma proteins was 0.78 for males (n = 40) and 0.68 for females (n = 40; Table 1). The higher difference in variability between white sucker and fathead minnow may reflect variation introduced when moving from the controlled laboratory environment to the field. However, as reported by Parsons et al. [13], variability in the metabolome as measured by the CV among field-captured and then laboratory-maintained organisms can be significantly greater than for field-captured organisms of the same species, so the relationship between field- and laboratory-reared organisms and variation in omics endpoints may not be so straightforward. Moreover, it is not entirely clear whether transcripts exhibit more (or less) variability compared with proteins or metabolites.

**Males and females**

In the white sucker experiment, an increased variability between the male and female plasma proteomes was demonstrated (~10% greater in males; Table 1). In a study using laboratory-exposed South American catfish, Rhamdia quelen, the global mean CV across 520 plasma proteins was 0.52 for control males and 0.41 for control females [14], similar to the observed difference in variation between the male and female white sucker plasma proteomes [12]. That estimate is closer to the molecular variation reported in Martyniuk and Houlanah [10]. Thus, there is some evidence that laboratory-reared fish may introduce less variation in molecular responses compared with field-sampled individuals. The difference in variance between males and females also exists in the metabolome. A recent comparison of metabolite concentrations in wild Sockeye salmon (Oncorhynchus nerka) livers revealed CV values ranging between 0.11 and 0.78 for females and 0.12 and 0.45 for males sampled from the same location. Males and females contained highly contrasting concentrations for some, but not all, metabolites [15]. Intraspecies variability also appears to be a function of the type of biological sample under analysis. Biofluids (e.g., plasma and urine) generally show higher variability than tissues (e.g., liver, brain, gonad) [13]. Indeed, the proteomic data from R. quelen [14] demonstrate a similar trend to the metabolomics data of Parsons et al. [13], where plasma variation was greater than that for liver but not for head kidney (Table 1). Difference in variation between the sexes was more apparent in the plasma proteome of R. quelen and white sucker compared with organ tissues, which is to be expected as the plasma contains proteins from the entire organism, including sex-specific organs, whereas the head kidney and liver are involved more exclusively in non–sex-specific functions. Regardless, these proteome and metabolome examples demonstrate the importance of subclassifying results based on sex prior to interpretation of omic data, as others have discussed [16,17].

Table 1. Variation in mean global protein expression in male and female fish plasma, liver, and head kidney, sampled from the field and assessed in the laboratory

<table>
<thead>
<tr>
<th>Setting</th>
<th>Tissue</th>
<th>Number</th>
<th>Sex</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catostomus commersonii&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Field</td>
<td>Plasma</td>
<td>40</td>
<td>Male</td>
</tr>
<tr>
<td></td>
<td>Field</td>
<td>Plasma</td>
<td>40</td>
<td>Female</td>
</tr>
<tr>
<td>Rhamdia quelen&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Laboratory</td>
<td>Plasma</td>
<td>40</td>
<td>Male</td>
</tr>
<tr>
<td></td>
<td>Laboratory</td>
<td>Plasma</td>
<td>40</td>
<td>Female</td>
</tr>
<tr>
<td></td>
<td>Laboratory</td>
<td>Liver</td>
<td>10</td>
<td>Male</td>
</tr>
<tr>
<td></td>
<td>Laboratory</td>
<td>Liver</td>
<td>10</td>
<td>Female</td>
</tr>
<tr>
<td>Pimephales promelas&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Laboratory</td>
<td>Liver</td>
<td>3</td>
<td>Male</td>
</tr>
<tr>
<td></td>
<td>Laboratory</td>
<td>Liver</td>
<td>3</td>
<td>Female</td>
</tr>
</tbody>
</table>

<sup>a</sup>Study was conducted using label-free quantitative shotgun proteomics. The coefficient of variation (CV) was determined on total mean intensity counts for each protein excluding replicates with non-detects [11,13].

<sup>b</sup>Study was conducted using quantitative tIRAQ labeling. Criteria for CV for mean protein fold change (n = 186) were that proteins were quantified with unique peptides and that proteins were quantified in all three biological replicates. Quantification of proteins included peptides that ranged in number from 2 to >20 in each biological replicate [59].
Microhabitats

As omics approaches are increasingly applied to field-based ecotoxicology, careful attention must be given to the characterization of microhabitats. In the aquatic environment, there are many potential sources of variation that can complicate biological responses in wild fish species. In the field, for example, concentrations and potencies of contaminants vary both temporally and spatially in complex mixtures and waste effluents [7,18]. The chemical speciation of elemental contaminants in the aquatic environment can change as a result of variables such as pH, dissolved oxygen, and the presence of other dissolved ions. All of these factors can vary depending on the geochemistry of the environment and will affect the bioavailability and toxicological effects of the contaminant in the organism [19,20]. It is well documented that temperature, salinity, pH, and dissolved oxygen can influence biological responses [21–23]. In ecotoxicology the variability that is a result of environmental changes can be an important aspect of, or even the focus of, a study. Therefore, experimental design should capture those sources of variation that are important to the hypothesis being tested and should minimize (or at the very least describe) those sources of variation that could obscure interpretation of the results. Currently, there is a lack of understanding about how rapidly omic profiles respond to contaminant exposure and how such responses are influenced by other environmental factors that contribute to microhabitat effects.

Summary

Together, the aforementioned examples demonstrate the complexity of biological variability in omic datasets. Not surprisingly, the levels of biological variation differ between laboratory and field studies, among different tissue types, between vertebrate sexes, and in relation to physiological status. These differences can have a profound effect on statistical power, false discovery rate, and consequently the ability to draw informed conclusions regarding the meaning of biological responses that are revealed by omics datasets. Thus, when one is designing a field study, which would reasonably be expected to have inherently greater variability than a laboratory study, the effect size, the selection of the biological matrix, the sex to be sampled, and the statistical power need to be given careful consideration.

TECHNICAL VARIATION

Technical variation reflects the imprecision of the analytical method and tends to increase when the precision is low because of instrumental error, sample handling error, and operator error. A common criticism of most omic methods is that biological variation could be masked by high technical variation. Even though many omics approaches have demonstrated that technical variation is often more than 10 times lower than biological variation [13,16,24], the perception that omic methods are imprecise persists. Nonetheless, it is important to address technical variation at the beginning of method development by using appropriate quality control measures.

Transcriptomics

Minimum information about a microarray experiment (MIAME) was developed to address technical variation and to standardize the microarray dataset comparison [25]. There are a number of quality control steps for microarray software packages and data analysis methods. Those include scatter plots of outliers, statistical assessment of global signal intensity (i.e., kernel density and box plots), spike-in report probes, and cluster dendrograms, to name a few. Spike-in control probes such as those for Agilent and Affymetrix protocols provide valuable information on the limit of detection of the microarrays, and can be used to estimate background signal. Expression levels for probes failing below this background level should be considered noise, and expression data should be analyzed as such. A preliminary step in the reduction of technical variation is the measurement of total RNA integrity [26]. Reports show that measurements of messenger RNA (mRNA) abundance can be reflected by the quality of the total RNA [27]; however, in our experience, high-quality total RNA can still contain poor-quality mRNA. The inclusion of housekeeping (or reference) genes serves to normalize mRNA expression among samples, regardless of whether the source of mRNA variation is low abundance, loss, or deterioration [26]. Thus, there are a number of quality control metrics and steps in the microarray hybridization process that can ensure high-quality data.

Similar considerations apply to the generation and analysis of high-quality RNA-sequencing data. The sequencing of an expressed transcriptome currently involves multiple steps in a complex and intricate workflow, parts of which can be automated at a cost. The reverse transcription (RT) of complementary DNA (cDNA) from mRNA and its subsequent amplification by polymerase chain reaction (PCR) in current workflows both depend on enzyme activity, which can introduce technical variability and bias. The data analysis pipeline relies on quality control and filtering of the reads from the sequencer, the assembly of a reference transcriptome if an annotated genome is not available, the alignment of sequences to the reference transcriptome or genome, estimations of abundances for isoforms, and quantitative analysis of transcript abundance and differential expression [28–30]. The optimization, evaluation, and eventually standardization of the various strategies for each of those critical steps are the focus of much current research. At the time of this writing, however, there are few agreed-on best data analysis pipelines or practices. To a large extent, the choice of pipeline depends on the sequencing platform, the target species, and the hypotheses. In addition to the influence of biases (as a result of RT, PCR, and gas chromatography) and sequence quality on the quality of RNA-sequencing data, there are other considerations that should influence the choice of sequencing and data analysis strategy adopted, depending on the question posed. For example, if one wishes to assess the effect of an experimental treatment on low-abundance transcripts, then the influence of overabundant transcripts must be mitigated. The depth of coverage of the sequenced transcriptome must be adequate to permit reliable counts of the measured transcripts; there have been several in silico and experimental estimations of the influence of sequencing depth, read length, and number of replicates on the detectability of transcripts and the identification of differential expression transcripts [31,32]. The current era is one of rapid advances in sequencing technologies and platforms, which might soon, for example, make possible the sequencing of mRNA on nanopore platforms [33] without prior conversion to cDNA. The redundancy of the RT step from RNA-sequencing workflows would eliminate a major source of technical variability and uncertainty. Such advances are enhancing the quantity and quality of RNA-sequencing data at ever decreasing costs. Concurrently, there have been impressive advances and continuing developments in the algorithms, alignment tools, and data reduction packages for
identifying transcripts and assessing whether they or their associated genes and/or isoforms are differentially expressed [34,35]. Moreover, many of these data reduction tools are open source, albeit also associated with steep learning curves for those with limited experience in Linux and/or the R environment.

**Proteomics**

In high-throughput shotgun proteomics, technical precision is not always consistent across the proteome. Variation in expression of the individual white sucker plasma proteins ranged from \( CV = 0.17 \) to 1.3 (mean values are presented in Table 1) [12], which could reflect that the expression of some proteins is more tightly regulated than that of others, but alternatively could reflect technical variation introduced by the shotgun method. The overall technical variability introduced by the label-free liquid chromatography–mass spectrometry (LC-MS) shotgun method (described in Silva de Assis [36] and Simmons et al. [37]) ranges from \( CV = 0.04 \) to 0.31 (9 replicates) for higher abundance plasma proteins (>100 pmol). In other shotgun approaches, fold change values are determined from multiplexed samples labeled with different isobaric tags (i.e., iTRAQ labeling). Precision tends to be better for labeled compared with label-free shotgun methods [38], but it comes at the expense of dynamic range and protein sequence coverage [38,39]. Recent advances in automated spectral processing software have allowed accurate batch normalization, total ion count extraction, and relative quantification of label-free data such that intensity counts obtained from label-free analyses are currently considered equal to relative isobaric labeling [39]. Indeed, in our experience with shotgun proteomics, the digestion and labeling steps introduce more than 50% of the technical variability, whereas replicate LC-MS injections tend to be more consistent (<10%). Ultimately, targeted and multiple reaction monitoring approaches have the potential to reduce technical variation by using labeled internal standards and are capable of accurately quantifying dozens of protein targets simultaneously within a single instrumental analysis [40]. Technical variability in targeted analyses for white sucker vitellogenin (Vtg) using a stable-isotope labeled peptide standard was low (\( CV = 0.016 \)) compared with proteins quantified using the shotgun method (Table 1). The stable-isotope method also had greater sensitivity and lower detection limits than the shotgun method [37].

**Metabolomics**

Although metabolite analysis has been conducted for decades, the increasing application of nuclear magnetic resonance and MS-based platforms to this area has vastly improved sensitivity, selectivity, and analytical throughput compared with historical techniques. Nonetheless, these platforms have their own challenges, which need to be considered when one is conducting metabolomics experiments. For example, nuclear magnetic resonance is nondestructive, rapid (for measurement), and highly precise, but is relatively insensitive and often requires considerable time to interpret complicated spectra. In contrast, MS is superior in sensitivity and speed of data processing for uncomplicated data sets, but is hampered by challenges associated with peak alignment and matrix-induced ionization effects, which can contribute significant variability to the data. Depending on whether the experiment is untargeted or targeted, different quality control protocols are utilized to decrease data variability in MS data. For example, in untargeted metabolomics and proteomics, novel strategies have been developed involving the use of quality control samples that are injected regularly throughout a batch. The quality control samples typically consist of combined portions of unknowns used in the study, or a reference material [39,41]. This approach has led to the correction of sample results using a correction factor obtained from analysis of quality control samples. Although the advantage of untargeted metabolomics is clearly in the breadth of coverage (relative to targeted methods), there has been considerable concern over variation in untargeted data. This is exacerbated by the fact that fold-change data generated by this approach are not directly comparable among laboratories or studies.

In contrast, targeted metabolomics (and targeted proteomics) methods utilize a suite of native and isotopically labeled standards (internal standards), which can be used in conjunction with reference materials to assess accuracy, precision, and instrumental dynamic range simultaneously controlling for inter- and intrabatch variability. The quantitative nature of this approach facilitates intercomparison of data between studies, between laboratories, or over time. However, even in targeted methods, full complements of isotopically labeled internal standards are not always available, and thus a structurally similar surrogate is used. This can lead to variability if the internal standard responds differently than the native it is used to track. To illustrate this, we reexamined taurine data from 70 wild sockeye salmon livers (\( n = 31 \) female, \( n = 39 \) male) measured by liquid chromatography-tandem mass spectrometry (LC-MS/MS) [15]. To compare the variability introduced from different approaches with metabolite measurement, we quantified CVs for taurine using 1) absolute peak areas; 2) relative responses using \( d_4 \)-putercine, \( ^{13}C_2^{15}N \)-glycine, or \( ^{13}C_2 \)-taurine as an internal standard; and finally, 3) concentrations using a quadratic, 1/x calibration curve and \( ^{13}C_2 \)-taurine as an internal standard. The results are provided in Table 2. Not surprisingly, CVs are quite high in the absence of an internal standard (0.44 CV for females and 0.48 CV for males for peak area alone). This is decreased by half when the response was normalized to that of \( d_4 \)-putrescine, which is structurally dissimilar and elutes >1.5 min after taurine in this particular method. Further improvements in CV were made by switching to \( ^{13}C_2 \)-glycine, likely because of its structural similarity and elution time within 0.5 min of taurine. Finally, CVs were decreased by an additional 5% when \( ^{13}C_2 \)-taurine (a structurally identical internal standard) was used, and by an additional 1% to 2% when relative responses to concentration were converted using a quadratic calibration curve. The latter effect arises from nonlinearity in analyte response over the measured concentration range, which can only be accounted for by using a calibration curve.

In the above example, a conversion of relative response to concentration produced only a minor improvement in precision. However, it is conceivable that if a large proportion of data was obtained at concentrations close to detector saturation, where response is clearly nonlinear, then improvements in precision from the use of a quadratic calibration curve would be considerably more dramatic. This point is particularly salient for untargeted MS-based methods, which by their very nature do not use calibration curves, and are therefore susceptible to increased variability at either end of the instrument’s linear dynamic range. In the absence of an exact-matched internal standard, normalization of response to the internal standard of closest retention time can provide a considerable improvement in results relative to absolute peak areas. It is also worth noting that the 0.11 to 0.12 CV observed in taurine concentrations in Table 2 represents error from both instrumental analysis and
interindividual variability in male and female salmon. When error from sample preparation and analysis alone is considered, CVs decrease to 5%, which is consistent with metabolites from a range of metabolite classes (alanine, eicosapentaeonic acid, hexose, carnitine, and lysophosphatidylcholine C18:0) [15]. Not surprisingly, the use of an exact match internal standard and calibration curve can reduce analytical variability to far less than biological variability.

Validation

Future inclusion of omics data in risk assessment and management will depend on well-defined standard protocols that include validity testing and quality controls [16]. Interlaboratory comparisons have been performed previously for metabolomics, proteomic, and transcriptomic samples (reviewed by van Aggelen et al. [16] Garcia-Reyero and Perkins [1]). Although data obtained from metabolomic methods currently appear to be the most consistent between laboratories and instruments, transcriptomic and proteomic approaches are still highly comparable. The gold standard to ensure quality control and consistency across different laboratories is the use of certified reference materials. Certified reference materials are under development in metabolomics [42], proteomics [43], transcriptomics [44], and genomics [45]. Human biofluids are currently the most commonly banked specimen for use as standard reference materials. For aquatic systems toxicology applications, standard reference materials for biofluids and tissues from model laboratory organisms such as zebrafish and fathead minnows might be the most reasonable starting point. Interlaboratory studies may also provide insight when one is creating standard protocols and methods. Interlaboratory studies are often performed to validate methods and determine method aspects in need of strengthening. At the time of writing, efforts are under way to conduct interlaboratory transcriptomics studies for fathead minnows (A. Feswick, University of New Brunswick, St. John, NB, Canada, personal communication) and have been completed for the aquatic amphipod *Eohaustorius estuarius* (D.E. Vidal-Dorsch, Southern California Coastal Water Research Project, Costa Mesa, CA, USA, personal communication).

**EXPERIMENTAL DESIGN TO ADDRESS VARIATION**

The principles of experimental design in aquatic toxicology are fundamentally the same as the principles of experimentation introduced by Fisher in the first half of the 20th century [46]. Toxicological studies typically draw multiple comparisons among blocked treatment and control groups with tightly controlled variables using a balanced design that includes randomized sampling. Variation is a subtle but important concept in experimental design. The investigator must decide what type of information is required to be able to make an informed decision regarding the validity of the hypothesis and then design an experiment that will provide the required dataset. However, such experimental principles are challenged in omics-based approaches, which are not focused on a single hypothesis and include variation from hundreds (to thousands) of response variables within a single experiment [1,47]. Unexpectedly high variability in the data can confound interpretation of data obtained from even a well-planned experiment, and thus it is important to design experiments with variability in mind.

**Functional data**

One way to address biological variability is to focus on molecular function. Functional datasets are less concerned with absolute changes in expression because of exposure, and are more focused on how the function of the observed system changes. Functional datasets tend to be based on qualitative observation and normally contain discrete values and count data. For example, if a greater proportion of the molecules detected from organisms exposed to a toxicant are related to detoxification compared with organisms that have not been exposed, this would demonstrate biological response. Thus, fluctuations in expression and the presence of nondetects become less relevant and do not greatly affect the interpretation of data. However, this approach to analyzing omic data was not a viable option for aquatic ecotoxicologists until recently. An increased functional annotation for fish and invertebrate genes and proteins in publicly available databanks has advanced the field of aquatic ecotoxicology by allowing the functional interpretation of omic data, although annotation remains a significant challenge [1,16]. Recently, Simmons and Sherry [12] applied principle component analysis to demonstrate differences in the functional proteome of white sucker collected from the Athabasca River (AB, Canada) upstream of oil sands extraction compared with fish collected downstream of extraction facilities. That particular study was challenging because there were few suitable sampling locations along the Athabasca River within the Athabasca Oil Sand Bitumen Deposit that represented an appropriate reference and because there was high year-to-year sampling variation. Using only functional annotations of the large proteome dataset and data-reduction methods, those authors provided evidence that fish downstream of oil sands facilities were facing different environmental challenges from fish upstream. Another advantage of using functional datasets is their particular relevance to studies addressing epigenetic and posttranslational modification [48–51]. Ultimately, a systems toxicology study that has functional perspective gains relevance and meaning, and for that reason most omic datasets contain at least some functional information.
Differential expression

Differential or relative expression data are common in omics datasets, where fold-changes in expression between a negative reference and a treatment group are extracted from the data. Differential datasets demonstrate what aspects of the biological system have changed in response to a treatment in a highly efficient manner and result in a smaller dataset because of the removal of data not directly related to the effect. However, there are potential drawbacks to the differential approach. For instance, interindividual variability can reduce statistical power, and selection of a suitable negative reference group outside the laboratory environment can become complex. Additional statistical questions can arise in the differential approach, for which there are few straightforward answers. For example, at what point in the data analysis should the differential comparison be employed, which reference individuals should be compared with which treatment individuals, and should individuals be pooled before comparison? Oberg and Vitek [52] address many of these issues with a well-described review on how to apply the principles of statistical design to quantitative proteomic experiments; their concepts can easily be applied to transcriptomics and metabolomics approaches as well. Specifically, they discuss how to include randomization, replication, blocking, and pooling into different proteomics workflows (labeled and label-free study designs) and the benefits and drawbacks for each scenario. Zhang et al. [53] also present various statistical tests (the \( t \) test, \( G \) test, \( LPE \) test, and \( AC \) test) that can be used for various scenarios to identify significant differences in expression. In contrast to the analysis of relative expression data from shotgun methods, a targeted approach measures and compares absolute expression. Classical regression analyses and multiple comparison statistics are more applicable in targeted approaches because the data are closer to the expected normal distribution.

Site selection

Traditional exposures are best suited to the laboratory because they allow the investigator to control variables such as dose, diet, temperature, and light cycle. The goal of most laboratory experiments in ecotoxicology is to make inferences regarding what might happen under similar conditions in the field. Because it is difficult to control exposures in the field, comparisons between biological responses are rarely as straightforward for field-obtained organisms as they are in laboratory experiments. A possible solution to the lack of control one has in the field would be to forgo the quest for a suitable reference site, which is rarely met, and instead use a gradient design. Comparisons can then be made based on severity of exposure and may even provide excellent data for regression analyses. The exposure gradient design also adds more weight of evidence to the results and thus greater decision-making power. Berninger et al. [54] applied transcriptomics to evaluate effluent gradients in Minnesota (USA) using a functional approach and principle component analysis, and the data demonstrated the utility of the gradient approach for addressing biological variability. Simmons et al. [37] applied proteomics tools in a gradient design using both functional analysis with contingency statistics and relative expression with analyses of variance to demonstrate which aspects of the functional proteome were responding to remediation efforts within the Thunder Bay Area of Concern on Lake Superior, Ontario, Canada. Thus, the gradient approach may provide a dataset that is similar in design to the traditional laboratory regime, but removes the necessity for an appropriate and ideal or nonconfounded reference site, which is increasingly difficult to find in the natural environment [55–57].

Repeated measures

Selection of biofluid or tissue type is a crucial aspect of experimental design. Tissues are excellent for targeting a specific type of molecular response, with the added benefit that biological variation and sex-specific differences might be reduced. Indeed, the majority of field applications reported to date in which omic tools have been employed with fish have been tissue based [24, 58–63]. However, despite the greater variability in biofluid characteristics (discussed above), the ability to collect samples nonlethally offers some considerable advantages over tissue-based studies. With care, it can be possible to sample the same individual repeatedly over time, thus allowing it to serve as its own control (i.e., repeated measures) [1, 47]. If the exposure has a temporal component, repeated measures offers a potential solution to the selection of a nonconfounded reference for field applications, in which relevant sites are often difficult to identify. For example, the makeup of wastewater treatment plant effluents is often dynamic in nature [7, 18]. Thus, greater statistical power can be achieved in a repeated measures design, somewhat offsetting the greater variation inherent in biofluids. Second, it is important that the principle of the 3Rs (replacement, refinement, and reduction) is respected and that the impact of sampling is minimized to the greatest extent possible [64]. Because biofluid samples (e.g., plasma and urine) can be collected nonlethally, greater numbers of samples can be collected with minimal impact. However, for repeated sampling of internal biofluids, fish may need to be fitted with catheters, the use of which might need to be supervised by a veterinarian. Finally, the use of biofluids can provide a holistic picture of the health (or lack of health) of the organism because of the systemic role of biofluids [65–67]. Unfortunately, repeated captures can be difficult in the field, and thus the repeated measures concept may be limited to caging studies in practice.

In vitro studies

The use of cell cultures, whether primary or cell lines, is an option for the control of genetic and phenotypic variability [68]. Cell cultures derived from several ecotoxicologically relevant organisms (e.g., zebrafish [\( Danio rerio \)], rainbow trout [\( Oncorhynchus mykiss \)], and top minnow [\( Poeciliopsis lucida \)]) are commercially available and typically display lower biological variability than populations of whole organisms [69]. In a typical test, either the environmental matrix (e.g., surface waters) or an extract of the sample can be applied to the cultured cells, usually in a multiwell plate, after which one can measure changes in transcripts, proteins, or metabolites. In vitro tests are readily miniaturized, automated, and scaled up to facilitate high-throughput analyses [70–72]. Thus, large numbers of environmental samples can be assessed for biological impact relatively quickly and inexpensively. As proof of concept, Teng et al. reported the use of nuclear magnetic resonance to measure metabolomics in a zebrafish liver cell line (ZFL) that was exposed to the model synthetic estrogen 17\( \alpha \)-ethynylestradiol (EE2) [73]. Their results showed a clear dose response to increasing concentrations of EE2 as well as metabolite changes in keeping with those previously reported in vivo [74, 75]. Thus, even though cell cultures have their own complexities, they can be an attractive alternative to in vivo systems when low biological variability is a necessity (e.g.,
the detection of subtle responses to exposures) and when the focus is on responses to exposures within specific organs (e.g., liver). The low biological variability of cell cultures, however, can be disadvantageous if the goal is to understand or predict the ability of a toxicant to affect an inherently variable population of wild organisms. Moreover, if valid predictions are to be made from the results of cell-line-based exposures to whole organ or organism responses, it will also be important to understand the toxicokinetics of the test substance in the in vitro system, as recently demonstrated for the toxicity responses of cultured fish cells [76].

Replication

Statistical power is a measure of the probability of rejecting the null hypothesis when it is false, and it is a function of sample size and variance. One shortfall of transcriptomics studies, and omics in general, has been the lack of consideration of statistical power in many analyses. Few microarray studies in any taxa conduct a priori power analyses, primarily because of the lack of information regarding how the transcriptome responds to specific biological perturbations. Other emerging omic approaches, such as shotgun proteomics and RNA-sequencing, suffer from the same lack of information across all endpoints; thus, decisions regarding experimental design and method development are mostly made on an ad hoc basis. So a question quickly arises: How many replicate samples are enough to make an informed decision regarding a chemical exposure?

To investigate further, we used the free online program microarray sample size [77] to generate information regarding power and biological sample sizes in the context of the use of microarrays in ecotoxicological studies. Two power analyses were performed that investigated the relationships among false positives, power, and the number of microarrays required to meet specified criteria. The assumptions for both analyses were as follows: 1) The microarray contains 15-K gene probes; it is important to note that 60-K microarrays may require higher samples sizes. 2) The average fold change for a gene between control and treatment was ±2-fold (100% change in mRNA steady-state levels); based on our experience, most mRNA abundance changes tend to fall between 1.1- and 3.0-fold. 3) The standard deviation of a fold change of a gene is 0.7, a value recommended by the software that is realistic based on previous microarray experiments. 4) A power of 0.80 is reasonable for detecting departure from the null hypothesis. 5) The tolerated number of false positives was 750 (or 5% for a 15-K microarray). Based on these criteria, the number of microarrays required to reach a power of 0.8 was determined, as well as how many microarrays were required to reduce the number of false positives to <100. Moreover, we calculated how many microarrays were required to detect average fold changes <2.0. Based on our power analyses, at least 8 microarrays were required to obtain a power of 0.8 (Figure 1A), which is generally accepted as sufficient for rejecting the null hypothesis when there is a statistical difference among genes. Reducing the number of false positive to <100 required 13 microarrays per treatment group or site (Figure 1B). Lastly, to have sufficient power to detect fold changes approximating 1.5-fold, >23 microarrays were required, and 5 microarrays were required to detect changes in genes >2.5-fold (Figure 1C). A traditional power analysis performed using an online statistical power calculator [78] and data obtained from label-free LC–MS/MS shotgun proteomics data [12] determined that samples size should be 12 to 25 individuals, depending on the protein, to detect changes in expression that are approximately 25% different, which is similar to the numbers suggested by others for environmental effects monitoring programs [79]. These 2 examples clearly demonstrate that a trade-off must occur if the number of experimental groups is increased, such as multiple doses and time points, and selecting replicate numbers so that the experiment remains cost-effective and maintains adequate statistical power.

Figure 1. The relationship between the number of microarrays required and considerations for a well-designed transcriptomics experiment: (A) relationship between microarray number and experimental power; (B) relationship between microarray number and the number of false positives (false discovery rate of 5%); (C) relationship between microarray number and average fold change.
Timing

Finally, the importance of sampling time point when designing an omics experiment is worth mentioning. Responses at the metabolomic, proteomic, or transcriptomic level do not necessarily occur concurrently and cannot be assumed to be permanent, particularly if the response is compensatory or adaptive. This presents a considerable challenge, as a response could be missed entirely if sampling occurs at an inappropriate time point. For example, Bowman et al. [80] dosed largemouth bass with a single i.p. injection of estradiol (E2; 2 mg/kg) and observed a peak of Vtg and estrogen receptor mRNAs after 2 d. In another study, Villeneuve et al. [81] reported that E2 and estradiol-17β (E17β) caused a peak of Vtg mRNA after 24 h in fathead minnow, but recovery of plasma E2 was noted after 8 d of exposure in both dose groups. Although there are no specific guidelines as to the appropriate duration of exposure to measure a specific omics response, the aforementioned studies demonstrate the importance of including time course experiments to capture compensatory or homeostatic responses. It is also critical to match the exposure time to the question or hypothesis that is most relevant in the environment. In some scenarios, such as pulse exposures to agricultural runoff, the short-term response is of most value. In many other scenarios, such as exposure to historically contaminated sediments, the consequences of chronic exposure are most relevant.

ADDRESSING VARIABILITY WITH STATISTICAL ANALYSES

Multivariate statistics

The large quantity of data produced from omics approaches has necessitated the use of novel data analysis tools for the classification of sample groups. Among the most common are multivariate statistical models, such as principle component analysis or partial least squares discriminant analysis. The application of these statistical models with omics datasets is ubiquitous in the peer-reviewed literature, made possible by convenient, freely available data processing packages, such as MetaboAnalyst, MeltDB, metaP-Server, and others [82]. However, there is a growing concern that such tools, if not understood, could lead to gross misinterpretation or overfitting of data.

Principle component analysis is a convenient tool for reducing the dimensionality of a dataset, facilitating classification of sample groups (e.g., cases vs controls) based on observed variables. This is particularly useful for the definition of groups based on large datasets. However, because principle component analysis classifies samples based on the main source of variability in a dataset, it can lead to misinterpretation of data, in particular if the main source of variability is not relevant to the hypothesis being tested. As highlighted recently by Kjedhal and Bro [83], this problem is increasingly pronounced in untargeted datasets, which include thousands of individual variables, many of which are irrelevant for the particular study. The implementation of a blocking design (the arrangement of samples into similar groups) during analytical runs, as suggested by Oberg and Vitek [52], would also address this particular concern. In addition, we suggest that the application of principle component analysis should be performed after spectral data have been processed and matched to identifiable biological targets (i.e., metabolites, proteins, and transcripts) so that the variable loadings onto the components reflect true biological variation and not technical artifacts that could arise from batch-to-batch variation. The next step in a typical omics workflow often involves the proper identification of outliers, which can contribute a great deal of uninformative variation that can confound the identification of meaningful and often subtle responses.

Because outliers often contribute a considerable amount of variation to a dataset, principle component analysis is an obvious choice for effectively identifying them. In addition, objective statistical tools can be applied to principle component analysis score plots (e.g., Hotellings T²) such that outliers can be identified and excluded in a systematic and unbiased fashion, thus minimizing any subjectivity in the process of outlier identification. This step in the analysis of omics datasets is critical for the effective use of supervised approaches (e.g., partial least squares discriminant analysis) later in the analyses, which can be significantly hampered by the presence of outliers. The partial least squares discriminant analysis approach has even been shown to achieve excellent group separation on entirely randomized datasets [60]. When between-group similarity is not dominant in the data, partial least squares discriminant analysis can rank the influence a particular variable has on treatment-group separation, and is robust to multicollinearity. Thus partial least squares discriminant analysis is often considered superior to the unsupervised principle component analysis [84]. Increasing numbers of articles have been published highlighting the scarcity of proper validation among multivariate statistical models. Permutation testing, training datasets, and cross-model validation are required to prevent overfitting of data [17]. Finally, using functional classifications as variable inputs and genetic algorithms can increase biological relevance [12,17].

Post hoc testing

Because of the large numbers of comparisons among treatments, post hoc tests that correct for false discovery rates are commonly used to control for type I error (false positives) in omic datasets [84]. Post hoc tests controlling for false discovery rates (such as the Benjamini and Hochberg procedure) have been adopted instead of post hoc tests that adjust for family-wise error rates (such as the Bonferroni correction) in omic studies [84]. Based on a literature survey of 200+ transcriptomic studies in teleost fishes over a decade (2002–2011), (C. Martyniuk and L. Fanjoy, University of New Brunswick, St. John, NB, Canada, unpublished data) determined that the relative frequency of use of post hoc tests was the following; false discovery rate (Benjamini and Hochberg) > Bonferroni > Tukey’s range test > Fisher’s least significant difference > Newman–Keuls. Less frequently used post hoc tests were Sidak, Duncan’s new multiple range test, Dunnett’s test, and Dunn’s tests. Five studies stated that a multiple testing correction was not used, and approximately 50 manuscripts did not specify the post hoc test. An interesting approach is the proposed use of an optimal alpha [85], which could potentially be applied to large-scale datasets. The optimal alpha adjusts alpha to minimize the combination of type I and II errors by calculating the minimum of the average of alpha and beta at the critical effect size. Another solution to the post hoc problem is likely to emerge from the use of Bayesian statistics, which offers hierarchical and empirical models that can address false discovery rates (q-values) [86], differential expression of isoforms [87], and nonparametric datasets [88] without affecting type I and II errors [89]. Bayesian models have been applied to the analysis of RNA-sequencing [90], transcriptome array [91], label-free
proteome [92], and metabolomics [93] datasets, and we expect to see them used more frequently as omics technologies continue to develop.

**Nondetects**

When global statistical analyses such as those used for omics data are performed, the proper handling of nondetects and analytical incompleteness needs to be addressed [84]. Nondetects are an issue in statistical analyses because the true value is unknown; it could lie anywhere between zero and the detection limit of the instrument. Most commonly, analysts have simply deleted nondetects from their statistical analysis or have substituted the nondetect with one-half of the method detection limit. However, both of these methods create bias in the data [94]. Instead, it is recommended that either maximum likelihood estimation or nonparametric score testing be used to statistically describe the data [94]. In general, nonparametric score testing such as rank-sum, contingency tables, and survival analyses lend themselves well to analyzing classical toxicological datasets, and their usefulness in assigning significance to omic data is becoming increasingly apparent in functional and network analyses [37] and also for feature comparisons [53].

**Phenotypic anchoring**

Finally, phenotypic anchoring and linking responses to adverse outcome pathways offers the systems toxicologist the ability to draw powerful conclusions based on weight of evidence, particularly when data are combined from multiple approaches [3]. Phenotypic anchoring demonstrates that changes in molecular response can be correlated with physiological outcomes. Good examples of phenotypic anchoring in systems toxicology include changes to secondary sexual characteristics such as masculinization [59,95] and changes in organ tissue histopathology [18]. In another example, transcriptomics and proteomics studies were surveyed to identify common biological themes affected by androgenic hormones and chemicals [96]. What is needed is a weight of evidence approach [84]; meta-analyses identifying candidate biomarkers (genes, proteins, metabolites) that are responsive to classes of environmental chemicals should be more frequently conducted in ecotoxicology similar to that performed by Garcia-Reyero et al. [97].

**CHALLENGES AND RECOMMENDATIONS**

Over the past decade, the application of omic approaches in ecophysiology and ecotoxicology has increased exponentially and has generated mechanistic insights into biological responses associated with disease, toxicant exposures, hormones, and environmental stressors. Recent developments include next-generation sequencing for transcriptomics (RNA-sequencing) [98], as well as epigenetic studies and those characterizing genetic variation in aquatic toxicology. Targeted methods tend to have less technical variation, however, and shotgun methods facilitate discovery-based and exploratory research. Regardless of the method, omics approaches generate massive amounts of information from multiple genes, transcripts, proteins, and metabolites in a way that was not possible with traditional molecular methods. It is important to be aware, however, that each transcript, protein, and metabolite has its own inherent variability in expression that can be related to tissue, sex, physiological status, and age among other factors. Initial efforts have focused on characterizing individual variability in a limited number of scenarios (i.e., 1 sex, 1 treatment, or a single tissue). It is not entirely clear how that variability changes with age or physiology. The incorporation of gradient designs and repeated measures might help to more clearly identify biological effects that are the result of environmental exposures in the midst of high biological variation.

It is clear that the measurement of multiple endpoints leads to statistical challenges that require innovative solutions. Multivariate statistics such as principle component analysis and partial least squares discriminant analysis currently provide a good option for initial visualization of complex interactions in omics data because they provide intuitive plots. However, the biological meaning of separations among groups in a principle component analysis can be difficult to interpret, particularly if the treatment effect is small. Moreover, partial least squares discriminant analysis is susceptible to overfitting. Multiple comparisons within untargeted datasets have been a challenge for systems toxicologists because of heteroscedasticity among thousands of variables, high numbers of nondetects, and low sample sizes. Increases in sample size will increase statistical power, but can be a costly and time-consuming prospect. A suggested alternative to the use of conventional frequentist multiple comparisons tests for untargeted data is the use of Bayesian procedures. To facilitate the use of molecular endpoints in environmental monitoring programs and assessments, the effect size should be carefully evaluated. Thus, there is a need to collect baseline information on gene, protein, and metabolite responses in complex environments over time. Collecting baseline reference data is clearly not a trivial task. Once a baseline is established, however, a normal response range can then be generated for each molecule of interest. Responses outside the effect size can be considered a reliable indicator of biological effect.

**Acknowledgment**—The authors acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC) and STAGE-GDRI for funding various fellowships and projects that have collectively generated the information presented. We also thank H. Parikh for his technical support.

**Disclaimer**—The present study has been subject to review by the US Environmental Protection Agency (USEPA). Views expressed in the present review are those of the authors and do not necessarily represent views or policies of the USEPA.

**Data availability**—Data, associated metadata, and calculation tools are available by contacting the corresponding author (nina.simmons@ec.gc.ca).

**REFERENCES**


