Ecotoxicogenomics: Microarray interlaboratory comparability

Doris E. Vidal-Dorsch a, *, Steven M. Bay a, Shelly Moore a, Blythe Layton a, Alvine C. Mehinto a, Chris D. Vulpe b, Marianna Brown-Augustine b, Alex Loguinov b, Helen Poynton c, Natália Garcia-Reyero d, e, Edward J. Perkins e, Lynn Escalon e, Nancy D. Denslow f, Colli-Dula R. Cristina f, g, Tri Doanh h, Shweta Shukradash h, i, Joy Bruno j, Lorraine Brown j, Graham Van Aggen l, Paula Jackman k, Megan Bauer k

a Southern California Coastal Water Research Project, Costa Mesa, CA, USA
b University of California, Berkeley, Berkeley, CA, USA
c University of Massachusetts Boston, Boston, MA, USA
d US Army Engineer Research and Development Center, Vicksburg, MS, USA
e University of Florida, Gainesville, FL, USA
f Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional, Mérida, Mexico
g Agilent Technologies, Santa Clara, CA, USA
h Strand Scientific Intelligence Inc., San Francisco, CA, USA
i Environment Canada, North Vancouver, B.C., Canada
j Environment Canada, Moncton, NB, Canada

H I G H L I G H T S
- Environmental transcriptomic data are reproducible when following similar protocols.
- Subsets of genes can be consistently identified as differentially expressed.
- Different factors affect reproducibility including RNA quality and expertise.
- Variability can be reduced by training and following consistent protocols.

A R T I C L E   I N F O

Article history:
Received 22 January 2015
Received in revised form 27 July 2015
Accepted 2 August 2015
Available online xxx

Keywords:
Ecotoxicogenomics
Microarray
Gene expression
Intercalibration
Inter- and intra-laboratory reproducibility
Amphipod

A B S T R A C T

Transcriptomic analysis can complement traditional ecotoxicology data by providing mechanistic insight, and by identifying sub-lethal organismal responses and contaminant classes underlying observed toxicity. Before transcriptomic information can be used in monitoring and risk assessment, it is necessary to determine its reproducibility and detect key steps impacting the reliable identification of differentially expressed genes. A custom 15K-probe microarray was used to conduct transcriptomics analyses across six laboratories with estuarine amphipods exposed to cyfluthrin-spiked or control sediments (10 days). Two sample types were generated, one consisted of total RNA extracts (Ex) from exposed and control samples (extracted by one laboratory) and the other consisted of exposed and control whole body amphipods (WB) from which each laboratory extracted RNA. Our findings indicate that gene expression microarray results are repeatable. Differentially expressed data had a higher degree of repeatability across all laboratories in samples with similar RNA quality (Ex) when compared to WB samples with more variable RNA quality. Despite such variability a subset of genes were consistently identified as differentially expressed across all laboratories and sample types. We found that the differences among the individual laboratory results can be attributed to several factors including RNA quality and technical expertise, but the overall results can be improved by following consistent protocols and with appropriate training.

Published by Elsevier Ltd.
1. Introduction

Most established environmental assessment methods rely on whole-animal exposures and their survival, growth, development and reproductive effects evaluations (Ankley et al., 2006); yet, decision makers require screening programs that consider other adverse effects (Corvi et al., 2006). Hence, there is a rising need for innovative methods that can thoroughly investigate a wide array of biological effects after exposure to complex environmental contaminant mixtures. Transcriptomic responses could complement traditional ecotoxicology data, among other ways, by providing mechanistic information, identifying sub-lethal effects and helping in the identification of environmental contaminants. Recent studies have demonstrated the potential of transcriptomics as a toxicity identification evaluation tool (Antczak et al., 2013; Biales et al., 2013; Martinovic-Weigelt et al., 2014).

However, before gene expression data can be used in a decision-making context, it is necessary to determine its reproducibility and to identify key issues for standardization. Collaborative studies are needed to obtain the information required to characterize methods and differential expression data reproducibility (Brazma et al., 2001). Previous studies such as the Microarray Quality Control Project (MAQC) using mammalian cell lines demonstrated that microarray technology can be reproducible with careful consideration of technical issues (MAQC-Consortium, 2006). Environmental assessments or studies investigating the reproducibility of microarray analyses are uncommon. Data reproducibility can be affected by factors such as genetic variability in field collected organisms, use of custom-designed microarrays (which may lack some of the quality assurance/quality control present in standardized arrays for model organisms), varying sample preparation and preservation methods, and differing techniques for RNA extraction and quality assurance procedures.

The goal of this study was to determine the reproducibility of custom microarray data and to identify analysis steps that can impact the reproducibility of gene expression results. The current study investigated if extracted RNA quality, labeling and microarray hybridization success, and differential gene expression data were comparable among six laboratories with varying degrees of expertise. The amphipod species Eohaustorius estuarius, commonly used in coastal marine sediment assessments, was chosen for the present study. Standardized toxicity methods have been developed for this species and it is commonly used in regulatory programs (USEPA, 1994). An Agilent custom gene microarray was previously created for this species and it is widely used in regulatory programs. A summary of the amphipod exposure conditions is shown in Table SI-1.

Environmental assessments or studies investigating the reproducibility of microarray analyses are uncommon. Data reproducibility can be affected by factors such as genetic variability in field collected organisms, use of custom-designed microarrays (which may lack some of the quality assurance/quality control present in standardized arrays for model organisms), varying sample preparation and preservation methods, and differing techniques for RNA extraction and quality assurance procedures.

The study investigated if extracted RNA quality, labeling and microarray hybridization success, and differential gene expression data were comparable among six laboratories with varying degrees of expertise. The amphipod species Eohaustorius estuarius, commonly used in coastal marine sediment assessments, was chosen for the present study. Standardized toxicity methods have been developed for this species and it is commonly used in regulatory programs (USEPA, 1994). An Agilent custom gene microarray was previously created for this species and was used to develop molecular toxicity identification evaluations (Bay et al., 2012). An assessment of microarray reproducibility using standardized methods and data analysis will help to determine the potential to transfer microarray technology from purely research applications into monitoring and assessment applications.

2. Methods

2.1. Study design

Six laboratories analyzed replicate samples of E. estuarius amphipods exposed to cyfluthrin-spiked sediment and controls. Five of them ( Labs A, B, D, E and F) used the same commercially printed custom microarray and similar protocols for sample processing and microarray analysis. The sixth laboratory (Lab C) analyzed a subset of the samples using an older version of the microarray (Bay et al., 2012). Two sample types were used to evaluate the variability of array results due to various steps in the analysis (Fig. 1). First, each laboratory was given aliquots of RNA Extracts (Ex) from cyfluthrin-exposed amphipods and from control amphipods. Second, all laboratories were sent cyfluthrin-exposed and control whole body (WB) amphipods to extract the RNA themselves. The sample analysis results from each laboratory were compiled by a single laboratory and statistically analyzed to compare several aspects of RNA quality and differential gene expression within and between laboratories.

2.2. Exposure

The E. estuarius were collected from Yaquina Bay (Oregon, USA) and acclimated for three days under laboratory conditions. The amphipods were exposed to cyfluthrin-spiked and control sediments (spiked with carrier solvent) for ten days according to standard procedures (USEPA, 1994). A nominal cyfluthrin concentration of 1.6 μg/kg (1 μg/kg actual) was used, the measured amount corresponds to an LC20 value estimated from previous data (Bay et al., 2013). A summary of the amphipod exposure conditions is shown in the supplemental information (Table SI-1).

Sediment spiking methods are described in the supplemental information (SI). Tests were conducted at 15 ± 1°C, salinity of 20 g/kg, and under constant illumination. The amphipods did not receive supplemental feeding during the exposure. Water quality (total ammonia (NH3), dissolved oxygen, conductivity, pH, and temperature) was recorded at the beginning and at the end of the experiment for each experimental treatment. Water quality conditions were similar between control and cyfluthrin-exposed treatments. Dissolved oxygen was 8.4 ± 0.5 mg/L (average ± standard deviation), NH3 averaged 1.8 ± 0.5 mg/L, and pH averaged 8.1 ± 0.3.

A total of 33 exposure chambers were used for the cyfluthrin treatment and 28 chambers for the controls. A larger number of

![Fig. 1. Steps taken to prepare and analyzed each replicate. One laboratory (Lab A) generated all RNA extract samples sent to participating laboratories. All sample results were statistically analyzed by the same group.](image-url)
replicates was used for the cyfluthrin exposure since some mortality was expected for the pesticide concentration chosen. Each chamber contained 20 amphipods at the start of the exposure.

After 10 days, surviving amphipods were removed from the test chambers of each replicate group and pooled together. Pooled animals were then divided into seven sets and preserved for later analysis. One set was used for preparation of a common pooled RNA extract and the rest were distributed to the individual laboratories for extraction and analysis. A total of 32 samples were sent to each laboratory: eight replicates of cyfluthrin-exposed RNA extracts (Ex), eight control Ex replicates, eight replicates of whole body cyfluthrin-exposed amphipods (WB), and eight replicates of WB controls. Each WB sample contained 3–5 amphipods and each Ex sample represented the RNA mass equivalent of 3 animals. Mortality in both sample types was 25%. The animal’s individual weight varied, but using 3 or more amphipods ensured that each laboratory had a minimum RNA sample of 50 ng/µl.

2.3. RNA extraction

Amphipod total RNA was extracted using the RNeasy plus Mini Kit (QiagenEN, Valencia, CA) in most cases. Manufacturer protocols were followed. The RNA Ex samples were prepared by Lab A and sent to all laboratories. The RNA extracts received by each laboratory were originally extracted from a single group of amphipods, which were then aliquoted into separate samples for each laboratory. Eight replicate extracts were prepared; each represented a biological replicate. Preliminary testing was done to verify high RNA quality in the extracts. Lab E used RNA STAT-60 reagent (Tel-Test, Friendswood, TX) to extract its WB samples because the recommended protocol did not work in their case and they experienced low RNA yields using QIAGEN kits. Consultation with the manufacturer suggested that humidity conditions in the laboratory were causing improper filtrations in the kit’s columns.

The quality of the RNA was compared across laboratories using metrics for purity and integrity. The RNA was quantified using a Nanodrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE), and the absorbance ratios at A260/280 and A260/230 were used to assess purity. The A260/280 ratio reflected protein presence, while A260/230 indicated the presence of organic compounds such as carbohydrates and phenols, which absorb at 230 nm; ~2.0 values are generally accepted as “pure” for RNA for both ratios (Robert and Farrell, 2010). In this study we used values ≥1.8 as acceptable values.

Laboratories A, B, D, and E determined RNA integrity using an Agilent 2100 BioAnalyzer (Agilent, Santa Clara, CA). This technique uses the ratio between 18S and 28S RNA to calculate a RNA integrity number (RIN) number (Agilent report publication number 5989-1165EN). However, for E. estuarius it is difficult to obtain single 28S RNA peaks during extraction since the amphipod 28S RNA contains subunits, thus the RIN numbers obtained are generally low or in a few occasions cannot be calculated. In consultation with Agilent and based on our previous experience with this organism, we have determined that E. estuarius RNA samples with a RIN above 4 were of acceptable quality for microarray analysis. The electropherograms produced were also qualitatively assessed to ensure that RNA was not degraded. The RIN values were available only for Laboratories A, B, D, and E. Laboratory F used a different instrument, a Bio-Rad Experion Automated Electrophoresis System which also measures RNA quality but does not provide a RIN number, thus the values were not used for comparisons.

2.4. Microarray procedures

The microarray used for the present study is the second version of the custom E. estuarius microarray. The design was updated from the original by eliminating redundant probes, adding Agilent quality control probes, and including gene probes with a base composition or “BC” score of <4; BC scores indicate how well a particular probe is computationally predicted to perform, this prevents inappropriate signals such as those obtained by non-specific hybridization. The current version of the array (Design ID: 042708) contains 8507 probes (50 of them are replicated 10 times on the array) and 536 Agilent control probes. The previous version of the microarray (Design ID: 029610) contained 8610 probes.

Each laboratory applied their amphipod samples (Ex or WB) to Agilent’s 8 × 15,000 oligonucleotide microarray slides, except laboratory C that used a 4 × 44,000 slide format. The sample analysis was conducted with 4 microarray slides per laboratory. Each slide had an 8 × 15,000 design. All of these slides came from a single production batch. For both formats the probes were replicated. Lab C also used 4 slides for a total of 16 arrays to process samples. All the slots in the microarray slides, which were not used for control probes were used with probe replicates. After total RNA extractions, one-color Agilent Low Input Quick Amp Labeling Kits (Agilent, Santa Clara, CA) were used for cDNA synthesis, cRNA labeling, amplification, and hybridization following Agilent’s protocols (protocol #G4140-90040). The Cy3 dye incorporation or labeling efficiency values were compared across laboratories. After hybridization, the microarray slides were scanned on Agilent scanners (Table SI-2). Agilent’s Feature Extraction software was used to extract probe intensity data, as well as to document quality control results for each sample. The resulting data were compared among laboratories to determine probe hybridization quality across all laboratories. All microarray analysis data were added to the NCBI Gene Expression Omnibus (GEO) database (GSE64213).

2.5. Participating laboratories and their data

Four academic institutions and two government agencies participated in this study. All these laboratories had conducted gene expression analyses prior to those conducted for the current study. Three of them had conducted gene expression analysis with E. estuarius or other amphipods and the other three had not conducted studies with amphipods.

Five laboratories provided data on RNA extraction procedures and RNA quality. The sixth laboratory (Lab C) used a different microarray version and only analyzed previously extracted samples (Ex). Hence, Lab C results are only used in comparisons of differential gene expression of extracted samples. Five laboratories conducted the RNA quality data analysis, although no RIN data were available for Lab F (used a different instrument to assess RNA quality). Lab E analyzed fewer replicates due to difficulties with homogenization and RNA extraction of WB samples (Table SI-2). Analyses were conducted on data from 171 arrays that were successfully hybridized.

2.6. Statistical analysis

All gene expression data were collated and analyzed by a single facility using the same statistical methods. Data were resolved from microarray images using Agilent Feature Extraction (V.11). The raw intensity data were imported into GeneSpring GX (Agilent Technologies), background subtracted, quantile normalized, and Log2 transformed. Low quality data were flagged and removed before subsequent analysis using the software default parameters. Low quality data were filtered out by using a percent value shift filtering. Genes <20 percentile were removed from each treatment group. These genes were removed because they were considered not
differentially expressed. Differential gene expression profiles were calculated for each laboratory. Differential gene expression was calculated as follows: first the geometric mean was calculated across the replicate samples for both control and treated sample types; then t-test analysis was conducted with Benjamini–Hochberg FDR of p < 0.05; followed by fold change analysis (using Agilent GeneSpring GX V.12.6).

The data were normally distributed and one tailed analysis of variance (ANOVA) models were used to evaluate variability due to RNA extraction procedures (including tissue homogenization method and duration), RNA purity (A260/230 and A260/280 ratios) and integrity (RIN scores). Similarly, ANOVA was used to determine which factors (RNA extraction methods, RNA quality, or laboratory) contributed variance to Cy3 labeling efficiency. One sided Tukey tests were performed on statistically significant ANOVA models using R (http://www.R-project.org/; Wickham, 2009; Lawrence, 2012).

To compare interlaboratory results, the differentially expressed data (from treated replicate samples) were averaged within each laboratory. Kendall concordance coefficients (W) were computed using R. For this study, W values > 0.5 were considered to identify concordance, W values < 0.5 and > 0.3 were considered inconclusive, while W values < 0.3 were considered as no concordance (Legendre, 2005; http://www.statstodo.com/KendallW_Exp.php). The concordance coefficient was used to provide a measure of comparison based on covariation, agreement, and correspondence. Principal component analysis (PCA) was used to visualize the distribution of the data and to further investigate variability among laboratories control samples (using JMP Genomics V.6). Percent agreement was calculated as the percent of probes that were differentially expressed and in the same direction, in five or more laboratories for Ex samples. For this study 85% was considered as high agreement. Additionally, Pearson correlations were used to investigate intra-laboratory associations among all replicates for each laboratory. To incorporate the information from Lab C into the analysis the data were first filtered to ensure that only genes in common to both microarray designs were used for analysis.

3. Results

3.1. RNA quality for whole body samples

All laboratories extracting RNA from whole body (WB) samples had low RNA purity measurements with respect to A260/230 absorbance ratios. A minimum of 1.8 is often used as a quality acceptance threshold. However, none of the laboratories had median A260/230 values above 1.5, and the ANOVA results revealed similarities in this ratio between laboratories (Fig. SI-1). Percent contaminants from the RNA extraction procedure are known to absorb at 230 nm, in particular, guanidino thiocyanate, which is present in the Qiagen Lysis buffer. According to a Qiagen Newsletter (Qiagen, 2010) although concentrations of guanidino thiocyanate of 1 mM can reduce the 230/260 ratio below 1.0, concentrations of up to 100 mM do not impact downstream applications involving reverse transcription. Another potential contributor to the absorbance at 230 nm and the low 260/230 ratio is carbohydrates, which could result from the high chitin content in the amphipod exoskeleton. Despite the low A260/230 ratios, all labs were able to achieve cRNA yield and labeling efficiency that met or exceeded Agilent’s recommendation for microarray hybridization. This suggests that the contaminant(s) that absorbed at 230 nm did not impact RNA quality or inhibit downstream reactions.

All laboratories extracting RNA from WB samples had high removal of proteins during RNA purification; the A260/280 ratios were above 1.8. The ANOVA results showed that the individual laboratory conducting the RNA extraction and the homogenization method significantly affected the A260/280 ratios (p < 0.05). These ANOVA results indicated that during RNA purification the presence of variability was introduced by two factors, the personnel conducting the analysis at each laboratory (since the same RNA extraction method was used) and the different tissue homogenization methods. Tukey analysis revealed that laboratories A, B and D had comparable results while laboratories E and F were significantly different from each other and from the other laboratories; Lab E had the lowest A260/280 ratio (Fig. SI-1).

For WB samples, the RNA integrity scores were within the amphipod expected range (RIN > 4; see Methods). The score for most samples ranged from 4 to 6.8 (82%). Only 11 WB samples (18%) had RIN scores < 4. Subsequently, Tukey analyses showed two groupings in which laboratories A, D and E had similar RIN values, while D, E and B values were similar (Fig. SI-2). Lab B had the highest RIN values and Lab D the highest variability among laboratories. According to the ANOVA models the homogenization method and time did not significantly affect RNA integrity (p > 0.05).

3.2. Labeling efficiency

All laboratories achieved a sufficient level of Cy3 incorporation, although labeling efficiency varied among laboratories (Fig. SI-3). The ANOVA results indicated that the labeling efficiency was significantly different among laboratories and sample type (ANOVA p < 0.05). Tukey test results indicated differences among laboratories A, B and D, with A having the highest labeling efficiency and D the lowest. According to the ANOVA models, the RNA integrity (based on RIN values) significantly contributed to labeling efficiency differences among laboratories, but RNA purity was not a significant factor.

3.3. Gene expression

There was a higher percent of genes in common across the laboratories in the Ex samples (which had similar RNA extraction procedures and quality), compared to WB samples (Fig. 2). When comparing the Ex sample results, 1% of the total number of genes were identified as differentially expressed by most laboratories (5 of 6) and 95% not differentially expressed by all laboratories (Fig. 2).

Of the Ex sample-genes that were differentially expressed in one or more laboratories, 20% were found in common among all six laboratories and 80% were differentially expressed in at least one to five laboratories (4% of the total). A total of 60 genes were differentially expressed in Ex samples from all laboratories. Lab A had 162 genes differentially expressed, Lab B had 272, Lab C had 148, Lab D had 183, Lab E had 243 and Lab F had 219.

For the WB samples the data were only available for laboratories A, B, D, E and F. Among these laboratories, 83% of the genes were consistently identified as either differentially or not differentially expressed. Only 0.3% were identified as differentially expressed in all laboratories. A total of 24 genes were differentially expressed in WB samples from all laboratories. Lab A had 178 genes differentially expressed, Lab B had 264, Lab D had 48, Lab E had 91 and Lab F had 1260.

Further analysis of the differentially expressed genes in common among laboratories for Ex samples revealed that the number varied among laboratories. Laboratories A, B, D and E had a higher number of genes in common (95 of 795 total differentially expressed genes) than did the results when including Labs C and F (only 65 genes in common). Furthermore, if we were to exclude Labs C and F from the analysis presented in Fig. 2, the number of differentially expressed genes at labs A, B, D, E would change to 12%
instead of 1%. The two laboratories with the highest number of differentially expressed genes in common were Labs D and E (135 genes of 251 differentially expressed or 54%).

The lists of genes that were differentially expressed across laboratories were compared with a concordance analysis, which revealed a high degree of similarity in Ex samples. Their concordance was higher than that observed in WB samples. The Kendall coefficient (W) for Ex samples was 0.7 for Labs A, B, D, E and F, which used the same microarray version. When Lab C was included in the analysis the W decreased to 0.5 indicating a lower degree of concordance. For WB samples W = 0.4, and it could not be concluded that there was concordance in identification of differentially expressed genes.

Principal component analysis (PCA) was conducted to visualize the distribution of the control data using normalized, log transformed intensity data. Data visualization by PCA showed good intra- and inter-reproducibility of the microarray analyses for several laboratories. For the Ex samples individual control sample replicates of Labs A, B, D, E showed good reproducibility. While data from Labs C and F showed similarities among their replicates, but they were distinctive from those of the other laboratories (Fig. 3). Laboratory C analyzed the samples with a different custom microarray version. Laboratory F used a different scanning procedure when compared to the other laboratories. The analysis of WB samples also revealed variations in data similarity; Labs A, D and E were most similar, and Labs B and F were the least similar (Fig. 3).

Intra-laboratory reproducibility (within laboratories) was measured using correlation statistics. The results showed that differential gene expression data between the replicates analyzed by each laboratory are reproducible, particularly in Ex samples. Positive Pearson correlations for Ex samples ranged from 0.6 to 0.9 for laboratories A, B, D, and E; between 0.3 and 0.8 for Lab F and 0.5 to 0.8 for Lab C (Fig. SI-4).

The analysis results also showed positive correlations among replicates for WB samples, however there was a higher variability in these associations. The highest correlations were found among Lab A replicates (r = 0.6 to 0.9). The correlation results for Lab D, E and F indicated that differential gene expression patterns were less consistent among replicates (Fig. SI-5). For example, Labs D and E had inconsistent differential gene expression patterns in replicate samples 4 and 5. Lab F also had inconsistencies among replicates; in addition, their data also had a higher number of genes with statistically significant differential gene expression when compared to the other laboratories for WB samples.

The level of interlaboratory agreement in lists of differentially expressed genes for Ex samples was proportional to the magnitude of fold change. As the magnitude of differential expression increased so did the percent agreement among the laboratories (Table 1). The great majority of the up-regulated genes, which had more than 85% agreement among all laboratories, also had a magnitude of expression equal or greater to a 2 fold change (>1 log 2). In the case of the down-regulated genes, the threshold for high agreement was not as clear; only Labs A, B and D seemed to follow the pattern observed in the up-regulated genes.

Several genes showed high agreement in differential expression between both sample types. A total of 23 genes were differentially expressed in both sample types among all laboratories of 1539 differentially expressed in either sample (in one or more laboratories). While the magnitude of differential expression differed, there was agreement in the relative differential expression (increased or decreased; Fig. 4).

4. Discussion

The results of this study showed that microarray gene expression data are reproducible across and within laboratories; however, we identified several steps, which influence reproducibility such as variability in conducting analysis procedures, protocols interpretation, and sample quality. Differential gene expression results for splits of the same RNA extract (Ex) had greater reproducibility among laboratories than results obtained from samples that were extracted by different laboratories (WB). As all the samples were drawn from the same exposure, these results suggest that RNA extraction procedures, although using similar protocols between all labs (with one exception), can introduce variability in results. Similarly, the variability in the results using the same Ex samples suggests additional sources of variability, such as differences in implementation of the hybridization protocols, instrumentation, and laboratory specific experimental error. Despite the variability caused by these factors, a subset of differentially expressed genes with the highest degree of expression (>2 fold) was consistently identified across laboratories and sample types and with the same

![Fig. 2. Percent of genes and their expression across all six laboratories for extract samples and across five laboratories for whole body samples.](image-url)
The RNA extraction procedures and the resulting RNA quality are important factors to consider during analysis. During this study, it was challenging to fully assess its significance due to the characteristics of the amphipods and to the little information available to judge the RNA quality in these organisms. In this study we observed that Lab B and Lab A had significantly different RIN numbers in the five laboratories. No data is available for Lab D (Ex) except for the RIN scores) to the intra-laboratory variability suggesting that variability within a laboratory is affected by several factors, which are not easily measurable. For example, it is likely that technical expertise or differences in equipment may have impacted comparability among the laboratories.

In contrast, the observed differences in labeling efficiency did not appear to impact the resulting differential gene expression, perhaps because all of the values were above Agilent recommended values. In general, all laboratories achieved acceptable Cy3 labeling (>9 pmol/μg). Since all laboratories followed the same protocol, Cy3 labeling efficiency differences may be due to factors stemming from the type of equipment used (e.g., thermocycler efficiency differences). However, while it is necessary to obtain a minimum level of labeling for a successful hybridization, higher labeling efficiency was not necessarily predictive of more comparable gene expression results. For example, the labeling efficiencies of Lab A and B were significantly different, but these laboratories obtained similar gene expression results for Ex samples. Our results show that a specific labeling efficiency value is not necessary, as long as the value is above the recommended cutoff.

Slide scanning and data processing also influenced reproducibility. For instance, Lab F used an Agilent microarray scanner model that didn’t have an eXtended Dynamic Range (XDR) function. Without the XDR function the scanner automatically scanned the microarray slide twice and at two different PMT sensitivity levels. Scanners with XDR generate two linked images that Feature Extraction can co-process to yield a single unified set of extracted intensity data covering up to 6 – orders of magnitude. Without XDR function, gene signals max out at around 65,536 Fluorescent Units (FU), often this value becomes a limiting factor for gene expression studies because some genes are known to be expressed at higher levels. To overcome this issue for Lab F, the Agilent informatics team created a custom patch for Feature Extraction to combine the separate scans from lab F for the analysis but it was not exactly the same as other scanners, which used XDR. This might have been one of the factors that caused differences in the magnitude of the genes analysis and higher variability observed in the gene expression among WB replicates and decreased similarity of results with other groups.

Together our results suggest that differences in RNA extraction affected the reproducibility of differential gene expression data between the RNA samples extracted at multiple laboratories (WB) compared with those samples with RNA extracted at a single site (Ex). However, we could not directly connect the RNA quality (e.g., RIN scores) to the intra-laboratory variability suggesting that variability within a laboratory is affected by several factors, which are not easily measurable. For example, it is likely that technical expertise or differences in equipment may have impacted comparability among the laboratories.

The RNA extraction step has been previously identified as an important factor influencing differential gene expression results (Fleige and Pfafli, 2006). For example, Lab D had the highest variability in RNA integrity for their WB replicate samples which can be linked to lower RNA integrity scores for a couple of their samples. Subsequently, RNA quality variability could have affected differential expression results and caused the lower intra-laboratory agreement observed for Lab D (Fig. SI-5). Lab E used a different RNA extraction method for WB samples which may account for more variable RNA quality (as shown in the intra-correlation analysis) and higher variability observed in the gene expression among WB replicates and decreased similarity of results with other groups.

Together our results suggest that differences in RNA extraction affected the reproducibility of differential gene expression data between the RNA samples extracted at multiple laboratories (WB) compared with those samples with RNA extracted at a single site (Ex). However, we could not directly connect the RNA quality (e.g., RIN scores) to the intra-laboratory variability suggesting that variability within a laboratory is affected by several factors, which are not easily measurable. For example, it is likely that technical expertise or differences in equipment may have impacted comparability among the laboratories.

In contrast, the observed differences in labeling efficiency did not appear to impact the resulting differential gene expression, perhaps because all of the values were above Agilent recommended values. In general, all laboratories achieved acceptable Cy3 labeling (>9 pmol/μg). Since all laboratories followed the same protocol, Cy3 labeling efficiency differences may be due to factors stemming from the type of equipment used (e.g., thermocycler efficiency differences). However, while it is necessary to obtain a minimum level of labeling for a successful hybridization, higher labeling efficiency was not necessarily predictive of more comparable gene expression results. For example, the labeling efficiencies of Lab A and B were significantly different, but these laboratories obtained similar gene expression results for Ex samples. Our results show that a specific labeling efficiency value is not necessary, as long as the value is above the recommended cutoff.

Slide scanning and data processing also influenced reproducibility. For instance, Lab F used an Agilent microarray scanner model that didn’t have an eXtended Dynamic Range (XDR) function. Without the XDR function the scanner automatically scanned the microarray slide twice and at two different PMT sensitivity levels. Scanners with XDR generate two linked images that Feature Extraction can co-process to yield a single unified set of extracted intensity data covering up to 6 – orders of magnitude. Without XDR function, gene signals max out at around 65,536 Fluorescent Units (FU), often this value becomes a limiting factor for gene expression studies because some genes are known to be expressed at higher levels. To overcome this issue for Lab F, the Agilent informatics team created a custom patch for Feature Extraction to combine the separate scans from lab F for the analysis but it was not exactly the same as other scanners, which used XDR. This might have been one of the factors that caused differences in the magnitude of the genes

### Table 1

<table>
<thead>
<tr>
<th>Fold change</th>
<th>% Genes agreeing (5 of 6 Laboratories)</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ 1.4 (N = 44–70)</td>
<td>A 89 B 63 C 72 D 68 E 66 F 66</td>
</tr>
<tr>
<td>≥ 1.7 (N = 7–43)</td>
<td>94 86 88 87 87 100</td>
</tr>
<tr>
<td>≥ 2 (N = 0–39)</td>
<td>100 94 87 100 100 No data</td>
</tr>
<tr>
<td>≥ 2.4 (N = 0–33)</td>
<td>100 92 91 100 100 No data</td>
</tr>
<tr>
<td>≥ 2.8 (N = 0–18)</td>
<td>100 91 100 100 100 No data</td>
</tr>
<tr>
<td>Down-regulated</td>
<td>A 76 B 60 C 49 D 57 E 57 F 62</td>
</tr>
<tr>
<td>≥ 0.7 (N = 19–69)</td>
<td>89 90 59 86 77 75</td>
</tr>
<tr>
<td>≥ 0.6 (N = 4–48)</td>
<td>100 100 69 92 67 0</td>
</tr>
<tr>
<td>≥ 0.5 (N = 1–29)</td>
<td>100 100 69 100 50 No data</td>
</tr>
<tr>
<td>≥ 0.4 (N = 0–16)</td>
<td>No data 100 56 100 67 No data</td>
</tr>
<tr>
<td>≥ 0.35 (N = 0–9)</td>
<td>No data 100 56 100 67 No data</td>
</tr>
</tbody>
</table>
The use of different microarray versions also impacted result reproducibility. The original amphipod microarray version lacked some of the recommended control probes and 6276 probes present on the original were removed in second version. Several genes were differentially expressed only in Lab C, which used the original design; this lowered the concordance of results between laboratories. In general, the data from this laboratory was the least similar to that from the other laboratories. The origin of the differences between the results with the original and second array design is not clear but may be the result of batch differences in array preparation. Laboratories with higher number of genes differentially expressed were less comparable to those with lower numbers, since the number of genes expressed at background levels was higher.

Despite the observed discrepancies, comparable differential gene expression results were obtained for all laboratories. A subset of genes was differentially expressed in all of the laboratories and in the same direction of expression. Many of the genes differentially expressed were related to stress responses and to metabolic processes (e.g., heat shock proteins; Fig. 4). Other studies have highlighted the importance of having a common list of differentially expressed genes that distinguish biological conditions in exposed samples (Shi et al., 2008; Opitz et al., 2010). These types of consistent biological responses, despite the presence of factors that contribute to variability (e.g., expertise, methods selected, etc.), are needed to demonstrate that microarray technology is reproducible. It is encouraging that the level of agreement for a subset of genes across laboratories can be high despite variations within laboratories.

The outcomes of this study may have resulted in more variability that observed by previous intercalibration projects (e.g., MAQC). However, our study was fundamentally different in the sense that we used field-collected animals and evaluated biological replicates, which are intrinsically variable. The design of this study represents real word application of the “omics” technology for ecotoxicology analysis. The results of this study are very encouraging because despite all the variability expected from biological replicates and the potential variability added through the analysis steps, the microarrays provided a subset of common genes among different types of laboratories.

The present custom microarray reproducibility evaluation revealed that standardized methods for sample and data analysis are needed before this technology is transferred from research into monitoring and assessment. Multiple lines of evidence in this study revealed that it is possible to obtain reproducible microarray data (>50% differentially expressed genes in common were observed between the two laboratories D and E). However, the results of this study also indicate that there are opportunities to increase data reproducibility, for example through the implementation of standard operating procedures and training programs for tissue homogenization, RNA extraction, image acquisition, etc., as well as by establishing quality assurance and quality control parameters.
operators with high technical proficiency which ensure that proper quality control and assurance measures are followed throughout the analyses. In addition, standardized methods will need to be tailored to the specific organisms that will be investigated. For example, in the present study a commercially available RNA extraction kit was used. However, other methods such as phenol–chloroform procedures may be more suitable to extract RNA from amphipods, since these methods may isolate RNA from organisms with an exoskeleton more efficiently. Furthermore, some of the quality metrics and thresholds typically used to evaluate vertebrate sample quality (e.g., RIN) may not be appropriate for amphipod tissue due to unique characteristics of the 28S DNA (unpublished data).

The future research direction of the collaboration created for the present project will focus on investigating how the use of different statistical methods influences differential gene expression reproducibility. Previous studies have found that a great deal of variability can be introduced through differences in the methods to calculate gene differential expression (Abdullah-Sayani et al., 2006). Choosing the appropriate statistical analysis methods can be complicated since several approaches are available. Thus, the selection of the method to use may be based on factors such as statistical expertise or cost, and may not be the best for a given study. The current study used a t-test and FDR (5%) multiple test correction to determine differential gene expression, but many other methods such as the significance analysis of microarrays (SAM) are also commonly used. Such approaches will be investigated in future work to understand their impact on the results.

This study examined the interlaboratory reproducibility of analyses of samples obtained from a single experiment and showed encouraging results. Yet, other aspects of the toxicity testing process may impact the outcome of gene expression analyses and influence interlaboratory variability. An intercalibration study where each laboratory conducts the entire toxicity test is needed to provide an assessment of data comparability under conditions that are more representative of application in monitoring and assessment programs.

Acknowledgments

The authors wish to acknowledge the efforts of their staff and those of collaborating organizations, which have been instrumental to the success of this project. We thank SCCWRP staff Darrin Greenstein, Patricia Gonzalez, Fernando Vargas and Mary-Caitlin Jordan for conducting the laboratory exposures and coordinating sample storage and transfer. The authors would like to thank Agilent staff for all their support in this project particularly, Sharon Henne and the Genomics staff. This study was funded in part by Environment Canada for the Minister of the Environment (Contribution No. 1301230). In kind support for this project was also provided by all the participating laboratories.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.chemosphere.2015.08.019.

References


