Comparing microarrays to RNA-seq for transcriptomic analysis of whole fathead minnow larvae

Mitchell Kostich[1], David Bencic[2], Robert Flick[2], John Martinson[2], Weichun Huang[2], Greg Toth[2], and Adam Biales[2]

[1] The Jackson Laboratory for Genomic Medicine, Farmington, CT

[2] Molecular Indicators Branch, Great Lakes Toxicology and Exposure Division, Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. EPA, Cincinnati, OH

David Bencic | Bencic.david@epa.gov | 513-569-7201

www.epa.gov

Introduction

The fathead minnow (FHM; *Pimephales promelas*) is widely used in toxicological research. In the past two decades, modern 'omics approaches have been applied to this species in order to elucidate toxicological mechanisms of action and to develop the organism as a detection system for toxic chemicals in water. Recent work has suggested that microarray analysis of whole FHM larvae provides a viable and much less resource intensive alternative to transcript profiling of isolated tissues from adult FHM for detection of toxicants. Toxicant detection using transcript profile classification in whole animals can succeed even if measurements miss many low-expressed genes, if there are distinctive responses in highly expressed genes. A distinctive signature may also be based on genes indirectly affected by toxicant action, without temporal data, and without understanding the function of affected genes. Advances in sequencing technology have made RNA-sequencing (RNA-seq) an attractive alternative to microarrays for transcript profiling. In the present work, FHM larvae were exposed to sublethal concentrations of the toxicant bifenthrin, a neurotoxic pesticide, whose environmental concentrations are suspected of causing adverse effects in aquatic ecosystems. RNA-seq experimental parameters required to achieve similar performance to microarrays in the context of developing transcript profile-based classifiers for detection of toxic chemicals in water are presented.

Methods

FHM eggs were collected within 4-8 hours of fertilization, larvae hatched after 4 days, and exposures initiated at two days after hatching (6 days post fertilization). Static exposures were conducted for 48 h in an incubator maintained at 25 ± 1 ° C with a 16:8 h light:dark cycle. Exposure vessels were 250 mL glass beakers, filled with 100 mL of exposure water, with approximately 80% volume renewal after 24 h. Three sets of exposures, each including positive and negative controls, were performed on three consecutive weeks. For each exposure, larvae were exposed to moderately hard reconstituted water (MHRW) as a negative control or to (nominally) 1.6 µg/L bifenthrin. DMSO, used to dissolve bifenthrin and prepare a stock solution, had a final concentration of 0.002% in both treatment groups. Five beakers with 10 larvae per beaker were used for both negative control and bifenthrin exposure groups. From each beaker, two larvae at a time were transferred to a 1.5 ml centrifuge tube, water was removed, the tube was snap-frozen in liquid nitrogen and stored at -80 ° C until RNA isolation. Five pairs of larvae were collected from each beaker. Total RNA was isolated from samples using a protocol combining Tri Reagent with RNeasy Micro kit (Qiagen) and DNase digestion. Sixty-four total RNA (100 ng) samples were used for microarrays analysis, with at least 2 samples from each beaker for each treatment group from all three exposure experiments (n=2 samples or more per beaker, n=10 or more per treatment, and n=20 or more per experiment). Samples were randomly distributed within and across custom Agilent microarray slides for FHM (Agilent-036574 8x60k). Total RNA (1 µg) from the same 64 RNA samples were used for RNA-seq analysis. The TruSeq Stranded mRNA Library Prep kit (Illumina) was used for library preparation, with all libraries appearing as a single band ranging in size from 240-290 bp. Libraries were normalized and pooled for multiplex sequencing into 8 sets of 8, using the same sample combination as used for each microarray slide. Each pool was sequenced on an Illumina HiSeq 2500 Rapid Run flow cell as a single read 1 x 100 bp format, using Illumina HiSeq Rapid SBS reagents v2. Base calling was done by Illumina Real Time Analysis (RTA) v1.18.64 and output of RTA was demultiplexed and converted to FastQ format with Illumina Bcl2fastq v1.8.4.

Results & Discussion

For mapping efficiency, four different mapping programs and five different read lengths were tested, with and without trimming. STAR worked best for read lengths of 50 bases and less, while Bowtie2 worked best for longer read lengths. Trimming typically resulted in worse scores and mapping rates (Table 1). Six classifier algorithms were evaluated for both data types. Point estimates of the misclassification rate suggested that the best performing classifiers for RNA-seq and microarrays were random forest and elastic net, respectively (Table 2). Each of these classifiers was used for their respective data type for the rest of the comparisons. The effects of RNA-seq read lengths and sequencing depths were determined. Classifier performance at read lengths of 75 and 100 bases were significantly better than lower read lengths (Table 3) and all sequencing depths greater than 250k resulted in misclassification rates between 0-2% (Table 4). The effects of the number of biological replicates and batches on classifier performance were determined and while there was no difference with microarray training set size, with RNA-seq, as training set size increased, so did performance (Table 5). There was no significant difference between performance of classifiers based on expression microarray data versus RNA-seq data (Table 6).

Table 1: Sequence mapping parameters. The best scoring result for each combination of sequence length, trimming strategy and mapping program. PTotal: proportion of reads that were mapped. PRight: proportion of reads that mapped to the expected strand. PWrong: proportion of reads mapping to the wrong strand. Score: heuristic score = PRight - 2 x PWrong.

Length	Trimmed	Method	PTotal	PRight	PWrong	Score
25	FALSE	star	0.76772	0.74426	0.02346	0.69733
25	TRUE	star	0.76800	0.74372	0.02428	0.69515
25	FALSE	bbmap	0.79878	0.75635	0.04243	0.67149
25	TRUE	bbmap	0.80586	0.75951	0.04635	0.66680
25	FALSE	bowtie2	0.69736	0.67680	0.02056	0.63569
25	TRUE	bowtie2	0.68102	0.66097	0.02005	0.62086
35	FALSE	star	0.78192	0.75733	0.02459	0.70816
35	TRUE	star	0.78220	0.75738	0.02482	0.70773
35	FALSE	bowtie2	0.77636	0.75043	0.02593	0.69856
35	TRUE	bowtie2	0.77535	0.74944	0.02591	0.69762
35	FALSE	bbmap	0.77938	0.74942	0.02997	0.68948
35	TRUE	bbmap	0.77962	0.74948	0.03014	0.68919
35	FALSE	bwa	0.74114	0.72019	0.02096	0.67828
35	TRUE	bwa	0.72254	0.70233	0.02021	0.66190
50	TRUE	star	0.77672	0.75507	0.02165	0.71177
50	FALSE	star	0.77624	0.75474	0.02150	0.71174
50	FALSE	bowtie2	0.78823	0.76121	0.02702	0.70716
50	TRUE	bowtie2	0.78790	0.76089	0.02702	0.70686
50	FALSE	bwa	0.77821	0.75380	0.02441	0.70499
50	TRUE	bwa	0.77754	0.75321	0.02433	0.70455
50	FALSE	bbmap	0.78163	0.75326	0.02837	0.69652
50	TRUE	bbmap	0.78171	0.75328	0.02843	0.69642
75	FALSE	bowtie2	0.79937	0.77128	0.02810	0.71508
75	FALSE	bwa	0.79431	0.76785	0.02647	0.71492
75	TRUE	bowtie2	0.79926	0.77113	0.02813	0.71488
75	TRUE	star	0.77303	0.75358	0.01945	0.71468
75	FALSE	star	0.77276	0.75337	0.01939	0.71459
75	TRUE	bwa	0.79403	0.76754	0.02648	0.71457
75	TRUE	bbmap	0.79372	0.75975	0.02697	0.70280
75	FALSE	bbmap	0.78569	0.75806	0.02763	0.70279
100	FALSE	bowtie2	0.80774	0.77904	0.02871	0.72162
100	TRUE	bowtie2	0.80695	0.77836	0.02860	0.72116
100	TRUE	bwa	0.80402	0.77630	0.02772	0.72086
100	FALSE	bwa	0.80427	0.77645	0.02782	0.72081
100	TRUE	star	0.76859	0.75067	0.01792	0.71484
100	FALSE	star	0.76528	0.74769	0.01759	0.71251
100	TRUE	bbmap	0.78287	0.75779	0.02508	0.70763
100	FALSE	bbmap	0.78262	0.75718	0.02544	0.70629

Table 2: Classifier algorithm performance. Type: type of test data; 'Mix' is a balanced mix of positive and negative control samples; 'Neg' is negative control samples; 'Pos' is positive controls samples. L01: the misclassification rate (mean 0/1 loss). Low: nominal lower 95% confidence bound on L01. High: nominal upper 95% confidence bound on L01.

Platform	Algorithm	Type	L01	Low	High
RnaSeq	Random Forest	Mix	0.022	0.019	0.026
RnaSeq	Random Forest	Neg	0.016	0.013	0.019
RnaSeq	Random Forest	Pos	0.028	0.023	0.034
RnaSeq	Elastic Net	Mix	0.032	0.028	0.040
RnaSeq	Elastic Net	Neg	0.026	0.023	0.029
RnaSeq	Elastic Net	Pos	0.039	0.031	0.052
RnaSeq	Naive Bayes	Mix	0.035	0.032	0.039
RnaSeq	Naive Bayes	Neg	0.040	0.036	0.046
RnaSeq	Naive Bayes	Pos	0.030	0.025	0.036
RnaSeq	Partial Least Squares	Mix	0.027	0.022	0.033
RnaSeq	Partial Least Squares	Neg	0.012	0.010	0.014
RnaSeq	Partial Least Squares	Pos	0.041	0.034	0.051
RnaSeq	Support Vector Machine	Mix	0.061	0.052	0.073
RnaSeq	Support Vector Machine	Neg	0.084	0.079	0.090
RnaSeq	Support Vector Machine	Pos	0.038	0.026	0.063
RnaSeq	Gradient Boosting	Mix	0.046	0.042	0.052
RnaSeq	Gradient Boosting	Neg	0.035	0.032	0.039
RnaSeq	Gradient Boosting	Pos	0.057	0.051	0.066
Microarray	Random Forest	Mix	0.018	0.007	0.033
Microarray	Random Forest	Neg	0.018	0.001	0.045
Microarray	Random Forest	Pos	0.018	0.004	0.031
Microarray	Elastic Net	Mix	0.004	0.001	0.011
Microarray	Elastic Net	Neg	0.004	0.001	0.013
Microarray	Elastic Net	Pos	0.004	0.001	0.013
Microarray	Naive Bayes	Mix	0.013	0.004	0.025
Microarray	Naive Bayes	Neg	0.009	0.001	0.022
Microarray	Naive Bayes	Pos	0.013	0.001	0.027
Microarray	Partial Least Squares	Mix	0.011	0.001	0.029
Microarray	Partial Least Squares	Neg	0.000	NA	NA
Microarray	Partial Least Squares	Pos	0.022	0.001	0.054
Microarray	Support Vector Machine	Mix	0.051	0.031	0.067
Microarray	Support Vector Machine	Neg	0.009	0.000	0.027
Microarray	Support Vector Machine	Pos	0.094	0.063	0.112
Microarray	Gradient Boosting	Mix	0.045	0.025	0.067
Microarray	Gradient Boosting	Neg	0.009	0.000	0.022
Microarray	Gradient Boosting	Pos	0.080	0.045	0.116

Results & Discussion

Table 3: Read length versus performance. Random forest classifier performance at different RNA-seq read lengths. Type and L01 are as in Table 2. Entropy: mean cross-entropy loss. Suffixes 'low' and 'high' indicate the lower and upper bounds of nominal 95% confidence intervals.

Length	Type	L01	L01.low	L01.high	Entropy	Entropy.low	Entropy.high
25	Mix	0.029	0.024	0.036	0.188	0.156	0.234
25	Neg	0.022	0.016	0.027	0.152	0.108	0.238
25	Pos	0.038	0.028	0.047	0.226	0.186	0.274
35	Mix	0.029	0.024	0.035	0.178	0.146	0.218
35	Neg	0.029	0.021	0.037	0.164	0.119	0.225
35	Pos	0.030	0.023	0.038	0.193	0.153	0.250
50	Mix	0.037	0.030	0.043	0.138	0.117	0.168
50	Neg	0.027	0.020	0.035	0.113	0.088	0.154
50	Pos	0.046	0.036	0.055	0.165	0.135	0.210
75	Mix	0.006	0.003	0.011	0.048	0.038	0.063
75	Neg	0.000	NA	NA	0.019	0.017	0.023
75	Pos	0.013	0.006	0.021	0.077	0.060	0.102
100	Mix	0.008	0.004	0.013	0.044	0.036	0.058
100	Neg	0.002	0.000	0.004	0.022	0.019	0.026
100	Pos	0.014	0.007	0.024	0.068	0.053	0.089

Table 4: Sequencing depth versus performance. Random forest classifier performance at different RNA-seq read depths per sample. Rest of columns are as in Table 3.

Depth	Type	L01	L01.low	L01.high	Entropy	Entropy.low	Entropy.high
125k	Mix	0.107	0.093	0.121	0.469	0.411	0.537
125k	Neg	0.098	0.082	0.120	0.458	0.379	0.564
125k	Pos	0.115	0.096	0.134	0.480	0.404	0.573
250k	Mix	0.018	0.011	0.028	0.105	0.094	0.121
250k	Neg	0.008	0.003	0.014	0.078	0.069	0.087
250k	Pos	0.029	0.016	0.048	0.134	0.118	0.160
500k	Mix	0.020	0.015	0.028	0.107	0.082	0.144
500k	Neg	0.005	0.002	0.009	0.083	0.047	0.145
500k	Pos	0.036	0.027	0.048	0.132	0.103	0.183
1M	Mix	0.002	0.000	0.004	0.049	0.039	0.062
1M	Neg	0	NA	NA	0.017	0.014	0.022
1M	Pos	0.004	0.001	0.007	0.081	0.067	0.101
2M	Mix	0.003	0.000	0.006	0.038	0.028	0.051
2M	Neg	0	NA	NA	0.011	0.008	0.014
2M	Pos	0.006	0.001	0.012	0.066	0.051	0.086
4M	Mix	0	NA	NA	0.030	0.021	0.040
4M	Neg	0	NA	NA	0.006	0.004	0.010
4M	Pos	0	NA	NA	0.054	0.041	0.070
8M	Mix	0.004	0.001	0.008	0.038	0.027	0.054
8M	Neg	0	NA	NA	0.006	0.004	0.008
8M	Pos	0.007	0.002	0.016	0.072	0.054	0.095

Table 5: Replication versus performance. Performance of random forest (for RNA-seq) or elastic-net (for microarray) classifiers for different biological replicate configurations. NTrain: number of samples used for classifier training. NBatch: number of batches from which training samples were drawn. Rest of columns are as in Table 3.

Platform	NTrain	NBatch	Type	L01	L01.low	L01.high	Entropy	Entropy.low	Entropy.high
RnaSeq	10	1	Mix	0.0357	0.0311	0.0413	0.191	0.170	0.215
RnaSeq	10	1	Neg	0.0270	0.0228	0.0326	0.157	0.132	0.195
RnaSeq	10	1	Pos	0.0447	0.0373	0.0532	0.226	0.200	0.252
RnaSeq	20	1	Mix	0.0050	0.0025	0.0082	0.033	0.024	0.045
RnaSeq	20	1	Neg	0.0018	0.0004	0.0036	0.013	0.010	0.018
RnaSeq	20	1	Pos	0.0083	0.0037	0.0138	0.053	0.038	0.075
RnaSeq	20	2	Mix	0.0024	0.0011	0.0048	0.022	0.017	0.032
RnaSeq	20	2	Neg	0.0027	0.0011	0.0051	0.020	0.014	0.037
RnaSeq	20	2	Pos	0.0021	0.0005	0.0071	0.024	0.017	0.039
RnaSeq	40	2	Mix	0.0005	0	0.0014	0.006	0.004	0.010
RnaSeq	40	2	Neg	0	NA	NA	0.005	0.003	0.008
RnaSeq	40	2	Pos	0.0009	0	0.0028	0.008	0.004	0.015
Microarray	10	1	Mix	0.0078	0	0.0195	0.030	0.015	0.083
Microarray	10	1	Neg	0.0078	0	0.0234	0.037	0.011	0.135
Microarray	10	1	Pos	0.0078	0	0.0234	0.025	0.014	0.057
Microarray	20	1	Mix	0	NA	NA	0.014	0.010	0.022
Microarray	20	1	Neg	0	NA	NA	0.012	0.008	0.018
Microarray	20	1	Pos	0	NA	NA	0.018	0.009	0.037
Microarray	20	2	Mix	0	NA	NA	0.012	0.005	0.028
Microarray	20	2	Neg	0	NA	NA	0.016	0.006	0.039
Microarray	20	2	Pos	0	NA	NA	0.002	0.001	0.007
Microarray	40	2	Mix	0	NA	NA	0.004	0.002	0.009
Microarray	40	2	Neg	0	NA	NA	0.006	0.003	0.012
Microarray	40	2	Pos	0	NA	NA	0.001	0.001	0.004

Table 6: Platform comparison. Comparison of classifier performance using different RNA-seq read depths (using random forest and 100-base reads) and microarrays (using elastic-net). L01.delta: difference in average misclassification rate between RNA-seq and microarrays; positive values suggest microarrays were better, while negative values suggest RNA-seq was better. FDR: adjusted p-value for the null hypothesis that the difference is zero.

Depth	L01.delta	FDR
125k	0.023	0.240
250k	0	1
500k	0.005	1
1M	0.005	1
2M	-0.005	1
4M	-0.005	1
8M	0	1

Conclusions

- Classification may not require read lengths and read depths needed for comprehensive enumeration of differentially expressed genes.
- Microarrays and RNA-seq provide similar levels of classifier performance if reasonable read lengths and sequence depths are employed for RNA-seq.
- RNA-seq did not demonstrate a significant performance advantage over microarrays for toxicant detection.
- Lower cost, ease of use, additional data benefits, and the ability to work with a incomplete genome make RNA-seq a suitable substitute for microarrays.