Original Article
Environmental Health and Toxicology
Volume: 26, Article ID: e2011010: 7 pages
http://dx.doi.org/10.5620/eht.2011.26.e2011010 eISSN 2233-6567

EHT
Open Access

Analyses of Expressed Sequence Tags from *Chironomus riparius* Using Pyrosequencing : Molecular Ecotoxicology Perspective

Prakash M Gopalakrishnan Nair, Sun Young Park, Jinhee Choi

School of Environmental Engineering and Graduate School of Energy and Environmental System Engineering, University of Seoul, Seoul, Korea

Objects: *Chironomus riparius*, a non-biting midge (Chironomidae, Diptera), is extensively used as a model organism in aquatic eco-toxicological studies, and considering the potential of *C. riparius* larvae as a bio-monitoring species, little is known about its genome sequences. This study reports the results of an Expressed Sequence Tags (ESTs) sequencing project conducted on *C. riparius* larvae using 454 pyrosequencing.
Method: To gain a better understanding of *C. riparius* transcriptome, we generated ESTs database of *C.ripairus* using pyrosequencing method.
Results: Sequencing runs, using normalized cDNA collections from fourth instar larvae, yielded 20,020 expressed sequence tags, which were assembled into 8,565 contigs and 11,455 singletons. Sequence analysis was performed by BlastX search against the National Center for Biotechnology Information (NCBI) nucleotide (nr) and uniprot protein database. Based on the gene ontology classifications, 24% (E-value $\leq 1^{-5}$) of the sequences had known gene functions, 24% had unknown functions and 52% of sequences did not match any known sequences in the existing database. Sequence comparison revealed 81% of the genes have homologous genes among other insects belonging to the order Diptera providing tools for comparative genome analyses. Targeted searches using these annotations identified genes associated with essential metabolic pathways, signaling pathways, detoxification of toxic metabolites and stress response genes of ecotoxicological interest.
Conclusions: The results obtained from this study would eventually make ecotoxicogenomics possible in a truly environmentally relevant species, such as, *C. riparius*.

Key words: *Chironomus riparius*, Pyrosequencing, Ecotoxicogenomics

INTRODUCTION

Chironomus riparius (Chironomidae, Diptera), is widely used in aquatic exotoxicological studies for assessing acute and sub-lethal toxicities of contaminated sediments and for water monitoring due to their widespread occurrence, short life-cycle, easy to be reared in the laboratory, physiological tolerance to various environmental conditions [1,2]. To date, the endpoints used for monitoring such effects in *C. riparius* are based on a small number of specific biomarkers and measurements of organism level effects, such as survival and reproduction. Genomic-based techniques based on expression analysis of genes are important tools for investigating molecular level effects caused by exposure to environmental pollutants, which will provide the ability to detect mechanisms of action and subsequent adverse cellular level effects and associated with different types of toxicity [3,4]. As a pre-requisite for genomic based eco-toxicological studies knowledge of the *C. riparius* transcriptome is important but despite its eco-toxicological importance, no large scale transcriptome analysis of *C. riparius* has been done so far.

In a previous report Arvestad et al. [5] reported the transcriptome analysis of *C. tentans* midgut and an epithelial cell line using cDNA sequencing using conventional cDNA synthesis and sequencing method. However, with the advent of new high throughput pyrosequencing technologies using several genome sequencers, such as GS-FLX, GS-FLX-

Correspondence: Jinhee Choi, PhD
90 Jeonnong-dong, Dongdaemun-gu, Seoul 130-740, Korea
Tel: +82-2-2210-5622, Fax: +82-2-2244-2245
E-mail: jinhchoi@uos.ac.kr

Received: Apr 29, 2011, Accepted: Jul 06, 2011, Published Online: Aug 08, 2011
This article is available from: http://e-eht.org/

Titanium and SOLEXA, extensive cDNA sequence information can be obtained in a short period of time [6]. The GS-FLX pyrosequencer, from 454 Life Science/Roche, is well-suited for de novo transcriptome sequencing for the rapid production of sequence data with reduced time, labor and cost, and generates the longest reads [6,7]. Moreover, well characterized reference genomes of insects [8-10] could provide platforms for comparative genome analyses of non model organisms like *C. riparius*.

In this study, we present the first comprehensive characterization of the transcriptome of *C. riparius* 4th instar larvae using 454 pyrosequencing. Based on data corresponding to one single run on the FLX Gene Sequencer from 454 Life Science, almost 49,774,676 bases were assembled into transcripts and the majority of these have been annotated and functionally classified. In light of limited genomic and transcriptomic information, these data would significantly enrich the molecular aspects of *C. riparius* and its role in genomics based ecotoxicological studies.

MATERIALS AND METHODS

I. Insect Rearing and RNA Isolation

C. riparius strains were obtained from Korea Institute of Chemical Technology (Daejon, Korea). The larvae were reared on an artificial diet of fish flake food (Tetramin, Tetrawerke, Melle, Germany) in glass chambers containing dechlorinated tap water and acid washed sand, with aeration under a 16-8 h light-dark photoperiod at room temperature (20±1℃). The larvae were collected and total RNA samples were isolated using TRIZOL Reagent (Invitrogen Life Technology, USA). The RNA samples (A260/A280>1.8) were collected and mRNA was purified from the pooled total RNA (500 ㎍) by binding to oligo (dT) cellulose twice (Poly (A) Purist, Ambion).

II. cDNA Synthesis

For first strand synthesis, 10 ㎕ of the purified mRNA (5 ㎍), denatured at 65℃ for 10 minutes in a RNase free tube, rapidly chilled on ice, mixed with 5 ㎕ of 10X first-strand buffer, 5 ㎕ of 100 mM DTT, 5 ㎕ of dNTPs (2.5 mM each), 5 ㎕ of Oligo dT_{20} (50 ㎛), 2.5 ㎕ of Strata Script Reverse Transcriptase (200 U/㎕) in a 50 ㎕ reaction volume. First strand cDNA was synthesized at 42℃ for 60 minutes and cDNA synthesis and heat inactivated at 70℃ for 15 minutes and the tubes were placed on ice. For second strand cDNA synthesis, H_2O (61 ㎕), 100 M Tris-HCl, 20 ㎕ of second-strand buffer, dNTPs (6 ㎕, 10 mM each), DNA polymerase I (4 ㎕, 10 U/㎕), and RNase H (2 ㎕, 1.5 U/㎕) were mixed with the first strand synthesis reaction and incubated at 16℃ for 150 minutes. For end blunting, 23 ㎕ of blunting dNTP mix, 2 ㎕ of cloned pfu DNA polymerase (2.5 U/㎕) was incubated with the second strand synthesis reaction at 16℃ for 5 minutes. The cDNA was purified using QIAquick® PCR Purification Kit (QIAGEN, CA, USA) in a final elution volume of 50 ㎕.

III. Library Preparation

Approximately, 1㎍ of the final PCR product DNA was used to generate DNA library for Genome Sequencer FLX Titanium (Roche, Mannheim, GE). The fragments ends were polished (blunted), and short two adapters were ligated onto both ends. The adapters provide priming sequences for both amplification and sequencing of the sample library fragments, as well as the "sequencing key", a short sequence of 4 nucleotides used by the system software for base calling and, following repair of any nicks in the double-stranded library, release of the unbound strand of each fragment (with 5,-Adaptor A). Finally the quality of the library of single-stranded template DNA fragments (ssDNA library) was assessed using 2100 BioAnalyzer (Agilent, Waldbronn, GE), and the library was quantitated, including a functional quantitation to determine the optimal amount of the library to use as input for emulsion-based clonal amplification.

IV. Emulsion PCR

Single "effective" copies of template species from the DNA library to be sequenced were hybridized to DNA Capture Beads. The immobilized library was then resuspended in the amplification solution, and the mixture is emulsified, followed by PCR amplification. After amplification, the DNA-carrying beads were recovered from the emulsion and enriched. The second strands of the amplification products were melted away as part of the enrichment process, leaving the amplified single-stranded DNA library bound to the beads. The sequencing primer is then annealed to the immobilized amplified DNA templates.

V. Sequencing Run, Assembly, Annotation and Functional Categorization

After amplification, the DNA-carrying beads were set into the wells of a PicoTiterPlate device (PTP) such that wells contain single DNA beads. The loaded PTP was then inserted into the Genome Sequencer FLX Instrument, and sequencing reagents were sequentially flowed over the plate. Information from all the wells of the PTP is captured simultaneously by the camera, and can be processed in real time by the onboard computer. The reads were assembled using the GS De Novo Assembler (http://www.454.com/products-solutions/analysis-tools/gs-de-novo-assembler.asp). After assembly, the resulting contigs and singlets were

Table 1. Summary statistics of pyrosequencing of *C. riparius*

Pyrosequencing data	Total
Total number of high-quality reads	138,091
Total number of bases	49,774,676
Contigs	8,565
Singletons	11,455

Table 2. Length distribution of assembled contigs and singletons

Nucleotides length (bp)	Contigs	Singletons
100 - 199	245	1,659
200 - 399	2,630	4,825
400 - 599	3,481	4,963
600 - 799	807	8
800 - 999	510	-
1000 - 1999	733	-
> 2000	150	-
Maximum length	5,609	635
Average length	924	351
Total	8,565	11,455

Table 3. Summary of annotation of the *C. riparius* larval transcriptome

	Contigs	Singleton
Total number of sequences	8,565	11,455
Sequences with BLASTX matches in 'nr' database	5,114	4,710
Sequences with BLASTX matches in Uniprot data base	5,102	4,697
Number of sequences matching with known function gene in GenBank (e-value<=1.00E^{-5})	4,804 (24%)	
Number of sequences matching with unknown or unclassified function gene (e-value>1.00E^{-5})	4,708 (24%)	
Number of sequences with no significant homology (e-value>1.00E^{-5})	10,508 (52%)	

BlastX searched [11] against the protein databases "nr" and "Uniprot" (The UniProt Consortium, 2008). Functional categorization was done using database (http://mips.helmholtz-muenchen.de/projects/funca). The sequences were annotated using the Gene Ontology (GO) terms where possible according to molecular function, biological process and cellular component using database (http://www.geneontology.org/).

VI. KEGG Pathway Assignments

Pathway assignments according to Kyoto Encyclopedia of Genes and Genomes (KEGG) mapping was carried out using unique sequences that had BlastX scores with a cut off value of E = 10^{-5} The sequences were mapped to different KEGG biochemical pathways according to the EC distribution in the pathway database (http://www.genome.ad.jp/kegg/).

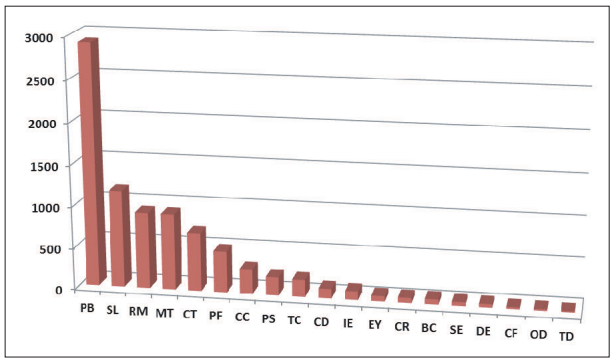

Figure 1. Top-ranked GO categories (molecular function) of assembled pyrosequencing ESTs.

PB: protein with binding function or cofactor requirement (structural or catalytic), SL: subcellular localization, RM: regulation of metabolism and protein function, M: metabolism, CT: cellular transport, transport facilities and transport routes, PF: protein fate (folding, modification, destination), CC: cellular communication/signal transduction mechanism, PS: protein synthesis, TC: transcription, CD: cell cycle and DNA processing, IE: interaction with the environment, IE: energy, CR: cell rescue, defense and virulence, BC: biogenesis of cellular components, SE: systemic interaction with the environment, DE: development (systemic), CF: cell fate, OD: organ differentiation, TD: tissue differentiation, GO: gene ontology, ESTs: expressed sequence tags.

RESULTS

I. Sequencing, Assembly and Sequence Analysis of Pyrosequencing ESTs

To get an overview of *C. riparius* transcriptome, total RNA was isolated from fourth instar larvae, and mRNA purification, cDNA synthesis, and sequence determination was done. These were then pyrosequenced, and in total, 138,091 reads were obtained, constituting a total of 49,774,676 bases of cDNA. Following the assembly of the sequences, a total of 8,565 contigs and 11,455 singletons were obtained (Table 1). Among the 8,565 contigs, 3,131 and 5,434 had lengths more than 500 and 100 base pairs, respectively (Table 2).

After assembly, the contigs and singletons were BlastX searched [11] against the protein databases "nr" and "Uniprot" (The UniProt Consortium, 2008). Of the 8,565 *C. riparius* contigs, 5,102 matched proteins in "Uniprot" and 5,114 matched proteins in "nr", while the numbers for the 11,455 singlets where 4,697 and 4,710, respectively (Table 3). After removing all redundant sequences, 9,512 sequences were obtained, representing a significant part of the *C. riparius* transcriptome.

II. Database Searching and Functional Annotation

Gene Ontology (GO) has been widely used to characterize gene function, annotation and classification [12]. As a whole, 24% of the sequences matched with known function

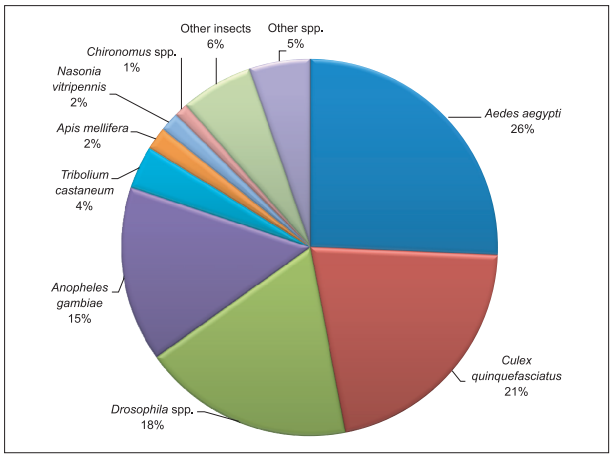

Figure 2. Percentage of *C. riparius* pyrosequence ESTs showing significant similarity with sequences in GenBank searched against NCBI.

NCBI: National Center for Biotechnology Information, EST: expressed sequence tag.

genes to existing gene models in BlastX searches (E-value $\leq 1^{-5}$), 24% showed no significant match and 52% of the pyrosequencing assemblies (E-value $\leq 1^{-5}$) did not match with any known sequences in the existing Genbank database and; thus, are likely to represent novel transcripts identified in this study (Figure 1). Similarities based on the results of BlastX searches showed the highest percentage of sequences match with *Aedes aegypti* (26%), followed by *Culex quinquefasciatus* (21%), *Drosophila spp*. (18%), *Anopheles gambiae* (15%), *Tribolium castaneum* (4%), *Apis mellifera* (2%) and *Nasonia vitripennis* (2%). A small number of sequences (1%) showed similarities with previously characterized genes of different *Chironomus* species and 6% of sequences matched with other insects and rest of the 5% matched with other (human, chicken, mouse, zebrafish, *C. elegans* or other organisms) species (Figure 2).

III. Pathway Analysis Based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) Classification

KEGG pathway analyses are widely used as a reference for the systematic interpretation of sequence data by linking individual genes to components of the KEGG biochemical pathways [13]. The pyrosequenced transcriptome of *C. riparius* was searched for the number of annotated gene sequences involved in shared specific KEGG pathways among animal phyla, using the unique sequences that had BlastX scores with a cut off value of E = 10^{-5}. It was found that 2908 (34.73%) genes were involved in proteins with a binding function or cofactor requirement, 1,165 (13.91%) genes were involved in sub-cellular localizations, 915 (11%) in metabolism, 917 (10.95%) in regulation of metabolism and protein function

Table 4. KEGG pathway mapping for *C. riparius* ESTs

KEGG pathway	No of transcripts
Metabolism	
Amino acid metabolism	81
Nitrogen, sulfur and selenium metabolism	17
Nucleotide/nucleoside/nucleobase metabolism	12
Phosphate metabolism	1
C-compound and carbohydrate metabolism	170
Lipid, fatty acid and isoprenoid metabolism	146
Metabolism of vitamins, cofactors, and prosthetic groups	2
Secondary metabolism	30
Energy	
Glycolysis and gluconeogenesis	27
Pentose-phosphate pathway	7
Pyruvate dehydrogenase complex	1
Tricarboxylic-acid pathway (citrate cycle, Krebs cycle, TCA cycle)	16
Respiration	7
Energy conversion and regeneration	7
Cell cycle and DNA processing	
Cell cycle and DNA processing	10
DNA processing	63
Cell cycle	55
Transcription	
Transcription	65
RNA synthesis	144
RNA processing	62
RNA modification	2
Protein synthesis	
Protein synthesis	143
Translation	41
Aminoacyl-tRNA-synthetases	46
Protein fate (folding, modification, destination)	
Protein fate (folding, modification, destination)	204
Protein folding and stabilization	64
Protein targeting, sorting and translocation	71
Protein modification	339
Assembly of protein complexes	4
Protein/peptide degradation	22
Protein with binding function or cofactor Requirement (structural or catalytic)	
Protein binding	792
Nucleic acid binding	666
Motor protein binding	52
Structural protein binding	63
Lipid binding	9
Amino acid/amino acid derivatives binding	2
C-compound binding	42
Metal binding	763
Nucleotide/nucleoside/nucleobase binding	1,019
Complex cofactor/cosubstrate/vitamine binding	411
Regulation of metabolism and protein function	
Regulation by	862
Regulation of protein activity	57
Cellular transport, transport facilities and transport routes	
Cellular transport, transport facilities and transport routes	466
Transported compounds (substrates)	273
Transport facilities	108
Transport routes	126

KEGG: kyoto encyclopedia of genes and genomes, TCA: tricarboxylic acid, EST: expressed sequence tag.

Table 4. Continued

KEGG pathway	No of transcripts
Cellular communication/signal transduction mechanism	
Cellular communication/signal transduction mechanism	136
Cellular signalling	152
Transmembrane signal transduction	74
Cell rescue, defense and virulence	
Stress response	51
Detoxification	33
Interaction with the environment	
Membrane excitability	2
Cell motility	6
Cell adhesion	86
Cellular sensing and response to external stimulus	10
Systemic interaction with the environment	
Plant / fungal specific systemic sensing and response	2
Animal specific systemic sensing and response	43
Cell fate	
Cell growth / morphogenesis	2
Cell death	8
Development (Systemic)	
Development (Systemic)	36
Animal development	8
Biogenesis of cellular components	
Biogenesis of cellular components	5
Eukaryotic plasma membrane	1
Cytoskeleton/structural proteins	28
Endoplasmic reticulum	1
Golgi	4
Nucleus	2
Mitochondrion	1
Peroxisome	4
Endosome	1
Vacuole or lysosome	2
Extracellular / secretion proteins	19
Tissue differentiation	
Animal tissue	13
Organ differentiation	
Animal organ	20
Subcellular localization	
Cell wall	2
Cytoplasm	299
Cytoskeleton	78
Cell junction	12
Endoplasmic reticulum	25
Golgi	41
Nucleus	453
Mitochondrion	95
Peroxisome	12
Vacuole or lysosome	8
Plastid	1
Extracellular / secretion proteins	250
Flagellum	1

KEGG: kyoto encyclopedia of genes and genomes, TCA: tricarboxylic acid, EST: expressed sequence tag.

and 700 (8.36%) in cellular transport, transport facilities and transport routes representing the largest groups with putative function, indicating the important metabolic activities in *C. riparius* (Table 4). These results showed the effectiveness of our transcriptome analysis of *C. riparius*, and indicated many of the candidate genes involved in many pathways and cellular processes.

Table 5. Examples of biomarker genes of ecotoxicological interest

Gene	ESTs	
	Contig number	Length bases*
Catalase	contig20333	1284
Cytochrome P450	contig03118	3868
Glutathione peroxidase	contig03949	509
Glutathione S-transferase	contig07420	1740
Heat-shock protein 24.1	contig04855	612
Heat-shock protein 27	contig17028	286
Heat-shock protein 67B2	contig00784	486
Heat-shock protein 90	contig20375	2593
Heat-shock protein70	contig07969	2125
Metallothionein	contig20500	1036
Superoxide dismutase	contig13190	1399
Thioredoxin reductase	contig04953	1730
Vitellogenin	contig06693	1528

ESTs: expressed sequence tags.
*Length of longest contig.

IV. Genes of Ecotoxicological Importance

A closer examination of the annotations revealed several genes that are of particular interest for environmental monitoring and ecological research. Several proteins catalyzing biotransformation of many xenobiotic and a number of important biomarkers for a large number of different compounds were present. Among the sequenced *C. riparius* transcripts, 117 different cytochrome P450 variants from 13 different families could be identified. Other interesting classes of proteins such as heat shock protein (24.1, 27, 67B2, 70, 90), genes coding for oxidative stress such as catalase, glutathione peroxidase, glutathione s-transferase, superoxide dismutase, thioredoxin reductase and several other biomarker genes such as metallothionein, vitellogenin were also present (Table 5).

DISCUSSION

C. riparius has been studied extensively because of their importance as an ecologically important biomonitoring species. However, due to limited knowledge of genomic resources necessary for mechanistic study, the effect of toxicants at the genomic level was rarely studied. This work describes the first assessment of the use of pyrosequencing in *C. riparius* and we have obtained a significant portion of the *C. riparius* transcriptome using this approach. To facilitate identifying sets of genes involved in a broad range of processes we developed our ESTs set from a normalized whole-body library.

As compared to Sanger-based approaches, which require cDNA cloning and bacterial transformation, transcriptome sequencing using massively parallel pyrosequencing exhibits

high sensitivity and detection of low-abundant transcripts [14,15]. Transcripts that previously have been hard to sequence can therefore be detected as in the case of Arabidopsis transcriptome profiling using pyrosequecing as reported by Weber et al. [16]. Even though, the length of the sequence is shorter than as compared to Sanger sequencing, the FLX Gene Sequencer used in this study generated 3,131 contigs with an average of 924 bp length which is longer compared to previous studies [17].

In our study, sequence names to the assembled sequences were given based on the best blast match for that sequence available in the public sequence data base and almost 50% of the genes were assigned gene names. However, another 50% of the sequences not matching to known genes in public sequence databases. In our studies we obtained 9,512, non-redundant genes and thus, a major part of the transcriptome of *C. riparius* has been obtained. One of the limitations in non-model organisms lacking fully sequenced genome where the transcriptome pyrosequencing is based on the number of genes expressed and without a fully-sequenced genome no clear data is available.

Genome sequences for the insects *D. melanogaster* [8], *A. gambiae* [9] and *A. aegypti* [10] have been reported and many other species are nearing completion. In our studies 81% of the *C. riparius* transcriptome closely related to insects *A. aegypti, D. melanogaster, C. quinquefasciatus*, and *A. gambiae* and therefore will provide a rich source of information for further investigation using comparative genome analyses. The expression levels of unknown transcripts were also as high as those aligning in annotated regions and these transcripts are likely to represent novel transcripts, which offer possibility to study new genes which may be specific to *C. riparius*. In earlier reports, many unique genes are observed in transcriptome studies of *M. sexta* [18].

In data analysis, many genes involved in different pathways, cellular processes and genes involved in metabolism of toxic substances or well-known biomarker genes (Table 4) are identified. Since *C. riparius* is extensively used in ecotoxicological studies, gene expression analysis are needed for mechanistic studies to understand changes in aquatic environment and the large collection of annotated sequences produced in this study represents a reasonably complete description of the *C. riparius* larval transcriptome. By correlating morphological as well as physiological charecters with molecular-level responses, caused by exposure to various toxicants, the subtle effect of various toxicants could be studied.

CONCLUSIONS

Pyrosequencing offers rapid characterization of a large portion of the transcriptome and therefore provides a comprehensive tool for gene discovery. Pyrosequencing the ESTs of 4th instar *C. riparius* larvae resulted in the identification of many sequences of ecotoxicological relevance. Analysis of the *C. riparius* transcriptome has revealed several gene candidates of ecotoxicological interest and further functional characterization will identify genes with relevance to ecotoxicology. The obtained transcriptome offers the additional option to design microarrays to study transcript regulation to understand the effect of environmental stressors. Transcriptome comparison with well-studied organisms will facilitate further the understanding of how environmental stressors effect at higher organisms levels using *C. riparius* as a model system. The platform will allow correlation of molecular-level responses, caused by exposure to various toxicants, to the unique morphological and physiological characters of *C. riparius*.

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MEST) (No. 2010-0027722).

CONFLICT OF INTEREST

The authors have no conflict of interest to declare on this study.

REFERENCES

1. Bettinetti R, Cuccato D, Galassi S, Provini A. Toxicity of 4-nonylphenol in spiked sediment to three population of *Chironomus riparius*. Chemosphere 2002; 46(2): 201-207.
2. Crane M, Sildanchandra W, Kheir R, Callaghan A. Relationship between biomarker activity and developmental endpoints in *Chironomus riparius* Meigen exposed to an organophosphate insecticide. Ecotoxicol Environ Saf 2002; 53(3): 361-369
3. Snell TW, Brogdon SE, Michael MB. Gene expression profiling in ecotoxicology. Ecotoxicology 2003; 12(6): 475-483.
4. Ankley GT, Daston GP, Degitz SJ, Denslow ND, Hoke RA, Kennedy SW, et al. Toxicogenomics in regulatory ecotoxicology. Environ Sci Technol 2006; 40(13): 4055-4065.
5. Arvestad L, Visa N, Lundeberg J, Wieslander L, Savolainen P. Expressed sequence tags from the midgut and an epithelial cell line of Chironomus tentans: annotation, bioinformatic classification of unknown transcripts and analysis of expression levels. Insect Mol Biol 2005; 14(6): 689-695.
6. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005; 437(7057): 376-380.
7. Hudson ME. Sequencing breakthroughs for genomic ecology and evolutionary biology. Mol Ecol Resour 2008; 8(1): 3-17.

8. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, et al. The genome sequence of *Drosophila melanogaster*. Science 2000; 287(5461): 2185-2195.
9. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, et al. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 2002; 298(5591): 129-149.
10. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, et al. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science 2007; 316(5832): 1718-1723.
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215(3): 403-410.
12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000; 25(1): 25-29.
13. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 1999; 27(1): 29-34.
14. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 2007; 8(7): R143.
15. Rothberg JM, Leamon JH. The development and impact of 454 sequencing. Nat Biotechnol 2008; 26(10): 1117-1124.
16. Weber AP, Weber KL, Carr K, Wilkerson C, Ohlrogge JB. Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. Plant Physiol 2007; 144(1): 32-42.
17. Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr, Grattapaglia D, Sederoff RR, et al. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics 2008; 9: 312.
18. Zou Z, Najar F, Wang Y, Roe B, Jiang H. Pyrosequence analysis of expressed sequence tags for *Manduca sexta* haemolymph proteins involved in immune responses. Insect Biochem Mol Biol 2008; 38(6): 677-682.