SUPERCOMPUTER ASSEMBLY AND ANNOTATION OF TRANSCRIPTOMES FOR ASSESSING IMPACTS OF ARMY STRESSORS ON ECOLOGICAL RECEPTORS

X. Chen*, K.A. Gust, M. S. Wilbanks, and E. J. Perkins

Environmental Laboratory, Engineer Research and Development Center, US Army Corps of Engineers,

Vicksburg, MS 39180-6199

N. D. Barker

Badger Technical Services, Engineering Research and Development Center, US Army Corps of Engineers,

Vicksburg, MS 39180-6199

D. Pham, L. Scanlan, and C. Vulpe

Department of Nutritional Science and Toxicology, University of California at Berkeley, Berkeley, CA 94720-3104

ABSTRACT

High-throughput DNA sequencing technology was utilized to describe the protein coding regions of genomic DNA (the transcriptome) for both Western Fence Lizard (*Sceloporus occidentalis*, WFL) and Japanese Quail (*Coturnix coturnix*, JQ). 928,759 and 559,819 total transcriptomic sequences for WFL and JQ, respectively, were clustered and assembled. Assembled unigenes with lengths ≥200 base pairs were annotated using Basic Local Alignment Search Tool (BLAST) against 5 publicly available protein sequence databases using the DoD supercomputers, Diamond (SGI Altrix ICE) and Jade (Cray XT4). A total of 58,962 and 44,455 unigenes were identified for WFL and JQ, respectively. Annotation of unigenes via similarity search against known proteins in the NCBI NR.aa and Refseq, EMBL-EBI UniProt-SwissProt, Uniref90, and Uniref100 protein coding databases provided 44 and 33 % unigene characterization for WFL and JQ, respectively. Sequences with significant similarity to known proteins were used to design custom ultra-high density gene expression microarrays which are being used to develop innovative methods to pro-actively assess the impacts of Army activity on environmental quality on installations. Further, this effort has developed a cyber-infrastructure capability with web-based tools and data visualization capability for the ERDC Environmental Laboratory to rapidly develop genomic infrastructure and gene expression tools for any ecological receptors that become species of concern.

1. INTRODUCTION

The utility of genomic tools has been broadly documented for the advancement of the biological sciences. Although genomic tool development for ecologically-relevant non-model species has lagged relative to model species, advancements in sequencing technology, bioinformatics processing, and gene expression platforms have led to an increasing number of non-model species having deep-coverage and well-annotated transcriptomes from which high-quality genomic tools have been produced [Rawat et al 2010]. We have developed a bioinformatics infrastructure and data processing pipelines to transition raw sequence data to robustly annotated coding genes to support gene expression profiling and biological impact assessment of Army stressors on ecological receptors (http://www.ifxworks.com/EnvironmentalSystemsBiology.html). These tools are being used to assess gene expression signatures in response to environmental perturbations in ecological model and environmental sentinel species such as Northern Bobwhite (*Colinus virginianus*), Fathead Minnow (*Pimephales promulus*), Earthworm (*Eisenia fetida*), Western Fence Lizard, Japanese Quail and, Staghorn Coral (*Acropora formosa*) [Garcia-Reyero et al 2009, Gong et al 2008, Gong et al 2007, Gust et al 2009, Rawat et al 2010, http://jeff.ifxworks.com/EGGT/]. These gene expression and cyber-infrastucture tools are proving to be indispensable as the focus of biological research and regulatory decision frameworks continue to shift toward systems biology and predictive toxicology approaches

[Heckmann et al 2008, Kavlock et al 2008, Robbens et al 2007].

2. MATERIALS AND METHODS

Tissue samples used to construct the normalized cDNA libraries for WFL and JQ were collected from five control animals of each sex for each species. The RNA pool used to construct the cDNA library for WFL included RNA extracted from brain, bone marrow, gut, heart, liver, ovary and testes tissues. The library for JQ included RNA extracted from the adrenal gland, brain, bursa, duodenum, heart, kidney, liver, lung, ovary, pituitary gland, spleen, testes and thyroid. All protocols were conducted consistent with Good Laboratory Practices and approved by the Institutional Animal Care and Use Committee at the U.S. Army Center for Health Promotion and Preventative Medicine.

2.1 Tissue Fixation and RNA Extraction

Immediately following euthanasia by CO2 asphyxia, tissue samples were fixed in RNA Later™ (Ambion, Austin, TX) following manufacturers recommendations. RNA extraction was conducted using RNeasy Mini RNA extraction kits (Qiagen Inc., Valencia, CA). RNA quality was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Waldbronn, Germany) with RNA 6000 Nano LabChips® RNA. Only samples with a 28s/18s ratio ≥2.0 and RNA integrity number (RIN) ≥7.0 were used for downstream applications. The RNA compilation for each WFL and JQ included 500ng of total RNA from each of the 46 and 44 total RNA samples collected for WFL and JQ, respectively.

2.3 cDNA Library Construction and Normalization

The SMART™ PCR cDNA Synthesis Kit (Clonetech Laboratories Inc. Mountain View, CA) was utilized to reverse-transcribe 1.0 µg of the total RNA sample into full length cDNAs for each WFL and JQ. The cDNA libraries were normalized prior to sequencing to capture both high and low abundance transcripts using the Trimmer cDNA Normalization Kit (Evrogen JSC, Moscow, Russia).

2.4 cDNA Sequencing

The normalized cDNA libraries for WFL and JQ were individually sequenced using massively-parallel pyrosequencing on a GS-FLX sequencer using a protocol to resolve 400bp reads. Briefly, cDNAs were nebulized and size-selected for 500 to 800 base pair fragments. Two primer sequences, Adaptor A and Adaptor B, were ligated to the fragments. cDNAs containing both an A and a B adaptor were melted into single stranded DNA, immobilized onto DNA capture beads and emulsified in oil for polymerase chain reaction (emPCR). The PCR emulsion was titrated to determine the optimal amount of ssDNA needed to create a 1:1 DNA fragment to bead ratio. emPCR was performed and the amplified library was loaded onto a 70 x 75 mm PicoTiterPlate and sequenced. A full PicoTiterPlate was used for each library preparation.

2.5 DNA Sequence Processing and Annotation

Genome-scale transcriptomes of WFL and the JQ were used for EST-based clustering and assembly via The Gene Indices Clustering Tools (TGICL) (Pertea et al. 2003), which uses megablast (Altschul et al. 1990) for homology-based clustering and CAP3 (Huang and Madan 1999) for assembly. Unigenes consisting of contiguous sequences (contigs) and singlets longer than 200 base pairs were selected for blastx (Altschul et al. 1990) homology-based coding potential detection and annotation against 5 sets of public available protein sequence knowledge sets: NCBI NR.aa (10,606,545 proteins) and Refseq (6,392,535 proteins), EMBL-EBI (http://www.ebi.ac.uk/) UniProt-SwissProt (515,203 proteins), Uniref90 (6,544,144 proteins), and Uniref100 (9,865,668 proteins). The CPU intensive computational biology analysis pipelines of clustering, assembly, and annotation were run via Portable Batch System (PBS) (http://en.wikipedia.org/wiki/Portable_Batch_System) through the DoD supercomputers Diamond, a SGI Altrix ICE with 1920 nodes and 15,360 cores, and Jade, a Cray XT4 system with 2146 nodes and 8,584 cores (http://www.erdc.hpc.mil/hardSoft/Hardware/home).

We have implemented a mature bioinformatics and computational biology system which includes: (1) a Relational Database Management Systems (RDBMS) - Oracle (Oracle, Redwood Shores, CA) for quick data retrieval and integration; (2) public and private data and results access via network shared file servers (http://jeff.ifxworks.com/EGGT/Quail_Lizard.html); (3) data and result visualization, data retrieval, and data mining via a public accessible web server (http://www.ifxworks.com/EnvironmentalSystemsBiology.html) and; (4) High performance and throughput computational analysis pipelines for quick data loading, retrieval, analysis, processing, integration, and validation (Figure 1 and 2).

Apache web server (http://www.apache.org/), html (http://en.wikipedia.org/wiki/HTML), JavaScript (http://en.wikipedia.org/wiki/JavaScript), and Perl CGI (http://perldoc.perl.org/CGI.html) were used for data web display and development of web-based tools for unigene and transcriptome/EST DNA sequence blast and

retrieval for data mining and microarray chip design (Figure 1 and 2).

3. RESULTS

The sequencing effort produced over 328 million base reads for the WFL and 189 million base reads for JQ in 928,780 and 559,833 sequence reads, respectively (Table 1). Average sequence read length for WFL was 354 bases and 348 for JQ. The sequence data sets were used to drive the development of bioinformatics and computational biology analysis pipelines to cluster, assemble and annotate protein-coding sequence data and select functional probes for microarray design.

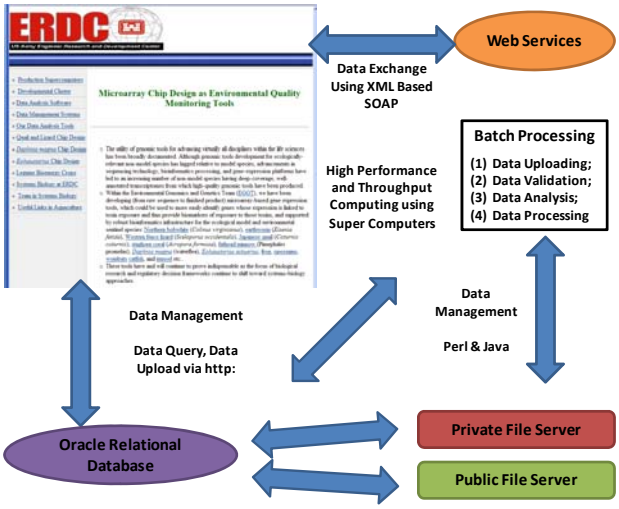

Figure 1. Bioinformatics system architecture used to implement data management, data security and web accessibility of publicly shared data and computational tools.

A total of 559,819 and 928,759 EST sequences out of 559,833 and 928,780 total DNA sequences after removing control sequences for sequencing runs, respectively, from JQ and WFL were selected for clustering and assembly. In all, 53,897 contigs and 5,065 singlets totaling 58,962 unigenes were identified for WFL, whereas 41,066 contigs and 3,389 singlets totaling 44,455 unigenes were identified for JQ (Table 2). Among the unigenes 33 to 44 % of singlets and contigs were annotated for protein-coding potential via homology-based annotation against NCBI NR.aa and Refseq, and EMBL-EBI UniProt-SwissProt, Uniref90, and Uniref100 protein sequence reference knowledgebases (Table 3).

The annotatable unigenes have been utilized to design custom ultra-high density microarrays via Agilent transcriptomic profiling technology. To facilitate access to the WFL and JQ datasets, web-based tools for DNA sequence manipulation such as blast and sequence retrieval (Figure 2) for both the raw and assembled

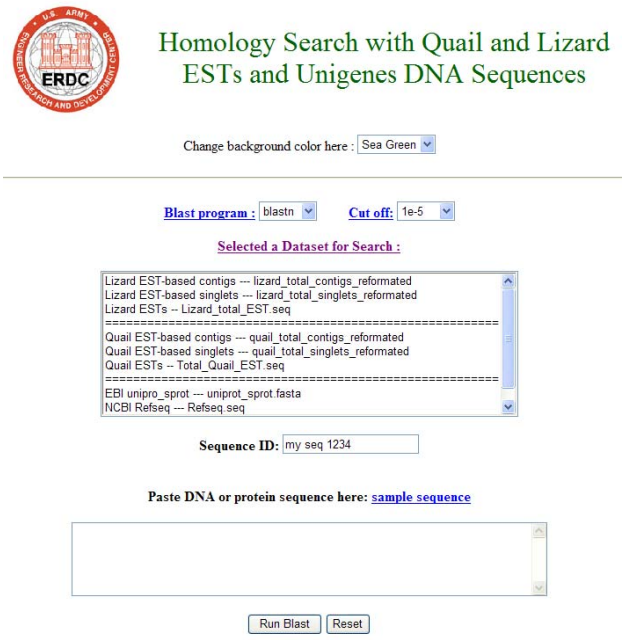

Figure 2. Sequence blast and retrieval web-based tools for ESTs / transcriptomes and unigenes of Western fence lizard and Japanese quail.

Table 1. Results of GS-FLX Pyrosequencing of normalized cDNA Libraries for Western fence lizard (WFL) and Japanese quail (JQ).

Sequencing Parameters	WFL	JQ
Raw Wells	2,125,263	1,157,019
Key Pass Wells	2,061,220	1,103,565
Passed Filter Wells	928,780	559,833
Total Bases	328,540,934	189,239,672
Length Average	354	338
Median Reads Length	397	388
Longest Reads Length	2,043	686
Shortest Reads Length	2	11

Table 2. Summary of sequence clustering and assembly for Western fence lizard (WFL) and Japanese quail (JQ).

Sequence Assembly	WFL	JQ
Total ESTs Available	928,759	559,819
Total Assembled Contigs	53,897	41,066
Total Singlets	5,065	3,389
Total Unigenes	58,962	44,455

Table 3. Unigenes homology-based coding potential detection and annotation against the following protein databases: NR.aa (10,606,545 proteins), Refseq (6,392,535 proteins), UniProt-SwissProt (515,203 proteins), Uniref90 (6,544,144 proteins), Uniref100 (9,865,668 proteins). WFL and JQ represent Western fence lizard and Japanese quail, respectively.

Unigene Dataset	Coding Detected	Non-Coding Detected	% Coding	Protein Database
WFL Contigs	23,385	30,512	43.39%	NR.aa
	23,173	30,724	43.00%	Refseq
	21,593	32,304	40.06%	UniProt-SwissProt
	23,463	30,434	43.53%	Uniref100
	23,508	30,389	43.62%	Uniref90
WFL Singlets	1,425	1,825	44.33%	NR.aa
	1,440	1,837	43.94%	Refseq
	1,457	1,820	44.46%	UniProt-SwissProt
	1,465	1,812	44.71%	Uniref100
	1,298	1,979	39.61%	Uniref90
JQ Contigs	17,873	23,193	43.52%	NR.aa
	17,732	23,334	43.18%	Refseq
	15,513	25,553	37.78%	UniProt-SwissProt
	18034	23,032	43.92%	Uniref100
	18,031	23,035	43.91%	Uniref90
JQ Singlets	1,208	2,181	35.65%	NR.aa
	1,195	2,194	35.26%	Refseq
	1,140	2,249	33.64%	UniProt-SwissProt
	1,217	2,172	35.91%	Uniref100
	1,211	2,178	35.73%	Uniref90

transcriptomes for each species have been developed for data mining and microarray chip design (http://jeff.ifxworks.com/EGGT/Analysis_Tools.html).

CONCLUSIONS

A computational biology analysis pipeline and web-based bioinformatics data visualization toolset for use in data retrieval, data mining, and genomics data processing have been established at the ERDC Environmental Laboratory for genome-scale transcriptome clustering, assembly, and annotation for the benefit of environmental sentinel species. The annotated unigenes and designed ultra-high density microarrays will be used in support of assuring environmental quality on installations. The development of this infrastructure has enabled a broad capability to develop highly robust monitoring tools to detect and assess the impacts of environmental stressors on Army ranges, to provide directed management for any target species found to be of concern on Army ranges. Specific examples of how tools developed for WFL and JQ are being used to support R&D for sustainment of the Army mission include: (1) Determination of corrected inter-species uncertainty factors for avian species that will improve the accuracy (and likely adjust highly conservative assumptions) of ecological risk assessment. (2) Determination of the influence of habitat degradation and global climate change on the toxicity and impacts of munitions compounds on the reptile model, WFL. Overall, this work has provided the infrastructure and tools to ensure population sustainability on Army ranges which in turn supports sustainability of the Army mission through enhanced and assured range access.

ACKNOWLEDGEMENT

We appreciate the assistance of Mr. David Dumas and Mr. Kenneth E. Lawrence at DoD Supercomputing Resource Center for kind support with operation of DoD supercomputers. This work was supported by the US Army Corps of Engineers Engineer Research and Development Center Program in Environmental Quality and Installations. Permission was granted by the Chief of the US Army Corps of Engineers to publish this information.

REFERENCES

Altschul, SF., Gish, W., Miller, W., Myers, EW., and Lipman, DJ., 1990: Basic local alignment search tool. *Journal of Molecular Biology*. 215:403-10.

Garcia-Reyero, N., Kroll, KJ., Li, L., Orlando, EF., Watanabe, KH., Sepúlveda, MS., Villeneuve, DL., Perkins, EJ., Ankley, GT., Denslow, ND., 2009:Gene expression responses in male fathead minnows exposed to binary mixtures of an estrogen and antiestrogen *BMC Genomics*. 10:308.

Gong, P., Guan, X., Inouye, LS., Deng, Y., Pirooznia, M., Perkins, EJ., 2008: Transcriptomic analysis of RDX and TNT interactive sublethal effects in the earthworm Eisenia fetida. *BMC Genomics.* 9 (Suppl 1):S15.

Gong, P., Guan, X., Inouye, LS., Pirooznia, M., Indest, KJ., Athow, RS., Deng, Y., Perkins, EJ., 2007: Toxicogenomic Analysis Provides New Insights into Molecular Mechanisms of the Sublethal Toxicity of 2,4,6-Trinitrotoluene in *Eisenia fetida*. *Environmental Science and Technology.* 41, 8195–8202.

Gust, KA., Pirooznia, M., Quinn, MJ., Jr, Johnson, MS., Escalon, L., Indest, KJ., Guan, X., Clarke, J., Deng, Y., Gong, P., Perkins, EJ., 2009: Neurotoxicogenomic investigations to assess mechanisms of action of the munitions constituents

RDX and 2,6-DNT in Northern bobwhite (*Colinus virginianus*). *Toxicological Sciences.* 110: 168–180.

Heckmann, L., Sibly, RM., Connon, R., Hooper, HL., Hutchinson, TH., Maund, SJ., Hill, CJ., Bouetard, A., Callaghan, A., 2008: Systems biology meets stress ecology: linking molecular and organismal stress responses in Daphnia magna. *Genome Biology.* 9: R40.

Huang, X. and Madan, A., 1999: CAP3: A DNA Sequence Assembly Program. *Genome Research*. 9: 868-877.

Kavlock, RJ., Ankley, G., Blancato, J., Breen, M., Conolly, R., Dix, D., Houck, K., Hubal, E., Judson, R., Rabinowitz, J., Richard, A., Setzer, RW., Shah, I., Villeneuve, D., Weber, E., 2008: Computational toxicology-a state of the science mini review. *Toxicological Sciences*. 103: 14–27.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lea, Y., White, J., Cheug, F., Parvizi, B., Tsai, J., and Quackenbush, J., 2003: TIGR Gene Indices Clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*. 19(5) : 651-2.

Rawat, A., Gust, KA., Deng, Y., Garcia-Reyero, N., Quinn, Jr. MJ., Johnson, MS., Indest, K., Elasri, MO., and Perkins, JE., 2010: From raw materials to validated system: The construction of a genomic library and microarray to interpret Systemic perturbations in Northern bobwhite. *Physiological Genomics*. 42:219-235.

Robbens, J., van der Ven, K., Maras, M., Blust, R., De Coen, W., 2007: Ecotoxicological risk assessment using DNA chips and cellular reporters. *Trends in Biotechnology.* 25: 460–466.