Rapid development of molecular resources for a freshwater mussel, *Villosa lienosa* (Bivalvia:Unionidae), using an RNA-seq-based approach

Author(s): Ruijia Wang, Chao Li, James Stoeckel, Gregory Moyer, Zhanjiang Liuand Eric Peatman
Source: Freshwater Science, 31(3):695-708. 2012.
Published By: The Society for Freshwater Science
DOI: http://dx.doi.org/10.1899/11-149.1
URL: http://www.bioone.org/doi/full/10.1899/11-149.1

Freshwater Science, 2012, 31(3):695–708

DOI: 10.1899/11-149.1
Published online: 24 May 2012

Rapid development of molecular resources for a freshwater mussel, *Villosa lienosa* (Bivalvia:Unionidae), using an RNA-seq-based approach

Ruijia Wang[1,4], Chao Li[1,5], James Stoeckel[2,6], Gregory Moyer[3,7], Zhanjiang Liu[1,8], AND Eric Peatman[1,9]

[1] *Fish Molecular Genetics and Biotechnology Laboratory, Department of Fisheries and Allied Aquacultures and Program of Cell and Molecular Biosciences, Aquatic Genomics Unit, Auburn University, Auburn, Alabama 36849 USA*

[2] *Molluscan and Crustacean Ecology Laboratory, Department of Fisheries and Allied Aquacultures, 203 Swingle Hall, Auburn University, Auburn, Alabama 36849 USA*

[3] *US Fish and Wildlife Service, Warm Springs Fish Technology Center, Conservation Genetics Laboratory, 5308 Spring Street, Warm Springs, Georgia 31830 USA*

Abstract. Freshwater mussels (Unionidae) are among the most endangered groups of organisms in the world, and their conservation and recovery are priorities throughout North America, especially the southeastern USA. We used a ribonucleic acid sequencing (RNA-seq)-based approach to develop molecular resources for *Villosa lienosa*, the little spectaclecase. We sequenced barcoded samples (Illumina HiSeq 2000 platform) and assembled (Trans-ABySS) 778,234 contigs (average length = 707.5 base pairs [bp]) from 162 million filtered reads. We identified 23,742 unigene hits against the National Center for Biotechnology Information nonredundant database and 36,582 microsatellites with sufficient flanking sequence for primer design. Microsatellite validation indicated a 36% polymorphic rate (16/44 tested markers) in the tested population (26 individuals; mean = 5 alleles/marker). Analysis of differentially expressed genes between heat-stressed and untreated controls enabled us to identify 604 genes involved in stress-response pathways. Real-time polymerase chain reaction validation of gene-expression results using individual samples confirmed RNA-seq patterns ($r = 0.847$, $p < 0.001$). RNA-seq is a powerful tool for rapid development of molecular resources in nonmodel species, and our study is the first large-scale transcriptome project in freshwater mussels. The validated microsatellite set and stress-associated genes are being used in parentage analysis and health-assessment surveys to support mussel conservation.

Key words: mussel, RNA-seq, transcriptome, heat stress, microsatellite, next-generation sequencing, Unionidae.

Freshwater mussels (Unionidae) are among the most endangered groups of organisms in the world (Lydeard et al. 2004). North America, particularly the southeastern USA, is a hotspot of freshwater mussel diversity and an area of special concern because of habitat degradation caused by changes in land use (stream alteration and impoundment), effects of introduced species, and losses of host fish necessary for mussel life cycles. The number of species and sizes of mussel populations have declined precipitously during the last 30 y to the extent that >½ of the freshwater mussel species in North America are now recognized as extinct or imperiled (Williams et al. 1993, Christian et al. 2007, Hoftyzer et al. 2008). Conservation and recovery of freshwater mussel species is a priority throughout North America, especially in the southeastern USA (Neves et al. 1997, Gidiere and Borden 2007, Strayer and Dudgeon

[4] E-mail addresses: rzw0010@tigermail.auburn.edu

[5] czl0020@tigermail.auburn.edu

[6] jas0018@auburn.edu

[7] greg_moyer@fws.gov

[8] liuzhan@auburn.edu

[9] To whom correspondence should be addressed. E-mail: peatmer@auburn.edu

2010). However, efforts aimed at captive propagation and reintroduction are hindered by significant gaps in our understanding of mussel assemblages, spatial arrangements of populations, naturally occurring sex ratios and male spawning contributions, and genetic variability within wild populations (Roe et al. 2001, Jones et al. 2006, Berg et al. 2007).

Molecular markers and tools can be useful for accurate identification of mussel species (Campbell et al. 2008, Boyer et al. 2011), assignment of parentage (Christian et al. 2007), surveys of population genetic diversity (Eackles and King 2002, Jones et al. 2004, Geist et al. 2009, Galbraith et al. 2010), and analysis of gene expression (Chaty et al. 2004). However, molecular resources for mussel research have developed slowly, and investigators have relied heavily on use of degenerate or conserved primers and adaptation of small marker sets from closely related species. As a result, molecular information beyond a handful of mitochondrial sequences is lacking for most mussel species (but see Inoue et al. 2010, Díaz-Ferguson et al. 2011, Williams and Moyer 2011). Such is the case with *Villosa lienosa*, the little spectaclecase, a mussel with a widespread distribution throughout Gulf of Mexico river drainages (Williams et al. 1993, Haag et al. 1998, Gangloff 2003). *Villosa lienosa* is common in small Coastal Plain streams. Thus, it is an ideal model species for studying mussel population and reproductive structure and responses to environmental perturbations. These studies require or can be significantly strengthened by molecular tools.

Next-generation sequencing combines nanotechnology, massive reaction parallelization, and microfluidics to achieve savings in reagent costs, labor, and time associated with nucleic acid sequencing. Previous techniques required construction of deoxyribonucleic acid (DNA) libraries, normalization, extensive cloning, and individual plasmid purification, but next-generation sequencing largely eliminates the need for these steps and has shortened genome projects by years. The result has been to open the door for cost-effective application of recent advances in genomic sequencing to critical ecological species and questions (see Ekblom and Galindo 2011). We used next-generation-based ribosomal nucleic acid sequencing (RNA-seq) to capture a significant portion of the *V. lienosa* transcriptome (expressed portion of the genome), thousands of potential marker loci, and stress-related expression signatures in a single lane of an Illumina HiSeq next-generation sequencing run. Our results are the first transcriptome sequence of a unionid mussel, significantly deepening the pool of molecular resources available for this taxon, and serve as a guide for similar studies in related taxa.

Methods

Sampling, RNA and DNA isolation

We collected 10 male freshwater mussels (*Villosa lienosa*) from Chewacla Creek (lat 32°32′9.50″N, long 85°29′48.19″W) near Auburn, Alabama (USA), in August 2010. We transported mussels to the Molluscan and Crustacean Ecology laboratory and held them at 10 ± 2°C for 4 mo. In late December, we brought 5 of the 10 mussels up to 29 ± 2°C over a 6-d period. We collected tissues including adductor muscle, mantle, and gill from all 10 mussels, placed them individually in 5-mL RNA later™ (Ambion, Austin, Texas), held them at 4°C for 1 d, and stored them at −80°C until analysis. Immediately before analyses, we ground the tissue samples separately to a fine powder in the presence of liquid N_2. We extracted total RNA with the RNeasy Lipid Tissue Mini Kit (Qiagen, Valencia, California) with DNase I (Invitrogen, Grand Island, New York) treatment according to the kit protocol. Following quantification and quality checking, we selected 2 individuals from each treatment group and pooled equal amounts of RNA from their component tissues. We used these 4 samples to prepare and sequence cDNA libraries. Libraries were individually barcoded (indexed) and combined within a single lane of an Illumina HiSeq 2000 100-base pair (bp) paired-end (PE) run (see *Illumina sequencing* below).

We collected 26 *V. lienosa* mussels from a 0.5-km reach of Chewacla Creek and preserved them in 97% alcohol. We collected ~20 mg of tissue from the foot of each mussel and transferred it into 600 µL of digestion buffer and 10 µL of 100 mg/mL Proteinase K. We isolated DNA with the Gentra Puregene Tissue Kit (Qiagen) following manufacturer's instructions. DNA concentration and purity were estimated on an Ultrospec 1100 Pro spectrophotometer (GE Sciences, Pittsburgh, Pennsylvania) and by electrophoresis on a 1.5% agarose gel.

Illumina sequencing

We prepared sequencing libraries with 2.14 to 3.25 µg of starting total RNA and processed them with the Illumina TruSeq RNA Sample Preparation Kit according to the TruSeq protocol. We amplified the libraries with 15 cycles of polymerase chain reaction (PCR), and TruSeq indices 1 to 4 were contained in the Illumina adaptors. The final, amplified library yields were 30 µL of double-stranded product (19.8–21.4 ng/µL) with an average length of 268 bp, indicating a concentration of 110 to 140 nM. After quantitation with KAPA Library Quant Kits (Kapa Biosystems, Woburn, Massachusetts) and dilution, the libraries

were clustered 4 per lane and sequenced on a HiSeq 2000 instrument with 100-bp PE reads at the Hudson-Alpha Genomic Services Laboratory (Huntsville, Alabama). We processed the image analysis, base calling, and quality-score calibration with Illumina Pipeline Software (version 1.5; Illumina, San Diego, California) and exported FASTQ-read files containing the sequencing reads, quality scores, and PE-reads information for trimming and assembly. We processed raw reads for initial trimming in CLC Genomics Workbench (version 4.7.2; CLC bio, Aarhus, Denmark) by removing adaptor sequences, ambiguous nucleotides ("N" at the end of reads), and low-quality sequences (quality scores < 20 or read length < 15 bp).

De novo assembly with various assemblers

We used the de Bruijn graph method as the primary algorithm in RNA-seq assembly. Briefly, reads are broken into smaller DNA sequences (k-mers, k denotes the sequence length in bases) (Zerbino and Birney 2008). Assembly is done by capturing overlaps of length $k - 1$ between these k-mers and generating contigs (series of overlapping DNA sequences). Several sequence contig assembly algorithms and software programs have been developed to assemble RNA-seq reads (Martin and Wang 2011). Given the importance of assembling long, accurate contigs to capture mussel genes and molecular marker loci and to identify differential expression correctly, we compared 3 de Bruijn graph-based assembler software packages for de novo transcriptome assembly: ABySS (version 1.2.5; Birol et al. 2009, Simpson et al. 2009), Velvet (version 1.1.04; Zerbino and Birney 2008), and CLC Genomics Workbench (CLC) (version 4.7.2). ABySS and Velvet are the 2 dominant assemblers for RNA-seq and produce multiple-k-mer assemblies. CLC is a commercial software program that uses an optimized high-speed single k-mer assembly module.

Forty-seven ABySS assemblies were produced with k-mer sizes from 50 to 96, erode-strand E = 0, and scaffold option off. The minimum number of pairs needed to consider joining 2 contigs was set to 10. All k-mer assemblies were scaffolded and merged into 1 assembly by Trans-ABySS (version 1.2.0; Robertson et al. 2010). Seven Velvet assemblies were produced with k-mers equal to 50, 55, 61, 67, 75, 85, and 97 when the range of k-mer size was set from 50 to 99 and insert length was set to 268. Contigs from all 7 assemblies were used as input to produce a final assembly using the AssemblyAssembler (version 1.3) module of Velvet. One assembly was produced with CLC (version 4.7.2) with k-mer set to 24 (maximum value is 31) automatically based on the total bp of the input data. Additional trimming was done in the 3 final assemblies from various assemblers, and contigs <200 bp in length were discarded because of a low annotation rate (Gotz et al. 2008). CD-HIT (version 4.5.4; Li and Godzik 2006) and CAP3 (Huang 1999) were used to remove assembly redundancy by setting global sequence identity in CD-HIT to 1, and the minimal overlap length and % identity in CAP3 to 100 bp and 99% (Fig. 1).

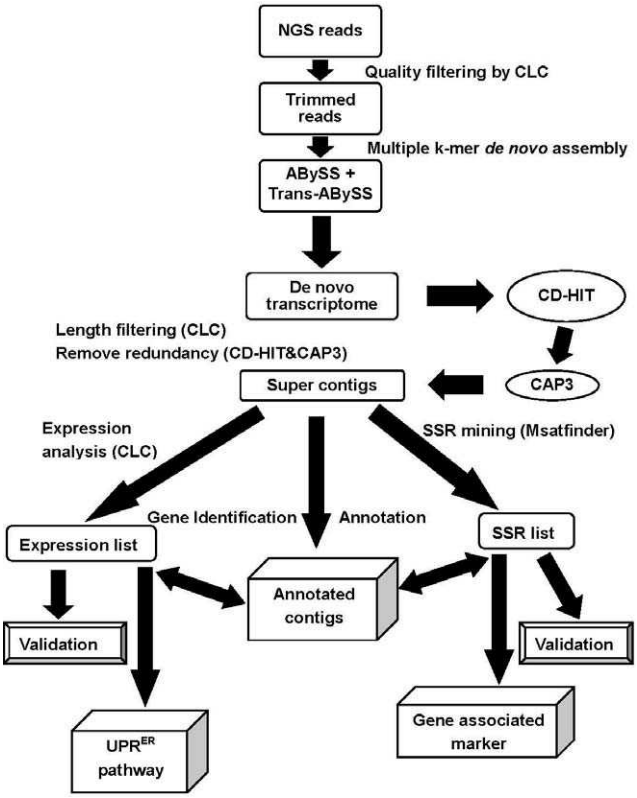

FIG. 1. Schematic presentation of mussel transcriptome analysis. NGS = next generation sequencing, SSR = simple sequence repeats, UPR^{ER} = endoplasmic reticulum unfolded protein response. CLC, ABySS, Tran-ABySS, CD-HIT, and CAP3 are molecular analysis tools (see Methods for details).

Gene identification and annotation

We used the final Trans-ABySS assembly contigs as queries against the National Center for Biotechnology Information (NCBI) nonredundant (nr) protein database, the UniProtKB/SwissProt database, and the RefSeq Invertebrates protein database. We used Basic Local Alignment Search Tool (BLASTX) for this process. We set the cut-off Expect value (E = the likelihood that the matching sequence is obtained by chance) to 1*e*–5 and returned only the top 10 hit

results of each query. Initially, we assigned the top gene identifications (gene ID) and names to each contig. We replaced hypothetical or uncharacterized top BLAST results with more informative hits from the top 10 list when available.

We implemented gene ontology (GO) annotation analysis with the UniProtKB/SwissProt and nr BLAST results in Blast2GO (version 2.5.0; Gotz et al. 2008), which is an automated tool for the assignment of GO terms. We imported the UniProtKB/SwissProt BLAST result and the unique, additional part of the nr BLAST result into Blast2GO. The final annotation file was produced after completion of gene-ID mapping, GO term assignment, annotation augmentation, and generic GO Slim processing. We categorized the annotation result with respect to functional terms by biological process, molecular function, and cellular component.

Transcript-level gene expression analysis of heat stress

We used the modified assembly of ABySS/Trans-ABySS as the pseudo-reference against which trimmed reads were mapped for analysis of gene expression. We calculated reads per kilobase of exon model per million mapped reads (RPKM) (Brockman et al. 2008) as the original expression value, normalized to ensure that samples were comparable (Allison et al. 2006, Brockman et al. 2008). We calculated the expression fold change based on the modified expression value between the heat and control groups. We used the Baggerly test (a weighted *t*-type test statistic; Baggerly et al. 2003) to test the significance of the expression fold change. We used the RNA-seq module and the expression analysis module in CLC to run this analysis. We set the threshold of gene expression selection to $p < 0.05$, mapped reads > 5, weighted proportions of fold change $\geq |2|$, and we divided the significantly differentially expressed genes into up- and down-regulated groups (Fig. 1).

Functional enrichment, pathway analysis, and real-time PCR validation

We did functional enrichment analysis in Ontologizer (Alexa et al. 2006, Falcon and Gentleman 2007) with the UniProtKB gene-association file (http://www.geneontology.org/). We used the gene identification list produced by the UniProtKB/SwissProt BLAST results as the population data set input to the Ontologizer program. We linked the up- and down-regulated results from expression analysis to the UniProtKB/SwissProt BLAST results and input the gene identification list of up- and down-regulated genes as the study set. We used the Parent–Child–Union algorithm (Grossmann et al. 2007) to identify the overrepresented GO terms in the study set. Briefly, we treated each GO term independently and took into account parent–child relationships. We measured the overrepresented GO terms with respect to the presence of their parental terms. We used a Benjamini–Hochberg false discovery rate test (FDR; Benjamini and Hochberg 1995) to calibrate the *p*-value and manually filtered uninformative upper-level GO terms.

In total, we selected 22 significantly expressed genes based on functional enrichment and pathway results and designed primers with Primer3 (available from: http://frodo.wi.mit.edu/primer3/input.htm). We converted 6 muscle RNA samples (derived from 6 mussels), including the 4 samples used in sequencing and 1 each from the heat and control groups, to cDNA using the iScript™ cDNA Synthesis kit (Bio-Rad Laboratories, Hercules, California). We diluted all cDNA products to 250 ng/μL and used them in the quantitative real-time PCR (q-RT-PCR) reaction with the SsoFast™ EvaGreen® Supermix on a CFX96 RT-PCR Detection System (Bio-Rad Laboratories). The thermal-cycling profile consisted of an initial denaturation at 95°C (30 s), followed by 40 cycles of denaturation at 94°C (5 s), and an appropriate annealing/extension temperature (58°C, 5 s). We used an additional temperature-ramping step to produce melting curves of the reaction from 65 to 95°C. We set the housekeeping gene 18S as the reference gene and calculated relative fold changes with the Relative Expression Software Tool (version 2009; Pfaffl et al. 2002) based on the cycle threshold (Ct) values generated by q-RT-PCR.

Microsatellite-marker identification and verification

We implemented the microsatellite-marker mining process in Msatfinder (version 2.0.9; Thurston and Field 2005) with a repeat threshold of 8 dinucleotide repeats or 5 tri-, tetra-, penta-, or hexanucleotide repeats. We randomly selected 47 markers with primer sequences from the mining result. We did the PCR validation on 26 individuals from which DNA had been isolated previously. We added an M13 forward tail sequence, 5′-GAGTTTTCCCAGTCAC-GAC3′, to the 5′ end of all forward primers and labeled them with IRD700 or IRD800 fluorescent label, which can be detected by the automated DNA sequencer (model 4200L, LI-COR Inc., Lincoln, Nebraska). The PCR cycling process included 2 annealing cycles, one consisting of 20 cycles with annealing temperature 57°C followed by another consisting of

TABLE 1. Summary of Illumina sequencing data of the freshwater mussel transcriptome.

Result	Heat stress 1	Heat stress 2	Control 1	Control 2	Total
Number of reads	39,489,838	39,230,928	42,040,158	42,406,428	163,167,352
Average read length (base pairs [bp])	100	100	100	100	
Number of reads after trimming	39,264,927	39,029,227	41,746,326	42,218,892	162,259,372
% kept after trimming	99.4	99.5	99.3	99.6	
Average read length after trimming (bp)	97.8	97.3	96.2	97.6	

15 cycles with annealing temperature 53°C. We analyzed and scored the PCR products following standard protocols and previous research (Reece et al. 2004). We used POPGENE (version 1.3.2; Yeh and Boyle 1999) to estimate number of alleles/locus, heterozygosity (observed and expected), and conformance to Hardy–Weinberg equilibrium (HWE).

Results

Illumina RNA-seq sequencing

Illumina sequencing generated ~163.2 million 100-bp reads, including ~78.7 million reads from the heat-stress group and ~84.4 million reads from the control group. After initial trimming to remove any adaptor sequences, low-quality bases, and short reads, a total of ~162.2 million reads (99.45% of total reads) encompassing ~15.8 billion bp were generated (Table 1). Raw reads were archived at NCBI Sequence Read Archive (SRA) under Accession SRP009061.

De novo assembly and redundancy filtration of different assemblers

Trans-ABySS.—Forty-seven assemblies (k-mer sizes 50–96) were generated in ABySS totaling 25 million contigs. Assemblies ranged from 45,000 contigs to 1.03 million contigs with N50 (statistical index of length N where 50% of all bases in the sequences are in a sequence of length $L > N$) sizes from 475 to 774 bp. The combined assembly using Trans-ABySS generated ~1.8 million contigs with average length of 619 bp and N50 size of 1191 bp (Table 2). The Trans-ABySS assembly contained 64.3% of contigs >200 bp and 17.2% of contigs >1000 bp. All contigs <200 bp were removed from further analysis because they are known to give poor downstream results (Gotz et al. 2008, Feldmeyer et al. 2011, Garg et al. 2011). Multiple-k algorithms have the potential to produce redundant assemblies (Robertson et al. 2010) that, if not removed, may cause errors in subsequent analyses. To alleviate this problem, we first relied on

TABLE 2. Summary of de novo assembly results of Illumina sequence data using various assemblers. N50 = statistical index of length N where 50% of all bases in the sequences are in a sequence of length $L > N$.

Result	Trans-ABySS	Velvet	CLC
Contigs (≥100 base pairs [bp])	1,806,746	501,764	1,311,097
Large contigs (≥1000 bp)	309,815	39,261	43,127
Maximum length (bp)	20,534	20,315	33,548
Average length (bp)	619	440	266
N50 (bp)	1191	1198	305
Contigs after length filtering (≥200 bp)	1,193,893	356,346	438,104
% contigs after length filtering	64.3%	71.0%	33.4%
Average contig length after length filtering (bp)	860.3	552.7	521.1
Contigs[a]	1,177,801	229,188	438,070
Average length (bp)[a]	856.4	612.0	521.2
Contigs[b]	778,234	193,016	437,283
Average length (bp)[b]	707.5	627.9	521.8
Reads mapped to final reference (%)	148,378,497 (91.4%)	120,921,555 (74.5%)	138,638,637 (85.4%)

[a] After CD-HIT-EST
[b] After CD-HIT-EST + CAP3

TABLE 3. Mussel contig annotation based on Basic Local Alignment Search Tool (BLAST) homology searches against various protein databases. Putative gene matches were at $E \leq 1e{-}5$. Hypothetical gene matches denote those BLAST hits with uninformative annotation. * indicates quality unigene hits with more stringent parameters, including score ≥ 100, $E \leq 1e{-}20$. nr = National Center for Biotechnology Information nonredundant database, UniProt = UniProtKB/SwissProt database, Invertebrate = RefSeq Invertebrates protein database, GO = gene ontogeny, bp = base pair.

Database	Putative gene matches	Annotated contigs ≥ 500 bp	Annotated contigs ≥ 1000 bp	Unigene matches	Hypothetical gene matches	Quality unigene matches*
UniProt	111,049	73,569	43,369	22,785	0	12,791
Invertebrate	213,755	146,899	88,517	32,537	2007	17,564
nr	248,059	180,224	118,812	46,043	2875	23,742
GO terms	98,419	65,681	38,276	26,328	N/A	N/A

Trans-ABySS's built-in redundancy elimination solutions, but found that additional steps (CD-HIT and CAP3) were required to identify a unique reference. These additional steps resulted in a reduced Trans-ABySS average contig size (707.5) and total number of contigs (778,234) (Table 2).

Velvet.—In total, 1.7 million contigs were produced in 7 assemblies by Velvet (7 k-mers between 50 and 99 bp), ranging from 3717 contigs to 0.57 million contigs with N50 sizes ranging from 432 to 839 bp. The combined assembly, using Velvet's AssemblyAssembler, generated ~0.5 million contigs with an average length of 440 bp and N50 of 1198 bp. Contig length distributions differed from the Trans-ABySS assembly. Seventy-one percent of contigs were >200 bp, but only 7.8% of contigs were >1000 bp. Approximately 0.3 million contigs were removed during the length and redundancy filtration steps, resulting in a final average contig size and contig number of 627.9 bp and 229,188, respectively (Table 2).

CLC Genomics Workbench.—A single k-mer (k = 24; automatically selected) procedure generated 1.3 million contigs with average length 266 bp and N50 size 305 bp (Table 2). Only 33.4% of contigs were >200 bp and 3.3% of contigs were >1000 bp. Because of the nonredundant single-k approach, only 821 contigs were subsequently removed during the CD-HIT/CAP3 procedure. Final average contig size and contig number were 521.8 bp and 437,283, respectively (Table 2).

Best assembly selection.—In a comparison of the assemblies resulting from the 3 approaches (Table 2), Trans-ABySS stood out for its larger percentage of large contigs, longest final contig size, and largest number of contigs after filtering. In addition, 91.4% of the initial reads mapped to the final Trans-ABySS reference assembly, compared to 74.5% in Velvet and 85.4% in CLC, illustrating the more comprehensive nature of the Trans-ABySS assembly. Therefore, we selected the Trans-ABySS assembly for subsequent gene discovery, differential expression analysis, and microsatellite-marker mining.

Gene identification and annotation

In total, 14.3% (111,049) of the Trans-ABySS contigs had a significant BLAST hit against the UniProt database and matched 22,785 unique protein accessions (Table 3). More contigs could be putatively annotated based on hits against the Invertebrate (27.5%) and nr (31.9%) databases, albeit with more uninformative hypothetical protein hits (Table 3). To evaluate the quality of the assembled genes further, the number of unique genes (unigenes) matching protein sequences in public databases were identified with the more stringent criteria of a BLAST score ≥100 and *E*-value ≤ 1*e*–20. These criteria allowed identification of 12,791, 17,564, and 23,742 genes from the UniProt, Invertebrate, and nr databases, respectively. Analyses of species with the highest number of quality matches (above criteria) with the assembled mussel transcriptome in the NCBI nr database were revealing on several counts (Fig. 2). Close to 14% of quality matches came against the Florida lancelet (*Branchiostoma floridae*), an ancient chordate, followed by the acorn worm (~8%), and zebrafish (~4%). The Pacific oyster (*Crassostrea gigas*) and the Mediterranean mussel (*Mytilus galloprovincialis*) were the only representative bivalves accounting for ≥1% of quality hits, with ~3.4% of hits cumulatively. That >30,000 proteins from lancelet are available in the nr database probably explains the abundance of hits to this species (cf. *C. gigas* and *M. galloprovincialis* with 1704 and 1817 entries, respectively). Contigs with unique nr database identities were submitted to the NCBI Transcriptome Shotgun Assembly (TSA) database and are available under accession numbers JR494687–JR540729. Corresponding putative annotations are available in Table S1 (available online from: http://dx.doi.org/10.1899/11-149.1.s1).

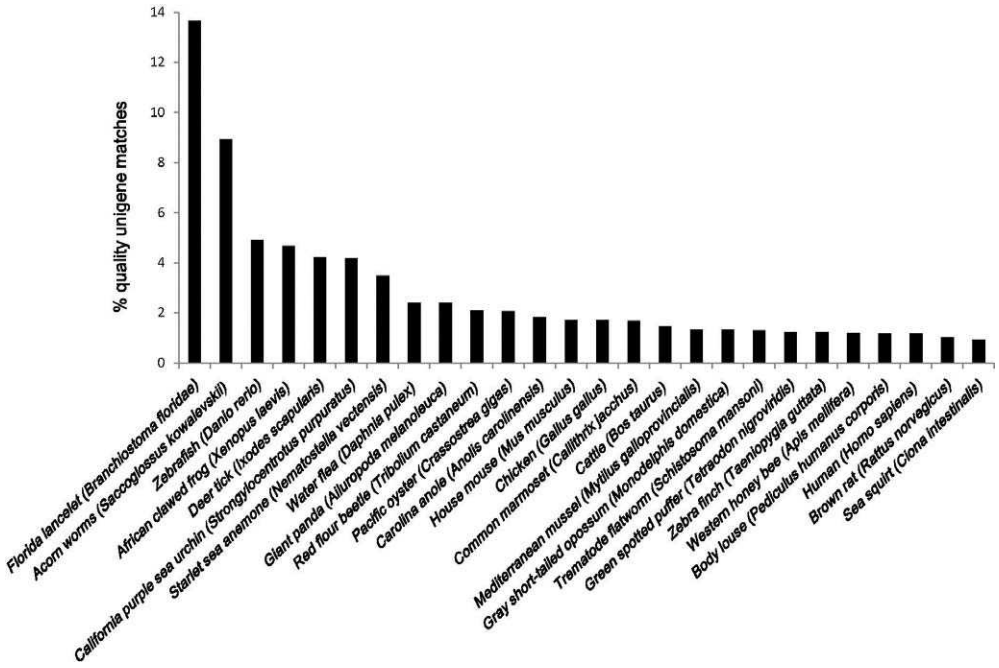

FIG. 2. Quality unigene match (National Center for Biotechnology Information nonredundant database) distribution by species. Species accounting for $\geq 1\%$ of total quality (score $\geq$ 100, $E \leq 1e{-}20$) unigene matches are listed by common names followed by scientific names.

In total, 244,427 GO terms including 67,845 (27.8%) cellular-process terms, 126,811 (51.9%) molecular-functions terms, and 49,711 (20.4%) biological-process terms were assigned to 26,328 unique gene matches using Blast2GO. Analysis of the level-2 GO term distribution showed that metabolic process (GO:0008152), binding (GO:0005488), and cell (GO:0005623) were the most common annotation terms in the 3 GO categories (Fig. S1; available online from: http://dx.doi.org/10.1899/11-149.1.s2). One or more GO terms could be assigned to 98,419 contigs by linking gene identity and GO terms to the broader BLAST results against nr and UniProt databases (Table 3).

Transcript-level gene expression analysis of heat stress

A total of 1934 of the 778,234 (0.2%) assembled contigs showed significant differential expression between heat and control groups (Table 4). Among the differentially expressed contigs, 442 showed fold changes $>|5|$. After linking the expression and BLAST annotation results, 604 unique genes were identified as responding to the elevated temperature regimen (Table S1). Functional enrichment analysis was carried out to identify GO categories overrepresented in the differentially expressed gene set in comparison with the overall contig (gene) population (Table S2; available online from: http://dx.doi.org/10.1899/11-149.1.s3). Functional categories involved in unfolded protein binding, response to stimulus, heat-shock protein binding, and endoplasmic reticulum (ER)-associated processes were all enriched among up-regulated genes indicating a classic stress response (Fig. S2; available online from: http://dx.doi.org/10.1899/11-149.1.s4). Categories involved in muscle development and use were enriched among down-regulated genes.

Based on the results of the enrichment and expression analysis and previous studies on heat-shock response (Brun et al. 2008, Liu and Chang 2008, Haynes and Ron 2010), the components of a relatively complete heat-shock response and ER unfolded protein response (UPR^{ER}) pathway responding to the treatment conditions, including caspases and ER chaperones, such as heat-shock proteins and calreticulin (Fig. 3), were identified.

Primers were designed for significantly expressed genes selected on the basis of the functional enrichment and pathway results (20 genes). Three heat-stressed mussels and 3 control mussels were used for the q-RT-PCR. Expression in muscle was analyzed separately to judge the applicability of the pooled tissue RNA-seq results to a more typical experiment on single-tissue gene expression. The fold change generated from the Relative Expression Software Tool (version 2009; Pfaffl et al. 2002) based on the Ct-values of the real-time reaction were compared with the results of the RNA-seq expression analysis. Fold changes of the 20 genes

TABLE 4. Statistics of ribonucleic acid sequencing (RNA-seq) analysis of differential gene expression in freshwater mussel after heat stress. Values indicate contigs/genes passing cutoff values of fold change ≥ 2 ($p < 0.05$) and read number/condition > 5.

Variable	Total	Up	Down	Average read number	Fold >5	Fold >10	Fold <−5	Fold <−10
Contigs	1934	1158	776	109.3	265	64	157	34
Unique genes	604	371	233	128.4	46	7	42	5

were significantly correlated with those generated with RNA-seq analysis ($r = 0.847$, $p < 0.001$; Fig. 4). Nineteen of the 20 genes showed the same direction (up or down) of differential expression as found by RNA-seq. In some cases, real-time results showed upregulation relative to gene expression in control samples but differed substantially from the RNA-seq results in magnitude of fold change (e.g., HSP70; Table S3; available online from: http://dx.doi.org/10.1899/11-149.1.s5), a result that may have reflected the effect of differing expression levels in the pooled tissues vs muscle alone. In general, the differentially expressed list generated from pooled-tissue RNA-seq contained good candidate genes for stress-related studies in individual tissues. All tested primers showed specific amplification of a single product, a result indicating the reliability and accuracy of the Trans-ABySS reference assembly.

Microsatellite markers identification and verification

After Trans-ABySS assembly, microsatellites or simple sequence repeats (SSRs) were mined from the transcriptome contigs with Msatfinder. From a total of 65,086 microsatellites, 56.2% (36,582) had sufficient flanking regions to allow design of primers (Table 5). These 36,582 microsatellites were distributed across 31,960 contigs. The microsatellite-bearing

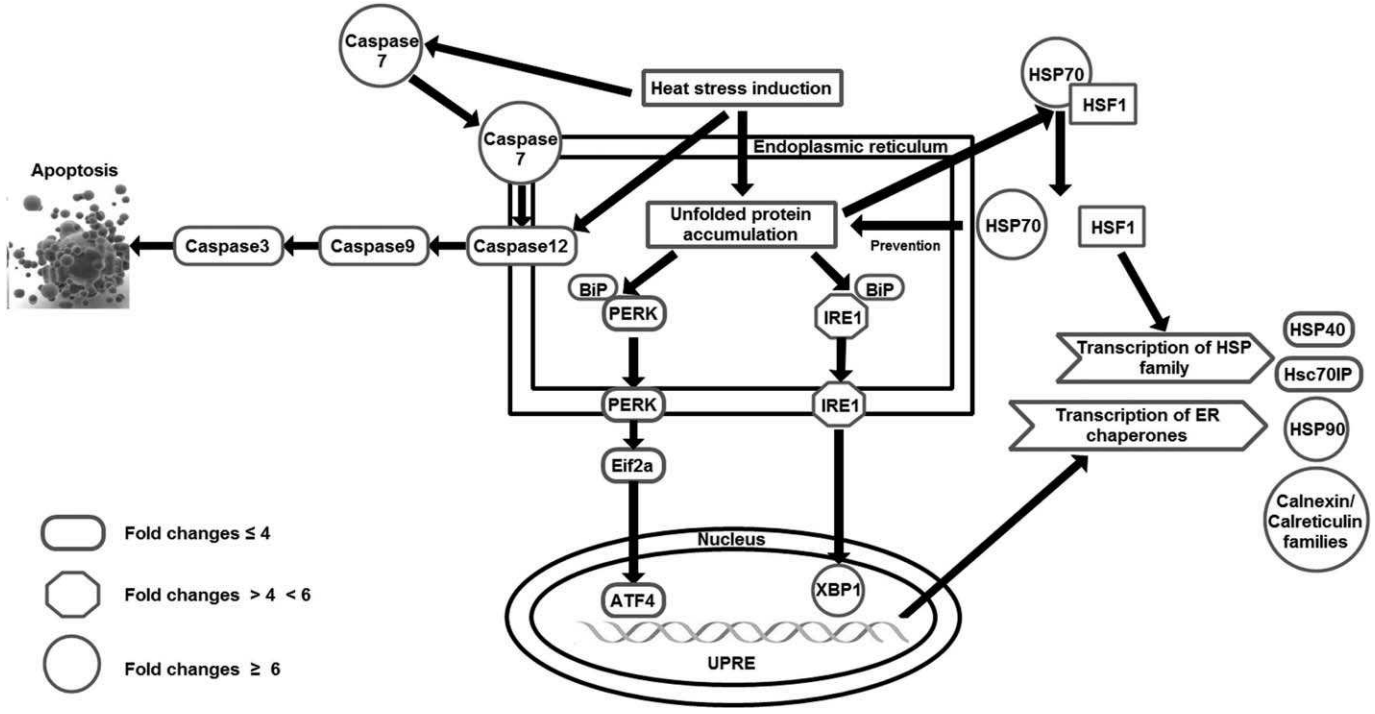

FIG. 3. Differential expression of the heat-shock response and endoplasmic reticulum (ER) unfolded protein response (UPR^{ER}) pathway components and putative functional signaling and relationships of differentially expressed genes in *Villosa lienosa* after elevated temperature challenge. Putative pathway is based on relevant literature (Schroder and Kaufman 2004, Haynes and Ron 2010) and indicates observed induction of mussel genes with shapes indicating fold change levels (key). ER = endoplasmic reticulum, HSP = heat-shock protein, HSF1 = heat shock transcription factor 1, Hsc70IP = heat shock cognate 70 kDa protein-interacting protein, BiP = glucose-regulated protein (GRP78), PERK = proteins proline-rich receptor-like protein kinase, Eif2a = eukaryotic translation initiation factor 2A, IRE1 = serine threonine-protein kinase endoribonuclease, XBP1 = x-binding protein 1, ATF4 = cyclic AMP-dependent transcription factor 4, UPRE = unfolded protein response elements.

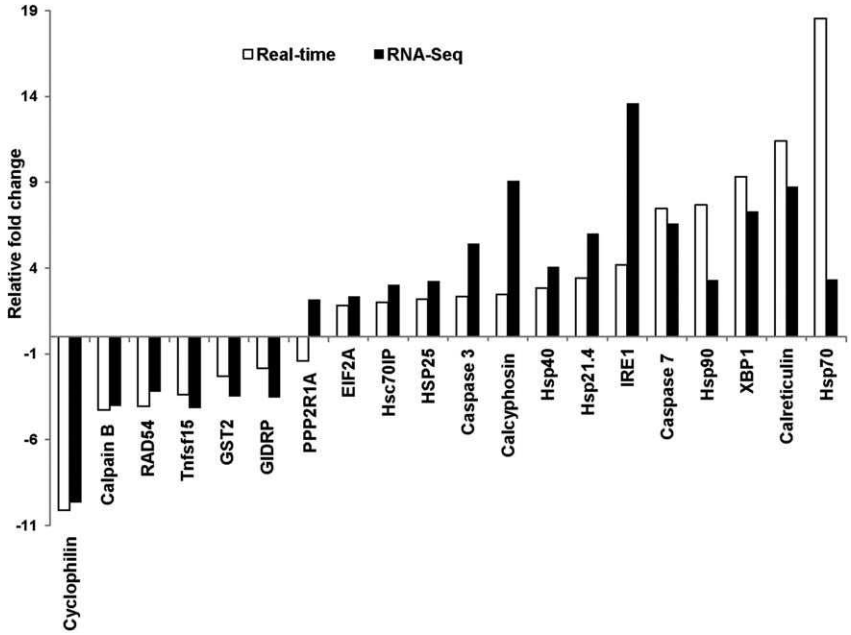

FIG. 4. Comparison of fold changes between ribonucleic acid sequencing (RNA-seq) expression analysis and real-time polymerase chain reaction (PCR) ($r = 0.847$, $p < 0.001$). RAD54 = deoxyribonucleic acid (DNA) repair and recombination protein, Tnfsf15 = tumor necrosis factor ligand superfamily member 15, GST2 = sigma class glutathione-s-transferase 2, GIDRP = growth inhibition and differentiation-related protein, PPP2R1A = medium tumor antigen-associated 61-kD protein, Eif2a = eukaryotic translation initiation factor 2A, Hsc70IP = heat shock cognate 70 kDa protein-interacting protein, HSP25 = heat shock protein 25, HSP40 = heat shock protein 40, IRE1 = serine threonine-protein kinase endoribonuclease, HSP90 = heat shock protein 90, XBP1 = x-binding protein 1.

contigs had 9,607 putative gene matches to the nr database from 2,607 unique genes and were usually found within 3′ untranslated regions (UTR) of captured genes. When these markers were categorized by repeat type (Table 5), 53.3% (34,719) had dinucleotide repeat patterns, 30.5% (19,834) had trinucleotide repeat patterns, and 16.2% (10,533) had a repeat motif of >3 bases.

To validate this large microsatellite set and to understand better its potential for future parentage studies, 47 of the putative markers were chosen for genotyping in a local population of 26 mixed-sex *V. lienosa* mussels. Forty-four of 47 (93.6%) of the markers could be amplified by PCR and 16 (34%) markers showed scoreable polymorphic patterns in the tested individuals after polyacrylamide gel electrophoresis (Table 5). Allele numbers ranged from 2 to 10, with an average of 5 alleles/marker. Primer information and other details of the 16 polymorphic markers can be found in Table S4 (available online from: http://dx.doi.org/10.1899/11-149.1.s6).

Discussion

Long time frames and the expense of conducting large-scale sequencing projects have constrained development of genome resources and tools to a few model species. Molecular research in model species has been critically important for advancing our basic understanding of genome structure, gene function, physiological development, and adaptive processes, but large gaps remain in application of these findings to ecological settings and species (Ekblom and Galindo 2011). The advent of next-generation sequencing techniques offers exciting opportunities to fill these gaps and to explore the complexities of population structure, community dynamics, and phenotypic variation on a molecular level.

The situation of freshwater unionid mussels, which are among the most endangered groups of organisms in the world (Lydeard et al. 2004), typifies the dearth of molecular information available for critical taxa. At the time of writing, <9000 nucleotide sequences for the family were publicly available. We focused on a relatively abundant member of *Villosa*, *V. lienosa*, but other members of the genus, including *Villosa fabalis*, *Villosa vanuxemensis umbrans*, *Villosa nebulosa*, and *Villosa choctawensis*, are considered imperiled and several are under consideration for listing as endangered under the federal Endangered Species Act (Center for Biological Diversity 2011). Before our

TABLE 5. Statistics of simple sequence repeats (SSRs) identified from the mussel transcriptome. nr = National Center for Biotechnology Information nonredundant database.

Variable	Number
SSR mining	
Total number of sequences examined	778,234
Total size of examined sequences (base pairs [bp])	550,600,555
Total number of identified SSRs	65,086
Total number of SSRs with primers	36,582
Contigs containing SSRs with primers	31,960
SSRs with primers associated with putative gene matches in nr	9,607
SSRs with primers associated with unique gene matches	2,607
Distribution of SSRs in different repeat types	
Dinucleotide	34,719
Trinucleotide	19,834
Tetranucleotide	10,285
Pentanucleotide	162
Hexanucleotide	86
Summary of SSRs validation	
Individuals	26
Total markers	47
Amplified markers	44
Polymorphic markers	16

study, only 104 nucleotide sequences were available to assist conservation efforts for *Villosa* species, including 88 mitochondrial sequences (many redundant) and 16 anonymous microsatellite loci. Only 8 mitochondrial sequences were available from *V. lienosa*. Rather than use universal or degenerate primers or assays with low reproducibility (e.g., amplified fragment length polymorphism [AFLP] or random amplification of polymorphic DNA [RAPD]) for downstream studies, we chose to use RNA-seq technology to rapidly expand molecular resources for *V. lienosa*.

Studies of nonmodel organisms pose challenges to genomic research. Funding and sample size are often limited, and sequence assembly in a novel species can be complicated by lack of reference genomes, low sequence coverage, and high rates of sequence polymorphism (Pop and Salzberg 2008, Parchman et al. 2010). We attempted to address these challenges while meeting our immediate need for: 1) microsatellite markers for parentage analysis of larval mussels (glochidia) and 2) gene sequences and preliminary expression profiles after heat stress. Generation of barcoded libraries from individual samples is a large component of next-generation sequencing costs. We minimized these costs by barcoding only 2 individuals from each treatment condition (baseline and elevated temperature) and combined all 4 barcoded tags into a single lane of Illumina Hi-Seq sequencing. In previous studies, we pooled multiple individuals within a single barcoded library to maximize single-nucleotide polymorphism (SNP) capture (Liu et al. 2011).

Reported rates of polymorphism are high in marine mussels and existing unionid references were lacking. Therefore, we barcoded samples representing single individuals to maximize our likelihood of a high-quality assembly. We pooled the representative tissues (muscle, mantle, and gill) from each individual to maximize the diversity of captured gene transcripts. Our approach yielded 15.8 billion bp of quality sequence in 162 million reads. Based on highly conservative criteria, we captured a minimum of 12,791 unique genes in *V. lienosa*. We captured >600 differentially expressed genes by comparing read counts between the 2 temperature treatments. Last, we sequenced 36,582 microsatellite loci with sufficient flanking sequence for primer design.

RNA-seq experimental protocols are still fluid, and assembly and analysis algorithms are evolving rapidly (Marguerat and Bahler 2010, Robertson et al. 2010). Therefore, we attempted to validate and confirm our results on several levels. Construction of a high-quality, nonredundant reference transcriptome is particularly critical for generation of accurate gene sequences and digital expression profiling. Therefore, we compared the ability of several popular software packages to assemble the 100-bp Illumina reads into a comprehensive contig set. Our analyses showed that the multiple-k approach is critical to transcriptome assembly. The larger number of k-mers (47) used by Trans-ABySS led to a superior assembly relative to Velvet (7 k-mers) and CLC (single k). However, additional steps were necessary to remove the redundancies generated by the multiple-k software (Fig. 1).

Mussel population declines have been linked with environmental conditions that exceed the organisms' normal tolerances (e.g., elevated temperatures associated with drought; Golladay et al. 2004, Pandolfo et al. 2010). Expression analysis of transcriptomic responses to heat stress can identify biomarkers useful for assessing mussel health in baseline and perturbed conditions. For example, investigators have used microarray studies of marine mollusks (Pacific oysters and blue mussels) to profile the effects of heat shock (Lang et al. 2009, Lockwood et al. 2010). Our results, although from a minimal sample set and a single time point, correspond well with results of these studies. We captured differentially expressed molecular chaperones, antioxidants, immune factors, cytoskeletal

components, and apoptosis mediators after exposing mussels to elevated temperature (Table S1). Unlike microarrays, RNA-seq does not restrict expression profiles to previously captured genes and is not subject to errors caused by cross-hybridization.

Most strikingly, our pathway analysis revealed that our differentially expressed gene set included the components of a nearly complete canonical UPR^{ER}. Stress induced by heat leads to the accumulation of unfolded proteins (Fig. 3). Preferential binding of heat shock protein (HSP) 70 with these unfolded proteins releases transcription factor heat shock factor 1 (HSF1) to further activate HSP molecular chaperones, such as HSP40, HSP70, heat shock cognate (Hsc) 70IP, and HSP90. The UPR^{ER} is regulated through transmembrane proteins proline-rich receptor-like protein kinase (PERK) and serine threonine-protein kinase endoribonuclease (IRE1), following disassociation from ER chaperone, 78 kDa glucose-regulated protein (GRP78/BiP). These factors ultimately lead to the transcription of ER chaperones including calreticulin and calnexin. The UPR also can regulate apoptotic pathways via IRE1, triggering a cascade of caspase activation leading to cell death (Schroder and Kaufman 2005, Haynes and Ron 2010). We validated the expression profiles of many of the described pathway components in individual tissue samples from additional mussels (Fig. 4), confirming their utility as targets for future studies of stress responses in Unionid species.

A key experimental goal was to generate a large microsatellite resource for use in studying *Villosa* species, particularly *V. lienosa*. We sought to identify enough polymorphic loci to allow assignment of parentage to glochidia (Christian et al. 2007). Microsatellites remain the marker of choice for such studies because the possibility of numerous alleles per loci reduces the number of markers needed relative to the number of SNP markers (Tokarska et al. 2009, Hauser et al. 2011). We succeeded in identifying many potentially polymorphic microsatellites with RNA-seq. A substantial subset of these markers is associated with known gene sequences. Thus, they will be a useful resource in future population genomic studies to identify candidate loci associated with adaptation and divergent selection (Vasemägi et al. 2005, Stinchcombe and Hoekstra 2008). Several of the validated microsatellites deviated significantly from HWE. Deviations may reflect the sample population used, an excess of null alleles, or trait-linked markers under selection. Further work will be needed to identify the source of deviations from HWE in these loci before they are used in population genetic surveys or parentage analysis.

We have described the use of a simple, economical approach for rapid generation of large-scale molecular resources for the freshwater mussel, *V. lienosa*. We captured a substantial portion (>46,000 unique gene matches) of the mussel transcriptome in high-quality contigs encoding tens of thousands of genes and microsatellites. In addition, use of mussels exposed to elevated temperatures allowed enrichment for stress-related genes and identification of differentially expressed pathways relevant to heat stress. Validation using q-RT-PCR confirmed the accuracy of expression measures based on read counts. The methods used will be applicable to studies in other nonmodel species. Ongoing research with the tools developed here will help to advance our understanding of mussel physiology and life histories.

Acknowledgements

We thank Huseyin Kucuktas, Ludmilla Kaltenboeck, Fanyue Sun, Luyang Sun, Jianguo Lu, and Shikai Liu for their technical assistance in various experiments and for critical reading of the manuscript. Ruijia Wang was supported by a scholarship from the China Scholarship Council (CSC) for studying abroad. Use of trade names throughout the manuscript does not constitute endorsement by the US government.

Literature Cited

ALEXA, A., J. RAHNENFUHRER, AND T. LENGAUER. 2006. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. Bioinformatics 22:1600–1607.

ALLISON, D. B., X. CUI, G. P. PAGE, AND M. SABRIPOUR. 2006. Microarray data analysis: from disarray to consolidation and consensus. Nature Reviews Genetics 7:55–65.

BAGGERLY, K. A., L. DENG, J. S. MORRIS, AND C. M. ALDAZ. 2003. Differential expression in SAGE: accounting for normal between-library variation. Bioinformatics 19:1477–1483.

BENJAMINI, Y., AND Y. HOCHBERG. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B: Methodological 57:289–300.

BERG, D. J., A. D. CHRISTIAN, AND S. I. GUTTMAN. 2007. Population genetic structure of three freshwater mussel (Unionidae) species within a small stream system: significant variation at local spatial scales. Freshwater Biology 52:1427–1439.

BIROL, I., S. D. JACKMAN, C. B. NIELSEN, J. Q. QIAN, R. VARHOL, G. STAZYK, R. D. MORIN, Y. ZHAO, M. HIRST, J. E. SCHEIN, D. E. HORSMAN, J. M. CONNORS, R. D. GASCOYNE, M. A. MARRA, AND S. J. JONES. 2009. *De novo* transcriptome assembly with ABySS. Bioinformatics 25:2872–2877.

BOYER, S. L., A. A. HOWE, N. W. JUERGENS, AND M. C. HOVE. 2011. A DNA-barcoding approach to identifying juvenile

freshwater mussels (Bivalvia:Unionidae) recovered from naturally infested fishes. Journal of the North American Benthological Society 30:182–194.

BROCKMAN, W., P. ALVAREZ, S. YOUNG, M. GARBER, G. GIANNOUKOS, W. L. LEE, C. RUSS, E. S. LANDER, C. NUSBAUM, AND D. B. JAFFE. 2008. Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Research 18:763–770.

BRUN, N. T., V. M. BRICELJ, T. H. MACRAE, AND N. W. ROSS. 2008. Heat shock protein responses in thermally stressed bay scallops, *Argopecten irradians*, and sea scallops, *Placopecten magellanicus*. Journal of Experimental Marine Biology and Ecology 358:151–162.

CAMPBELL, D. C., P. D. JOHNSON, J. D. WILLIAMS, A. K. RINDSBERG, J. M. SERB, K. K. SMALL, AND C. LYDEARD. 2008. Identification of 'extinct' freshwater mussel species using DNA barcoding. Molecular Ecology Resources 8:711–724.

CENTER FOR BIOLOGICAL DIVERSITY. 2011. 374 southeast species move toward endangered species act protection Center for Biological Diversity, Tucson, Arizona. (Available from: http://www.biologicaldiversity.org/news/press_releases/2011/southeast-freshwater-species-09-26-2011.html)

CHATY, S., F. RODIUS, AND P. VASSEUR. 2004. A comparative study of the expression of *CYP1A* and *CYP4* genes in aquatic invertebrate (freshwater mussel, *Unio tumidus*) and vertebrate (rainbow trout, *Oncorhynchus mykiss*). Aquatic Toxicology 69:81–94.

CHRISTIAN, A. D., E. M. MONROE, A. M. ASHER, J. M. LOUTSCH, AND D. J. BERG. 2007. Methods of DNA extraction and PCR amplification for individual freshwater mussel (Bivalvia: Unionidae) glochidia, with the first report of multiple paternity in these organisms. Molecular Ecology Notes 7:570–573.

DÍAZ-FERGUSON, E., A. WILLIAMS, AND G. MOYER. 2011. Isolation and characterization of microsatellite loci in the federally endangered fat threeridge mussel (*Amblema neislerii*). Conservation Genetics Resources 3:757–759.

EACKLES, M. S., AND T. L. KING. 2002. Isolation and characterization of microsatellite loci in *Lampsilis abrupta* (Bivalvia: Unionidae) and cross-species amplification within the genus. Molecular Ecology Notes 2:559–562.

EKBLOM, R., AND J. GALINDO. 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107:1–15.

FALCON, S., AND R. GENTLEMAN. 2007. Using GOstats to test gene lists for GO term association. Bioinformatics 23: 257–258.

FELDMEYER, B., C. W. WHEAT, N. KREZDORN, B. ROTTER, AND M. PFENNINGER. 2011. Short read Illumina data for the *de novo* assembly of a non-model snail species transcriptome (*Radix balthica*, Basommatophora, Pulmonata), and a comparison of assembler performance. BMC Genomics 12:317.

GALBRAITH, H. S., K. M. WOZNEY, C. M. SMITH, D. T. ZANATTA, AND C. WILSON. 2010. Isolation and characterization of microsatellite loci in the freshwater mussel *Lasmigona costata* (Bivalvia: Unionoida). Conservation Genetics Resources 3:9–11.

GANGLOFF, M. M. 2003. The status, physical habitat associations, and parasites of freshwater mussels in the upper Alabama river drainage, Alabama. PhD Dissertation, Auburn University, Auburn, Alabama.

GARG, R., R. K. PATEL, S. JHANWAR, P. PRIYA, A. BHATTACHARJEE, G. YADAV, S. BHATIA, D. CHATTOPADHYAY, A. K. TYAGI, AND M. JAIN. 2011. Gene discovery and tissue-specific transcriptome analysis in chickpea with massively parallel pyrosequencing and web resource development. Plant Physiology 156:1661–1678.

GEIST, J., J. GEISMAR, AND R. KUEHN. 2009. Isolation and characterization of the first microsatellite markers for the endangered swan mussel *Anodonta cygnea* L. (Bivalvia: Unionoidea). Conservation Genetics 11: 1103–1106.

GIDIERE, P. S., AND A. B. BORDEN. 2007. *Epioblasma penita*: the southern combshell—go forth and multiply. Environment and Natural Resources 22:18–22.

GOLLADAY, S. W., P. GAGNON, M. KEARNS, J. M. BATTLE, AND D. W. HICKS. 2004. Response of freshwater mussel assemblages (Bivalvia:Unionidae) to a record drought in the Gulf Coastal Plain of southwestern Georgia. Journal of the North American Benthological Society 23:494–506.

GOTZ, S., J. M. GARCÍA-GÓMEZ, J. TEROL, T. D. WILLIAMS, S. H. NAGARAJ, M. J. NUEDA, M. ROBLES, M. TALON, J. DOPAZO, AND A. CONESA. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Research 36:3420–3435.

GROSSMANN, S., S. BAUER, P. N. ROBINSON, AND M. VINGRON. 2007. Improved detection of overrepresentation of gene-ontology annotations with parent–child analysis. Bioinformatics 23:3024–3031.

HAAG, W. R., J. WARREN, AND L. MELVIN. 1998. Role of ecological factors and reproductive strategies in structuring freshwater mussel communities. Canadian Journal of Fisheries and Aquatic Sciences 55:297–306.

HAUSER, L., M. BAIRD, R. HILBORN, L. W. SEEB, AND J. E. SEEB. 2011. An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (*Oncorhynchus nerka*) population. Molecular Ecology Resources 11(Supplement 1):150–161.

HAYNES, C. M., AND D. RON. 2010. The mitochondrial UPR–protecting organelle protein homeostasis. Journal of Cell Science 123:3849–3855.

HOFTYZER, E., J. D. ACKERMAN, T. J. MORRIS, AND G. L. MACKIE. 2008. Genetic and environmental implications of reintroducing laboratory-raised unionid mussels to the wild. Canadian Journal of Fisheries and Aquatic Sciences 65:1217–1229.

HUANG, X. 1999. CAP3: a DNA sequence assembly program. Genome Research 9:868–877.

INOUE, N., R. ISHIBASHI, T. ISHIKAWA, T. ATSUMI, H. AOKI, AND A. KOMARU. 2010. Gene expression patterns and pearl formation in the Japanese pearl oyster (*Pinctada fucata*): a comparison of gene expression patterns between the pearl sac and mantle tissues. Aquaculture 308(Supplement 1):S68–S74.

JONES, J. W., M. CULVER, V. DAVID, J. STRUTHERS, N. A. JOHNSON, R. J. NEVES, S. J. O'BRIEN, AND E. M. HALLERMAN. 2004. Development and characterization of microsatellite loci in the endangered oyster mussel *Epioblasma capsaeformis* (Bivalvia: Unionidae). Molecular Ecology Notes 4: 649–652.

JONES, J. W., E. M. HALLERMAN, AND R. J. NEVES. 2006. Genetic management guidelines for captive propagation of freshwater mussels (Unionoidea). Journal of Shellfish Research 25:527–535.

LANG, R. P., C. J. BAYNE, M. D. CAMARA, C. CUNNINGHAM, M. J. JENNY, AND C. J. LANGDON. 2009. Transcriptome profiling of selectively bred pacific oyster *Crassostrea gigas* families that differ in tolerance of heat shock. Marine Biotechnology 11:650–668.

LI, W., AND A. GODZIK. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.

LIU, S., Z. ZHOU, J. LU, F. SUN, S. WANG, H. LIU, Y. JIANG, H. KUCUKTAS, L. KALTENBOECK, E. PEATMAN, AND Z. LIU. 2011. Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array. BMC Genomics 12:53.

LIU, Y., AND A. CHANG. 2008. Heat shock response relieves ER stress. EMBO Journal 27:1049–1059.

LOCKWOOD, B. L., J. G. SANDERS, AND G. N. SOMERO. 2010. Transcriptomic responses to heat stress in invasive and native blue mussels (genus: *Mytilus*): molecular correlates of invasive success. Journal of Experimental Biology 213(20):3548–3558.

LYDEARD, C., R. H. COWIE, W. F. PONDER, A. E. BOGAN, P. BOUCHET, S. A. CLARK, K. S. CUMMINGS, T. J. FREST, O. GARGOMINY, D. G. HERBERT, R. HERSHLER, K. E. PEREZ, B. ROTH, M. SEDDON, E. E. STRONG, AND F. G. THOMPSON. 2004. The global decline of nonmarine mollusks. BioScience 54:321–330.

MARGUERAT, S., AND J. BAHLER. 2010. RNA-seq: from technology to biology. Cellular and Molecular Life Sciences 67: 569–579.

MARTIN, J. A., AND Z. WANG. 2011. Next-generation transcriptome assembly. Nature Reviews Genetics 12: 671–682.

NEVES, R. J., A. E. BOGAN, J. D. WILLIAMS, S. A. AHLSTEDT, AND P. W. HARTFIELD. 1997. Status of aquatic mollusks in the southeastern United States: a downward spiral of diversity. Pages 43–85 *in* G. W. Benz and D. E. Collins (editors). Aquatic fauna in peril: the southeastern perspective. Tennessee Aquarium, Chattanooga, Tennessee.

PANDOLFO, T. J., W. G. COPE, C. ARELLANO, R. B. BRINGOLF, M. C. BARNHART, AND E. HAMMER. 2010. Upper thermal tolerances of early life stages of freshwater mussels. Journal of the North American Benthological Society 29: 959–969.

PARCHMAN, T. L., K. S. GEIST, J. A. GRAHNEN, C. W. BENKMAN, AND C. A. BUERKLE. 2010. Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics 11:180.

PFAFFL, M. W., G. W. HORGAN, AND L. DEMPFLE. 2002. Relative expression software tool (REST©) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucleic Acids Research 30:e36.

POP, M., AND S. L. SALZBERG. 2008. Bioinformatics challenges of new sequencing technology. Trends in Genetics 24: 142–149.

REECE, K. S., W. L. RIBEIRO, P. M. GAFFNEY, R. B. CARNEGIE, AND S. K. ALLEN. 2004. Microsatellite marker development and analysis in the eastern oyster (*Crassostrea virginica*): confirmation of null alleles and non-mendelian segregation ratios. Journal of Heredity 95:346–352.

ROBERTSON, G., J. SCHEIN, R. CHIU, R. CORBETT, M. FIELD, S. D. JACKMAN, K. MUNGALL, S. LEE, H. M. OKADA, J. Q. QIAN, M. GRIFFITH, A. RAYMOND, N. THIESSEN, T. CEZARD, Y. S. BUTTERFIELD, R. NEWSOME, S. K. CHAN, R. SHE, R. VARHOL, B. KAMOH, A. L. PRABHU, A. TAM, Y. ZHAO, R. A. MOORE, M. HIRST, M. A. MARRA, S. J. JONES, P. A. HOODLESS, AND I. BIROL. 2010. *De novo* assembly and analysis of RNA-seq data. Nature Methods 7:909–912.

ROE, K. J., P. D. HARTFIELD, AND C. LYDEARD. 2001. Phylogeographic analysis of the threatened and endangered superconglutinate-producing mussels of the genus *Lampsilis* (Bivalvia: Unionidae). Molecular Ecology 10: 2225–2234.

SCHRODER, M., AND R. J. KAUFMAN. 2005. The mammalian unfolded protein response. Annual Review of Biochemistry 74:739–789.

SIMPSON, J. T., K. WONG, S. D. JACKMAN, J. E. SCHEIN, S. J. JONES, AND I. BIROL. 2009. ABySS: a parallel assembler for short read sequence data. Genome Research 19:1117–1123.

STINCHCOMBE, J. R., AND H. E. HOEKSTRA. 2008. Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity 100:158–170.

STRAYER, D. L., AND D. DUDGEON. 2010. Freshwater biodiversity conservation: recent progress and future challenges. Journal of the North American Benthological Society 29: 344–358.

THURSTON, M. I., AND D. FIELD. 2005. Msatfinder: detection and characterisation of microsatellites. CEH Oxford, Oxford, UK. (available from: http://www.genomics.ceh.ac.uk/msatfinder/)

TOKARSKA, M., T. MARSHALL, R. KOWALCZYK, J. M. WÓJCIK, C. PERTOLDI, T. N. KRISTENSEN, V. LOESCHCKE, V. R. GREGERSEN, AND C. BENDIXEN. 2009. Effectiveness of microsatellite and SNP markers for parentage and identity analysis in species with low genetic diversity: the case of European bison. Heredity 103:326–332.

VASEMÄGI, A., J. NILSSON, AND C. R. PRIMMER. 2005. Expressed sequence tag-linked microsatellites as a source of gene-associated polymorphisms for detecting signatures of divergent selection in Atlantic salmon (*Salmo salar L.*). Molecular Biology and Evolution 22:1067–1076.

WILLIAMS, A., AND G. MOYER. 2011. Isolation and characterization of 21 microsatellite loci for the federally threatened yellowfin madtom (*Noturus flavipinnis*) with cross species amplification in *N. baileyi.* Conservation Genetics Resources 3:1–3.

WILLIAMS, J. D., M. L. WARREN, K. S. CUMMINGS, J. L. HARRIS, AND R. J. NEVES. 1993. Conservation status of freshwater mussels of the United States and Canada. Fisheries 18(9):6–22.

YEH, F., AND T. BOYLE. 1999. POPGENE version 1.3.2: Microsoft window-based freeware for population genetic analysis. (Available from: http://www.ualberta.ca/~fyeh/popgene_download.html)

ZERBINO, D. R., AND E. BIRNEY. 2008. Velvet: algorithms for *de novo* short read assembly using *de bruijn graphs*. Genome Research 18:821–829.

Received: 7 November 2011
Accepted: 5 March 2012