Comparative Biochemistry and Physiology, Part D 7 (2012) 94–99

Contents lists available at SciVerse ScienceDirect

Comparative Biochemistry and Physiology, Part D

journal homepage: www.elsevier.com/locate/cbpd

Characterizing short read sequencing for gene discovery and RNA-Seq analysis in *Crassostrea gigas*

Mackenzie R. Gavery, Steven B. Roberts *

School of Aquatic and Fishery Sciences, University of Washington, 1122 NE Boat Street, Seattle, WA 98105, USA

ARTICLE INFO

Article history:
Received 12 September 2011
Received in revised form 16 December 2011
Accepted 20 December 2011
Available online 29 December 2011

Keywords:
Next-generation sequencing
Environment
Gene expression
Pacific oyster
Crassostrea gigas

ABSTRACT

Advances in DNA sequencing technology have provided opportunities to produce new transcriptomic resources for species that lack completely sequenced genomes. However, there are limited examples that rely solely on ultra-short read sequencing technologies (e.g. Solexa, SOLiD) for transcript discovery and gene expression analysis (i.e. RNA-Seq). Here we use SOLiD sequencing to examine gene expression patterns in Pacific oyster (*Crassostrea gigas*) populations exposed to varying degrees of anthropogenic impact. Novel transcripts were identified and RNA-Seq analysis revealed several hundred differentially expressed genes. Gene enrichment analysis determined that in addition to biological processes predicted to be associated with anthropogenic influences (e.g. immune response), other processes play important roles including cell recognition and cell adhesion. To evaluate the effectiveness of restricting characterization solely to short read sequences, mapping and RNA-Seq analysis were also performed using publicly available transcriptome sequence data as a scaffold. This study demonstrates that ultra-short read sequencing technologies can effectively generate novel transcriptome information, identify differentially expressed genes, and will be important for examining environmental physiology of non-model organisms.

1. Introduction

High-throughput DNA sequencing technologies are providing new opportunities to generate genomic resources for non-model organisms. A widely used approach is transcriptome sequencing, which has the benefit of providing increased coverage as a result of the reduced representation of the genome. A primary platform being used to generate transcriptomic resources in non-model species is the Roche 454 GS-FLX (454) followed by de novo assembly of sequence reads. This approach has been used to characterize transcriptomes of diverse taxa including plants (e.g. Novaes et al., 2008), insects (e.g. Vera et al., 2008), corals (e.g. Meyer et al., 2009), molluscs (e.g. Craft et al., 2010) and fish (e.g. Fraser et al., 2011). One benefit of using the 454 platform is that reads are longer compared to other common high-throughput sequencing systems, such as the Illumina Genome Analyzer IIx (Solexa) and Applied Biosystems SOLiD (SOLiD). Compared to the approximately 350 bp read length from the 454 platform, Solexa and SOLiD provide 'ultra-short reads' that are commonly less than 75 bp. The benefits of the ultra-short read platforms include increased number of reads and decreased cost. Sequencing on these platforms can be up to 30 times less expensive compared to 454 sequencing (Shendure and Ji, 2008). Recently, researchers have begun to examine the applicability of using Solexa and SOLiD for generating de novo transcriptomes in non-model species. For example, a transcriptome was generated for the snail (*Radix balthica*) using Solexa (Feldmeyer et al., 2011). A study in sockeye salmon (*Oncorhynchus nerka*) used SOLiD to compare results of de novo assembly versus mapping to public expressed sequence tag (EST) databases (Everett et al., 2011). Everett et al. (2011) determined that assemblies using public EST databases had a higher percentage of mapped reads and higher coverage than de novo assemblies. These studies demonstrate that current sequence assembler performance is sufficient for producing accurate and functionally informative transcriptomes generated from ultra-short read platforms.

In addition to assembling transcriptomes, high-throughput sequencing can also be used to directly examine gene expression levels, a method referred to as RNA-Seq. In RNA-Seq, high throughput sequencing reads generated from cDNA libraries are aligned to a common reference sequence or scaffold (e.g. whole genome) to produce a transcriptome map that includes transcript abundance for each gene. RNA-Seq provides similar information as hybridization based microarray analysis, however, RNA-Seq has an increased dynamic range compared to hybridization-based methods (Wang et al., 2009). Furthermore, RNA-Seq is not limited to analysis of known sequences like qPCR and microarray technology, which makes RNA-Seq especially appropriate for non-model species.

The RNA-Seq approach has been primarily used in organisms with sequenced genomes, but very recently RNA-Seq has been applied in non-model organisms. For example, RNA-Seq was used to investigate

* Corresponding author. Tel.: +1 206 685 3742; fax: +1 206 685 7471.
E-mail addresses: mgavery@uw.edu (M.R. Gavery), sr320@uw.edu (S.B. Roberts).

1744-117X/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.cbd.2011.12.003

the basis of phenotypic variation between lake trout (*Salvelinus namaycush*) ecotypes using the 454 platform (Goetz et al., 2010). RNA-Seq was also used to identify genes expressed in guppies (*Poecilia reticulata*) in response to predator cues using Solexa sequencing (Fraser et al., 2011). SOLiD transcriptome sequence reads have been used to investigate genes involved in response to temperature and settlement cues in coral larvae (Acropora millepora) (Meyer et al., 2011). In the latter two studies, Solexa or SOLiD short reads were mapped to a scaffold consisting of contigs generated from other sources (i.e. 454, ESTs). These studies conclude that this approach is effective in generating accurate and informative gene expression results. RNA-Seq analysis using one set of ultra-short read data as both the scaffold and individual reads for expression analysis would be the most cost efficient, especially for those organisms where genomic resources are limited. To date, a thorough evaluation of the effectiveness of this approach has not been performed.

The primary goal of this study was to evaluate the effectiveness of utilizing the SOLiD platform to both characterize the transcriptome and analyze gene expression patterns in the Pacific oyster (*Crassostrea gigas*). As part of this study, gene expression patterns between oyster populations exposed to varying degrees of anthropogenic impact were compared. RNA-Seq was performed using only the ultra-short read consensus sequences generated from de novo assembly as a scaffold. In order to evaluate the effectiveness of using solely ultra-short read data, RNA-Seq was also performed using publicly available transcriptome data as a scaffold and the results were compared. This work not only evaluates the use of limited ultra-short read sequence data for characterizing transcriptomes in non-model organisms, but also offers insight into the physiological responses of aquatic invertebrates in natural environments.

2. Materials and methods

2.1. Site selection

Oysters were collected from two locations in Puget Sound, Washington, USA. The sites were selected based on a difference in perceived degree of anthropogenic impact. The mouth of Big Beef Creek (BBC) in Hood Canal is a low impact site, and Drayton Harbor (DH), located in North Puget Sound, is an elevated impact site. The level of impact refers to water quality as determined by the Washington State Department of Ecology and Puget Sound Assessment and Monitoring Program (Newton et al., 2002). BBC has a relatively low population density compared to DH and routine monitoring by Washington State Department of Health shows low bacterial loads. DH is ranked as the number one shellfish growing area impacted by fecal coliform pollution (WSDOH, 2006). Additionally, the density of commercial dairies and animal keeping areas in the region surrounding DH is significantly higher than BBC (WSDOH, 2006), and a municipal wastewater treatment plant discharges in proximity to DH.

2.2. Sampling and library construction

Oysters were collected from both sites in April of 2009. At each site, gill tissue was immediately sampled from 16 oysters using sterile procedures and stored in RNAlater (Ambion). RNA was isolated from individual gill tissue samples (~50 mg) using Tri-Reagent (Molecular Research Center). To eliminate possible DNA carryover, total RNA was DNase treated using the Turbo DNA-free Kit (Ambion) according to the manufacturer's “rigorous” protocol. RNA from all individuals at a site (n = 16) was pooled in equal quantities (650 ng) to provide template for SOLiD libraries. Pooled samples were enriched for mRNA using the Ribominus Eukaryote Kit for RNA-Seq (Invitrogen) and MicroPolyA Purist Kit (Ambion). Libraries were prepared using the SOLiD Whole Transcriptome Analysis Kit (Applied Biosystems) and sequencing was performed using the SOLiD3 System (Applied Biosystems).

2.3. Sequence analysis

All sequence analysis was performed with CLC Genomics Workbench version 4.0 (CLC Bio). Initially, sequences were trimmed based on quality scores of 0.05 (Phred, Ewing and Green, 1998; Ewing et al., 1998) and the number of ambiguous nucleotides (>2 on ends). Sequences smaller than 20 bp were also removed. De novo assembly was carried out using the following parameters: limit = 8, mismatch cost = 2 and a minimum contig size of 200 bp. For comparison purposes, quality trimmed reads were also mapped to the 82,312 contigs in GigasDatabase (version 8) (Fleury et al., 2009). Parameters used for this reference mapping included: limit = 8 and mismatch cost = 2. Sequences and corresponding annotations from GigasDatabase were downloaded from the *C. gigas* Public Sigenae Contig Browser (http://public-contigbrowser.sigenae.org:9090/Crassostrea_gigas). Reference mapping, using the same parameters, was used to distinguish mitochondrial transcripts using the *C. gigas* mitochondrian genome (GenBank: AF177226).

Consensus sequences from the de novo assembly were compared to the UniProtKB/Swiss-Prot database (http://uniprot.org) in order to determine putative descriptions. Comparisons were made using the BLAST algorithm (Altschul et al., 1990). A cutoff E-value of 1E-05 was used for annotations. Associated GO terms (Gene Ontology database: http://www.geneontology.org) were used to categorize genes into parent categories and were assigned a functional group based on the MGI GO Slim database (URL: http://www.informatics.jax.org). The MGI GO Slim terms for ‘other biological processes’ and ‘other metabolic processes’ are not included in this analysis.

For RNA-Seq analysis, expression values were measured as RPKM (reads per kilobase of exon model per million mapped reads) (Mortazavi et al., 2008) with an unspecific match limit of 10 and maximum number of 2 mismatches. Statistical comparison of RPKM values between the BBC and DH libraries was carried out using Baggerly's test (Baggerly et al., 2003), and multiple comparison correction was performed using a false discovery rate. Genes were considered differentially expressed in a given library when 1) the p-value was less than or equal to 0.05 and 2) a greater-than-or-equal-to two-fold change in expression across libraries was observed. Galaxy was used for analysis (i.e. table joins) during annotation and RNA-Seq analysis (Blankenberg et al., 2010; Goecks et al., 2010). RNA-Seq analysis was performed using two different scaffolds including 1) the consensus sequences from de novo assembly of SOLiD reads and 2) contigs in GigasDatabase.

In order to identify enriched biological themes and GO terms, the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 was used (Huang et al., 2009a,b). Specifically, corresponding UniProt accession numbers for differentially expressed genes were used as the gene list, and compared to a complete list of the corresponding UniProt accession numbers of the respective transcriptome (i.e. results of de novo assembly or reference mapping) for the background. Biological Process terms (DAVID ‘BP Level 2’ categories) were considered significantly enriched when the p-value was less than 0.05.

3. Results

3.1. C. gigas SOLiD sequencing

After quality trimming, 20.7 and 24.6 million reads (average length: 40.6 bp) remained from the BBC and DH cDNA libraries, respectively. A majority of the reads (98%) corresponded to nuclear transcripts with the other 2% mapping to mitochondria protein coding genes. The quality trimmed reads from each library were combined for de novo assembly and reference mapping. All sequence data has been submitted to the NCBI Short Read Archive under accession number: SRP007621.

3.1.1. De novo assembly

De novo assembly of reads from the combined libraries resulted in 18,510 consensus sequences with an average length of 276 bp. Twenty three percent of the reads assembled using this approach. The average number of assembled reads per consensus sequence was 454 and the mean coverage was 61.7x (Fig. 1).

3.1.2. Reference mapping (GigasDatabase)

SOLiD reads were also mapped to publicly available *C. gigas* transcriptomic resources (GigasDatabase v8). Reads from the combined libraries mapped to 64,645 of the 82,314 contigs in the database. The average number of reads per contig was 376 and the mean coverage was 15.8x (Fig. 1). See Table 1 for a full comparison of results of the de novo assembly compared to reference mapping.

3.1.3. De novo assembly: annotation

A total of 7724 consensus sequences could be annotated, 3931 of which could be classified using GO Slim terms. The most highly represented biological process was transport, followed by protein metabolism (data not shown). Of those consensus sequences associated with transport a majority were involved in protein and ion transport. Comparatively, 7296 of the GigasDatabase contigs with mapped reads were annotated with biological process GO terms. When the associated GO terms were evaluated, two of the most highly represented biological processes identified after binning into broader GO Slim terms included protein and RNA metabolism (data not shown).

3.1.4. De novo assembly: identification of novel transcripts

Short read consensus sequences generated from de novo assembly were compared to GigasDatabase v8 to identify novel sequences. Approximately 10% of the sequences (1776) did not have a significant match (E-value > 1.0E-01). Of these, 742 could be annotated (see Supplementary Table 1) and 690 could be classified using GO Slim. The 4 most highly represented biological processes included: transport, developmental processes, cell organization and biogenesis, and cell adhesion (Fig. 2).

Table 1
Summary of assembly and RNA-Seq statistics for de novo assembly and reference mapping (GigasDatabase v8).

		De novo assembly	Reference mapping
Assembly	Mapped reads	8,407,963	29,107,760
	Unmapped reads	36,944,698	16,244,901
	Contigs	18,510	77,433
	Average contig length	276	554
	Average contig coverage	62	16
	Contigs annotated to GO Slim	3931	7296
RNA-Seq	Differentially expressed genes	2991	427
	Enriched GO biological process	15	3

3.2. RNA-Seq analysis

3.2.1. De novo-based RNA-Seq

RNA-Seq analysis using the de novo assembled short read consensus sequences as the scaffold identified 2991 differentially regulated features. Most of these features represented moderately expressed transcripts (100–10,000 total reads), but 20% were rare transcripts (<100 total reads). Six consensus sequences were expressed uniquely in the BBC library and 5 were expressed only in the DH library. None of the uniquely expressed features could be annotated. Of differentially expressed features with reads in both libraries, 1200 were expressed higher in the BBC library and 1791 were expressed higher in the DH library. A subset of the differentially expressed features (751 in BBC and 313 in DH, respectively) could be annotated (see Supplementary Table 2). A majority of these annotated features represented a twofold difference, but overall differences ranged between 2 and 409 fold.

Functional enrichment analysis identified 15 biological processes that were overrepresented in the differentially expressed gene set (Fig. 3). The most significantly enriched process was cell adhesion (p-value = 8E-15), followed by cell recognition (p-value = 5E-5).

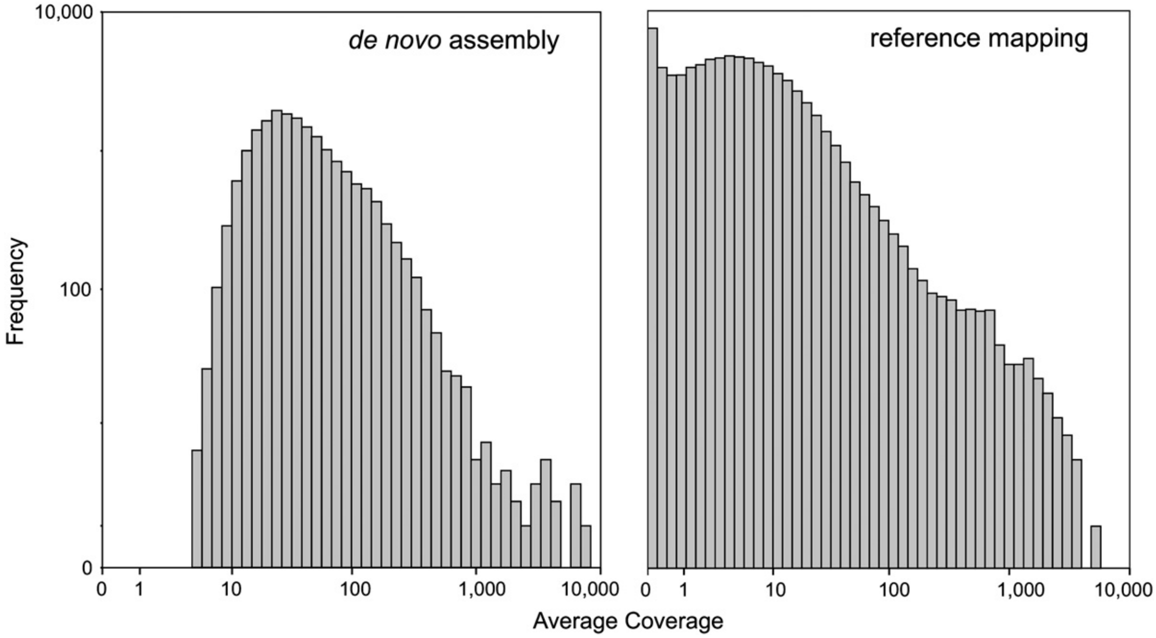

Fig. 1. Coverage distribution for de novo assembly and reference mapping. Histograms showing average read coverage for de novo assembly and reference mapping to GigasDatabase v8 for the combined *C. gigas* SOLiD transcriptome libraries.

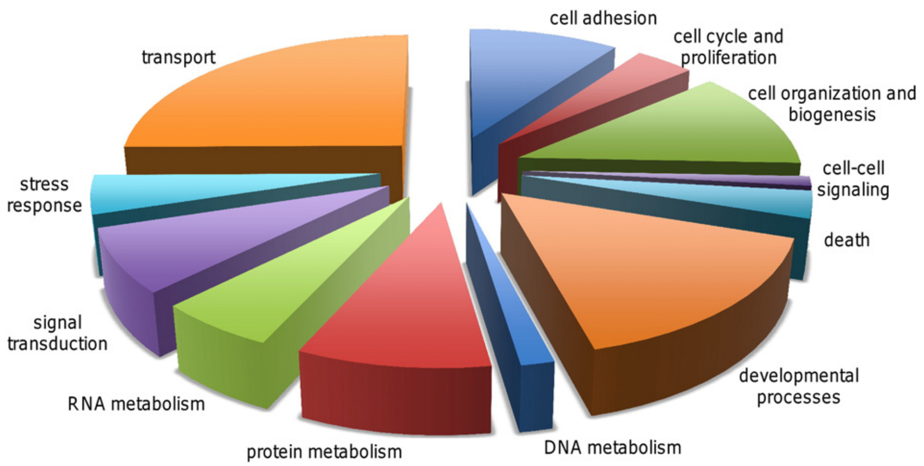

Fig. 2. Functional classification of novel transcripts identified by de novo assembly of the combined SOLiD transcriptome libraries. Categories represent 'GO Slim' terms associated with Biological Processes.

3.2.2. Reference-based RNA-Seq

For comparison, RNA-Seq was also performed using GigasDatabase v8 as the scaffold. In total, 427 differentially expressed features were identified. Of those, 239 were expressed higher in the BBC library and 189 were expressed higher in the DH library. Of these, 216 contigs could be annotated. Table 1 provides a comparison of data from both RNA-Seq procedures.

Functional enrichment analysis identified three biological processes that were enriched in the differentially expressed gene set. The most significantly enriched process was microtubule-based processes followed by oxidation reduction and cell recognition. One term, cell recognition (p value = 6E-3), overlapped between the de novo based and reference based RNA-Seq analysis. The other terms were unique to each analysis.

4. Discussion

This study evaluates the effectiveness of using high-throughput, short read sequencing technology to characterize the transcriptome of taxa with limited genomic resources. Specifically, SOLiD sequencing was carried out on cDNA libraries from Pacific oysters from two locations with differing anthropogenic influence. Sequence assembly and RNA-Seq analysis were carried out using resources generated solely as part of this study and compared to respective analyses using a publicly available transcriptome database. We found that limited ultra-short read sequence data can provide valuable information about transcriptome activity. Furthermore, we provide new genomic resources for *C. gigas* and have identified differences in oysters from areas that have experienced different degrees of human impact. These combined data significantly contribute to what we know about oyster biology but also offer a framework for efficiently characterizing transcriptomic differences in species lacking sequenced genomes. Advantages and limitations of using short read sequencing technology for gene discovery and RNA-Seq analysis are discussed.

4.1. Gene discovery

The number of Pacific oyster consensus sequences generated from de novo assembly is comparable to similar studies in sockeye salmon (Everett et al., 2011) and *R. balthica* (Feldmeyer et al., 2011). However, as expected, mean contig length (276 bp) was shorter than transcriptome

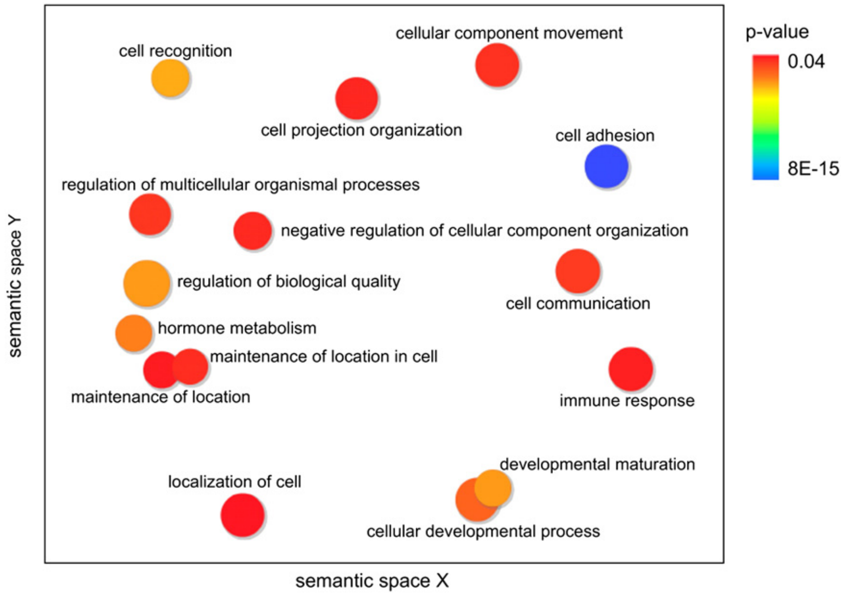

Fig. 3. Gene ontology categories overrepresented in the differentially expressed gene set. Color indicates p-value of the enrichment and size is proportional to the number of genes in the category. Spatial arrangement reflects a grouping of categories by semantic similarity.

characterizations that use 454 pyrosequencing. Recent studies in guppies (Fraser et al., 2011) and chum salmon (Seeb et al., 2010) produced mean contig lengths of 464 bp and 412 bp, respectively. In the current study our average coverage was 62x compared to 5x reported by Seeb et al. (2010). Dohm et al. (2008) have indicated greater than 20x coverage is sufficient to minimize effects of sequencing errors. We were able to annotate 42% of the consensus sequences generated from the de novo assembly. This included a large number of transcripts (742 contigs) not present in public databases. The number of novel sequences identified is slightly higher than reported in studies using Sanger sequencing for gene discovery in *C. gigas* (Gueguen et al., 2003; Roberts et al., 2008). The functional classification of the novels transcripts identified using SOLiD sequencing was highly diverse with a large proportion being involved in transport, developmental processes, stress response, and cell adhesion.

Several genes of interest were identified in the novel contigs, many of which are associated with response to stress. A number of these transcripts have been shown to be involved specifically in the immune response. For instance, a sequence with similarity to dual oxidase 2 was identified. In *Drosophila melanogaster* this protein regulates the production of reactive oxygen species in response to infectious and commensal microbes (Ha et al., 2009). The mitogen-activated protein kinase (MAPK) signaling pathway is involved in phagocytosis and the prophenoloxidase cascade in invertebrates (Lamprou et al., 2007). A subset of genes involved in this pathway has been previously identified in a *C. gigas* (Roberts et al., 2008). Here we identified a novel sequence in this pathway, mitogen-activated protein kinase kinase kinase 7 (M3K7). Another important component of the invertebrate immune system are bactericidal enzymes. A transcript similar to myeloperoxidase (MPO), which functions as a bactericide by generating hypochlorous acid (Harrison and Schultz, 1976), was present in the de novo consensus sequences. While this protein has been identified in molluscs based on its catalytic activity (Schlenk et al., 1991), this is the first time the nucleotide sequence has been reported in oysters. An additional sequence of interest possesses homology to a SAM domain and HD domain-containing protein, which has been shown to be involved in antiviral responses in humans (Rice et al., 2009).

Oysters and other coastal invertebrates are frequently exposed to xenobiotics. One of the first steps involved in the metabolism and subsequent exclusion of xenobiotics is binding of a ligand (i.e. aromatic hydrocarbon) to the aryl hydrocarbon receptor. As part of this sequencing effort we identified a transcript similar to aryl hydrocarbon receptor nuclear translocator (ARNT). ARNT encodes a protein that forms a complex with the ligand-bound aryl hydrocarbon receptor, and is required for receptor function (Hoffman et al., 1991). Activation of the aryl hydrocarbon receptor initiates transcription of cytochrome p450 oxidases. Several genes in this family have been previously reported in *C. gigas* (Roberts et al., 2008). Xenobiotic conjugates and metabolites are eventually excreted from the cell by membrane transporters in the multidrug resistance protein family. A contig generated as part of the de novo sequencing effort identified a transcript similar to multidrug resistance protein 1. Together the new sequences identified here demonstrate that limited ultra-short read sequencing provides an important resource for gene discovery.

When reference mapping was carried out, the proportion of reads that could be putatively annotated increased. While we have demonstrated that the sole use of a limited short read sequencing data set can provide cost-effective, valuable, novel genomic information, an available scaffold (i.e. EST contigs, genome) can provide benefits with respect to number of mapped reads and subsequent ability to annotate.

4.2. RNA-Seq

Using limited short read data we were able to effectively perform RNA-Seq analysis in the Pacific oyster. This is one of the first studies describing RNA-Seq analysis using solely ultra-short read data, along with other very recent publications in the crustacean *Pandalus latirostris* (Kawahara-Miki et al., 2011) and insect *Plutella xylostella* (Etebari et al., 2011). A similar approach is Tag-Seq, which utilizes short (<30 bp) tags, generally from the 3′ ends of transcripts to characterize differentially expressed genes. A recent study by de Lorgeril et al. (2011)) utilized Tag-Seq to identify approximately 4000 unique, immune responsive genes in *C. gigas*. In the current study, we were able to identify and annotate 1064 differentially expressed transcripts in *C. gigas* populations exposed to varying degrees of anthropogenic impact. Tag-Seq can be relatively less expensive than RNA-Seq with respect to coverage, however a reference scaffold is required. In addition, because tags are usually generated from a single end of a transcript, RNA-Seq analysis, as described here, has the advantage of identifying and quantifying novel transcripts (Cullum et al., 2011). In our RNA-Seq study, 18% of the differentially expressed transcripts were novel, representing a significant contribution to genomic resources. Together these studies demonstrate how advances in sequencing technology will continue improve our ability to characterize physiological responses in non-model organisms.

When comparing differentially expressed genes in oysters from the two sites, there was a large difference in the number of differentially expressed genes depending on whether the RNA-Seq was based on de novo or reference based assembly. Specifically, RNA-Seq performed using the de novo assembled consensus sequences reported seven-times as many differentially expressed genes as the RNA-Seq analysis using GigasDatabase v8. One possible explanation for this discrepancy is that using the de novo assembly as a scaffold may result in multiple sequences representing the same gene. In other words, the consensus sequences are relatively short and fragments representing different regions of the same gene may not overlap. As a majority of these differentially expressed genes could not be annotated, it is difficult to determine the precise impact of this possibility. However, 889 of the 1064 annotated, differentially expressed genes were deemed unique based on the protein identification code of the UniProt ID, suggesting there may be other factors contributing to this difference. As would be expected, based on the proportion of differentially expressed genes, the number of enriched GO biological processes identified was also different between the two analyses. It is likely that this difference is related to the scaffold itself, as all genes making up the scaffold are used as the “background” for the enrichment analysis. Therefore, it is possible that the de novo based enrichment analysis is more biologically relevant, as the background is a better representation of the genes expressed under similar conditions.

RNA-Seq analysis revealed that the set of transcripts differentially expressed between BBC and DH was most significantly enriched in genes associated with cell adhesion. In general, cell adhesion can be divided into to two general types. The first is a stable cell–cell adhesion that is critical for the organization of tissues. The second is a transient cell adhesion involved in processes such as cell adhesion between hemocytes and cell adhesion to pathogens. This transient type of cell adhesion is a critical part of invertebrate innate immunity by way of recognition of non-self particles, as well as chemotaxis and aggregation of hemocytes (reviewed by Johansson, 1999). The specific genes that are contributing to the difference between the two libraries include integrins, laminins and cadherins, which are expressed approximately 2–4 times higher in the DH library. While the precise biological role for this increased expression cannot be determined from this study, it could indicate the presence of specific contaminants in the environment. For instance, integrin expression increased in response to pathogen exposure in white shrimp (*Litopenaeus vannemai*) (Lin et al., 2010). In addition, estrogen exposure stimulates hemocyte binding to laminin 1 and collagen IV in mussels (*Mytilus galloprovincialis*) (Koutsogiannaki and Kaloyianni, 2011). While we can only speculate on the functional role, it is interesting to note that it is consistent with the environmental data from this locale, as DH is a site close to urban wastewater discharge and intensive agriculture exposure. However, additional research is required to determine the role of genes associated with cell adhesion and environmental exposures in oysters.

5. Conclusions

Ultra-short read sequencing technology, such as SOLiD, provides a powerful and effective means for gene discovery and expression analysis in organisms with limited genomic resources. We have shown that it is technically possible and efficient to use this approach to 1) generate transcriptomic resources, 2) identify novel genes, and 3) perform RNA-Seq analysis. In terms of gene expression, de novo based RNA-Seq analysis does not rely on previous transcriptome information and results can be annotated at the biological process level. As high-throughput sequencing platforms continue to improve, they will serve as important tools for examining environmental physiology of non-model organisms.

Supplementary materials related to this article can be found online at doi:10.1016/j.cbd.2011.12.003.

Acknowledgments

This work was partially funded by University of Washington, School of Aquatic and Fishery Sciences. MRG's graduate appointment is funded in part by the Victor and Tamara Loosanoff Endowed Fellowship and the John G. Peterson Scholarship. The authors would also like to thank Samuel J White and Caroline G Storer for comments on earlier versions of this manuscript.

References

Altschul, S.F., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Baggerly, K., Deng, L., Morris, J., Aldaz, C., 2003. Differential expression in SAGE: accounting for normal between-library variation. Bioinformatics 19, 1477–1483.

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., Taylor, J., 2010. Galaxy: a web-based genome analysis tool for experimentalists. Curr. Protoc. Mol. Biol. 19 (Unit 19.10.1-21).

Craft, J.A., Gilbert, J.A., Temperton, B., Dempsey, K.E., Ashelford, K., Tiwari, B., Hutchinson, T.H., Chipman, J.K., 2010. Pyrosequencing of Mytilus galloprovincialis cDNAs: tissue-specific expression patterns. Plos One 5, e8875.

Cullum, R., Alder, O., Hoodless, P.A., 2011. The next generation: using new sequencing technologies to analyse gene regulation. Respirology 16, 210–222.

de Lorgeril, J., Zenagui, R., Rosa, R.D., Piquemal, D., Bachère, E., 2011. Whole transcriptome profiling of successful immune response to *Vibrio* infections in the oyster *Crassostrea gigas* by digital gene expression analysis. PLoS One 6 (8), e23142.

Dohm, J.C., Lottaz, C., Borodina, T., Himmelbauer, H., 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 36, e105.

Etebari, K., Palfreyman, R., Schlipalius, D., Nielsen, L., Glatz, R., Asgari, S., 2011. Deep sequencing-based transcriptome analysis of *Plutella xylostella* larvae parasitized by *Diadegma semiclausum*. BMC Genomics 12, 446.

Everett, M., Grau, E., Seeb, J., 2011. Short reads and non-model species: exploring the complexities of next generation sequence assembly and SNP discovery in the absence of a reference genome. Mol. Ecol. Resour. 11, 93–108.

Ewing, B., Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8, 186–194.

Ewing, B., Hillier, L., Wendl, M.C., Green, P., 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8, 175–185.

Feldmeyer, B., Wheat, C.H., Krezdorn, N., Rotter, B., Pfenninger, M., 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.

Fleury, E., Huvet, A., Lelong, C., de Lorgeril, J., Boulo, V., Gueguen, Y., Bachère, E., Tanguy, A., Moraga, D., Fabioux, C., Lindeque, P., Shaw, J., Reinhardt, R., Prunet, R., Davey, G., Lapègue, S., Sauvage, C., Corporeau, C., Moal, J., Gavory, F., Wincker, P., Moreews, F., Klopp, C., Mathieu, M., Boudry, P., Favrel, B., 2009. Generation and analysis of a 29,745 unique Expressed Sequence Tags from the Pacific oyster (*Crassostrea gigas*) assembled into a publicly accessible database: the GigasDatabase. BMC Genomics 10, 341.

Fraser, B.A., Weadick, C.J., Janowitz, I., Rodd, H., Hughes, K.A., 2011. Sequencing and characterization of the guppy (*Poecilia reticulata*) transcriptome. BMC Genomics 12, 202.

Goecks, J., Nekrutenko, A., Taylor, J., 2010. The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 1, R86.

Goetz, F., Rosauer, D., Sitar, S., Goetz, G., Simchick, C., Roberts, S., Johnson, R., Murphy, C., Bronte, C., MacKenzie, S., 2010. A genetic basis for the phenotypic differentiation between siscowet and lean lake trout (*Salvelinus namaycush*). Mol. Ecol. 19, 176–196.

Gueguen, Y., Cadoret, J.-P., Flament, D., Barreau-Roumiguière, C., Girardot, A.-L., Garnier, J., Horeau, A., Bachère, E., Escoubas, J.-M., 2003. Immune gene discovery by expressed sequence tags generated from hemocytes of the bacteria-challenged oyster, *Crassostrea gigas*. Gene 303, 139–145.

Ha, E., Lee, K., Seo, Y.Y., Kim, S., Lim, J., Oh, B., Kim, J., Lee, W., 2009. Coordination of multiple dual oxidase–regulatory pathways in responses to commensal and infectious microbes in *Drosophila* gut. Nat. Immunol. 10, 949–957.

Harrison, J.E., Schultz, J., 1976. Studies on the chlorinating activity of myeloperoxidase. J. Biol. Chem. 251, 1371–1374.

Hoffman, E.C., Reyes, H., Chu, F.F., Sander, F., Conley, L.H., Brooks, B.A., Hankinson, O., 1991. Cloning of a factor required for activity of the Ah (dioxin) receptor. Science 252, 954–958.

Huang, D.W., Sherman, B.T., Lempicki, R.A., 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37, 1–13.

Huang, D.W., Sherman, B.T., Lempicki, R.A., 2009b. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nat. Protoc. 4, 44–57.

Johansson, M.W., 1999. Cell adhesion molecules in invertebrate immunity. Dev. Comp. Immunol. 23, 303–315.

Kawahara-Miki, R., Wada, K., Azuma, N., Chiba, S., 2011. Expression profiling without genome sequence information in a non-model species, pandalid shrimp (*Pandalus latirostris*), by next-generation sequencing. PLoS One 6 (10), e26043.

Koutsogiannaki, S., Kaloyianni, M., 2011. Effect of 17β-estradiol on adhesion of *Mytilus galloprovincialis* hemocytes to selected substrates. Role of alpha2 integrin subunit. Fish Shellfish Immunol. 31, 73–80.

Lamprou, I., Mamali, I., Dallas, K., Fertakis, V., Lampropoulou, M., Marmaras, V.J., 2007. Distinct signalling pathways promote phagocytosis of bacteria, latex beads and lipopolysaccharide in medfly haemocytes. Immunology 121, 314–327.

Lin, Y.C., Tayag, C.M., Huang, C.L., Tsui, W.C., Chen, J.C., 2010. White shrimp *Litopenaeus vannamei* that had received the hot-water extract of *Spirulina platensis* showed earlier recovery in immunity and up-regulation of gene expression after pH stress. Fish Shellfish Immunol. 29, 1092–1098.

Meyer, E., Aglyamova, G., Wang, S., Buchanan-Carter, J., Abrego, D., Colbourne, J.K., Willis, B.L., Matz, M.V., 2009. Sequencing and de novo analysis of a coral larval transcriptome using 454 GS-Flx. BMC Genomics 10.

Meyer, E., Aglyamova, M.V., Matz, G.V., 2011. Profiling gene expression responses of coral larvae (*Acropora millepora*) to elevated temperature and settlement inducers using a novel RNA-Seq procedure. Mol. Ecol. 20, 3599–3616.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5, 585–587.

Newton, J.A., Albertson, S.L., Voorhis, K.V., Maloy, C., Siegel, E., 2002. Washington State Marine Water Quality, 1998 through 2000, Publication #02-03-056. (111pp.).

Novaes, E., Drost, D.R., Farmerie, W.G., Pappas, G.J., Grattapaglia, D., Sederoff, R.R., Kirst, M., 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics 9, 312.

Rice, G.I., Bond, J., Asipu, A., Brunette, R.L., Manfield, I.W., Carr, I.M., Fuller, J.C., Jackson, R.M., Lamb, T., Briggs, T.A., Ali, M., Gornall, H., Couthard, L.R., Aeby, A., Attard-Montalto, S.P., Bertini, E., Bodemer, C., Brockmann, K., Brueton, L.A., Corry, P.C., Desguerre, I., Fazzi, E., Cazorla, A.G., Gener, B., Hamel, B.C.J., Heiberg, A., Hunter, M., van der Knaap, M.S., Kumar, R., Lagae, L., Landrieu, P.G., Lourenco, C.M., Marom, D., McDermott, M.F., van der Merwe, W., Orcesi, S., Prendiville, J.S., Rasmussen, M., Shalev, S.A., Soler, D.M., Shinawi, M., Spiegel, R., Tan, T.Y., Vanderver, A., Wakeling, E.L., Wassmer, E., Whittaker, E., Lebon, P., Stetson, D.B., Bonthron, D.T., Crow, Y.J., 2009. Mutations involved in aicardi-goutieres syndrome implicate samhd1 as regulator of the innate immune response. Nat. Genet. 41, 829–832.

Roberts, S.B., Goetz, G., White, S., Goetz, F., 2008. Analysis of genes isolated from plated hemocytes of the Pacific oysters *Crassostrea gigas*. Mar. Biotechnol. 11, 24–44.

Schlenk, D., Garcia Martinez, P., Livingstone, D.R., 1991. Studies on myeloperoxidase activity in the common mussel, Mytilus edulis. Comp. Biochem. Physiol. 99C, 63–68.

Seeb, J.E., Pascal, C.E., Grau, E.D., Seeb, L.W., Templin, W.D., Roberts, S.B., Harkins, T., 2010. Transcriptome sequencing and high-resolution melt analysis advance SNP discovery in duplicated salmonids. Mol. Ecol. Resour. 11, 335.

Shendure, J., Ji, H.L., 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26, 1135–1145.

Vera, J.C., Wheat, C.W., Fescemyer, H.W., Frilander, M.J., Crawford, D.L., Hanski, I., Marden, J.H., 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol. Ecol. 17, 1636–1647.

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63.

Washington State Department of Health: Office of Shellfish and Protection, 2006. A Report for the Puget Sound Assessment and Monitoring Program: Atlas of Fecal Coliform Pollution in Puget Sound: Year 2005 (Publication #332-061, 80 pp.).