Opportunities in Functional Genomics: A Primer on Lab and Computational Aspects

Author(s): Steven B. Roberts and Mackenzie R. Gavery
Source: Journal of Shellfish Research, 37(4):747-754.
Published By: National Shellfisheries Association
https://doi.org/10.2983/035.037.0406
URL: http://www.bioone.org/doi/full/10.2983/035.037.0406

Journal of Shellfish Research, Vol. 37, No. 4, 747–754, 2018.

OPPORTUNITIES IN FUNCTIONAL GENOMICS: A PRIMER ON LAB AND COMPUTATIONAL ASPECTS

STEVEN B. ROBERTS* AND MACKENZIE R. GAVERY
School of Aquatic and Fishery Sciences, University of Washington, 1122 NE Boat Street, Seattle, WA 98105

ABSTRACT As computational and sequencing technologies continue to flourish, the barrier for those interested in complementing traditional ecological and physiological studies with functional genomics is easier to overcome. Here, an overview of transcriptome sequencing and DNA methylation analyses in shellfish is provided, primarily for those with fundamental interests and training in a different domain. The approaches covered here can provide valuable information on how organisms respond to their environment and also be used to evaluate evolutionary relationships. First, biological and technological background is provided, highlighting studies in shellfish that have applied these approaches. This is followed by practical methods and tools for conducting this work at the laboratory bench and in front of a computer. In an effort to provide educational resources and the ability to update computational and analytical resources as they become available, a supplementary repository for this work has been created and is available at https://github.com/sr320/fun-gen.

KEY WORDS: genomics, bioinformatics, gene expression, DNA methylation, DNA sequencing

BACKGROUND

When considering marine invertebrates whether in an ecological, aquaculture, or culinary perspective, one does not often consider the expression of genes and associated regulatory mechanisms; however, for some disciplines this could be a valuable direction to expand and complement scientific endeavors. Even for those interested in molecular biology, there is often a steep learning curve from a clear fundamental understanding to practical application. This is particularly the case for functional genomics, a field that is increasingly shedding light on numerous aspects of shellfish biology. In terms of ecology, areas that could be studied using functional genomics approaches include adaptation, population structure, and species interaction. From the physiological perspective, the focus of this review, this would include understanding how fundamental biology works (e.g., growth, reproduction, and immune response). Combined, this information could be effectively applied in numerous ways including shellfish aquaculture, resource management, and ecological modeling. Here, a short introduction to gene expression and DNA methylation in shellfish is provided, as well as practical advice on how to get started.

This work focuses on ''functional genomics'' which are approaches associated with the ''expression'' of the genome in some manner. Briefly, there are portions of the genome (genes) that can be transcribed, often then translated to proteins to carry out fundamental processes. Specifically, there is a focus on messenger RNA (mRNA), which is the RNA that is transcribed from the genome where introns are excluded and a polyadenylated tail is appended. There are numerous scientific questions that can be addressed by examining mRNA expression patterns, including providing a snapshot into environmental response, identifying early signals of phenotypes, and characterizing mechanisms of disease resistance. Gene expression analysis can be used to address specific hypotheses, but also can be a powerful tool for discovery and descriptive science. The latter can easily form a basis for hypotheses that can be tested in later experiments using genomic and nongenomic approaches.

There are a myriad of processes that can influence transcriptional activity, including epigenetic mechanisms. Epigenetics is a burgeoning field in nonmodel organisms that refers to mechanisms that can alter expression in a heritable manner without altering the underlying nucleotides themselves. Unlike the nucleotides themselves, epigenetic mechanisms, including DNA methylation, histone modifications, and noncoding RNAs, are sensitive and responsive to environmental signals. As such, epigenetic mechanisms have been proposed to be an important link between the phenotype and environment, and could transform how we consider adaptation depending on the degree of heritability.

One could encapsulate functional genomics into the foundation of what controls organismal physiology, including the response to environmental stress. An additional component of genomics is how alteration of nucleotides can alter a phenotype. We will not explore aspects of genetic variation here, although there are several reviews covering the topic in relation to shellfish (Dégremont et al. 2015, Robledo et al. 2017). There are also several comprehensive reviews on the application of genomics in shellfish (Suárez-Ulloa et al. 2013, Gómez-Chiarri et al. 2015a, Sanmartín et al. 2016).

CURRENT STATE OF RESOURCES AND TECHNOLOGY

Ultimately, having a completely sequenced genome of an organism can greatly increase the ability to study the expression and variation; however, the number of species where the genomes have been fully assembled to this level of complete chromosomes is limited. One notable example where this has been accomplished is *Crassostrea virginica* (GenBank accession GCA_002022765, Gómez-Chiarri et al. 2015b). Given the complex nature of genome arrangement in many marine invertebrates, genome assembly is not a simple task and requires significant resources in both computing hardware and bioinformatic expertise. The technology is advancing at such a rate that it would be challenging to comment on ideal sequencing and associated bioinformatic approaches for genome assembly.

*Corresponding author. E-mail: sr320@u.washington.edu
DOI: 10.2983/035.037.0406

Generally, sequencing the genome of a shellfish species will often involve identifying an individual with limited genome complexity (e.g., inbred), generating sequencing libraries constituted of variable fragment lengths, and relying on software to assemble the many pieces of the genome together in the correct order and orientation. If genomes are available for your species of interest, this could certainly improve downstream analysis and increase options for interpretation.

Some of the first mollusc genomes to be sequenced were the California sea hare, *Aplysia californica*, and the owl limpet, *Lottia gigantea*. The *L. gigantea* genome is 348 Mb, has a substantial number of repetitive elements (21% of the assembled genome), and about 23,800 genes (Simakov et al. 2013). The Pacific oyster, *Crassostrea gigas*, was the first bivalve genome published with 11,969 scaffolds (Zhang et al. 2012). A total of 28,027 genes were predicted. In early 2017, a number of new bivalve genomes were published including *Patinopecten yessoensis* (Wang et al. 2017) and genomes of a deep-sea vent/seep mussel (*Bathymodiolus platifrons*) and a shallow-water mussel (*Modiolus philippinarum*) (Sun et al. 2017). Another popular bivalve, the eastern oyster (*Crassostrea virginica*), has also been sequenced (Gómez-Chiarri et al. 2015b) with data available ahead of publication (*C. virginica* isolate RU13XGHG1–28, GenBank accession: GCA_002022765).

A primary entry into genomics and examining the organismal response is describing the set of coding (mRNA) and noncoding products expressed in the genome. Sequencing technology is at a place where generating the full set of expressed mRNA can be readily accomplished. The approach, termed RNA-Seq, refers to the use of high-throughput or next-generation sequencing (e.g., Illumina HiSeq, PacBio Sequel) to sequence the full complement of mRNA in a given cell/tissue type. The precursor to this approach would be sequencing Expressed Sequence Tags using Sanger sequencing technology. It is the evolution of sequencing technologies beyond the Sanger-based that has made all aspects of genomics more accessible.

In practical terms, to identify expressed portions of a genome, one would isolate this fraction from tissue or cells based on attributes of eukaryotic coding transcripts (e.g., polyadenylated 3′ end) and/or size fractionation. Details of this are provided in the following section. The limits of what can be derived from a single transcriptome should be considered; however, there are a number of valuable products that could be generated from an RNA-seq experiment including targets for quantitative polymerase chain reaction analysis, basis for comparative evolutionary studies, nucleotide diversity characterization, and motif analyses. Examples of early transcriptome studies using Expressed Sequence Tags include those focusing on immune function in *Crassostrea gigas* (Gueguen et al. 2003) and *Crassostrea virginica* (Jenny et al. 2002). In one of the early uses of ''next-generation'' sequencing efforts in shellfish, Hedgecock et al. (2007) used massively parallel signature sequencing to examine gene expression patterns in two partially inbred and two hybrid larval populations. There are numerous studies in shellfish that provide valuable transcriptomic resources using RNA-seq technology (Timmins-Schiffman Friedman 2012, Zhang et al. 2012, Núñez-Acuña & Gallardo-Escárate 2013, Moreira et al. 2014, Pauletto et al. 2014, Teaniniuraitemoana et al. 2014, Valenzuela-Muñoz et al. 2014, Valenzuela-Miranda et al. 2015).

Epigenetic mechanisms, such as DNA methylation, are key to our understanding of how environment impacts phenotype, as they can contribute to gene regulatory activity and be influenced by the environment. DNA methylation refers to the enzymatic addition of a methyl group to a cytosine residue in DNA, which occurs almost exclusively at CpG dinucleotides (i.e., a cytosine located 5′ of a guanine) in animals. The enzymatic machinery supporting DNA methylation includes a family of DNA methyltransferases (DNMT) including the maintenance methyltransferase DNMT1 and the *de novo* methyltransferases DNMT3A/3B. In plant and mammalian studies, it has been shown that DNA methylation is sensitive to external factors including maternal behavior (Weaver et al. 2004), nutrition (Dolinoy et al. 2007), and photoperiod (Azzi et al. 2014).

Most of the studies on DNA methylation in shellfish have focused on *Crassostrea gigas*, as the available genome significantly improves effectiveness of data analysis. One primary means to identify DNA methylation in any given cell or tissue is to use bisulfite treatment of DNA, which converts unmethylated cytosines to uracil, whereas methylated cytosines are maintained as cytosine. Informatically, this sequence information would be compared with a reference genome to determine which cytosines are methylated. This approach, generally termed ''bisulfite sequencing,'' can be performed on whole genomes (whole genome bisulfite sequencing) or reduced representation genomes (reduced representation bisulfite sequencing). Genome reduction techniques can be cost effective in that more individuals can be sequenced for the same cost because only a small fraction of the genome is being sequenced. DNA methylation can be characterized at the fragment level using antibody approaches without bisulfite treatment, whereas bisulfite treatment will offer nucleotide-level resolution. Alternatives to bisulfite sequencing approaches that apply high-throughput sequencing include methylated DNA immunoprecipitation (MeDIP-Seq) and methylation-sensitive cut counting techniques, such as EpiRADseq (Kurdyukov & Bullock 2016, Schield et al. 2016, Dimond et al. 2017). Techniques such as EpiRADseq are particularly useful for nonmodel species lacking genomic resources.

DNA methylation was examined genome-wide in the Pacific oyster where it was reported that 15% of CpGs were methylated in a somatic tissue (Gavery & Roberts 2013), compared with 60%–70% in mammals. In oysters, as in other invertebrates, the methylated fraction of the genome tends to consist of gene bodies, whereas other genomic regions exhibit less methylation. In the Pacific oyster, high levels of methylation in gene bodies (and putative promoter regions) are associated with high levels of expression (Gavery & Roberts 2013, Olson & Roberts 2014). Interestingly, genes with limited methylation in oysters have variability in exon-specific expression across tissue types, indicating that hypomethylation may be associated with increased plasticity (Gavery & Roberts 2014). Similar results have been observed in scleractinian reef corals (Dixon et al. 2014, Dimond & Roberts 2016). Whereas more studies are needed to quantify this relationship, there are significant implications for improving resilience in shellfish—particularly if DNA methylation patterns are heritable. There are few studies on heritability of methylation patterns, although Rondon et al. (2017) have recently shown parental herbicide exposure influences progeny DNA methylation patterns in oysters.

PRACTICAL ASPECTS

In this section, practical considerations for getting into functional genomic analyses are provided. To reiterate, the target audience here is the student and scientist with primarily ecology-based training and limited training in molecular biology. This section is organized starting with living cells and ending with data visualization, with a focus on transcriptomic sequencing and sequence-based DNA methylation analyses.

Experimental Considerations

As with any experiment, replication, number of individuals, effect size, and resources available need to be considered. One resource to consider might be the availability of current genomic resources. For instance, is the genome for your species of interest available? This is not a necessity for many of the applications, as RNA-Seq data can be assembled *de novo*; however, for some DNA methylation characterization approaches (e.g., bisulfite sequencing) a draft or complete genome is critical for analysis.

For functional genomics studies, the tissue and/or cell type also needs to be carefully considered for biological relevance. Unlike the DNA, where all cell types contain the same nucleotide information, gene expression and associated gene regulatory process provide a biological snapshot in time and space for a particular cell type. For example, to determine how a shellfish responds to a pathogen, one option would be to isolate the primary immune cells, hemocytes, at multiple time points post infection and compare expression and/or methylation differences in exposed organisms versus controls. Likewise, if the goal of the study were to characterize how water conditions alter regulatory mechanism associated with reproduction, gonad tissue would be a primary tissue to target for analysis.

Once cells or tissue(s) from an organism are identified for functional genomic analysis, samples will need to be preserved to limit degradation. RNA is more labile than DNA, thus proper extraction and preservation are critical. Acceptable options, often dictated by proximity to resources, include immediate –80°C freezer storage, liquid nitrogen, or a commercial preservative such as RNAlater (Ambion). Total RNA can be extracted using commercial products (e.g., TriReagent; Sigma-Aldrich, Trizol; ThermoFisher) followed by further purification of mRNA using a technology that targets the poly-A 3′ region of eukaryotic genes. Extraction of DNA for sequence-based analysis of DNA methylation would also need similar attention to target cells or tissue(s), with extraction of DNA possible with commercial kits (e.g., E.Z.N.A. Mollusc DNA kit; Omega Bio-Tek). Once DNA or mRNA is isolated, it can be used for construction of libraries that can be sequenced on high-throughput sequencing platforms. The number of companies and associated sequencing technologies has seen a turnover in the past decade with new technologies in development. As of the writing of this review, Illumina is the primary company involved in sequencer manufacturing and for simplicity will be the basis for technology and sequence data formats discussed here.

Library Preparation and Sequencing

Regardless of application, libraries generated for high-throughput sequencing have the same basic architecture; DNA or cDNA fragments (often referred to as "inserts") are ligated to adapters on both ends. The adapters (typically around 60 bp) contain regions to bind primers for the sequencing reaction and oftentimes have index regions so that multiple samples can be pooled (and later demultiplexed) based on unique indices. Inserts can be generated by sonication to produce random fragments (typically used for whole genome sequencing), or via digestion (e.g., heated, cationic digestion of RNA in the case of RNA-seq). Enyzmatic digestion of DNA enables high coverage of a reduced representation of the genome (e.g., reduced representation bisulfite sequencing). In these cases, a subsequent size-selection step typically follows to only include a small range of fragments (~200 bp), which is important to reduce sequencing bias due to variability in insert size. Commercial kits (e.g., Illumina TruSeq) can be purchased to generate libraries in-house. In addition, university-based core facilities or companies can be contracted to construct libraries and perform sequencing. It is advisable to contact the sequencing facility and/or experienced bioinformatition to determine important parameters such as sequence read coverage, sample size, pooling options, and budget considerations.

There are a few common variations on the type of sequence data that is generated, usually designated by read length (e.g., 100 bp is a common read length) and either paired-end or single-end reads (sequencing both ends of the insert or just one, respectively). These different options can have impacts on sequencing costs, turnaround time, and data yields. These parameters should be considered during the experimental design process. Again, it is recommended to consult with a sequencing facility to understand the differences in these options as they apply to your specific experiment or goals.

Bioinformatics

Once the sequencing run is finished, there will be a large amount of data in the form of text files (e.g., fastq files) that will need to be processed to gain biological information. Data management is one of the most common issues encountered in bioinformatics, which includes moving, storing, opening, and documenting files. Some minimal computational proficiency is beneficial before delving into analysis of large amounts of sequence-based data. Foremost is a basic understanding of the command line and manipulating files in this manner. Some great resources to gain proficiency in this area include Software Carpentry Courses (software-carpentry.org) and a number of other publications (Wilson et al. 2014, Buffalo 2015). Other valuable considerations for any type of computational analysis are reproducibility and version control, both of which are covered well in the resources mentioned. Reproducibility is valuable for the benefit of others, but more importantly for your own workflow. A hallmark of data analysis in this field, particularly when just starting out, is the need to rerun analysis with minor adjustments in parameters. This often requires time and computational resources given the size of the data files. Fully documenting your work will allow for precise comparisons of the output of similar approaches, which in turn will allow you to assess proper workflows. The nature of the data and tools associated with genomics lends itself to sharing, and as a field, genomics has been in the forefront of open science practices. One primary reason for this is the realization that within these massive sequence files are answers to questions and

unrealized discoveries that often the investigator might not even consider. In addition, there is great value in building on prior work to advance our understanding of physiological responses. For instance, a basic transcriptome generated from gonad tissue of a shellfish at different maturational stages could be used by other laboratory groups to identify genetic markers, develop quantitative polymerase chain reaction assays, design primers to identify homologs in other species, and create an *in silico* reference for shotgun proteomic analysis. There are now several exciting avenues for early publishing of genomic data products including online notebooks, preprint servers (e.g., biorxiv), and data descriptor publications (e.g., Scientific Data—Nature Publishing Group). Jupyter Notebooks (jupyter.org) and GitHub (github.com) are excellent means to properly document analysis and facilitate open science.

The next section covers the computational aspects of sequence data analysis from the raw data to visualization. Some examples of useful software are provided, however a more complete list of software and online platforms for data analysis are provided in Table 1, along with the respective citations.

Sequence Read Quality and Trimming

An initial sequence file will be in the fastq format. Fastq files are text based with nucleotide and quality information. Software packages, including FastQC, are good for initial quality assessment of the sequence data. It is important to understand if the sequencing reads have been pretrimmed by the sequencing facility or if they are raw sequences. Trimming typically includes trimming low quality bases and adapter sequences and removal of short sequences. Some sequencing facilities may perform these trimming functions before providing the data, whereas others will provide untrimmed sequences. It is important to know if/how the sequences you receive from the sequencing facility have been trimmed or not as it is very important to trim sequences before downstream analyses. Table 1 includes packages that perform these trimming functions.

Sequence Read Assembly

In instances where there is no genome available for your target species, a first step for gene expression–based projects

TABLE 1.

Software and resources useful in functional genomic data analyses.

Raw data quality control
FastQC (Andrews 2010)—quality metrics for high-throughput sequencing runs
Sequence trimming
Trimmomatic (Bolger etal. 2014)—adapter and quality trimming
TrimGalore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore)—wrapper for cutadapt (Martin 2011)—adapter and quality trimming, extra functionality for bisulfite sequencing
Transcriptome sequence assembly and characterization
Trinity (Grabherr et al. 2011)—*de novo* assembly with downstream analysis capability
Trinotate (Haas et al. 2013)—annotation suite designed for automatic functional annotation of transcriptomes
Dammit! (Scott 2017)—a simple *de novo* transcriptome annotator
Transrate (Smith-Unna et al. 2016)—*de novo* transcriptome assembly quality analysis
Sequence read mapping
Bowtie(2) (Langmead et al. 2009)—alignment tool for high-throughput sequencing
BWA (Li & Durbin 2009)—alignment tool for high-throughput sequencing
STAR (Dobin et al. 2013)–alignment tool for high-throughput sequencing to a reference genome
TopHat(2) (Kim et al. 2013)—splice junction mapper for RNA-Seq reads (uses Bowtie)
BSMAP (Xi & Li 2009)—bisulfite sequencing specific alignment
Bismark (Krueger & Andrews 2011)—bisulfite sequencing specific alignment (uses Bowtie)
Samtools (Li et al. 2009)—SAM/BAM manipulations: conversion, sorting, indexing etc.
Differential gene expression and differential methylation analysis
RSEM (Li & Dewey 2011)—alignment based transcript abundance estimation
Kallisto (Bray et al. 2016)—alignment-free transcript abundance estimation
EdgeR (Robinson et al. 2010)—R package for differential expression analysis
DeSeq (Anders & Huber 2012)—R package for differential expression analysis
Methylkit (Akalin et al. 2012)—R package for analysis of DNA methylation profiles
MethylExtract (Barturen et al. 2013)—generates methylation maps and detects sequence variation
MACAU (Lea et al. 2015)—differential methylation analysis for bisulfite sequencing data
Platforms for genomic analyses and visualization
Galaxy (Goecks et al. 2013, Afgan et al. 2016)—web-based platform for accessible, reproducible, and transparent analysis and visualization
Integrated genome viewer (Robinson et al. 2011, Thorvaldsdóttir et al. 2013)—visualization tool for interactive exploration of large, integrated genomic datasets
CoGe (Lyons & Freeling 2008, Lyons et al. 2008)—online system for making the retrieval and comparison of genomic information and sequence data
Cyverse (Merchant et al. 2016)—cyberinfrastructure for enabling data to discovery for the life sciences
Bedtools (Quinlan & Hall 2010)—suite of utilities for comparing genomic features
Samtools (Li et al. 2009)—suite of programs for interacting with high-throughput sequencing data
Cytoscape (Shannon et al. 2003)—software platform for visualizing complex networks
Revigo (Supek et al. 2011)—web service that visually summarizes gene ontology terms
SQLShare—(Howe et al. 2011)—database-as-a-service platform designed to facilitate data sharing and collaborative analysis

would be to assemble a reference transcriptome using software such as Trinity (Grabherr et al. 2011). Assembly of the reads is required to confidently assign each of the short sequencing reads to a particular gene. The output of an assembly is a fasta format file with nucleotide information contained in "contigs." Contigs are longer, contiguous lengths of sequences that have been compiled from the short sequencing reads. A primary challenge once a transcriptome is produced is attempting to assign identification (annotation) to the putative genes. There are a number of robust software packages for annotation and assessment of transcriptome assemblies (see Table 1). These packages rely on algorithms that compare unknown sequences to known sequences. A widely used algorithm for this type of comparison is BLAST (Altschul et al. 1990) and is available online and as a stand-alone version. When annotating a new transcriptome, BLAST would need to be run locally (stand-alone version) using the command line. A tutorial on installing and using BLAST is available in the supplemental online repository.

Sequence Read Mapping

When performing a comparative RNA-Seq experiment, that is looking at gene expression differences under specific conditions, the quality-trimmed, short sequence reads are typically aligned to a "reference," which could be the newly assembled transcriptomes, a reference transcriptome or a genome if available. There are many alignment tools designed specifically to handle high-throughput sequencing reads (see Table 1). Some programs are designed specifically if a reference genome is available (e.g., TopHat, STAR). For particular applications, such as bisulfite sequencing for DNA methylation, bisulfite-specific aligners (e.g., Bismark and BSMAP) need to be used as the sequences are heavily cytosine depleted. There are a number of parameters that need to be considered for mapping, and unfortunately, there is no simple recipe that can be followed. Read the manual of the software you are working with to understand the parameters. It is important to note that default parameters are not always best particularly for nonmodel species. Parameters that need to be considered include how many mismatches between a read and a reference are acceptable, what should happen to reads that map to more than one location in the reference, and what should be carried out with paired-end reads when the pairs are not identified.

A primary output of mapping sequence reads is a sequence alignment/map (SAM) or binary alignment/map (BAM) file. These files contain alignment information about each read, location of alignment to the reference, and quality and uniqueness of the alignment. Basic data that you would extract from alignment files include expression information, which for RNA-Seq, is typically "counts" (or number of reads) that map per gene. Note that there are also alignment-free methods for transcript quantification (e.g., Kallisto). For bisulfite sequence data, percent methylation for each CpG dinucleotide can be extracted from the alignment file by counting the number of times a cytosine or thymine is observed at a particular locus. There are a number of packages designed to take an alignment file as input to extract counts or proportion methylation (see Table 1). In addition, Samtools is a useful suite of tools for sorting, indexing, and converting between SAM/BAM format, some or all of which may be required depending on downstream applications.

Differential Gene Expression and Differential Methylation Analysis

Commonly, an end goal of a gene expression or DNA methylation experiment is to perform differential expression or differential methylation analysis where data are statistically modeled to identify biologically relevant differences between experimental groups. There are a number of tools designed to identify differences in high-throughput sequencing data, and examples of commonly used software are listed in Table 1. As in all aspects of informatics, care should be taken to understand the way the data are being normalized, modeled, and statistically interpreted.

Beyond identification of differentially expressed genes or differentially methylated cytosines (or regions) are possibilities to annotate or summarize genes and regions in a broader biological context. For differential methylation analysis, it is common practice to annotate differentially methylated cytosines or regions according to genes or other functional genomic elements in close proximity in the genome. This type of spatial analysis is relatively easy to perform using genome feature file formats (e.g., bed, gff) and tools such as Bedtools. Another commonly used approach to look at broader biological functions of gene set lists is referred to as gene set enrichment analysis. Enrichment analysis includes when differentially expressed or differentially methylated gene lists can be statistically analyzed to see if genes within a biological function or pathway are enriched in response to the experimental variable of interest (see Table 1 for software suggestions). For additional information regarding considerations for differential gene expression analysis using RNA-Seq, see Conesa et al. (2016), and for differential methylation analysis using bisulfite sequencing see Lea et al. (2017).

Visualizations

Often, there is value in taking the large amount of nucleotide data and revealing biologically meaningful results in a visual manner. Visualization of the data should happen at different steps in the analysis pipeline. For instance, it is helpful to see sequence quality information (e.g., FastQC) and visualize sequence reads mapping back to a reference sequence (e.g., Integrated genome viewer). Several software packages mentioned previously have visualization built in (e.g., Trinity, TopHat, Methylkit). There are numerous general use visualization packages in R (R Core Team 2014) [e.g., ggplot2 (Wickham 2010), superheat (Barter & Yu 2015)] that can be used to visualize gene expression/DNA methylation data in the form of principal component analyses and heatmaps. Heatmaps, in particular, can also be valuable for visualizing the amount of variation within groups. Functional genomics has many tools for looking past the single gene level to networks of genes that may be regulated in response to a particular condition. Resources such as The Database for Annotation, Visualization, and Integrated Discovery can provide information on regulatory networks that may be associated with a set of differentially expressed or differentially methylated genes, and allow for better biological interpretation of the data. Particularly, working with shellfish, caution should be taken when relying on pathway analysis as this will commonly be based on model species and make assumptions regarding similarity of gene function across species.

CONCLUSIONS

Sequencing technology, software, and consequently, an understanding of functional genomics is constantly changing. Thus, although this review attempts to outline practical aspects along with the current knowledge, it is likely that both will be different in the next decade. The primary goal here was to outline potential value in exploring the functional genomics side of shellfish, specifically expression of genes along with the epigenetic mechanisms associated with gene regulation. For shellfish particularly, the epigenetic component has significant potential for expansion as there is a firm foundational knowledge of primary processes, with several outstanding questions. In addition, many shellfish experience heterogeneity in environmental conditions, and having a better understanding of epigenetic processes will provide insight into organismal- and population-level responses. Recent examples of this include the role of DNA methylation in oyster development (Riviere et al. 2017) and a study describing how DNA methylation status influences invasive species success (Ardura et al. 2017). Going forward, our understanding of biological and ecological processes will improve at the molecular level, which will ultimately allow scientist to better predict ecosystem impacts in our changing world.

LITERATURE CITED

Afgan, E., D. Baker, M. van den Beek, D. Blankenberg, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, C. Eberhard, B. Grüning, A. Guerler, J. Hillman-Jackson, G. Von Kuster, E. Rasche, N. Soranzo, N. Turaga, J. Taylor, A. Nekrutenko & J. Goecks. 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res*. 44: W3–W10.

Akalin, A., M. Kormaksson, S. Li, F. E. Garrett-Bakelman, M. E. Figueroa, A. Melnick & C. E. Mason. 2012. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol*. 13:R87.

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

Anders, S. & W. Huber. 2012. Differential expression of RNA-Seq data at the gene level—the DESeq package. Heidelberg, Germany: European Molecular Biology Laboratory (EMBL).

Andrews, S. 2010. FastQC: a quality control tool for high throughput sequence data.

Ardura, A., A. Zaiko, P. Morán, S. Planes & E. Garcia-Vazquez. 2017. Epigenetic signatures of invasive status in populations of marine invertebrates. *Sci. Rep*. 7:42193.

Azzi, A., R. Dallmann, A. Casserly, H. Rehrauer, A. Patrignani, B. Maier, A. Kramer & S. A. Brown. 2014. Circadian behavior is light-reprogrammed by plastic DNA methylation. *Nat. Neurosci*. 17:377–382.

Barter, R. L. & B. Yu. 2015. Superheat: an R package for creating beautiful and extendable heatmaps for visualizing complex data. arXiv [stat.AP].

Barturen, G., A. Rueda, J. L. Oliver & M. Hackenberg. 2013. MethylExtract: high-quality methylation maps and SNV calling from whole genome bisulfite sequencing data. *F1000 Res*. 2:217.

Bolger, A. M., M. Lohse & B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120.

Bray, N. L., H. Pimentel, P. Melsted & L. Pachter. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol*. 34:525–527.

Buffalo, V. 2015. Bioinformatics data skills: reproducible and robust research with open source tools. Sebastopol, CA: O'Reilly Media, Inc.

Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang & A. Mortazavi. 2016. A survey of best practices for RNA-seq data analysis. *Genome Biol*. 17:13.

Dégremont, L., C. Garcia & S. K. Allen, Jr. 2015. Genetic improvement for disease resistance in oysters: a review. *J. Invertebr. Pathol*. 131:226–241.

Dimond, J., S. Gamblewood & S. Roberts. 2017. Genetic and epigenetic insight into morphospecies in a reef coral. *Mol. Ecol*. 26:5031–5042.

Dimond, J. L. & S. B. Roberts. 2016. Germline DNA methylation in reef corals: patterns and potential roles in response to environmental change. *Molecular Ecology* 25:1895–1904.

Dixon, G. B., L. K. Bay & M. V. Matz. 2014. Bimodal signatures of germline methylation are linked with gene expression plasticity in the coral *Acropora millepora*. *BMC Genomics* 15:1109.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson & T. R. Gingeras. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21.

Dolinoy, D. C., D. Huang & R. L. Jirtle. 2007. Maternal nutrient supplementation counteracts bisphenol A-induced DNA hypomethylation in early development. *Proc. Natl. Acad. Sci. USA* 104:13056–13061.

Gavery, M. R. & S. B. Roberts. 2013. Predominant intragenic methylation is associated with gene expression characteristics in a bivalve mollusc. *PeerJ* 1:e215.

Gavery, M. R. & S. B. Roberts. 2014. A context dependent role for DNA methylation in bivalves. *Brief. Funct. Genomics* 13:217–222.

Goecks, J., C. Eberhard, T. Too; Galaxy Team, A. Nekrutenko & J. Taylor. 2013. Web-based visual analysis for high-throughput genomics. *BMC Genomics* 14:397.

Gómez-Chiarri, M., X. Guo, A. Tanguy, Y. He & D. Proestou. 2015a. The use of -omic tools in the study of disease processes in marine bivalve mollusks. *J. Invertebr. Pathol*. 131:137–154.

Gómez-Chiarri, M., W. C. Warren, X. Guo & D. Proestou. 2015b. Developing tools for the study of molluscan immunity: the sequencing of the genome of the eastern oyster, *Crassostrea virginica*. *Fish Shellfish Immunol*. 46:2–4.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman & A. Regev. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol*. 29:644–652.

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

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. Macmanes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. Leduc, N. Friedman & A. Regev. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc*. 8:1494–1512.

Hedgecock, D., J.-Z. Lin, S. DeCola, C. D. Haudenschild, E. Meyer, D. T. Manahan & B. Bowen. 2007. Transcriptomic analysis of growth heterosis in larval Pacific oysters (*Crassostrea gigas*). *Proceedings of the National Academy of Sciences of the United States of America* 104:2313–2318.

Howe, B., G. Cole, E. Souroush, P. Koutris, A. Key, N. Khoussainova & L. Battle. 2011. Database-as-a-service for long-tail science. In:

Scientific and Statistical Database Management. Presented at the International Conference on Scientific and Statistical Database Management. Berlin, Germany: Springer. pp 480–489.

Jenny, M. J., A. H. Ringwood, E. R. Lacy, A. J. Lewitus, J. W. Kempton, P. S. Gross, G. W. Warr & R. W. Chapman. 2002. Potential indicators of stress response identified by expressed sequence tag analysis of hemocytes and embryos from the American oyster, *Crassostrea virginica*. *Mar. Biotechnol. (NY)* 4:81–93.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley & S. L. Salzberg. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36.

Krueger, F. & S. R. Andrews. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27:1571–1572.

Kurdyukov, S. & M. Bullock. 2016. DNA methylation analysis: choosing the right method. *Biology (Basel)* 5:3.

Langmead, B., C. Trapnell, M. Pop & S. L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25.

Lea, A. J., J. Tung & X. Zhou. 2015. A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data. *PLoS Genet.* 11:e1005650.

Lea, A. J., T. P. Vilgalys, P. A. P. Durst & J. Tung. 2017. Maximizing ecological and evolutionary insight in bisulfite sequencing data sets. *Nature Ecology & Evolution 1:*1074–1083.

Li, B. & C. N. Dewey. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323.

Li, H. & R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin & 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.

Lyons, E. & M. Freeling. 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* 53:661–673.

Lyons, E., B. Pedersen, J. Kane, M. Alam, R. Ming, H. Tang, X. Wang, J. Bowers, A. Paterson, D. Lisch & M. Freeling. 2008. Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.* 148:1772–1781.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17:10–12.

Merchant, N., E. Lyons, S. Goff, M. Vaughn, D. Ware, D. Micklos & P. Antin. 2016. The iPlant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. *PLoS Biol.* 14: e1002342.

Moreira, R., M. Milan, P. Balseiro, A. Romero, M. Babbucci, A. Figueras, L. Bargelloni & B. Novoa. 2014. Gene expression profile analysis of Manila clam (*Ruditapes philippinarum*) hemocytes after a *Vibrio alginolyticus* challenge using an immune-enriched oligo-microarray. *BMC Genomics* 15:267.

Núñez-Acuña, G. & C. Gallardo-Escárate. 2013. Identification of immune-related SNPs in the transcriptome of *Mytilus chilensis* through high-throughput sequencing. *Fish Shellfish Immunol.* 35:1899–1905.

Olson, C. E. & S. B. Roberts. 2014. Genome-wide profiling of DNA methylation and gene expression in *Crassostrea gigas* male gametes. *Front. Physiol.* 5:224.

Pauletto, M., M. Milan, R. Moreira, B. Novoa, A. Figueras, M. Babbucci, T. Patarnello & L. Bargelloni. 2014. Deep transcriptome sequencing of *Pecten maximus* hemocytes: a genomic resource for bivalve immunology. *Fish Shellfish Immunol.* 37:154–165.

Quinlan, A. R. & I. M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

Robinson, J. T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz & J. P. Mesirov. 2011. Integrative genomics viewer. *Nat. Biotechnol.* 29:24–26.

Robinson, M. D., D. J. McCarthy & G. K. Smyth. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–140.

Robledo, D., C. Palaiokostas, L. Bargelloni, P. Martínez & R. Houston. 2017. Applications of genotyping by sequencing in aquaculture breeding and genetics. *Rev. Aquacult.* pp. 1–13.

Rondon, R., C. Grunau, M. Fallet, N. Charlemagne, R. Sussarellu, C. Chaparro, C. Montagnani, G. Mitta, E. Bachère, F. Akcha & C. Cosseau. 2017. Effects of a parental exposure to diuron on Pacific oyster spat methylome. *Environ. Epigenet.* 3:1–13.

Sanmartín, R. M., S. Roberts & A. Figueras. 2016. 9—Molluscs. In: MacKenzie, S. & S. Jentoft, editors. Genomics in aquaculture. San Diego, CA: Academic Press. pp. 223–245.

Schield, D. R., M. R. Walsh, D. C. Card, A. L. Andrew, R. H. Adams & T. A. Castoe. 2016. EpiRADseq: scalable analysis of genomewide patterns of methylation using next-generation sequencing. Methods in ecology and evolution/British. *Ecol. Soc.* 7:60–69.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski & T. Ideker. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13:2498–2504.

Simakov, O., F. Marletaz, S.-J. Cho, E. Edsinger-Gonzales, P. Havlak, U. Hellsten, D.-H. Kuo, T. Larsson, J. Lv, D. Arendt, R. Savage, K. Osoegawa, P. de Jong, J. Grimwood, J. A. Chapman, H. Shapiro, A. Aerts, R. P. Otillar, A. Y. Terry, J. L. Boore, I. V. Grigoriev, D. R. Lindberg, E. C. Seaver, D. A. Weisblat, N. H. Putnam & D. S. Rokhsar. 2013. Insights into bilaterian evolution from three spiralian genomes. *Nature* 493:526–531.

Smith-Unna, R., C. Boursnell, R. Patro, J. M. Hibberd & S. Kelly. 2016. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res.* 26:1134–1144.

Suárez-Ulloa, V., J. Fernández-Tajes, C. Manfrin, M. Gerdol, P. Venier & J. M. Eirín-López. 2013. Bivalve omics: state of the art and potential applications for the biomonitoring of harmful marine compounds. *Mar. Drugs* 11:4370–4389.

Sun, J., Y. Zhang, T. Xu, Y. Zhang, H. Mu, Y. Zhang, Y. Lan, C. J. Fields, J. H. L. Hui, W. Zhang, R. Li, W. Nong, F. K. M. Cheung, J.-W. Qiu & P.-Y. Qian. 2017. Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nat. Ecol. Evol.* 1:0121.

Supek, F., M. Bošnjak, N. Škunca & T. Šmuc. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6:e21800.

Scott, C. 2016. dammit: an open and accessible de novo transcriptome annotator. Available at: www.camillescott.org/dammit.

R Core Team. 2014. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Riviere, G., Y. He, S. Tecchio, E. Crowell, M. Gras, P. Sourdaine, X. Guo & P. Favrel. 2017. Dynamics of DNA methylomes underlie oyster development. *PLoS Genetics* 13:e1006807.

Teaniniuraitemoana, V., A. Huvet, P. Levy, C. Klopp, E. Lhuillier, N. Gaertner-Mazouni, Y. Gueguen & G. Le Moullac. 2014. Gonad transcriptome analysis of pearl oyster *Pinctada margaritifera*: identification of potential sex differentiation and sex determining genes. *BMC Genomics* 15:491.

Thorvaldsdóttir, H., J. T. Robinson & J. P. Mesirov. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14:178–192.

Timmins-Schiffman Friedman, M. W. R. 2012. Genomic resource development for shellfish of conservation concern. *Mol. Ecol. Resour.* 13:295–305.

Valenzuela-Miranda, D., M. A. Del Río-Portilla & C. Gallardo-Escárate. 2015. Characterization of the growth-related transcriptome

in California red abalone (*Haliotis rufescens*) through RNA-Seq analysis. *Mar. Genomics* 24:199–202.

Valenzuela-Muñoz, V., M. A. Bueno-Ibarra & C. G. Escárate. 2014. Characterization of the transcriptomes of *Haliotis rufescens* reproductive tissues. *Aquacult. Res.* 45:1026–1040.

Wang, S., J. Zhang, W. Jiao, J. Li, X. Xun, Y. Sun, X. Guo, P. Huan, B. Dong, L. Zhang, X. Hu, X. Sun, J. Wang, C. Zhao, Y. Wang, D. Wang, X. Huang, R. Wang, J. Lv,, Y. Li, Z. Zhang, B. Liu, W. Lu, Y. Hui, J. Liang, Z. Zhou, R. Hou, X. Li, Y. Liu, H. Li, X. Ning, Y. Lin, L. Zhao, Q. Xing, J. Dou, Y. Li, J. Mao, H. Guo, H. Dou, T. Li, C. Mu, W. Jiang, Q. Fu, X. Fu, Y. Miao, J. Liu, Q. Yu, R. Li, H. Liao, X. Li, Y. Kong, Z. Jiang, D. Chourrout, R. Li & Z. Bao. 2017. Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nat. Ecol. Evol.* 1:0120.

Weaver, I. C. G., N. Cervoni, F. A. Champagne, A. C. D'Alessio, S. Sharma, J. R. Seckl, S. Dymov, M. Szyf & M. J. Meaney. 2004. Epigenetic programming by maternal behavior. *Nat. Neurosci.* 7:847–854.

Wickham, H. 2010. ggplot2: elegant graphics for data analysis (use R!), 4th edition. 2009. Corr. 3rd printing 2010 edition. Gewerbestrasse, Switzerland: Springer.

Wilson, G., D. A. Aruliah, C. T. Brown, N. P. Chen Hong, M. Davis, R. T. Guy, S. H. D. Haddock, K. D. Huff, I. M. Mitchell, M. D. Plumbley, B. Waugh, E. P. White & P. Wilson. 2014. Best practices for scientific computing. *PLoS Biol.* 12:e1001745.

Xi, Y. & W. Li. 2009. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* 10:232.

Zhang, G., X. Fang, X. Guo, L. Li, R. Luo, F. Xu, P. Yang, L. Zhang, X. Wang, H. Qi, Z. Xiong, H. Que, Y. Xie, P. W. Holland, J. Paps, Y. Zhu, F. Wu, Y. Chen, J. Wang, C. Peng, J. Meng, L. Yang, J. Liu, B. Wen, N. Zhang, Z. Huang, Q. Zhu, Y. Feng, A. Mount, D. Hedgecock, Z. Xu, Y. Liu, T. Domazet-Lošo, Y. Du, X. Sun, S. Zhang, B. Liu, P. Cheng, X. Jiang, J. Li, D. Fan, W. Wang, W. Fu, T. Wang, B. Wang, J. Zhang, Z. Peng, Y. Li, N. Li, J. Wang, M. Chen, Y. He, F. Tan, X. Song, Q. Zheng, R. Huang, H. Yang, X. Du, L. Chen, M. Yang, P. M. Gaffney, S. Wang, L. Luo, Z. She, Y. Ming, W. Huang, S. Zhang, B. Huang, Y. Zhang, T. Qu, P. Ni, G. Miao, J. Wang, Q. Wang, C. E. Steinberg, H. Wang, N. Li, L. Qian, G. Zhang, Y. Li, H. Yang, X. Liu, J. Wang, Y. Yin & J. Wang. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* 490:49–54.