BONUS CHAPTER

Molecular Tools for Marine Biology

Introduction and Definitions

- **Molecular techniques use DNA, RNA, and related molecules to understand a wide range of biological processes.**

DNA is the fundamental molecule that determines the structure of life, and in recent years DNA technology has become a necessity in understanding natural systems including the biology of individuals, marine biological species, communities, and ecosystems. It is the purpose of this chapter to introduce you to the basics required to understand much of the new revolution that is sweeping through marine biology through the use of **molecular techniques** (DNA-based or DNA-related). We will discuss how molecular data is collected (e.g., through DNA sequencing) and how related observations, such as the expression of genes, can be important in the understanding of biological processes and performance in marine systems. We will also show how DNA-based data can be used to assess biological diversity.

Here are some important terms that come up in molecular studies. **DNA** is the carrier of genetic information from one generation to the next. DNA is usually packaged in chromosomes found in the cell nucleus, which replicate during ordinary mitosis, or cell division, or during meiosis, in which gametes are produced to form the next generation. **DNA is a double-stranded molecule** held together by weak hydrogen bonds. The backbone of DNA is a repeated pattern of a sugar group, deoxyribose, alternating with a phosphate group. Each strand has a sequence of **nucleotides**, which serve as the monomers, which, when assembled are the crucial determinants of genetic identity on the DNA strand.

There are four nucleotides: adenine (A), guanine (G), cytosine (C), and thymine (T). Each strand of the double helix has a sequence of nucleotides, but binding between strands occurs only between A and T, or G and C. If you have a sequence of nucleotides on one of the two DNA strands, the nucleotide sequence on the other strand is therefore a sequence of the complementary nucleotides that can bind with those in the first strand. Eventually, each amino acid in a protein will be determined by three sequential nucleotides, or a **triplet**. A table relating the specific triplets to specific amino acids is the **genetic code**.

This is an oversimplification, but a **gene** is a stretch of nucleotides that eventually codes for a protein—a string of amino acids—synthesized through a series of steps. The DNA sequence of the entire gene is the template for **transcription** to produce a complementary strand of RNA, which is the eventual template for the process of **translation**, in which the synthesis of protein produces a defined polypeptide sequence of amino acids. The strand of amino acids is then folded into a three-dimensional structure that is crucial in the protein's function, for example, in catalyzing a cellular reaction, such as an enzyme.

I repeat: this is an oversimplification. For example, many genes (strings of nucleotides) are interrupted by noncoding sequences, usually called introns. For proper transcription to occur, the sequencing part of the DNA, known as exons, have to be read and connected to construct the entire gene sequence that will be transcribed to produce the complete transcribed sequence of RNA. The process of RNA splicing allows the intron sequences to not be read as part of the

Courtesy of Wikimedia Commons, user Lilly_M.

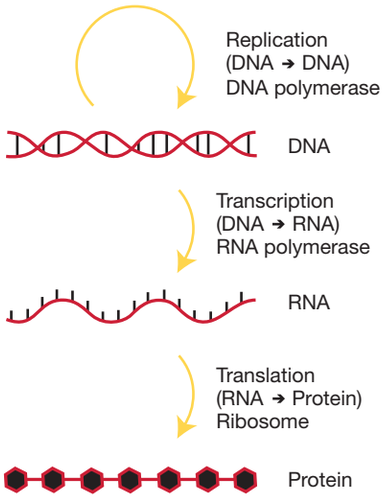

FIG. 1 The connection between DNA, RNA, and the final synthesized protein.

mRNA. An enzyme known as RNA polymerase II helps catalyze the transcription process to mRNA.

- **"Omics" is a key to understanding the use of DNA and related techniques.**

You may have already heard terms such as genomics, metagenomics, and proteomics. These words conjure up a family of "omics" types of molecular analyses that comprise the new approach to studying biological function and molecular diversity from DNA data. This overall structure is built around a basic relationship in biology, often termed "the central dogma":

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{Protein}$$

These three keystones (**Figure 1**) of the central dogma correspond to a series of "omics" approaches:

$$\text{Genomics} \rightarrow \text{Transcriptomics} \rightarrow \text{Proteomics}$$

These three approaches give us basic information on gene sequence and structure (**genomics**); expression of genes in cells (**transcriptomics**); and the abundance, diversity, and action of proteins in a cell (**proteomics**).

Genomics

- **Genomics is the field that uses DNA sequencing and statistical techniques to convert raw data on nucleotide sequences into genes, the assembly of genes, and associated sequences that regulate the expression of genes.**

Genomics is the field that considers the construction of nucleotide sequences into genes, assembly of genes, and associated sequences that help regulate the expression of genes. Genomics research includes studies producing and interpreting nucleotide sequences ranging from stretches of DNA responsible for all or part of a single gene to the whole **genome**, which is the entire DNA sequence of an organism. It is not our purpose here to go into all of the details but to give you an outline of what sorts of data are collected and for what purpose.

In recent years, methods of sequencing large parts of the entire genome have been developed, but much of our database still includes sequences of single genes or parts of them. These sequences may be derived from **mitochondrial genes** or **nuclear genes**. Mitochondria are relatively easy to isolate and have a small number of genes in their genomes. We have a large database of mitochondrial gene sequences. One mitochondrial gene, cytochrome coxidase 1 (CO1) was designated as the **barcode gene**, and has been used widely as an identifier for animal and plant species. However, most studies constructing evolutionary trees (see Chapter 4) use as many genes as possible, in order to maximize information from sequences with widely varying rates of evolutionary sequence change. Mitochondrial genes are maternally inherited, whereas nuclear genes come from both parents. While low copy number nuclear DNA is harder to isolate, it provides a large range of genes that often are unlinked and thus provide more historical information. Mitochondrial gene variants are all linked within the same mitochondrion, so you do not get as independent information by sequencing several mitochondrial genes. On the other hand, mitochondrial gene data can be very useful if you are tracing the migratory patterns of females, as in nesting turtles homing to beaches.

- **The polymerase chain reaction is an important means of amplifying specific areas of DNA for accurate sequencing.**

A common objective is to produce a DNA sequence for a single gene or part of a gene. Before sequencing is performed, it is necessary to have enough DNA of the required sequence to do the analysis. The **polymerase chain reaction (PCR)** was developed to amplify DNA of a specific sequence of interest. PCR is useful only when there is a good match between short nucleotide probes, known as **primers**, which will bind to known and invariant parts of the DNA of a gene of interest. After DNA is extracted from a biological sample, two short primer sequences bind to two opposing locations within a known DNA sequence that is assumed to be in the DNA sample.

In the reaction mixture, the sample DNA is first heated to denature and separate the DNA into its two strands (**Figure 2**). Then the annealing step involves the primers, binding to the complementary sequences of the sample as it cools. After that, with the aid of a thermostable DNA polymerase, the dissolved nucleotides in the mix rapidly bind to the remaining stretch of the sample DNA, extending the sequence and connecting the primers, producing a complete DNA product. Using a thermocycler, this process is repeated many times, amplifying large amounts of the sequences isolated by the primers. This DNA product can later be sequenced with an automatic DNA sequencer tailored to sequencing a single strand of DNA. The interested student should consult references on **Sanger sequencing**.

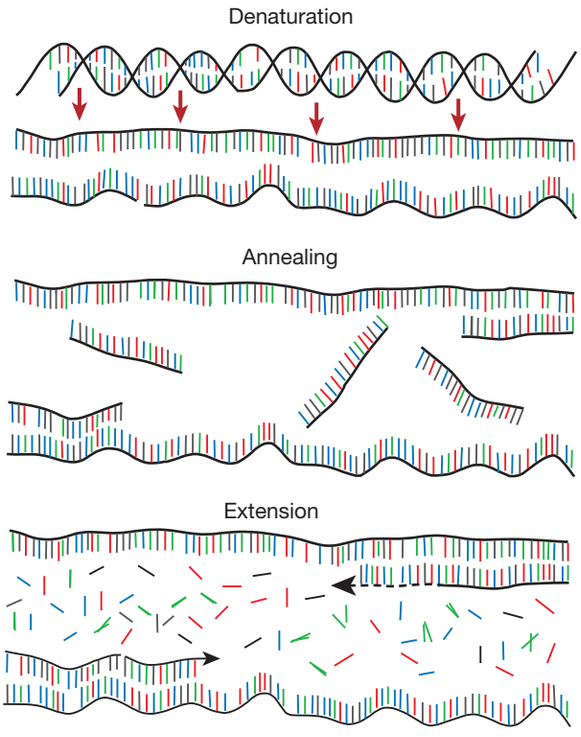

FIG. 2 Steps in the polymerase chain reaction. Heat is used to denature DNA into separate strands. Then primers in the annealing step bind to complementary parts of the sample DNA. Individual nucleotides bind to complementary sites with the aid of DNA polymerase, which helps to copy the DNA between the primers. cyclical temperature change causes this process to repeat and greatly amplify the production of DNA product.

Sequence variation is used to study natural selection, genetic variation within populations, and degrees of genetic differentiation between populations.

In studies designed to learn about genetic variation, researchers often look for spans of DNA sequence where the rate of evolution is high, so even closely related species will differ in sequence. Eventually a sample of sequences is obtained that can be used to study variation at the intraspecific and interspecific levels. Within-population sequence variation may be used to study the degree of genetic variation maintained within a population, or the degree of sequence difference along the geographic range of a species. On the interspecies level, sequence differences can be used to assemble a hierarchical tree that shows the evolutionary relationships among the species. We discuss the construction of an evolutionary tree from sequence data in Chapter 4.

An extremely useful application of sequencing is to identify DNA polymorphisms that are meaningful in relation to natural selection. Variation at a single site or a few sites might have functional significance that permits an understanding of phenotypic variation in the environment. For example, field mice are known to have different color coats that allow the mice to blend with their background environment and elude visual predators. Hoektra et al. (2005) identified a melanocortin-1 receptor in beach mice that decreases receptor function. Variation at one nucleotide site was associated with color differences, but it was also found that the path of natural selection on molecular polymorphisms differed at different sites to produce the same result of light-colored mice on light sand backgrounds and dark-colored mice on dark backgrounds at each location.

Next-generation sequencing includes a variety of technologies to produce very large DNA sequence data sets, which can be used to study genetic variation on a large scale.

In recent years, sequencing has been automated to allow processing of large portions of the genome. A series of methods, commonly known as **next-generation sequencing**, can produce, when assembled, very long DNA sequences. Next-generation sequencing is a catch-all phrase to describe a number of distinct techniques. They all share the high throughput methods needed to capture complete genomes. However, there is a level of error in the sequencing process so repetition, careful quality control, and appropriate statistical analysis are all important in getting accurate sequences. A major objective of this level of sequencing and data processing is to assemble whole **genomes** from a large number of segments acquired by the sequencing process.

Once a sequence is produced with confidence, it can be aligned with DNA sequences taken from other individuals of the same species. With a completely aligned set of sequences, one might discover some sites that have the same nucleotide location, or **site**, among DNA samples from many individuals. But at other aligned sites, there might be differences in the identity of the nucleotide at that particular site in the DNA. These **single nucleotide polymorphisms (SNPs)**, give a much wider estimate of genetic variation than was formerly possible. It is now possible to estimate genetic variability from a much larger portion of the genome than was possible with single DNA strand sequencing using PCR and Sanger sequencing usually of single genes. But variation derived from PCR studies of single genes also produce SNPs, as mentioned above, and can effectively be used to study single genes, especially when the function of the proteins are known.

SNPs derived from next-gen sequencing can be used to study traditional problems of regional genetic differentiation, but the potential access to more data is a great boon over studies of single genes using more traditional PCR approaches. For example, Williams et al. (2010) used a next-gen sequencing method known as 454 pyrosequencing to study genomes from different populations of the killifish *Fundulus heteroclitus* along the east coast of the United States. They identified 96 SNPs and found a differentiation between northern and southern populations, with evidence of increasing isolation by distance. Using SNPs, they were also able to relate sites of *F. heteroclitus* to a Gulf of Mexico species *F. grandis*.

A study of genetic variation at two levels shows the variable applications of sequencing to genetic problems. Two species of slipper shell snails in the genus *Crepidula*, both living in the intertidal and shallow subtidal of the east

coast of North America, have very different means of dispersal: *Crepidula fornicata* has a planktonic larva, whereas *C. convexa* has direct release of crawling juveniles from the mother. One might expect that the dispersal of *C. fornicata* might produce more genetic similarity of populations over a stretch of coastline than *C. convexa*, since planktonic larvae of *C. fornicata* would travel far and produce genetically homogeneous populations over greater stretches of coastline than in the species with direct release of young near the mother. Dispersal transports genes and tends to counter local processes such as genetic drift or natural selection. Using a study of variation at six microsatellite genetic loci, it was possible to examine genetic variation that was neutral relative to changing environmental conditions, and it was also found that the planktonic *C. fornicata* had less regional genetic differentiation than the nondispersing *C. convexa* (Cahill and Viard, 2014).

On a larger next-generation sequencing level, *Crepidula fornicata* and *C. convexa* have been used to compare genetic variation in the center of the geographic range, as compared with the northern edges of the ranges of these two species. By using a next-gen sequencing approach, it was possible to compare nearly 2,000 single nucleotide polymorphic (SNP) sites of *C. fornicata* to over 300 SNP sites of *C. convexa*. In both species, there was more allelic richness and unique alleles in the center of the range than at the northern edge (Cahill and Levinton, 2016). Local population extinctions, natural selection, or other processes might have reduced variability at the northern range edge.

- **For next-generation sequencing to be useful, we need methods to process raw data, translate sequences into known genes, and maximize accuracy. These steps comprise the field of bioinformatics.**

Next-generation sequencing can produce sequences of hundreds of thousands of nucleotides, which must be rendered into recognizable sequences that can be interpreted as a series of discrete genes. This must be done despite the fact that there is inherent error in recognizing and recording individual nucleotides, nucleotide sequences, and SNPs along with error in connecting large sequences that will result in a complete and accurate genome. The sum total of approaches, analytical and statistical, taken to maximize efficiency and minimize error and to finally identify sequences of genes is known as **bioinformatics**.

In order to maximize accuracy of sequence determination, we must increase the number of **reads**, or sequencing rounds. The number of nucleotides that are read, relative to the total number of nucleotides in a genome, is the estimate of the read **coverage** of a genome. Coverage is an estimate of redundancy, which will give better accuracy of a sequence, since there will be fewer gaps. **Sequencing depth** is a related estimate of how many times a given nucleotide has been sequenced, and higher numbers (e.g., 10×, 20×) predict greater probability of accuracy. Error can also be reduced by having multiple **contigs**, stretches of DNA sequenced that overlap and therefore give multiple estimates of specific DNA sequences that can be reliably connected. After reliable sequences are produced, a computational process of **gene annotation** begins in which specific stretches of DNA are identified as specific functioning genes. This process is much more effective when we have previous in-depth knowledge of gene location and function, as we do in a number of so-called model species such as mice, fruitflies, and the bacterium *Escherichia coli*.

Biodiversity Through Metagenomics

- **Metagenomics is the use of DNA technology to assay the range of biological diversity in a natural sample. There are many molecular probe techniques, and their usefulness varies depending on the research problem.**

One of the most important applications of genomics is to assay the **diversity of species in natural communities**. For example, we might want to know the number and identities of species of bacteria in a seawater plankton sample. We might also want to know the prey eaten by a predator and might want to identify the range of prey by examining DNA sequences that are preserved in the predator's feces. DNA assays might be the most efficient way of finding species that cannot be identified from morphology alone. While it might be possible to sequence all of the DNA in a sample and eventually identify all species, it is more efficient to use known **reference sequences** of species to develop assays to test if any or all known species are in a given location.

In order to sample for DNA sequences of microbial organisms in a sediment or in the water column, one must have a fast and appropriate technique. **DNA probes** are most useful. Diagnostic sequences for specific species or closely related species can be developed from laboratory-cultured representative species. Small segments of DNA, known as oligonucleotides, are then used as **probes** to bind to specific sequences of extracted DNA in from a natural sample. Sometimes successive dilutions are required to isolate single cells. These probes are then hybridized to the natural sample DNA, and they bind selectively to sequences of complementary DNA. Probes are now widely available for many species of marine microorganisms. Some probes are very specific, while others may bind to a number of related species, which can then be individually sequenced to get an idea of the number of species in a water body.

PCR can be used to extract a specific gene for sequencing. This is especially useful when there is a database of sequences, as for the gene 16S ribosomal RNA, which can be used to identify the range of microbial organisms in a sample of plankton. But less specific approaches can also be used. For example, **amplified fragment length polymorphisms (AFLPs)** allow the identification of a specific microbial group, but without having any prior sequence data. A series of restriction endonucleases are used to cut the DNA into a series of fragments at thousands of sites across the genome, and PCR is used to amplify the fragments. Then it is possible to use the set of these fragments to identify a species by the presence and absence of the

fragments. For more on these methods, the interested student should consult Burton (2009).

- **Microorganismal diversity of bacterioplankton and eukaryotic phytoplankton have been successfully assayed in surveys of the plankton.**

Many microbial species consist of cells in the picoplankton range, often less than 1 μm in diameter. In the water column we find a variety of species of bacteria of similar size, including normal saprophytic forms, cyanobacteria, and a wide range of nitrogen-processing bacteria. Along with these prokaryotic species, we find a large number of single-celled eukaryotes. It is not possible to identify all of these forms just on the basis of morphology obtained by means of microscopy. We discuss some applications of this technique in Chapter 8.

- **Next-gen sequencing techniques are now widely used to assay for oceanic microbial diversity.**

An early attempt at producing a complete censusing of oceanic microorganisms was initiated by Craig Venter, who funded a worldwide expedition on the sailing vessel *Scorpion*, designed to sample seawater and use modern DNA technology to sequence all of the microorganismal DNA from seawater samples. Many of these organisms are completely unknown, so conventional probes were not used. Instead, Venter's group employed a method known as shotgun sequencing, which breaks up DNA into many shorter stretches without any specificity, using Sanger sequencing methods, and then using sophisticated DNA sequence matching and alignment programs to link sequences together into genes and even whole genomes (Rusch et al., 2007). DNA was divided into small fragments, which were cloned and sequenced. These so-called reads were then aggregated into longer stretches, called contigs, which were connected by overlapping identical sequences. A statistical continuation of this process produced longer and longer sequences. Many of the sequences could be identified by statistical comparisons with the nearly 600 genomes already sequenced. Other techniques involve analyzing the DNA sequences to characterize the proteins that might be encoded by some of the DNA sequences.

In recent years, the development of next-gen sequencing platforms has greatly improved the sequence accumulation step, and a variety of statistical matching techniques can be used to search for those known genes that identify microbial species. This approach was applied to an intensive study of deep-sea water masses in the Atlantic Ocean with a focus on the V6 region of the 16S rDNA gene, which is universal in organisms. Bacterial communities were dominated by a small number of species, but there was also a very large number of rare species at very low abundance, which has come to be termed the "rare biosphere" (Sogin et al., 2006). Such dominance and a stretch of rare taxa have been found in many other systems, with strong evidence of seasonal variation in dominance. Next-gen sequencing in metagenomics is also being applied to studies of water quality and many other approaches, as next-gen platforms become less and less expensive (Staley and Sadowsky, 2016).

- **Prokaryotic assemblages of microorganisms can be assayed for biodiversity by ARISA, a powerful PCR-based technique.**

As we have discovered, microorganismal diversity is very difficult to survey, because of the difficulty of recovering and counting organisms that are less than a micron in size and that are often impossible to identify with morphological features alone under a microscope. Yet it is crucial to understand how microorganisms process materials in ecosystems, and this can only be done with an understanding of which microorganisms are present and in which proportions. For example, here is an interesting problem that would benefit from an enumeration of microbial diversity: Is organic matter in bottom sediments decomposed more rapidly when a high diversity of microbial organisms is present, as opposed to a very low diversity? There is a wide range of possible metabolic mechanisms to decompose organic matter (see Chapter 15). Sediments include bacteria that break down organic materials in the presence of oxygen (aerobic bacteria), reduce sulfate to gain energy (sulfate-reducing bacteria), break down organic material in the absence of oxygen to produce alcohols, and many other transformations.

Bacteria and Archaea both have two highly conserved genes: 16SrRNA and 23SrRNA. The **automated ribosomal intergenic spacer analysis (ARISA**) employs PCR to extract a highly sequence-variable internal transcribed spacer (ITS) sequence, found *between* the two genes. This ITS can be used to identify different bacterial groups. The primers used in the analysis can be labeled with fluorescent dyes, which makes it possible to visualize on a gel the DNA fragments attaching to the primers, producing a plot called an electropherogram. One problem is that a given ITS may produce more than one peak on an electropherogram, so this method can be used only in a very general way to assess microbial diversity.

Figure 3 shows pie diagrams that give an idea of the prokaryotic microbial diversity from sediments in three lagoons in Italy. As can be seen, the Goro lagoon has by far the greatest microbial diversity for the three sites, whereas the lagoon of Venice has the lowest microbial diversity. It is of great interest that this ARISA measure of

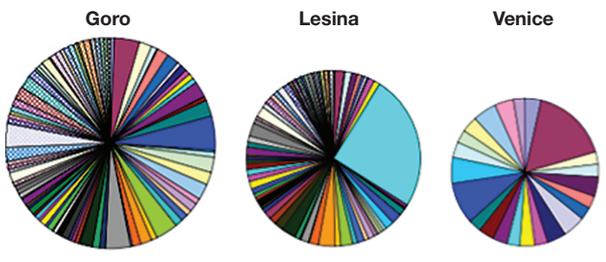

FIG. 3 Pie diagrams showing microbial diversity from three Italian soft sediment lagoons, based on an ARISA analysis. Diversity of bacterial species is strongly correlated with the rate of breakdown of organic matter in the sediment and correlated positively with the abundance of meiofaunal organisms. (From Danovaro and Pusceddu, 2007)

microbial diversity is closely associated with measures of the efficiency of breakdown of organic matter in sediments, probably because there are more ways that a given organic substance can be degraded when more microbial types are present (Danovaro and Pusceddu, 2007).

Transcriptomics and Studies of Gene Expression

- **Transcriptomics is the study of expression of genes by assaying RNA transcripts.**

As discussed in Chapter 5, it is very important to understand the physiological state of organisms. For example, as the summer progresses, high temperatures in an estuary may lead to high oxygen consumption rates by microorganisms and low wind, resulting in stagnant waters of low dissolved oxygen. How much stress are fish or crabs suffering as hypoxia sets in? Can we provide a measure of stress that directly assays the response at the cellular level? We could measure such responses if we understood the physiological responses to lowered oxygen in the water and the biological processes used in response to this environmental change. As the change occurs, many genes involved in physiological activities will change their degree of **gene expression**, which is the sum of processes that led to genes eventually producing proteins, such as enzymes. The process of gene response results in **gene transcription**, where the DNA sequence is copied into **messenger RNA (mRNA)** with the aid of RNA polymerase. This is the first step of **gene expression**, the sum total of processes that leads to the production of proteins, which function in various cellular processes, such an enzymatic acceleration of various cell reactions. The **transcriptome** is the total sum of expressed mRNAs in an organism at a given moment.

An important application of transcriptomics is to understand how an organism responds to stressful environmental changes. For example, in fish, the larval stage is well known as a time when physiological stress results in developmental abnormalities, slowdowns in growth and death. It would be of great value to know what biological functions are being disrupted the most. To understand this, it would be very useful to pinpoint physiological changes by identifying times when there are disruptions of cellular function, which might be registered by a strong change in the expression of genes important in certain metabolic functions. So, if environmental dissolved oxygen declines, we might want to assay the genes responsible for respiration in cells.

A first important step in studying gene expression is to have complete DNA sequences, for whole genomes, of some model species. One can use these models to find sequences for specific genes believed to be involved in a metabolic process. So, for example, in fish we have complete genome sequences for zebra fish *Danio rerio*, medaka *Oryzias latipes*, and the fugu *Fugu rubripes*. To assay for gene expression, we need a library of gene-specific sequences of messenger RNA transcripts, which are each synthesized as single-stranded **cDNA**. From these cDNA sequences, libraries of fragments of each sequence are produced, known as **expressed sequence tags (ESTs)**. These tags can be used to assay for the presence of messenger RNAs in an organism or in a specific tissue and therefore give us a global snapshot of the degree of gene expression of a wide range of genes, as the whole organism is subjected to some environmental change, such as the reduction of dissolved oxygen. ESTs have been used to study gene expression in a number of species of fish, including Atlantic salmon *Salmo salar* and Atlantic halibut *Hippoglossus hippoglossus*, and from invertebrates such as the blue crab *Callinectes sapidus* (see Coblentz et al., 2006).

- **Microarray analysis uses an EST library to simultaneously estimate the amount of gene expression across many genes. Rapid analysis can occur through the use of premanufactured cassettes of probe sequences.**

Microarrays are sets of sequences in prepared probe cassettes that use the EST library to simultaneously study the level of **gene expression** in many genes or the whole genome to get an idea of how an organism responds to an experimental condition or environmental change at the gene level. Microarrays are constructed of glass, silicon chips, or nylon membranes. Thousands of reference genes can be arrayed on a single silicon chip as spots. A sample of extracted messenger RNA, converted to cDNA, can be placed on this reference chip, and sequence-specific hybridization can be estimated by color-specific fluorochromes that indicate the degree of hybridization, which is proportional to the level of expression for each gene in the sample we have applied to the chip. Such approaches are rapidly being applied to the study of gene expression as a response to environmental stress (see Chapter 22). Large EST libraries have been developed for the oysters *Crassostrea virginica* and *C. gigas* (e.g., Jenny et al., 2007) and for the porcelain crab *Petrolisthes cinctipes* (Garland et al., 2015), which has become an important marine model for the study of temperature stress.

Suppose a phytoplankton culture of interest is exposed to increased dissolved nitrate. Which genes are responding to the change? If the genome of this species has been previously sequenced and annotated, it is possible to synthesize a large set of probe sequences complementary to the expressed DNA. After subjecting the culture to higher dissolved nitrate (an important nutrient for phytoplankton), the RNA in the culture is extracted and after synthesized back to DNA is exposed to a **microarray** of gene-specific sequences. The sequences that bind indicate which genes were expressed and their levels of expression. These results can be compared to the spectrum of expression of genes when a phytoplankton culture is not exposed to nitrate, which acts as an experimental control.

- **RNA-seq analysis allows the study of mRNAs without needing previous sets of cDNAs for hybridization as in microarray analysis.**

Microarrays have the great advantage of matching extracted products, mRNA, of expressed genes and hybridizing them to previously prepared cassettes of known gene sequences.

But in recent years, next-generation sequence methods have merged with the study of expressed mRNAs by using the new large-scale sequencing techniques to directly study mRNA levels. The mRNAs of an organism are extracted and then sequenced. This method does not require *a priori* information about gene sequences. Using statistical matching techniques, it is possible to identify the expressed mRNAs as products of specific genes and to see whether classes of genes are expressed more (or upregulated) or expressed less (or downregulated). This gives us an idea of what part of the genome is responding to environmental change.

Figure 4 shows a work chart for extracting RNA and using next-gen sequencing methods to produce samples that permit the study of levels of gene expression. RNA is extracted from tissue of an organism (e.g., the hemolymph of a crustacean) that has been exposed to a defined set of environmental conditions and purified and subjected to next-gen sequencing. The sequencing step produces large RNA sequences, which must be statistically overlapped and rendered into a series of contigs. The contigs are analyzed to produce stable sequences, whose exact genetic function are not known, and are termed unigenes. Then software from the BLAST family is used to annotate the sequences, in order to establish the identity of the genes that have been sequenced. This finally allows an analysis of crucial information such as the relative concentration of RNAs that have been produced, reflecting the degree of gene expression of a wide variety of genes. The expression data can be related to the function of the genes that have been identified.

- **Applications of transcriptomics include studies of nutrient transformations of phytoplankton and responses of animals to exposure to stress.**

Transcriptomic approaches can be used to study a wide array of processes and responses. For example, we discuss in Chapter 11 the many organisms that convert nitrogen in natural communities. Some bacteria gain energy by fixing nitrogen gas, eventually converting nitrogenous compounds to proteins. Other species of bacteria act on nitrogen-containing compounds that occur in decomposing organic matter, which, through a sequence of steps, results in the conversion of nitrogen eventually to the release of nitrogen gas to the water column and the atmosphere. Microarray techniques have been used effectively to study the identity of genes that function in nitrogen transformations. Water samples can be taken in an estuary, and the expression of different functional genes involved in a process such as denitrification can be studied as a function of location in the estuary, salinity, temperature, and other variables (Taroncher-Oldenburg et al., 2003).

Microarray-based approaches are also being used widely to study the response of marine animals to common pollutants and sources of human-induced stress such as hypoxia

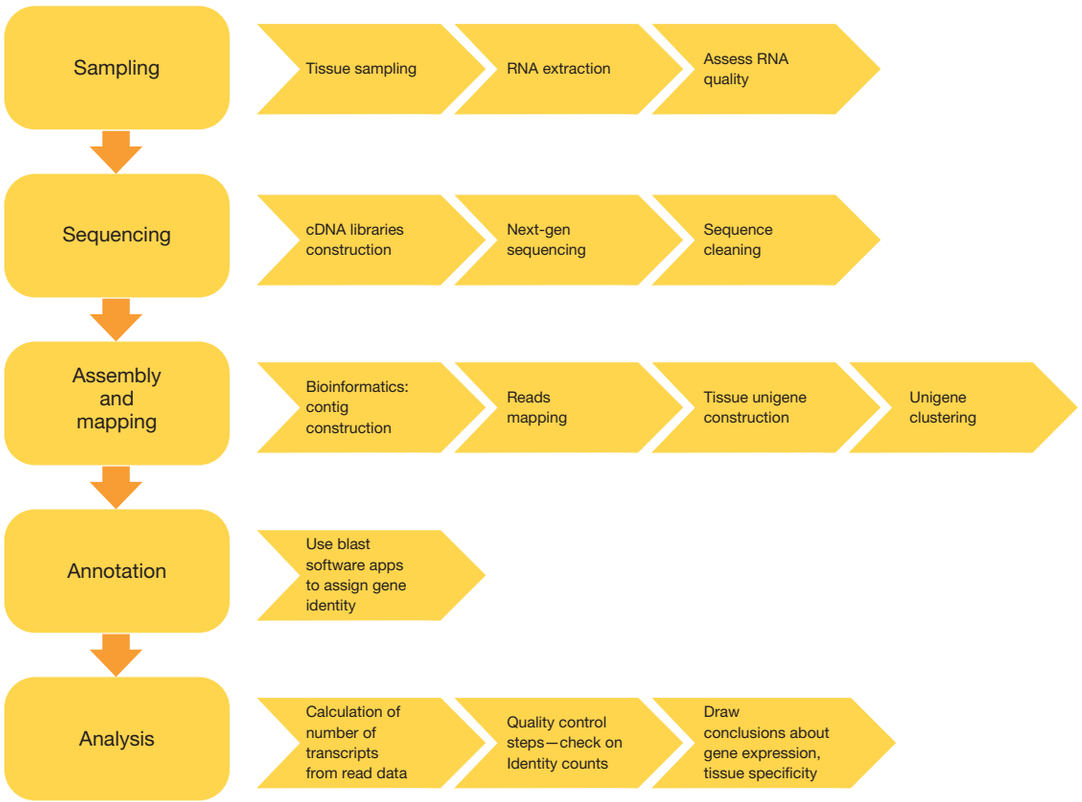

FIG. 4 Scheme of cDNA extraction, next-gen sequencing and data analysis to study gene expression for a large number of genes in an organism subjected to a defined set of environmental conditions. (After Moreira et al., 2015)

and thermal stress. These studies work on the premise that environmental stresses result in responses of gene expression, which increase the probability of survival through changes in physiological condition. For example, the delta smelt is an endangered fish species restricted to the Sacramento-San Joaquin estuary in California. It occurs in a watershed dominated by agriculture, which includes inputs of insecticides into the estuary. Insecticides may be useful in agriculture, but when they wash into coastal water, they have strong negative effects on many marine organisms. A microarray study was performed on the effect of the insecticide esfenvalerate on gene expression of a wide variety of functional parts of the genome. **Figure 5** shows the range of gene responses to exposure at very low concentrations. Responses were seen in over 100 genes in a wide array of biological functions, such as growth, neuromuscular function, and immune responses. The prominence of neuromuscular responses was especially interesting and could be related to swimming impairments that the fish suffered when being exposed to the insecticide in toxicity experiments (Connon et al., 2009). In many cases, evolutionary change has occurred where populations exposed to toxic pollutants have evolved resistance to the pollutant (see Chapter 22).

An important aspect of the study of gene expression is **tissue specificity**. Because different tissues and organs have widely different functions, we can test hypotheses that relate gene function to organismal function and even exposure to the environment. For example, Moreira et al. (2015) used next-gen sequencing techniques to study gene expression differences in different tissues of the mussel *Mytilus galloprovincialis*. Some groups of tissues, for example, mantle and muscle, were very similar in transcription pattern. But hemocytes had especially high transcript number for defense and immune-related proteins, which fits their function as defense tissue. Gill tissue had especially high concentration of transcripts relating to non-self-recognition genes, which fits the exposure of gill to ions that rapidly exchange with the external environment. Mantle is responsible for shell deposition and gonad formation, and transcripts relating to genes involved in calcium metabolism and reproduction were abundant, as might be expected.

- **Data from transcriptomics studies can be connected to population genetic studies to allow for focused studies on natural selection, for example, in the context of climate change.**

Using the diagram in Figure 4, we can develop a series of genes whose expressed proteins can be identified and related to physiological processes. We might find, for example, that exposure to ocean acidification results in changes of expression of genes that are associated with biomineralization. But this type of study can also be connected to regional studies of population genetics and natural selection. Suppose you are working in a region where regional ocean acidification is greater than average and this combines with lowering of dissolved oxygen. It would stand to reason that there might be strong natural selection on variation associated with genes whose gene products are connected to biomineralization

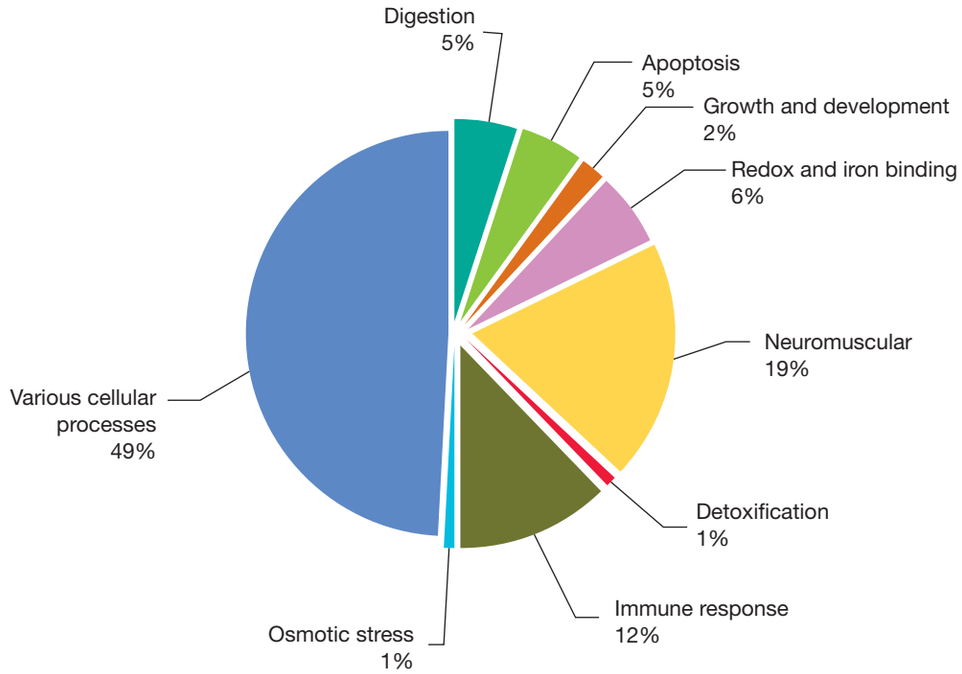

FIG. 5 Different classes of 118 identified gene-derived mRNAs of delta smelt in California that responded to exposure to the insecticide esfenvalerate.

and aerobic metabolism. The sequences that are collected in an RNA-seq analysis can be analyzed for single nucleotide polymorphisms (SNPs), as discussed earlier in this chapter. Then we can propose a hypothesis*: In a region with newly arisen acidification and lowered oxygen, there might be strong local natural selection for new alleles at genetic loci whose function is associated with the physiological parameters of biomineralization and aerobic metabolism.* This would lead to another hypothesis: In a local region with such a changed environment, *there would be stronger local variation in the frequency of changes of SNPs over geographic space*. As a control, we would predict that SNPs associated with genetic loci connected to proteins with very different function would show no geographic variation, since, for those loci, the environment has not changed. See Chapter 7 for a preliminary study in which this analysis has been done on regional variation in the red abalone (De Wit and Palumbi, 2012).

Proteomics

- **Proteomics is the analysis of the entire protein complement in a cell or other biological structure, but it may be combined with studies of transcription.**

Proteomics is the final logical extension of studies of the range of responses from the gene to the protein. It involves the identification, quantification, and amino acid sequencing of all of the proteins in a cell or a tissue. Why is this degree of information needed when we might already have information on gene expression via transcriptomics? First, these types of data provide more accurate assessments of the actual functioning of biological systems, since we quantify the abundances and structure of the final active biological molecules—the proteins—that are functioning and interacting within the cell. These types of data cannot always be directly predicted from the suite of mRNAs because there are a number of processes that may result in the final protein structure and abundance that is not predicted from RNA expression. Messenger RNAs function by being translated by the ribosomes into the polypeptides that constitute proteins. But in the process of **post-translational modification**, proteins may be altered in sequence and three-dimensional structure when forming the final functioning protein, which could not necessarily be predicted from the mRNA alone. As a result, proteomic studies might combine with transcriptomic studies to completely understand the process of protein production, from the gene to the final functioning protein. As a consequence the diversity of proteins can be much higher than genes, or mRNAs (**Figure 6**).

Proteomics also allows identification of molecules that are important in understanding the stability of proteins in the cell under stress. For example, cells often include a variety of **molecular chaperones**, or proteins that enable the proper folding of a newly synthesized protein or help in correcting proteins that have become unfolded, perhaps due to stressed cellular conditions. Proteomic studies allow the study of proteins and molecular chaperones to understand how stress might result in the increase of molecular chaperones to stabilize proteins that are becoming unfolded.

- **Proteomics investigations examine the response to a changing environment of protein complements in cells or tissues.**

Proteomics can be used in a similar way to transcriptomics to study biological processes, especially in response to

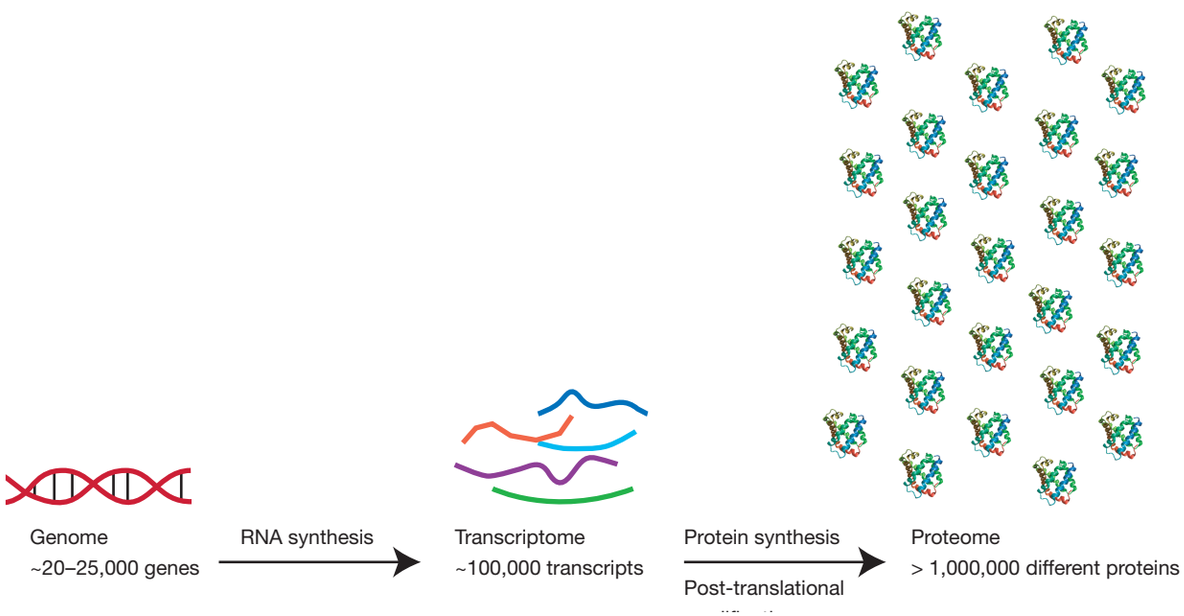

FIG. 6 The path of increasing diversity of proteins from original DNA sequences.

environmental change. For example, cellular concentration of proteins of marine organisms associated with aerobic metabolism suggest downregulation when the organism is subjected to high temperature stress, low dissolved oxygen, or low pH, but other proteins—for example, those associated with anaerobic metabolism—may be switched on under the same stress, suggesting an overall switch of metabolic style under stress (Tomanek, 2014). A higher-latitude mussel, *Mytilus trossulus*, on the west coast of the United States is more sensitive to thermal stress than a lower-latitude species *Mytilus galloprovincialis*, which is adapted to higher temperatures. When stressed by heat shock, *M. trossulus* shows a much greater production of tubulin-affinity chaperone proteins than *M. galloprovincialis*. Tubulins are important in ciliary function and mucus production—both crucial components of bivalve ciliary feeding—which suggests why tubulin chaperones increase in abundance when the more temperature-sensitive mussel *M. trossulus* is heat shocked (Tomanek, 2014). Such studies at the protein level give a complete picture of response and help explain why differently-adapted (e.g., high-versus low-temperature) species respond differently to the same stress.

CHAPTER SUMMARY

- Molecular techniques are those that employ studies of DNA or RNA to understand how an organism functions with regard to genetic structure, gene expression, and protein production.
- Genomics is the study of organisms by assaying their DNA sequences and patterns of gene number and arrangement, often at very large scales, from multiple genes to the entire size of the genome itself.
- Single genes can be efficiently sequenced if at least two short sequences, known as primers, can be combined with a thermal cycling method known as the polymerase chain reaction. The primers must be universal enough to hybridize with the DNA of the same gene type in a new organism.
- Larger-scale sequencing techniques, known as next-generation sequencing, allows massive amounts of sequence to be obtained. A series of multiple sequencing repeats allows increased coverage of the large stretches of DNA, which minimizes error of sequencing. Bioinformatics includes all of the statistical and mathematical methods used to take sequence raw data and translate it into finished sets of thoroughly vetted sequences of genes. Variation at specific sites within a population, as evidenced by single nucleotide polymorphisms (SNPs), allows us to study population changes owing, for example, to natural selection.
- Metagenomics is the use of DNA technology to assay the range of biological diversity in a natural sample. This method is especially useful in the study of microbial diversity, in which it is difficult to identify or to tell species apart from visual observation alone.
- Transcriptomics is the study of expression of genes by assaying RNA transcripts. An important application of transcriptomics would be to understand how an organism responds to stressful environmental changes.
- Proteomics approaches directly assay the range of proteins in a cell or tissue. These give a more direct estimate of functioning reactive molecules in a cell or tissue.

REVIEW QUESTIONS

1. If a gene is sequenced, what kind of information can we get?
2. Why is PCR so important in the steps leading to determining a gene's nucleotide sequence?
3. Describe a major problem in producing a large set of gene sequences from a next-generation sequencing approach.
4. What is the advantage of microarrays in the study of gene expression?
5. How could we use a transcriptomics approach to study the function of an organism in a changing environment? Give a hypothetical example.
6. Is there an advantage of using RNA-seq methods over previous transcriptomics methods? What would the advantage be?
7. What general experimental-environmental information about a marine organism would you want to learn from a transcriptomics project?
8. What is the difference in obtaining proteomic data, relative to transcriptomics data, when individuals of a species are subjected to some type of environmental stress?
9. What is the utility of a metagenomic method to assay the degree of microbial diversity in a planktonic environment? In an organic rich sediment?