Reference: *Biol. Bull.* **227:** 91–92. (October 2014)

Discovering Diversity with High-Throughput Approaches: Introduction to a Virtual Symposium in *The Biological Bulletin*

KENNETH M. HALANYCH[1] AND ANDREW R. MAHON[2]

[1]*Department of Biological Sciences, Auburn University, 101 Rouse Life Sciences, Auburn, Alabama 368490, ken@auburn.edu; and* [2]*Department of Biology, Institute for Great Lakes Research, Central Michigan University, Mount Pleasant, Michigan 48859, mahon2a@cmich.edu*

Amidst the current mass extinction caused by anthropogenic activity, we have new tools that can help us explore and understand biological diversity to a degree that was hard to imagine just 10 years ago. High-throughput genomic tools have enabled researchers to explore diversity within genomes and transcriptomes as well as within populations and between species. These technologies have perhaps been best exploited to discover bacterial and archeal diversity (*e.g.,* Simon and Rolf, 2011. *Appl. Environ. Microbiol.* 77: 1153–1161), but also play significant roles in understanding diversity of more complex organisms by providing robust phylogenetic hypotheses (*e.g.,* Kocot *et al.,* 2011. *Nature* 477: 452–456; Simon *et al.,* 2012. *Genome Biol. Evol.* 4: 1295–1309). High-throughput sequencing (HTS) is also at the forefront of screening samples for presence of rare, invasive, and threatened or endangered species (*e.g.,* Mahon *et al.,* 2014. *Conserv. Genet. Resour.* 6: 563–567; Thomsen *et al.,* 2012. *Mol. Ecol*. 21: 2565–2573). Importantly, diversity can be examined at multiple levels. For example, these approaches are revolutionizing our ability to assess variation within populations (*e.g.,* Davey *et al.,* 2011. *Nature Rev. Genet.* 12: 499–510), and they are even helping to assess diversity within a single individual (*e.g.,* Keane *et al.,* 2011. *Nature* 477: 292–294).

Technological advances can drive changes in several facets of the research enterprise. Perhaps the most tangible change in biodiversity research facilitated by HTS is the scale of data and analyses that are being undertaken. Whereas Sanger sequencing was limited to several hundred to thousands of sequences in a given biodiversity study, current technologies allow for millions of sequencing reads from a single sample. Compared to traditional sequencing, HTS approaches have facilitated an increase in the scale of data by a few magnitudes given similar budgets. This allows researchers to explore diversity of a sample, whether microbial sediment or the genome of an individual organism, in much greater depth and coverage. In general, if we have more data, we can draw more robust conclusions.

This virtual symposium seeks to encapsulate and highlight some of the ways in which HTS has helped explore, elucidate, and explain biodiversity in non-model systems. The symposium opens with three reviews showing how HTS approaches have helped address, and change, our views of biological diversity. Foraminifera are important components of marine ecosystems that scientists broadly use as indicator species in present-day oceans and as proxies for past oceanographic conditions. Pawlowski and colleagues (this volume) offer insights as to how HTS approaches have changed our understanding for this important group. Additionally, they discuss some pitfalls that influenced early HTS studies and outline how to avoid them, allowing our interpretation of data to be refined. Bik (this volume) offers more insight as to how metagenomic studies with HTS may be used to explore biological systems. Arguably, metagenomic surveys represent the primary use of HTS in studies of marine systems. Case studies using such metagenomic approaches are presented by Xu *et al.* (this volume) and Brannock *et al.* (this volume). The former examines ciliate diversity in the notoriously inhospitable McMurdo Dry Valleys of Antarctica, and the latter explores meiofaunal community structure in intertidal regions impacted by the *Deepwater Horizon* oil spill. The third review (Hemmer-Hansen *et al.,* this volume) focuses on how HTS approaches have addressed population genetic questions in

fishes. Given the ever-increasing demand on and rapid depletion of global marine fisheries, being able to more accurately determine population genetic parameters of exploited populations should lead to better management.

Often, significant changes in technology lead to new and different methodological approaches, as well as to changes in how theory is applied. Restriction-site associated DNA (RAD) tag or genotype by sequencing (GBS) approaches are examples of HTS protocols that are becoming widely popular because of their potential power for looking at population genetic issues. In particular, as one looks at more and more loci, better understanding of gene flow, population structure, and hybridization can be attained. Importantly, more extensive coverage of loci can increase the power of genetic/genomic analyses, thereby reducing the number of individuals from a given population that need to be sampled. Schweyen *et al.* (this volume) discuss new innovations with the RAD tag approach and offer ways to reduce some potential artifacts in data collection. In a similar vein, Hart (this volume) argues that HTS transcriptomic studies produce sufficient data to explore the interface of theoretical population genetic models (*e.g.,* isolation-with-migration) and codon usage models. This combination could provide unprecedented insight into the very early stages of reproductive isolation between populations. Both of these latter contributions demonstrate that HTS approaches to evolutionary questions provide dynamic areas of investigation and that researchers are actively refining approaches.

The symposium concludes with two other case studies of biodiversity revealed by HTS approaches. Because of the large volume of data produced with HTS, some previously labor-intensive undertakings are now much easier to complete. DePriest *et al.* (this volume) report on the complete mitochondrial genome of the red alga *Grateloupia taiwanensis*. The approximately 29,000-bp genome was compared to those of other red algal groups, allowing glimpses into the evolution of this important marine taxon. By comparison, Halanych and Kocot (this volume) repurposed several transcriptomic datasets generated in other studies to more fully assess the diversity of an important gene family. The Toll-like receptor (TLR) genes are part of the innate immunity system of animals that protects them from bacterial infection. Whereas TLR genes are well known from Arthropoda, Chordata, and a few other select groups, they were not previously reported from several other animal phyla. Halanych and Kocot find homologs in several lophotrochozoan taxa. Such studies are becoming more commonplace in single-investigator laboratories rather than large genome facilities.

One area that this symposium does not cover, data management and analysis, is perhaps the most significant challenge in using HTS approaches. Amounts of data being produced are unfamiliarly enormous for many researchers, and as such, employing HTS approaches can be initially daunting. Routinely, academic curricula (which are controlled by faculty) are slow to adapt to new methods and practices, and traditional classes allowing researchers to retool to their specific needs are limited. HTS research is forcing biologists to become more computationally savvy. Fortunately, an amazing diversity of open-source tools is available on the internet to allow researchers to self-educate. Learning a set of basic skills for data management and assessment is not difficult, but it does take time. Those looking for an entry in basic bioinformatics for HTS could consider *Practical Computing for Biologists* (Haddock and Dunn, 2011, Sinauer Associates, Inc.) or *Computational Biology* (Wünschiers, 2004, Springer-Verlag). There are also a number of commercially available software packages for dealing with HTS data. Many of these are graphical user interface (GUI) driven and easy to use, but can promote a "black box" approach to data assessment if one is not careful. We heartily advocate that investigators roll their sleeves up and familiarize themselves with command-line interfaces to more thoroughly understand how the software is analyzing the large volumes of HTS data generated rather than taking a "data in-data out" approach. A little familiarity with a Linux command line can go a long way.

Technology is improving at a phenomenal pace. High-throughput sequencing is being applied in new and innovative ways in multiple fields of biological and environmental sciences. Companies are developing HTS platforms that are not only faster and cheaper than the "last generation" but are also smaller and in some cases portable (*e.g.,* the MinION device by Oxford Nanopore Technologies). Truly we are in an age when yesterday's science fiction gadgets are becoming today's reality. Researchers need to be cognizant of their use of HTS. Just because it's available doesn't mean it is the best tool for the job at hand. Familiar approaches such as Sanger sequencing, quantitative PCR, *etc.,* are still useful, and in some cases, are more appropriate than the "throw it on a HTS machine and see what we get" approach. We have heard more than one story about an investigator getting a large amount of HTS data and not being able to interpret it meaningfully. Such stories are reminiscent of the supercomputer, named "Deep Thought," in *The Hitchhikers Guide to the Galaxy* by Douglas Adams. The data output was 42, but no one knew the question. Defining questions and sequencing approaches ahead of time is, of course, the best way to guarantee meaningful and cost-efficient outcomes with high-throughput approaches.