Syst. Biol. 61(5):713–715, 2012

DOI:10.1093/sysbio/sys050

The Promises and Pitfalls of Next-Generation Sequencing Data in Phylogeography

BRYAN CARSTENS[1,*], ALAN R. LEMMON[2], AND EMILY MORIARTY LEMMON[3]

[1]Department of Biological Science, 202 Life Sciences Building, Louisiana State University, Baton Rouge, LA 70808, USA; [2]Department of Scientific Computing, Dirac Science Library, Florida State University, Tallahassee, FL 32306-4120, USA; and [3]Department of Biological Science, 213 Biomedical Research Facility, Florida State University, Talahassee, FL 32306-4120, USA;

[*]Correspondence to be sent to: Department of Biological Science, 202 Life Sciences Building, Louisiana State University, Baton Rouge, LA 70808, USA; E-mail: bryan.c.carstens@gmail.com

The growth of phylogeography would not have been possible without the great advances in DNA sequencing technology achieved through the 1970s and 1980s. DNA sequencing based on primer extension with chain-terminating dideoxynucleotides (Sanger et al. 1977) represented a dramatic increase in the efficiency of DNA sequencing compared with previous chemical approaches. When coupled with the polymerase chain reaction (Saiki et al. 1985; Mullis and Faloona 1987), it became possible to efficiently generate sequence data from specific genes. The expansion of phylogeography coincided with increased availability of commercial machines such as the ABI 370 in the 1980s, making it feasible to efficiently assay genetic variation near the species boundary. Phylogeographic research exploded in popularity and now serves as an important bridge between phylogenetics and population genetics, promoting conceptual advances that straddle the boundary between these disciplines (e.g., species trees and related approaches to species delimitation). However, the vast majority of phylogeographic investigations to date have been limited to a relatively small number of genes. As researchers have recognized the limits of such data (e.g., Edwards and Beerli 2000; Hudson and Turelli 2003), they have become increasingly frustrated with the labor and expense associated with gathering data on a locus-by-locus basis.

The primary limitation associated with Sanger sequencing is its reliance on the visualization of the distribution of fluorescent dyes at the terminal ends of products for base calling. This limitation restricts Sanger technology to a single template, and to gathering data one locus at a time. In 1996, pyrosequencing was introduced (Ronaghi et al. 1996); in this approach, sequencing occurs as the complementary bases are added to ssDNA, initiating a chain reaction that releases a small flash of light that corresponds to a particular dNTP. This advance (i.e., sequence-by-synthesis) removed the restriction of collecting data one locus at a time. Next-generation sequencing (NGS), regardless of platform, shares the characteristic that it is libraries of template, rather than a single template isolated via PCR, that are sequenced. Data are gathered by the thousands (or millions) of reads, dramatically decreasing the cost per base and increasing the amount of data that can be included in research projects. Accompanying this change are several challenges that the discipline must meet before the promises of new sequencing technologies can be realized.

Unlike Sanger-generated data, the data generated by NGS platforms requires substantial processing before analysis, for reasons that are largely related to the scale on which the data are collected. With sequencing capacities up to ~3.0×10^9 reads (e.g., Illumina's HiSeq platform), large numbers of loci and individual samples can be included on a single run. Post-run processing includes the de-indexing of individual samples, quality control, alignment, and calling of single-nucleotide polymorphisms. Although resources exist, both in web based (e.g., Galaxy, DNAnexus) and pipeline form (e.g., Stacks, Catchen et al. 2011 and PRGmatic, Hird et al. 2011), the analysis of most data sets still requires a certain amount of custom scripting to process. For further details regarding next-generation sequencing, please see the recent reviews on this subject (Stapley et al. 2010; Rice et al. 2011; Glenn 2011; McCormack et al. 2012a).

Last year (2011), at the annual meeting of the Society of Systematic Biologists at the University of Oklahoma, we participated in a symposium titled "The Promises and Pitfalls of Next-generation Sequencing Data in Phylogeography". The symposium represented our attempt to bring together researchers who are making the transition from Sanger- to NGS-generated data in phylogeographic and phylogenetic research. Papers based on several of these talks are included in this special issue of *Systematic Biology*. In addition, other researchers presented work: Jeffrey Good, titled "Leveraging methods of targeted enrichment and high-throughput sequencing for comparative population genomic studies"; Alice Dennis, Luke Dunning, Shelly Myers, and Thomas Buckley, titled "Cold tolerance in New Zealand alpine stick insects: transcriptome variation within and among species"; and Justen Whitall, titled "New perspectives on cryptic divergence and ecological pleiotropy: from pine plastomes to Arctic mustard transcriptomes".

Four papers are included as part of this special issue:

Faircloth et al. (2012) report the discovery that ultraconserved elements are a new class of nuclear DNA markers, useful for large-scale phylogenetic ("phylogenomic") inference. This discovery is important because these molecular markers are an abundant, easily-collected, and easily-aligned source of genetic information shared across expansive taxonomic groups. By enriching these loci using sequence capture techniques and analyzing the enriched DNA with massively parallel sequencing, researchers can use this technique to interrogate hundreds to thousands of identical loci from diverse taxa (e.g., Amniotes) and address biological questions at a variety of time scales.

Lemmon et al. (2012) used a novel hybrid enrichment approach (termed *Anchored Enrichment*) coupled with Illumina HiSeq sequencing to target >500 loci in 10 taxa across the vertebrates for a phylogenetic study. Capture probes were designed from highly conserved regions of genomes available for five of the taxa (model organisms), and were then tested in these and five nonmodel taxa. They found that this approach allows rapid generation of phylogenetic data even for highly divergent nonmodel species at a fraction (<1%) of the per-base cost of Sanger sequencing. In addition, they found these loci can provide phylogenetic resolution at a variety of time scales in vertebrates and can resolve species trees despite gene-tree discordance. The present study, as well as the studies by Faircloth et al. (2012) and McCormack et al. (2012b), represent an extremely exciting advance for the field of systematics, providing a rapid and cost-effective approach to generating orders of magnitude more data for phylogenetics and phylogeography of nonmodel organisms than has previously been feasible.

Lemmon and Lemmon (2012) developed a reduced-representation library (RRL) approach combined with paired-end Illumina HiSeq sequencing to develop highly variable loci for studying the phylogeography of nonmodel systems; they tested this method in the chorus frogs (genus *Pseudacris*). The authors created RRLs to identify within- and between-species markers and tested these loci for amplification in the species from which the loci were derived as well as all other taxa in the genus. With this approach, they were able to rapidly and economically generate thousands of variable, single-copy nuclear loci. A subset of these loci with levels of sequence variation appropriate for phylogeography can be identified and used efficiently by using high-throughput amplicon sequencing techniques.

Zellmer et al. (2012) prepared RRL and collected sequence data using Roche 454 sequencing to infer the historical demography of the carnivorous plant *Sarracenia alata*. Samples were collected from 10 populations, and several analyses suggest that populations are isolated by the large river systems that divide the landscape. The sequence data provided by the NGS platform allowed Zellmer et al. to date the temporal divergence among populations, and to come to the conclusion that most diversification within this species occurred during the Pleistocene.

In one form or another, participants in the symposium largely agreed that any list of challenges for the next generation of phylogeography would include those associated with the generation of data from nonmodel species. As sequencing costs decrease, and small scale platforms such as the IonTorrent, 454 Jr, and Illumina MiSeq become more accessible, several specific questions need to be considered. For example, how should phylogeographers identify loci with an appropriate level of information for a particular research question? One approach is phylogenetic profiling, where potentially-informative loci are identified on the basis of patterns of site variation (Townsend et al. 2007). Related to this challenge is the targeting of specific loci for sequencing. There are a variety of enrichment approaches available, including Agilent SureSelect and MYcroarray MYselect. In addition, as multiplex PCR becomes more viable, approaches such as amplicon resequencing are becoming increasingly practical. Nearly all participants agreed that recent advances in indexing large numbers of individuals were promising (e.g., Meyer and Kircher 2010; Kircher et al. 2011; Neiman et al. 2011), and expressed optimism that library construction costs would continue to decline. In addition, many participants have successfully created "homebrew" kits that can further reduce costs associated with library preparation.

In the question-and-answer session that followed the symposium, the participants and audience came to a consensus that the discipline was quickly moving from one that was data-limited to one that was data-rich. As researchers who have only recently transitioned from postdoctoral to PI status, we are amazed at both the pace of change and the opportunities created by recent advances in DNA sequencing.

References

Catchen J.M., Amores A., Hohenlohe P., Cresko W., Postlethwait J.H. 2011. Stacks: building and genotyping loci de novo from short-read sequences. G3 (Bethesda) 1:171–182.

Edwards S.V., Beerli P. 2000. Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution 54:1839–1854.

Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61:717–726.

Glenn T.C. 2011. Field guide to next-generation DNA sequencers. Mol. Ecol. Res. 11:759–769.

Hird S., Brumfield R.T., Carstens B.C. 2011. PRGmatic: an efficient pipeline for collating genome-enriched second-generation sequencing data using a provisional reference genome. Mol. Ecol. Res. 11:743–748.

Hudson R.R., Turelli M. 2003. Stochasticity overrules the "Three-times rule": genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. Evolution 57:182–190.

Kircher M., Sawyer S., Meyer M. 2011. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 2011:1–8.

Lemmon A.R., Lemmon E.M. 2012. High-throughput identification of informative nuclear loci for shallow-scale phylogenetics and phylogeography. Syst. Biol. 61:745–761.

Lemmon A.R., Emme S., Lemmon E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61:727–744.

McCormack J.E., Hird S.M., Zellmer A.J., Carstens B.C., Brumfield R.T. 2012a. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogenet. Evol. (in press) doi:10.1016/j.ympev.2011.12.007.

McCormack J.E., Faircloth B.C., Crawford N.G., Gowaty P.A., Brumfield R.T., Glenn T.C. 2012b. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species tree analysis. Genome Research 22:746–754.

Meyer M., Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. 6:1–7. doi:10.1101/pdb.prot5548.

Mullis K.B., Faloona F.A. 1987. Specific synthesis of DNA invitro via a polymerase-catalyzed chain-reaction. Methods Enzymol. 155:335–350.

Neiman M., Lundin S., Savolainen P., Ahmadian A., Anderson M. 2011. Decoding a substantial set of samples in parallel by massive sequencing. PLoS One 6:e17785.

Rice A.M., Rudh A., Ellegren H., Qvarnstrom A. 2011. A guide to the genomics of ecological speciation in natural animal populations. Ecol. Lett. 14:9–18.

Ronaghi M., Karamohamed S., Pettersson B., Uhlén M., Nyrén P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. Anal. Biochem. 242:84–89.

Saiki R.K., Scharf S., Faloona F., Mullis K.B., Horn G.T., Erlich H.A., Arnheim N. 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle-cell anemia. Science 230:1350–1354.

Sanger F., Nicklen S., Coulson A.R. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA. 74: 5463–5467.

Stapley J., Reger J., Feulner P.G.D., Smadjia C., Galindo J., Ekblom R., Bennison C., Ball A.D., Bekerman A.P., Slate J. 2010. Adaptation genomics: the next generation. Trends Ecol. Evol. 25: 705–712.

Townsend J.P. 2007. Profiling phylogenetic informativeness. Syst. Biol. 56:222–231.

Zellmer A.J., Hanes M.M., Hird S.M., Carstens B.C. 2012. Deep phylogeographic structure and environmental differentiation in the carnivorous plant *Sarracenia alata*. Syst. Biol. 61:763–777.