MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 19–21

NEWS AND VIEWS

PERSPECTIVE

Next-generation phylogenetics takes root

JOHN E. MCCORMACK* and BRANT C. FAIRCLOTH†

*Moore Laboratory of Zoology, Occidental College, Los Angeles, CA 90041, USA; †Department of Ecology & Evolutionary Biology, University of California, Los Angeles, CA 90095, USA

It has been a tumultuous 5 years in phylogeography and phylogenetics during which both fields have struggled to harness the power of next-generation sequencing (NGS) (Ekblom & Galindo 2010; McCormack *et al.* 2012a). Fortunately, several methodological approaches appear to be taking root. In this issue of *Molecular Ecology*, O'Neill *et al.* (2013) employ one such method – parallel tagged sequencing (PTS) – to elucidate the phylogeography of a tiger salamander (*Ambystoma tigrinum*) species complex. This study demonstrates a practical application of NGS on a scale appropriate (and not overkill) for most biologists interested in phylogeography (~100 loci for ~100 individuals), and their results highlight several analytical challenges that lie ahead for researchers employing NGS techniques.

Received 20 August 2012; revised 28 August 2012; accepted 28 August 2012

At the heart of most next-generation sequencing (NGS) techniques, particularly when applied to phylogeography of nonmodel vertebrates such as the tiger salamander (Fig. 1), is the need to reduce the burden of data to a manageable and informative subset of the genome. O'Neill *et al.* (2013) accomplish this using parallel tagged sequencing (PTS), which is a system of tagging and pooling preamplified PCR products across individuals, such that amplicons from an entire data set can be sequenced in a single NGS run (Meyer *et al.* 2007, 2008). PTS is a highly targeted approach that uses prior knowledge about the loci of interest to collect data, and, in that way, it represents one of the few methods scaling traditional techniques to new sequencing technologies.

O'Neill *et al.* (2013) combined PTS with 454 sequencing because they did not simply want SNPs (Single Nucleotide Polymorphisms) mined from short reads, but also full loci featuring many linked SNPs – currently a necessary input for most coalescent-based analyses preferred by phylogeographers (e.g. species tree analysis). At 271 base pairs (bp), the average length of their loci was not particularly long by Sanger standards. However, these loci contained over 2600 SNPs. Their well-supported species trees suggest that the loci were long enough to generate a subset of informative gene trees.

Another benefit of PTS, highlighted in the paper, is the generation of a nearly complete data matrix across 100 individuals at 100 loci. The authors' final data set contained only 10% missing loci for a given individual. The completeness of the matrix allowed the authors wide latitude in their analytical methods by permitting both the analysis of SNPs with Structure (which is tolerant of missing data) and the analysis of full loci featuring linked SNPs in *BEAST (which is somewhat intolerant of missing data). Analytical flexibility is key to the study of young species complexes, like the tiger salamander, where the timescale of the research questions bridges the fields of population genetics and phylogenetics. O'Neill *et al.* (2013) discuss their results primarily in the context of the phylogeny – the history of lineage splitting and species delimitation. In addition to producing a species tree, their results suggest more evidence of fine-scale phylogeographic structure than previously thought. Presumably, further geographical sampling would permit the authors to drill into the geographical mosaic of current gene flow as well, hints of which are discernible in their Structure plots.

The highly targeted and nearly complete data sets of PTS contrast with those produced by a second suite of NGS approaches applied to phylogeography and phylogenetics of late: those using restriction digest to generate anonymous, but presumably orthologous, sets of loci across individuals. There are many variations on the basic approach (see Davey *et al.* 2011 for a review). Compared to PTS, the benefits include the number of loci interrogated

Fig. 1 *Ambystoma tigrinum*, one of 12 of the closely-related tiger salamander lineages included in the study. Photo credit: Kenneth Wray.

Correspondence: John E. McCormack
E-mail: mccormack@oxy.edu

(tens of thousands) and the independence of the method from existing genomic resources. The drawbacks include the occasional generation of incomplete data matrices, the inclusion of paralogous loci that are difficult to disentangle from orthologs and the narrow focus on SNPs, rather than the full sequence of each locus, which limits the analytical toolkit. These limitations may soon be moot due to the ever-increasing length of NGS sequencing reads and by methods that forego gene trees entirely and estimate coalescent parameters from SNP data alone (Bryant *et al.* 2012).

Another suite of genome reduction methods in widespread use involves targeted enrichment or 'sequence capture' of loci. Like PTS, sequence capture targets a distinct set of loci, and the sequence data collected can be used to create complete data matrices. Unlike PTS, sequence capture foregoes PCR amplification of targeted loci and instead uses a set of RNA or DNA probes as baits to hybridize and capture genomic DNA (Mamanova *et al.* 2009). The target loci and baits can then be enriched compared with nontarget DNA and sequenced *en masse* via NGS (Gnirke *et al.* 2009). Sequence capture is thus less laborious on the front end than PTS and offers the enticing ability to scale both enrichments and sequencing to many samples in multiplex. Targeted loci could include, for example, exons identified from genomes or transcriptomes. Ultraconserved elements are also desirable targets because they provide universal anchors for hundreds to thousands of loci spanning large portions of the tree of life (Crawford *et al.* 2012; Faircloth *et al.* 2012; McCormack *et al.* 2012b). The drawbacks of sequence capture include high library preparation costs, limited sequence tags for tracking libraries during NGS and few tools to accommodate analysis of hundreds to thousands of loci enriched from taxa without a reference genome. Changes in the marketplace (Illumina Nextera XT) combined with new tagging techniques (Faircloth & Glenn 2012; Meyer & Kircher 2012) and analytical tools should alleviate these concerns, and sequence capture may soon become so easy and affordable that it supplants other methods. Of course, as whole-genome sequencing costs continue to decline, all genome reduction approaches may eventually be supplanted. For now, the decision to use sequence capture or PTS probably rests with the availability of extant sequence data and the number of individuals and loci targeted, with smaller projects being more easily accomplished via PTS.

Each of these techniques removes the bottleneck that has prevented the application of NGS approaches to nonmodel taxa while creating a new speed bump along the way: the analysis of NGS data. O'Neill *et al.* (2013) traverse this issue by creating a freely available pipeline (NextAllele) for sequence analysis that combines splitting and sorting of multiplexed reads, identification and alignment of recovered loci, likelihood ratio validation of base calls, haplotype phasing and data export. This software package offers an easy-to-understand, integrated workflow that complements several excellent alternatives (McKenna *et al.* 2010; Catchen *et al.* 2011; Hird *et al.* 2011). What appears to set NextAllele apart is the ability to phase haplotypes directly from short sequence reads (physical phasing *sensu* Browning & Browning 2011), an advance that will take much of the pain and uncertainty out of haplotype determination.

Finally, O'Neill *et al.* (2013) provide interesting, if somewhat foreboding results from species trees generated from the subsets of their data. The authors discovered that when subsets of the most informative of their 94 loci were used to generate a species tree, Bayesian analyses converged quickly, the trees were highly supported, and the topologies were in agreement with one another and consistent with prior knowledge about the relationships of tiger salamander lineages. However, analysing additional data sets incorporating less informative loci eroded this phylogenetic stability – a finding that was also reflected in poor analytical convergence. Their results suggest that inclusion of less informative loci added so much noise to the signal that the analysis eventually broke down. This result lends an important cautionary note to the general excitement surrounding the era of 'big data'. We have worked under the mantra of 'more data are better' for so long that we sometimes forget that all data are not equal. What is the use of 1000 loci when the answer we are looking for can be provided by the 20 most informative loci, while the other 980 are merely running interference? It is a classic question (Hillis & Huelsenbeck 1992), but one we have perhaps forgotten, particularly with analytical advances that can accommodate so many sources of error and uncertainty. As it turns out, maybe noise is still noise. Whether the data come from PTS, sequence capture or whole-genome sequencing, tuning out the noise and honing in on the signal might return to the limelight as the key challenge of phylogenetics.

References

Browning SR, Browning BL (2011) Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, **12**, 703–714.

Bryant D, Bouckaert R, Felsenstein J, Rosenberg N, RoyChoudry A (2012) Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, **29**, 1917–1932.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, **1**, 171–182.

Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC (2012) More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters*, **8**, 783–786.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Ekblom R, Galindo J (2010) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.

Faircloth BC, Glenn TC (2012) Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. *PLoS One*, **7**, e42543.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC (2012) Ultraconserved elements anchor thousands of genetic markers for target enrichment spanning multiple evolutionary timescales. *Systematic Biology*, **61**, 717–726.

Gnirke A, Melnikov A, Maguire J *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, **27**, 182–189.

Hillis D, Huelsenbeck J (1992) Signal, noise, and reliability in molecular phylogenetic analyses. *Journal of Heredity*, **83**, 189–195.

Hird SM, Brumfield RT, Carstens BC (2011) PRGmatic: an efficient pipeline for collating genome-enriched second-generation sequencing data using a "provisional-reference genome". *Molecular Ecology Resources*, **11**, 743–748.

Mamanova L, Coffey AJ, Scott CE *et al.* (2009) Target-enrichment strategies for next-generation sequencing. *Nature Methods*, **7**, 111–118.

McCormack J, Hird S, Zellmer A, Carstens B, Brumfield R (2012a) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics & Evolution*. Advance Online Access. doi: 10.1016/j.ympev.2011.12.007.

McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC (2012b) Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Research*, **22**, 746–754.

McKenna A, Hanna M, Banks E *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Meyer M, Kircher M (2010) Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harb Protoc*, **2010**, pdb.prot5448, doi:10.1101/pdb.prot5448.

Meyer M, Stenzel U, Myles S, Prüfer K, Hofreiter M (2007) Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Research*, **35**, e97.

Meyer M, Stenzel U, Hofreiter M (2008) Parallel tagged sequencing on the 454 platform. *Nature Protocols*, **3**, 267–278.

O'Neill EM, Schwartz R, Bullock CT *et al.* (2013) Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. *Molecular Ecology*, **22**, 111–129.

J.E.M. is an Assistant Professor in the Biology department and Director and Curator of the bird and mammal collections at the Moore Laboratory of Zoology at Occidental College where he studies speciation in birds. B.C.F. is an Assistant Research Scientist at UCLA where he studies population and evolutionary genetics of non-model taxa.

---

doi: 10.1111/mec.12050