Surmounting the Large-Genome "Problem" for Genomic Data Generation in Salamanders

David W. Weisrock, Paul M. Hime, Schyler O. Nunziata, Kara S. Jones, Mason O. Murphy, Scott Hotaling, and Justin D. Kratovil

Abstract Salamanders have some of the largest genomes among all extant organisms, due in large part to the proliferation of repetitive elements and the expansion of intron size. This increased complexity and size has limited the application of genomic tools to the population genetic and phylogenetic study of salamanders, even as these methods have become common for most other organisms. However, the generation of genomic data in salamanders is not out of reach for most researchers. High-quality and informative data sets can be acquired for salamander-centric research projects with careful consideration of the genomic tool(s) most appropriate for the question at hand and how best to apply these to a salamander genome. Here, we review a range of genomic tools representing the current best options for use in the study of genome-wide variation within and between salamander species. This includes the use of

D. W. Weisrock (✉) · K. S. Jones
Department of Biology, University of Kentucky, Lexington, KY, USA
e-mail: david.weisrock@uky.edu

P. M. Hime
Department of Biology, University of Kentucky, Lexington, KY, USA

Biodiversity Institute, University of Kansas, Lawrence, KS, USA

S. O. Nunziata
Department of Biology, University of Kentucky, Lexington, KY, USA

Department of Entomology, University of Maryland, College Park, MD, USA

M. O. Murphy
Department of Biology, University of Kentucky, Lexington, KY, USA

Department of Biology, Miami University, Oxford, OH, USA

S. Hotaling
Department of Biology, University of Kentucky, Lexington, KY, USA

School of Biological Sciences, Washington State University, Pullman, WA, USA

J. D. Kratovil
Department of Biology, University of Kentucky, Lexington, KY, USA

Department of Entomology, University of Kentucky, Lexington, KY, USA

Paul Hohenlohe and Om P. Rajora (eds.), *Population Genomics: Wildlife*,
Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_36,

transcriptomics (RNAseq), restriction site-associated DNA sequencing (RADseq), sequence capture enrichment methods, and PCR-based parallel tagged amplicon sequencing. Each of these methods has a particular set of benefits, as well as limitations in the study of salamander genomics. We highlight their trade-offs and the factors that should be considered when choosing among them, and we provide descriptions of exemplar studies that illustrate their empirical applications. By making informed decisions about the choice and implementation of these subgenomic methods, we believe that they can be broadly and effectively applied as important resources for the study of salamander evolution and conservation.

Keywords Amphibian · Conservation genetics · Parallel tagged amplicon sequencing · Phylogenetics · Population genetics · RADseq · RNAseq · Targeted sequence capture

1 Introduction

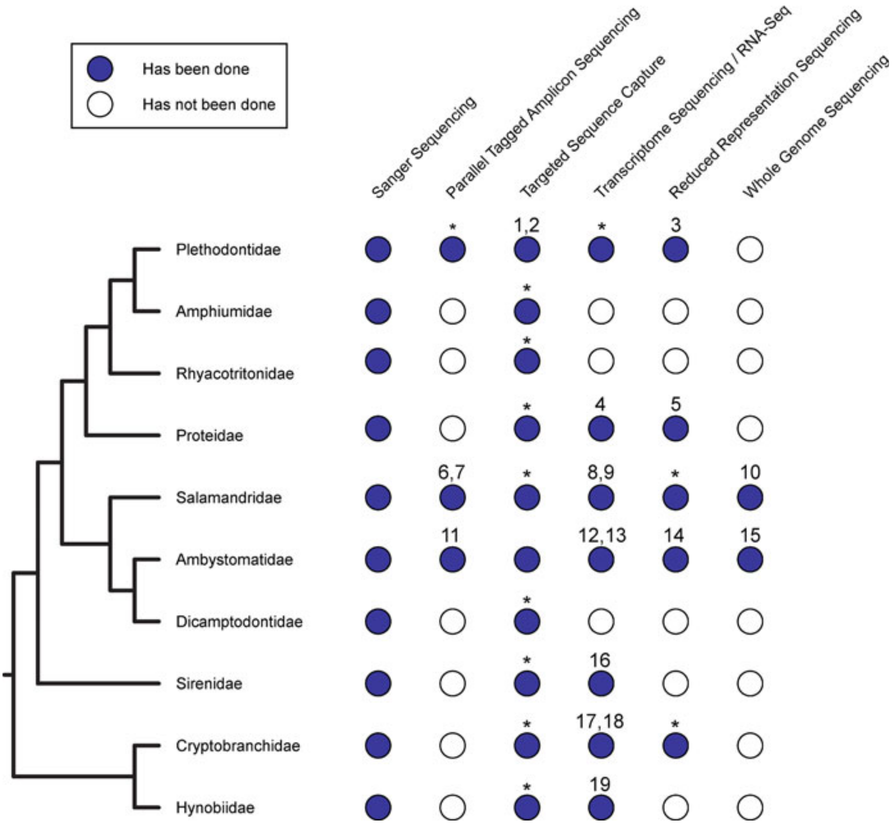
As next-generation sequencing (NGS) and the genomic revolution have swept forward the population genetic and phylogenetic study of non-model species, the use of new genomic tools for these pursuits in salamanders has lagged behind. The principal reason for this lag is their ridiculously large genomes – larger than almost all other vertebrate species (Sessions 2008). At their smallest, salamander genome sizes in the range of ~15 gigabases (Gb) can be found in many species of the genus *Desmognathus*, roughly five times the size of the human genome. At their largest, genomes have expanded to an astounding ~120 Gb in the Neuse River waterdog, *Necturus lewisi* (Gregory 2018). The approximately 700 salamander species fall somewhere in this range, typically around ~30–50 Gb. These absurdly large salamander genome sizes are particularly evident when put in the context of other major clades. For example, most mammals have ~3–4 Gb genomes, with a range of 1.6–6.3 Gb (Kapusta et al. 2017). Salamanders are exceptional even among other amphibians, with maximum haploid genome sizes within frogs of ~12 Gb (Olmo 1973) and within caecilians of ~14 Gb (Beçak et al. 1970). It is also worth noting that massive genome sizes in salamanders are not the result of polyploidization, as nearly all salamanders are diploid, with the exception of the unisexual members of the genus *Ambystoma* (Gibbs and Denton 2016).

Unsurprisingly, extremely high sequencing costs and the lack of availability for computational resources that can handle the inordinately large amount of data needed to produce a salamander genome have proved prohibitive in sequencing and assembling reference-quality salamander genomes. This has begun to change, as recent studies have produced genomic constructs for the axolotl (Ambystomatidae: *Ambystoma mexicanum*; Keinath et al. 2015; Nowoshilow et al. 2018) and the Iberian ribbed newt (Salamandridae: *Pleurodeles waltl*; Elewa et al. 2017). However, even these efforts have yielded highly fragmentary assemblies, highlighting some of the broader limitations and challenges in salamanders, ranging from the use of finite sequencing resources in an immense genome to the difficulty of placing subgenomic sequence data in the context of a whole-genome assembly.

It is likely that no single mechanism explains the evolution of large genome size in salamanders. Transposable elements (TEs) are common components of the genomes of most eukaryotes. However, studies of salamanders from the families Ambystomatidae, Cryptobranchidae, and Plethodontidae have revealed a disproportionately larger number of long terminal repeat retrotransposons, relative to other vertebrates, suggesting that the proliferation of these elements may be a driving factor in salamander genome gigantism (Sun et al. 2012; Sun and Mueller 2014; Nowoshilow et al. 2018). Introns are also substantially longer in salamanders relative to other vertebrate genomes and may contain greater numbers of regulatory regions (Smith et al. 2009; Nowoshilow et al. 2018). Salamanders also have very low metabolic rates relative to other vertebrates, and correlations between metabolic rate, cell volume size, and genome size have been proposed (Licht and Lowcock 1991). Given their vast size, it is likely that other notable aspects of salamander genomes will be discovered which set them apart from other vertebrates (e.g., Madison-Villar et al. 2016; Mohlhenrich and Mueller 2016; Elewa et al. 2017; Nowoshilow et al. 2018).

The challenge posed by large genomes varies across NGS tools, with each method posing its own suite of challenges. The use of PCR and capture-based approaches is constrained by the lack of baseline genome sequence information for most salamanders, limiting the generation of effective primers or capture baits. When these resources are available, large genome size does not seem to have a negative effect on PCR amplification of loci, but it does have an effect on capture-based enrichment methods, where capture baits are searching for "needles" in an extremely large "haystack." For anonymous locus methods, such as restriction site-associated DNA sequencing (or RADseq), challenges arise from the fact that larger genomes contain higher numbers of restriction enzyme recognition sites, and close attention is required to optimize the number of anonymous fragments for sequencing. RNA sequencing-based (RNAseq) approaches may be less hampered by large genome size, but certain analyses of the resulting data may be limited by the current lack of whole-genome resources for salamanders.

While these constraints have hindered the application of genomic data in the study of salamanders at the micro- and macroevolutionary levels, they are not insurmountable. Improvements in sequencing technologies continue to increase the amount of sequence data that can be generated while also decreasing costs. In addition, as researchers begin to take the plunge into the pool of available genomic tools and apply these to population genetic and phylogenetic questions in salamanders (Fig. 1), many of the kinks are beginning to be worked out of the data generation protocols, and a set of "best-practice" guidelines are emerging. This developing access to genome-wide data in salamanders brings with it a large genome upside: bigger genomes also harbor greater information about evolutionary history. For example, increased access to variable sites across the genome increases the probability of detecting recent coalescent events that can be informative of very recent population history. In addition, salamander genomes may contain a larger number of



● Has been done
○ Has not been done

	Sanger Sequencing	Parallel Tagged Amplicon Sequencing	Targeted Sequence Capture	Transcriptome Sequencing / RNA-Seq	Reduced Representation Sequencing	Whole Genome Sequencing
Plethodontidae	●	* ●	1,2 ●	* ●	3 ●	○
Amphiumidae	●	○	* ●	○	○	○
Rhyacotritonidae	●	○	* ●	○	○	○
Proteidae	●	○	* ●	4 ●	5 ●	○
Salamandridae	●	6,7 ●	* ●	8,9 ●	* ●	10 ●
Ambystomatidae	●	11 ●	●	12,13 ●	14 ●	15 ●
Dicamptodontidae	●	○	* ●	○	○	○
Sirenidae	●	○	* ●	16 ●	○	○
Cryptobranchidae	●	○	* ●	17,18 ●	* ●	○
Hynobiidae	●	○	* ●	19 ●	○	○

Fig. 1 A phylogenetic perspective of the variety of subgenomic and genomic methods that have been applied across the ten extant salamander families. Filled circles indicate that a particular method has been used in a given family. Open circles denote cases where a particular method has not been used for a family. Numbers above filled circles identify empirical examples for a particular method applied to a salamander family (1: Newman and Austin 2016; 2: Bryson et al. 2018; 3: Lucas et al. 2016; 4: Irisarri et al. 2017; 5: Murphy et al. 2018; 6: Zieliński et al. 2014a; 7: Wielstra et al. 2014a; 8: Czypionka et al. 2015; 9: Looso et al. 2013; 10: Elewa et al. 2017; 11; O'Neill et al. 2013; 12: Putta et al. 2004; 13: Nowoshilow et al. 2018; 14: Nunziata et al. 2017; 15; Nowoshilow et al. 2018; 16, 17: Irisarri et al. 2017; 18: Qi et al. 2016; 19: Matsunami et al. 2015). Filled circles marked with asterisks represent unpublished applications by the authors. Full references for these examples can be found in the literature cited

truly independent markers of species tree history, owing to the greater amount of recombinatorial decoupling of genetic variation over large chromosomal stretches. From these perspectives, salamanders may serve as unique systems for the study of evolutionary history.

Here, our primary goal is to review the many methods available for generating genomic data for studies in natural populations of non-model organisms and provide insight and guidance into their application in salamander genomes. While the spirit of this chapter lies within the context of conservation and wildlife genomics, many of the methods commonly used to study population-level genetic variation can be similarly applied at the phylogenetic level, and, when appropriate, we identify the strengths and weaknesses of each method at different scales of evolutionary divergence. We have written this review with the expectation that the reader will have a general familiarity with basic laboratory and sequencing methods, and we refer the reader to a number of reviews covering the new era of NGS in population and phylogenomics for more detail on sequencing methods (e.g., Davey et al. 2011; Lemmon and Lemmon 2013). Finally, we note that salamanders are not the only organisms with expanded genome sizes and the lessons learned in the application of genomic tools in salamanders can be leveraged in the study of other large-genome species.

2 Genomic Data Generation in Salamanders

Researchers interested in the study of genomic variation in natural populations now have a wide range of methods available for generating data that is appropriately targeted at their particular question (Fig. 2). While whole-genome sequencing is beginning to be a tractable approach for studying genetic variation within and among species with "normal" genome sizes, it is unlikely that this will become a reality for salamanders any time soon. However, other methods are available to comprehensively survey aspects of the genome. Deciding which method to use requires the consideration of factors that would apply to any taxonomic group, which largely revolve around the scale of divergence and what levels of genetic variation will be most informative for the questions at hand (Fig. 2). Salamanders, however, bring an extra set of genome-specific considerations. For example, the targeting of specific loci in the genome will require prior knowledge of genome sequence information and will likely require the availability of a relatively closely related genomic

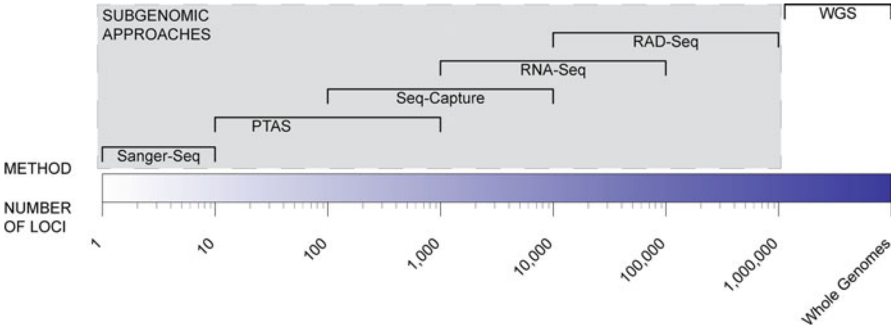


Fig. 2 A variety of subgenomic methods are available for data generation in salamanders. Different approaches will yield different numbers of loci, and researchers may select a method of data generation suited to the target numbers of loci for their study. Ranges of numbers of loci are approximate. WGS: whole-genome sequencing

resource from which to draw this information. In addition, the number of individuals that can be sequenced in parallel on an NGS platform will scale proportionally with the size of the genome under study. Each sequencing method can be affected differently by these different factors and, when coupled with the time and resources available to a project, will mean that different researchers may make different choices about the methods best suited to their project. As a note, all salamander genome size estimates presented here are taken from Gregory (2018).

3 Restriction Site-Associated DNA Sequencing (and Related Approaches)

3.1 General Overview

RAD sequencing (e.g., Miller et al. 2007), in its many varieties, provides one of the most straightforward ways to narrow down the number of genomic regions for sequencing. Through fragmentation of the genome with restriction enzymes and the subsequent reduction of this fragment pool to a particular size range, a substantially reduced subset of the genome can be created for sequencing. The use of the same restriction enzymes and size selection across multiple individuals provides the opportunity to recover a shared set of loci amenable to evolutionary analysis. These methods have been a boon to the ecological and evolutionary study of wild populations (Andrews et al. 2016), as well as for functional genomics (Baird et al. 2008), because they allow for the generation of large genome-wide data sets without the need for substantial prior information about the genomes under study. RAD sequencing is commonly used to uncover genetic variation, typically in the form of single-nucleotide polymorphisms (SNPs), and is most frequently used in studies at the population level or at the population-species interface. It can also be applied in a phylogenetic context across multiple species; however, increased evolutionary divergence reduces the shared overlap in orthologous loci among species. In this chapter, we do not review RAD-based protocols in detail and instead refer the reader to several original papers and reviews detailing their use (Miller et al. 2007; Elshire et al. 2011; Peterson et al. 2012; Andrews et al. 2016). In addition, we encourage readers to familiarize themselves with other complexities of these data that are not salamander-genome specific (e.g., allele and locus dropout; Gautier et al. 2013).

3.2 Salamander Genome Limitations

Larger genomes have a greater number of potential restriction enzyme cut sites and, as a result, more potential fragments to be sequenced. In addition, most researchers have a finite amount of sequencing effort that can be applied to a project. The sequencing of a single fragment (or locus) from an individual and the confident determination of its nucleotide composition and variation (i.e., homo- or heterozygous) require multiple independent sequence reads from the same fragment (the depth of sequencing coverage). Consequently, salamander researchers will need to consider how a RAD sequencing protocol can be optimized to reduce the overall set of fragments that can be sequenced and how this will intersect with their limited sequencing resources to permit the recovery of useful genomic data across multiple individuals.

The most important consideration for reducing the number of genomic fragments when working with large genomes is the choice of restriction enzymes, which has a large influence on the number of fragments that are produced. Restriction enzymes with longer, and rarer, recognition sites (e.g., 6 or 8 bp) will yield fewer fragments than those with smaller recognition sites. As a further step in reducing the number of fragments to be sequenced, researchers can use two restriction enzymes instead of one (i.e., a double digest, or ddRAD) and sequence only those fragments containing both cut sites (Peterson et al. 2012). It is likely that all salamander RAD sequencing studies will require a ddRAD-like approach to produce a library reduced enough to optimize sequencing efforts. Finally, the selection of a particular size range from the resulting fragment distribution provides yet another mechanism for reducing the number of fragments for sequencing.

Based on the study-specific requirements for numbers of loci, numbers of individuals, and per-locus depth of sequencing coverage, as well as the available sequencing resources, practitioners may optimize the restriction enzyme(s) and/or size selection window accordingly. Optimizing ddRAD approaches involves performing single- and double-digests of genomic DNA for multiple pairs of restriction enzymes and empirically estimating the number of sequenceable fragments within different size selection windows (as in Peterson et al. 2012, supplemental materials). While genomic resources are not required for this estimation, a best guess of genome size for the species under study can be used, which are available for all salamander families (Gregory 2018). Software is also available to perform in silico predictions of fragment numbers when a genome assembly is available (Lepais and Weir 2014). This can also be done using randomly generated sequence data as a proxy for an unknown genome, although we are unaware of any attempt to use this as a preparatory step for ddRAD in something as large as a salamander genome. Ultimately, after considering the constraints of the possible numbers of loci per individual, researchers should then select a restriction enzyme combination and size selection window best suited to their particular question and resources.

During the planning stage of a RAD sequencing project, it may be useful to quantify the interactions of important factors that will influence data generation. Based on the estimated number of loci per individual, the desired mean depth of sequencing coverage per locus, the number of individuals to be included in the study, and the estimated proportion of raw sequencing reads that can be assembled into loci, one can estimate the total amount of sequencing effort needed for the study according to Eq. 1.

$$\mathrm{SeqEffort} = (L \times C \times I)/R \tag{1}$$

SeqEffort is the number of total reads to be sequenced, L is the average number of loci (fragments) per individual, C is the mean per-locus depth of sequencing coverage, I is the number of individuals, and R is the proportion of sequencing reads passing all quality filters (e.g., sequencing quality scores, removal of PCR duplicates) and assembled into loci (on-target rate).

For example, in order to sequence 100,000 loci per individual to 30× mean coverage for 100 individuals, and an 85% on-target rate, 352,941,176 reads (or read pairs, for paired-end sequencing) would be required. This is, of course, idealized, and other factors will come into play. In our experiences with ddRAD in salamanders, empirical on-target rate was ~85% when we aimed for ~30× coverage, but numbers of individuals and loci varied by species and project.

When dealing with salamander genomes, restriction enzyme combinations may still yield an exceedingly high number of loci, and it may be difficult to optimize ddRAD protocols to produce fewer than tens of thousands to hundreds of thousands of loci per individual. These expectations are based upon our experience across five families of salamanders (Table 1). More recent modifications of the general RAD method may provide the means to further winnow down the numbers of loci produced per individual, either by performing an additional restriction enzyme digestion step (e.g., Graham et al. 2015) or by subsequently performing targeted enrichment on a subset of loci generated in an initial round of RAD sequencing [e.g., Rapture (Ali et al. 2016) or RADcap (Hoffberg et al. 2016)].

Once data are in hand, there are inherent complications with assembling and analyzing large numbers of loci that will be generated by a RAD sequencing protocol in a large genome. Perhaps the most obvious is that the computation time required to assemble sequencing reads into loci, and to compare loci across multiple individuals, scales with numbers of loci. Standard software packages (Catchen et al. 2011, 2013; Eaton 2014) work well for RAD locus assembly, but access to high-performance parallel computing resources is highly desirable. One special consideration in the assembly of salamander RAD sequencing data is the detection and filtering of potential paralogous loci. The large proportion of repetitive elements in salamander genomes greatly increases the probability that paralogs will be sequenced and that they might masquerade as orthologous loci. Paralogous loci should exhibit some characteristic patterns if assembled as a single locus, including extremely high sequence coverage and/or biologically implausible numbers of alleles (i.e., >2 for diploid species). Many assembly programs include functions that can filter according to these factors. Finally we note that while genome assemblies are now available for two salamander families, high rates of divergence across salamander families – and even between genera – are likely to limit the usefulness of

Table 1 Use of RAD-based sequencing in different salamander species

Family	Species	Genome size (Gb)	Target selection size (bp)	Est. # of sequence fragments	Maximum # loci recovered for an individual	Reference
Ambystomatidae	*Ambystoma opacum*	~30	300	273,120	300,869	Nunziata et al. (2017)
	Ambystoma talpoideum		300	133,450	131,201	
Cryptobranchidae	*Andrias*	~55	500	332,445	398,087	Hime et al. unpublished
	Cryptobranchus alleganiensis	~55	500	–	445,982	Hime et al. unpublished
Plethodontidae	*Desmognathus fuscus*	~15	300	259,376	67,724	Kratovil et al. unpublished
	Desmognathus marmoratus/ quadramaculatus	~15	500	166,085	85,520	Jones and Weisrock unpublished
	Eurycea sp.	~25	250–350	–	–	Lucas et al. (2016)
Proteidae	*Necturus maculosus*	~85	300	981,452	387,292	Murphy et al. (2018)
Salamandridae	*Laotriton laoensis*	~50	300	–	101,901	Jones et al. unpublished

All studies used a double-digest method with the enzymes *SphI* and *EcoRI* for library preparation except for Lucas et al. (2016), which used *MseI* and *EcoR1* salamander genome assemblies to a narrow range of closely related species. Hence RAD-based locus assembly will continue to be almost exclusively de novo for the foreseeable future.

3.3 Examples in Salamanders

RAD-based sequencing has been successfully implemented in multiple families of salamanders spanning a wide range of genome sizes (Table 1). Nunziata et al. (2017) provide a useful illustration of the application of ddRAD sequencing in salamanders, with a study of fine-scale population demographics in the ambystomatids *Ambystoma opacum* and *A. talpoideum* (genome sizes estimated between 24 and 36 Gb). After performing a series of test digests, *EcoRI* and *SphI* were identified as the best restriction enzyme pair, and a size selection window of 270 to 330 bp was estimated to contain ~130,000 and ~270,000 unique fragments in *A. talpoideum* and *A. opacum*, respectively. To sequence this fragment pool to a read depth of 10×, a maximum of 24 individuals were multiplexed per lane of an Illumina HiSeq 2500. After assembly, individuals of *A. talpoideum* had as many as 131,000 reconstructed loci, while *A. opacum* had ~300,000 loci, close to predictions based on their fragment distributions. A more important perspective is the number of loci recovered across multiple individuals, and here the number of shared loci is expected to drop. In *A. opacum*, when restricting the data to only include loci recovered from 95% of individuals, just 15,740 loci were retained. Increasing the allowed level of missing data to 15% bumped this up to 40,326 shared loci. While this is a substantially smaller number of loci than predicted for each individual, this reduction is typical of RAD sequencing studies. Furthermore, the data generated in Nunziata et al. (2017) were, nonetheless, highly informative, yielding important insights into the population demographics of rapidly changing salamander populations.

Lucas et al. (2016) used a genotyping-by-sequencing approach to estimate genetic diversity and gene flow in a wetland metapopulation of an undescribed species of *Eurycea*, which has a best-guess genome size estimate of ~25 Gb. A more limited sequencing effort was used in this study, but this still resulted in the assembly of ~6,200 unique loci and the identification of ~7,000 shared SNPs. While the complete details covering levels of missing data are not provided, this study still serves as an example of successful RAD sequencing in a salamander.

Near the upper end of the genome-size spectrum, ddRAD has been successfully used in a population structure study of the common mudpuppy, *Necturus maculosus* (~80–95 Gb genome; Murphy et al. 2018). Despite this truly massive genome size and corresponding massive number of fragments estimated per individual (~1 million, Table 1), this still resulted in a final data set of ~10,000 shared loci – with no missing data – across all sampled individuals (distributed across three river basins in Kentucky). Our lab has had similar success in phylogeographic studies of the hellbender salamander, *Cryptobranchus alleganiensis* (~55 Gb genome) and its

Asian sister genus, *Andrias* (~46–50 Gb genome; Hime et al. unpublished). While hundreds of thousands of loci were successfully assembled per individual, ~74,000 loci were still shared among ~100 individuals from across the Eastern United States. Maybe more surprisingly, we recovered ~43,000 shared loci between *Andrias* and *Cryptobranchus*, despite a divergence between these two clades of at least 15 million years (Kumar et al. 2017). The level of shared loci recovered between these genera is encouraging given the expectation of locus dropout due to the accumulation of substitutions in restriction enzyme recognition sites over time; however, we note that this level of shared recovery is not always found in interspecies comparisons. For example, our research in different plethodontid radiations has yielded low levels of shared ddRAD loci across species. Consequently, we emphasize that more empirical studies will be needed to know how generalizable patterns and levels of shared loci will be across interspecific salamander studies. While the increased divergence between species will lower the number of shared loci recovered, these loci will have higher levels of variability, relative to their patterns within species, and thus should still provide a large amount of information for interspecific questions (e.g., Lemmon and Lemmon 2012).

3.4 Guidelines for RAD-Based Sequencing in Salamanders

Given the wide range of genome sizes and compositions across salamanders, no single RAD-based sequencing protocol is expected to work for all species. Different species may require different restriction enzyme combinations, fragment size selection windows, and varying numbers of individuals that can be sequenced in parallel. It is likely that a ddRAD protocol will be required, as the fragment pool resulting from a single restriction enzyme digest will be too large and would dilute sequencing effort too much to produce useful results across multiple individuals. Beyond this one blanket recommendation, we recommend that the implementation of ddRAD-based studies in salamander species consider the following:

1. Following the protocol outlined in Peterson et al. (2012), researchers should evaluate the fragment distributions resulting from both single and double digests using an Agilent Bioanalyzer (or similar equipment). When combined with ballpark metrics of genome size, this allows for an estimation of the number of fragments within particular fragment-size windows. As a note, our lab has consistently found the enzyme combination of *SphI* and *EcoRI* to generate appropriate numbers of fragments, but this does not necessarily mean that these will be the best for all salamander RAD sequencing studies. Furthermore, not all studies will use the same fragment size selection window as applied in our studies in different families, and a thorough assessment of potential fragment numbers is encouraged to the identify particular fragment sizes to use and to be avoided.
2. Increase the amount of input genomic DNA above the ~50 ng range typically used in RAD studies. In our experiences, starting genomic DNA amounts in the

range of 1 μg has worked well for species with genomes in the range of 15–20 Gb (e.g., *Desmognathus*), and as much as 2.5 μg of genomic DNA was used for larger genomes (e.g., *Necturus maculosus*). These higher amounts of starting DNA ensure that sufficient quantities of genomic material remain after double digestion and size selection. Starting with larger quantities of DNA can also limit the number of PCR cycles required to reach desired final library concentrations (thus reducing the potential for PCR-induced errors).

3. Be cautious with the urge to increase the number of multiplexed individuals that are sequenced. A threshold exists that when passed will result in most sequenced loci being recovered at unacceptably low coverage to distinguish genuine SNP variation from sequencing error. What this coverage threshold is will depend on the study. For example, higher coverage will be necessary for population genetic questions and analyses where diploid genotypes for all individuals are important, and lower levels may suffice in studies where population-level estimates of allele frequencies are of interest. Here, we refrain from providing guidelines for levels of multiplexing on a "lane" of sequencing, as sequencing technologies – and RAD library protocols – continue to increase in efficiency and output. If the depth of coverage of initial rounds of sequencing is too low, be prepared to increase sequencing effort accordingly.

4 Transcriptomics and RNAseq

4.1 General Overview

Transcriptome sequencing (including RNAseq) involves the purification of transcribed RNA from a tissue or set of tissues, conversion to complementary DNA, and subsequent high-throughput sequencing. By focusing sequencing effort on transcribed regions of the genome, researchers are able to target coding regions, to the exclusion of other genomic content. Consequently, transcriptomics can be easily applied in salamanders, providing access to a large amount of genomic content with no more difficulty than its use in organisms with smaller genomes. These transcriptomic resources can then have multiple applications in salamander wildlife genomics. As perhaps its main application, transcriptomics is used to study differences in gene expression across tissues, individuals, or populations, to better understand the effects of spatial, environmental, or temporal factors on cellular processes (e.g., Trapnell et al. 2013). Transcriptomics also represents an effective method for directly targeting SNPs in coding regions, either as those segregating in a population or as fixed diagnostic markers between groups of interest (Zieliński et al. 2014a). As discussed above, it can be a direct and effective way to identify candidate loci to be developed into sequence capture-based or PTAS-based markers. Transcriptomics can also provide important context for many of the previously discussed anonymous loci generated through a RAD-based approach (Amores et al. 2011).

4.2 Salamander Genome Limitations

From a data generation perspective, the large salamander genome poses no significant challenge, relative to other taxa with smaller genomes. In this method, the RNA polymerase machinery does the important enrichment work for you. Researchers should be aware of the general challenges in employing transcriptomics in natural settings. This includes acquiring similar enough tissues from individuals under study to increase the probability of recovering the same set of expressed orthologous loci. Studies of speciation and local adaptation should also consider the particular tissue and developmental stage being sampled and whether their expressed genes will include the loci relevant to the study at hand. The need to rely on nondestructive sampling or challenges in field collecting (e.g., acquiring necessary permitting, or finding rare species) can all pose limitations to properly implementing a transcriptomic approach for population and evolutionary studies. In addition, with a lack of a whole-genome assembly, there are also likely to be many things that are unknown going into the study, including the number of potential loci to be expected. Finally, we point out that the computational overhead of transcriptome assembly for large salamander genomes is also not expected to be more burdensome than in other taxa.

4.3 Examples in Salamanders

The works of Putta et al. (2004) and Habermann et al. (2004) represent the earliest efforts in generating large-scale transcriptomic data from salamanders, with studies in the Mexican axolotl (*A. mexicanum*) and eastern tiger salamander (*A. t. tigrinum*). These studies predated current NGS technologies and were generated as ESTs that provided sequence data from one end of a transcribed exonic region. However, this still resulted in the identification of ~35,000 ESTs and >10,000 contigs with high sequence similarity to known human coding sequences. While many of the goals of this work were aimed at generating resources for the study of salamander regenerative developmental biology, these resources have also had substantial downstream applications in the generation of a genome-wide linkage map (Smith et al. 2005), the generation of PCR-based nuclear markers for the study of species boundaries in related Mexican species (Weisrock et al. 2006), and the study of hybridization and admixture between native and introduced species (Fitzpatrick et al. 2010).

In a more recent example, Keinath et al. (2017) provide an example for the use of transcriptomics in the generation of a high-quality linkage map for *Notophthalmus viridescens*. This work is particularly exciting in that it demonstrated a relatively simple and fast process for developing linkage maps from large-genome species without the requirement for tremendous sequencing resources (only one HiSeq2000 lane was used) or >F1 generations (a single mother and her 28 offspring were used). Given that whole-genome sequence assemblies for most salamanders are likely to be unavailable in the near future, transcriptome-based linkage maps will continue to serve as our best resources for studying genome structure and the placement of ecologically and functionally relevant loci (e.g., Voss and Smith 2005).

Transcriptomic data sets have also been recently used in phylogenomic studies at both shallow and deep evolutionary histories of salamanders. Rodríguez et al. (2017) used RNAseq data (along with ddRAD data) to resolve the recent history of divergence among species of the salamandrid genus *Salamandra*. Using rather modest sequencing effort on an Illumina MiSeq, the authors were still able to assemble a data set of 3,170 orthologous loci sampled from seven *Salamandra* species and two *Lyciasalamandra* outgroup species. In a study of deep phylogenetic relationships across jawed vertebrates, Irisarri et al. (2017) included sequence data generated using RNAseq from representatives of a number of salamander families. This study provides a good perspective on the sequencing effort required to recover known orthologous protein-coding genes. Using RNA sourced from multiple tissue types from the species *Andrias davidianus* (Cryptobranchidae), *Calotriton asper* (Salamandridae), *Proteus anguinus* (Proteidae), and *Siren lacertina* (Sirenidae), and a half of an Illumina MiSeq flow cell per species, they recovered between 59 and 81% of 233 core vertebrate genes (CVGs), a reference collection of one-to-one vertebrate orthologs that can be used to benchmark transcriptomic studies (Hara et al. 2015). Using a substantially greater sequencing effort in the salamandrid *Pleurodeles waltl* (381 million reads, or over 27× the number of sequence reads than in the above discussed species), recovery of CVGs approached 98%. Collectively, this demonstrated that standard transcriptomic sequencing approaches applied to diverse RNA pools in salamanders can lead to nearly complete recovery of the standard set of orthologous vertebrate genes but also that rather small sequencing efforts can still recover large sets of expressed genes.

A number of additional transcriptomic projects have been completed in salamanders to understand cellular responses in gene expression in an environmental context. Qi et al. (2016) used an RNAseq approach to study the immune response of the Chinese giant salamander, *Andrias davidianus*, when infected by a bacterial pathogen. This work yielded ~19,000 annotated coding genes and demonstrated the utility of RNAseq-based approaches in salamanders for identifying genes that potentially underlie functionally relevant pathways for immunity. Czypionka et al. (2015) used an initial round of transcriptome sequencing in *Salamandra salamandra*, coupled with the subsequent use of microarrays containing probes matching a set of ~22,000 assembled contigs identified as having open reading frames (ORFs), to study shared versus differential patterns of gene expression between nonlethally sampled tail clips and lethally sampled whole larvae. Interestingly, this work showed that a large proportion of genes (51%) had similar changes in expression among tail and whole-body tissues across different temperature treatments, suggesting that nonlethally sampled tail tissues may serve as a good proxy for environmentally influenced gene expression. Matsunami et al. (2015) used RNAseq to examine gene expression changes underlying phenotypic plasticity in *Hynobius retardatus* in response to different predators. This work generated ~740,000 assembled contigs, among which ~175,000 could be identified as protein coding based on the presence of

ORFs. Based on this large genomic resource, dozens of genes were identified that had differential expression under different predator regimes, and ultimately this led to new insights into the understanding of the evolution of phenotypic plasticity.

4.4 Guidelines for Applications in Salamanders

There is little salamander-specific advice that we can offer for the use of transcriptome sequencing, as there are inherent limitations to the use of this tool in species with large genomes. Standard laboratory and computational methods will apply. Perhaps the one relevant point to make is that the design and implementation of these projects can be done according to researchers' downstream goals. If the goal is marker development for subsequent use in sequence capture and PTAS studies, a single tissue source from an animal (e.g., a tail tip) can be sufficient to generate enough candidate loci. This may be ideal when nondestructive sampling is preferred, or when tissue sources are rare. Alternatively, when projects are aimed at identifying as many coding genes as possible, either in an attempt to uncover orthologous loci identified in other species or to study their expression differences across different treatments, multiple tissue sources from an animal are required.

5 Sequence Capture and Enrichment

5.1 General Overview

Sequence capture methods use synthetic oligonucleotide probes to target and enrich for genomic regions identified a priori. Biotinylated probes are annealed with fragmented and barcoded genomic DNA of a target species, with the probes finding their complementary match to target loci. These "captured" fragments are then sequestered by hybridization to streptavidin-linked beads and clonally amplified by high-fidelity PCR. The resulting enrichment products for multiple individuals are then sequenced in parallel on a NGS platform. By using the same set of probes across all individuals in a study, sequence capture methods provide an effective approach for generating data from shared orthologous loci. Sequence capture methods were kick-started with the generation of protocols to perform probe hybridization in solution (Gnirke et al. 2009) and since have been dominated by two general approaches, anchored hybrid enrichment (AHE; Lemmon et al. 2012) and ultraconserved elements (UCEs; Faircloth et al. 2012). They have also been implemented in a custom fashion in numerous taxonomic groups, typically in the form of exon and candidate locus capture (e.g., Bi et al. 2012; Linnen et al. 2013; Portik et al. 2016).

A sequence capture approach for genomic data generation can have many benefits over other genomic methods. First, it provides a methodologically efficient approach for sequencing known regions of the genome, as opposed to anonymous loci that are typically sequenced using RAD-based methods. Second, it can lead to the generation of highly complete data sets across individuals and species, as it does not suffer from allele dropout due to a single mutation or substitution in a restriction enzyme recognition site. Third, capture probes can be quite forgiving to mismatches with genomic templates, much more than can be tolerated in the annealing of PCR primers. Consequently, probes based on one species can be effectively used across a relatively wide range of divergent taxa; however, there are limitations to the level of divergence between probe taxa and capture taxa that we discuss below.

5.2 Salamander Genome Limitations

In any sequence capture reaction, capture probes must sift through a pool of genomic DNA to find complementary matches with their target loci. While this is an efficient method for enriching a sample with a desired set of loci for sequencing, the vast nature of a genome also leads to a large amount of "off-target" capture (Guo et al. 2012), or the enrichment of additional genomic regions that are not part of the specific set of targeted loci. This can occur for a number of reasons, including the promiscuous annealing of probes to nontargeted DNA under different reaction conditions and the carry through of high-copy regions of the genome (e.g., mitochondrial DNA). In salamander genomes, the expansion of many aspects of genomic content (e.g., larger introns and greater number of TEs) is expected to increase the amount of off-target enrichment. This leads to at least two complications: (1) capture probes are diluted across the genome in proportion to genome size, yielding a lower level of enrichment of targeted loci, and as a result, (2) greater sequencing effort will be required to recover targeted loci at a sufficient depth of read coverage.

The evolutionary divergence between probe taxa and target species also exerts a strong influence on capture success. For instance, in capture reactions applied to frogs, Hedtke et al. (2013) found that the numbers of recovered loci dropped precipitously with increasing divergence time between probe species and target species (also see Lemmon et al. 2012). Although a straightforward work-around is to design probes specifically from the taxon or taxa under study, the current scarcity of genomic resources for salamanders means that this is not likely to be a simple fix for many researchers.

There are at least two possible remedies that can be applied to mitigate off-target enrichment in salamanders. The first is to use capture probes with high specificity to the taxa under study. While targeting conserved stretches of DNA provides one mechanism for increasing the probability that probe sequences will have high complementarity to the template DNA, this can still lead to an exceptionally high level of off-target enrichment in salamanders. Even with high conservation between the probe taxon and the study taxon (e.g., when they are the same species), "on-target" sequence reads may at best only account for 20–30% of the total sequence reads (Bi et al. 2012; Faircloth et al. 2012), and this is expected to be substantially lower in salamanders. We have explored this approach in both hellbender salamanders (*Cryptobranchus*) and dusky salamanders (*Desmognathus*). Here we compared locus recovery using the original Lemmon et al. (2012) AHE method and probe set (in which the closest probe taxon was the frog *Xenopus* [*Silurana*] *tropicalis*) to a custom probe set that included capture probes designed specifically from genomic resources for *Cryptobranchus* and *Desmognathus* which we developed de novo. In the first set of tests, the evolutionary distance between the probe taxon and target species was ~300 million years, and we recovered 54.1% (277/512) and 72.8% (373/512) of loci from *Cryptobranchus* and *Desmognathus*, respectively. In contrast, use of the taxon-specific probe set in these two taxa increased locus recovery to 93.0% (319/343) in *Cryptobranchus* and 99.7% (342/343) in *Desmognathus*.

The other possible remedy to increase capture efficiency in salamanders is through decreasing the negative effect of the highly repetitive fraction of the genome (McCartney-Melstad et al. 2016). Using c_0t-1 DNA (Kallioniemi et al. 1992) developed from the species of interest, a large portion of the repetitive DNA in the genome can be "blocked," increasing the probability that individual capture probes will find their complementary match and reducing the amount of off-target capture. C_0t-1 DNA itself is the repetitive fraction of the genome and is created by fragmenting the genome, denaturing the DNA into single-strand form, and then slowly reannealing into double-strand form. Because high-copy, single-strand fragments should find their complementary match sooner than single-copy fragments, the collection of the early stages of fragment reannealing (i.e., c_0t-1) yields a collection of mostly repetitive DNA. Adding DNA from this c_0t-1 fraction to sequence capture reactions can be effective in blocking repetitive genomic DNA and increasing the probability that capture probes will find their on-target matches in the genome. In a test study applying exon-based capture probes based on transcriptomic resources for the Mexican axolotl, *Ambystoma mexicanum*, to the closely related tiger salamanders *A. californiense* and *A. mavortium*, McCartney-Melstad et al. (2016) demonstrated that the inclusion of c_0t-1 had a positive effect on on-target sequence reads. Furthermore, they found even more increased recovery of on-target sequence reads when using both c_0t-1 and increased concentrations of input DNA. While these effects did not push capture efficiency to very high levels of on-target sequencing, it did nearly double the rate of on-target sequencing, from ~10% without c_0t-1 and using standard DNA input concentrations to nearly 20% when using high amounts of both. The generation of twice as many on-target reads can substantially improve the number of recovered loci and provide the necessary read coverage needed to confidently identify SNP variation.

While the use of c_0t-1 blocker DNA can be an effective improvement to sequence capture in salamanders, it also has its limitations. First, it is not clear what level of divergence will be tolerated between the genomes of the c_0t-1 species and the targeted study species. Genomic repeat landscapes can be expected to change with increased divergence from a common ancestor, and it is likely that c_0t-1 may need to be derived from the study taxon or closely related species to be most effective. The generation of c_0t-1 DNA itself may pose a limitation in many taxa, particularly when tissue resources are small (e.g., when using nondestructive sampling) or rare (e.g., endangered or difficult to obtain species). Much remains to be explored with this method, including the potential for whole-genome amplification of the c_0t-1 fraction (as per the suggestion in McCartney-Melstad et al. 2016) to create large amounts of material from limited sources.

One additional caveat to consider in sequence capture is that the use of conserved genomic regions for probe development may prove limiting in the recovery of genetic variation for use in population-level studies. While targeted sequence capture based on conserved regions can perform well at shallow scales in terms of the proportion of target loci that are successfully captured, this performance may come at the cost of variation in loci. Conserved core regions of loci are less likely to contain informative sites at shallow scales, and it may be the more variable flanking regions of loci (e.g., introns or $3'$ untranslated ends of coding regions) that are of greatest utility for questions at the population genetic level. Salamander researchers targeting population-level questions may consider developing more species-specific capture markers for their projects (e.g., Lemmon et al. 2012), but given the lack of genomic resources for most families, these will probably need to be generated for the species or clade of interest.

5.3 Examples in Salamanders

To date, works of Newman and Austin (2016) and Bryson et al. (2018) represent the only published studies applying sequence capture – both in the form of UCEs – to address evolutionary questions in salamanders. Newman and Austin (2016) took steps to make the standard UCE approach more amphibian specific by restricting their UCE capture probe set to 2,064 probes covering 1,745 loci that had >85% sequence similarity to the *Xenopus* [*Silurana*] *tropicalis* genome. Excluding more divergent probes presumably has the effect of increasing the amount of on-target sequencing. A total of ~600 million sequence reads were generated from capture libraries for a set of 94 *Plethodon serratus* individuals sampled from across their range, along with two outgroup *P. cinereus* individuals. This resulted in the recovery of a large number of loci across the majority of individuals, but the exact number of loci retained for analysis varied across missing data filtering strategies. At the most individual inclusive level, a total of 321 loci (out of 1,745, or ~18%) were recovered for a set of 85 *P. serratus* and the two *P. cinereus*, where each locus was missing from no more than 20% of individuals. The number of retained loci increased to 1,327 (76%) when allowing for up to 40% missing individuals per locus. Newman and Austin (2016) also explored additional filtering strategies aiming at preserving the number of retained loci by removing individuals with greater numbers of missing loci.

Bryson et al. (2018) used a more standard tetrapod UCE probe set targeting 5,060 loci to capture loci for a range-wide study of the Mexican plethodontid, *Isthmura belli*. In an effort to increase capture efficiency, c_0t-1 blocker derived from chicken was used as part of the capture protocol, although it is unclear what effect this had given the tremendous evolutionary divergence between chicken and salamanders. While total sequencing effort was not described in this paper, it is clear that this approach yielded informative and useful data. Using a 50% missing data threshold, 1,094 loci were recovered across sequenced individuals. Increasing the missing data threshold to 30% reduced this to 796 loci.

Overall, the genomic takeaway from these studies is that sequence capture can work quite successfully in salamanders without major augmentations to standard protocols. However, it also provides another example of the effect of divergence between the probe taxon and target taxon. If highly complete data sets are desired, researchers will either have to decide between accepting a dropout of a high percentage of their loci, or individuals, when using divergent probes, or they may have to invest the effort in developing new probes for the taxa under study.

5.4 Guidelines for Applications in Salamanders

The primary piece of advice we can provide for the use of sequence capture-based methods in salamanders is to use probes with high sequence specificity to the taxa under study. The use of standard methods and probe sets such as AHE and UCE will likely result in successful data generation. But, the use of probes generated from genomic resources derived from the taxon or clade under study will more likely result in increases in capture efficiency, decreases in missing data, and reduction in the sequencing effort required. When projects are limited in their taxonomic scope, and when sufficient tissue samples are available, the use of c_0t-1 blocking DNA derived from the taxa under study will be greatly beneficial in improving capture efficiency and reducing the amount of sequencing effort (and corresponding fundage) required.

6 Parallel Tagged Amplicon Sequencing

6.1 General Overview

Parallel tagged amplicon sequencing (PTAS) emerged early in the transition from studies using one or a few loci to those using large numbers of loci. PTAS couples more traditional methods of sequence enrichment (i.e., PCR) with high-throughput sequencing to generate moderate- to large-scale data sets. The overall methodology is straightforward; given available PCR primers for multiple loci, amplicons from an individual are generated and then pooled and indexed either prior to or as part of library preparation. Multiple individuals are then sequenced in parallel on an NGS platform. Despite its simplicity, PTAS has not been as widely adopted for population genomic and phylogenomic studies, relative to sequence capture and RAD-based methods. This can be attributed to a number of reasons, including the much greater genomic sampling offered by the other methods, the lack of primer pairs for large suites of nuclear loci for most species, and the relatively larger laboratory effort required for PCR enrichment, although this latter issue can potentially be mitigated by merging multiplex PCR methods with PTAS (Campbell et al. 2015). Nonetheless, PTAS can still serve as an optimal data generation method for some labs and projects where the desired number of loci is modest and less than what would be more efficiently, and more cheaply, generated using a sequence capture approach (e.g., <100 loci; Shen et al. 2013). For many questions, data from a modest number of loci may be sufficient to produce well-supported results (e.g., Hime et al. 2016), and labs entertaining the idea of a genomic approach to their research may find the simple segue from PCR to NGS appealing, especially given the cost and equipment needs for other genomic methods. Generally, PTAS yields a low level of missing data, and because PCR is typically followed by confirmation with gel electrophoresis, researchers can have confidence in generating data across all loci that show positive amplification.

6.2 Salamander Genome Limitations

Genome size is not expected to have a direct negative impact on the use of PTAS. There is no evidence that PCR performance is influenced by genome size, and once amplicons have been generated, NGS is expected to perform just as well as with other enrichment strategies. PTAS may actually provide one of the most efficient sequencing strategies: as long as PCR does not amplify many secondary loci, amplicon pools should be highly enriched for the target loci, with little sequencing effort being squandered on off-target loci. As an example of this efficiency, we have included amplicon pools of different sets of loci amplified from two individuals (one *Ambystoma* and one *Desmognathus*) in the same indexed library and recovered all loci with high sequence coverage using a single multiplexed MiSeq run.

Perhaps the biggest limitation to the use of a PTAS approach in salamanders is the lack of developed PCR primer pairs for large numbers of loci. Few genomic resources are available, and primer development based solely on the currently available amphibian model genome (*Xenopus* [*Silurana*] *tropicalis*) is unlikely to yield primers that can be applied across salamanders with high levels of success. Developing primers from exons conserved across multiple vertebrate genomes has led to a toolkit of ~100 highly successful PCR primers that have been applied across salamander families (Shen et al. 2013) and across major clades within the Plethodontidae (Shen et al. 2016). One potential downside to relying on conserved exonic genomic regions for PCR primer development is that genetic variation may be limited when working at the population level. This is unlikely to be solved by anchoring primers in adjacent exons and amplifying across introns given the large intron size in salamanders (Smith et al. 2009; Nowoshilow et al. 2018). As discussed below, one of the best options for developing variable PCR-based markers is through the generation of transcriptomic data for the taxa under study, the identification of candidate orthologous and variable loci, and the development of species- or clade-specific PCR primers.

6.3 Examples in Salamanders

PTAS has been used most successfully in phylogeographic and population-level studies of two diverse salamander radiations. It was first applied in a systematic study of North American tiger salamanders (the *A. tigrinum* complex) aimed at understanding population structure and species boundaries across the range of the species complex (O’Neill et al. 2013). Using transcriptomic resources – generated as expressed sequence tags (ESTs) – for *A. t. tigrinum* and *A. mexicanum* (Putta et al. 2004), PCR primers were developed for a large suite of nuclear markers and then tested for successful PCR amplification, leading to a suite of primer pairs for 95 nuclear loci that amplified across the entire species complex (~5 million years of divergence). These loci were amplified from a total of 95 individuals, pooled and indexed by individual, and sequenced on a Roche 454 platform. Despite the use of this older NGS platform and a low number of sequence reads (~344,000) relative to current technology, this still resulted in a high proportion (~81%) of on-target sequences that could be successfully assigned to an indexed individual. This also resulted in a relatively high level of data completeness across individuals, with an average of just 11% missing data per individuals. The library preparation and NGS methods from this study represent older versions of genomic technology, and improvements in both will further improve the efficiency and recovery of PTAS data (Feng et al. 2016).

In a similarly designed set of independent studies in newts of the genera *Lissotriton* (Zieliński et al. 2014a) and *Triturus* (Wielstra et al. 2014a), PTAS markers were developed to study evolutionary patterns within and between species. Both studies used newer NGS platforms (Ion Torrent and Illumina), which yielded higher levels of read coverage per locus versus the 454-based approach employed by O’Neill et al. (2013). They also used a multiplexed approach toward their PCR work, bundling PCR primers for as many as 11 loci and greatly reducing the amount of laboratory effort required for amplicon generation. It should be clarified, however, that not all loci can be co-amplified in the same PCR reaction, as some primers can dimerize based on their nucleotide composition. Consequently, researchers interested in performing a similar multiplexed PCR strategy will need to invest some initial effort in determining locus pooling compatibility, and it is possible that given a limited set of PCR markers, not all loci will be able to be multiplexed.

While the data from these PTAS studies are on a much smaller scale than those typically generated with RAD sequencing or sequence capture, they have nonetheless been very informative about evolutionary patterns at shallow scales of divergence. The broadscale work of O’Neill et al. (2013) was transitioned into an informative study exploring the numbers of loci and information content required for species delimitation at shallow scales of divergence (Hime et al. 2016). The marker set and methods of Wielstra et al. (2014a) were used to gain important insights into the genetically admixed history of a recently described species (Wielstra and Arntzen 2014) and reconstruct the broader phylogeny of multiple species of *Triturus* (Wielstra et al. 2014b). Similarly, the work of Zieliński et al. (2014a) has been expanded to study the demographics of speciation between species of *Lissotriton* (Zieliński et al. 2016).

6.4 Guidelines for Applications in Salamanders

A PTAS strategy may be ideal for many researchers who (a) desire a multilocus approach to their work, (b) are addressing questions that don't require densely sampled genomic markers, and (c) prefer to not be deluged with massive data sets. This may be particularly true for many salamander researchers given the complexities discussed above for the use of RAD and capture-based sequencing protocols. Our best advice for the use of PTAS in salamander studies is to work from a pool of candidate markers developed from a closely related species. In the immediate future, this will probably require the generation of transcriptome resources to provide this pool of candidates, but given the modest number of loci that will go into a PTAS study, a large transcriptomic sequencing effort will not be required. We also recommend the selection of sets of primer pairs that generate amplicons of roughly the same length (within a few hundred base pairs of each other), which will maximize sequencing efficiency on NGS platforms, and we advocate the use of qPCR to quantify and normalize pooled amplicon concentrations between individuals in order to produce more even coverage across individuals.

7 Other Approaches

We have discussed four major genomic techniques that are commonly used in the study of population and interspecific genetic variation in non-model species. However, these do not represent the only genomic approaches to be used for these sorts of data. Many additional approaches have been devised which either augment one of these sets of tools to meet a particular need or which hybridize two of these tools together into a new method that overcomes some of the individual shortcomings. For example, the ddRAD method has been extended to include a third restriction enzyme digestion step, which has the effect of further reducing the fragment pool that goes into sequencing (Graham et al. 2015). RAD-based sequencing has also been paired with sequence capture methods to yield the perfectly named Rapture (Ali et al. 2016) and RADcap (Hoffberg et al. 2016) methods. In both cases, an initial round of RAD sequencing is used to develop candidate markers, which are then subsequently sequenced using a capture-based approach. This has the benefit of negating the effect of locus dropout due to recognition site mutations and decreasing the amount of missing data in a study. This may be particularly useful for RAD-based studies that aim to extend across species or previous projects.

One combination approach that has been used in salamanders is the identification of diagnostic SNPs between species of *Lissotriton montandoni* and *L. vulgaris* using comparative transcriptomics, followed by SNP-based genotyping of 192 loci sampled from >400 individuals using an Illumina GoldenGate assay to study admixture dynamics in contact zones between these species (Zieliński et al. 2014b). These assays do not actually yield sequence data and instead provide genotypes for each SNP based on patterns of fluorescence generated through PCR using allele-specific genotyping primers (Fan et al. 2006). Consequently, this component of the work may not be as “genomic” as other methods, but for salamander researchers interested in screening large numbers of individuals for population genetic study, it may actually be ideal to limit the genomic work (and its corresponding effort) to transcriptomic study of a relatively small amount of individuals, followed by mass genotyping of diagnostic or variable loci, especially given that these SNP genotype data can be highly informative about fine-scale population processes (Zieliński et al. 2014b). Alternatively, if genomic tools are not a limitation, this type of an approach can also be accomplished through the development of probes to the loci of interest and the use of sequence capture for locus recovery and genotyping.

8 Conclusions and Future Perspectives

Even with the large-genome constraints on the generation of genomic data for population genetic and phylogenetic analysis in salamanders, most current subgenomic methods can be effectively used when implemented properly. RADseq can be an effective tool for population-level and shallow-scale phylogenetic projects provided that researchers have enough sequencing depth for the numbers of fragments and individuals under study. Transcriptomics can serve multiple roles, from gene/locus discovery to studies of gene expression. Here, there is little genome-specific limitation in salamanders, with the greatest hurdle potentially being access to tissues for the extraction of relevant RNA pools. Sequence capture methods are a powerful source of phylogenetic markers from targeted regions of the genome but will work best in salamanders when capture probes are designed from relatively closely related taxa and when, again, enough sequencing depth is generated to adequately sequence the numbers of targeted fragments and individuals. Finally, parallel tagged amplicon sequencing has served as an effective method to generate moderate to large data sets in salamanders, without requiring the large sequencing effort needed with other genomic methods. However, it does require known and effective primers and more substantial laboratory effort to implement.

Other possibilities exist and we have not covered all of them here. Instead we emphasize that we foresee continued development of new methods and proposed variations on new and existing methods. Whole-genome sequencing and assembly in salamanders is just starting to take root, and the prospects of their broader application in salamander research loom large on the horizon. But, the era of and utility of subgenomic approaches in salamanders are here now. Successful implementations of these methods continue to be demonstrated, and it is likely that more studies and papers are in the works as we write this chapter. Our overarching piece of advice to those interested in adding a genomic perspective to their salamander research is to consider these methods in detail and choose the one that is best suited to your questions, your resources, and what you are personally willing to tackle.

Acknowledgments Our explorations of the use of genomic tools in salamanders have been generously funded by a range of organizations. We wish to thank the National Science Foundation for much of this support through awards DEB-0949532 and DEB-1355000 (to DWW), DEB-1601586 (to DWW and PMH), DEB-1601470 (to DWW and SON), DEB-1406876 (to DWW and JDK), a National Science Foundation Graduate Research Fellowship under Grant No. 3048109801 (to PMH), and a Blue Waters Graduate Research Fellowship from the National Center for Supercomputing Applications (to PMH). Our work in *Cryptobranchus* and *Necturus* has also been funded through the Kentucky Science and Engineering Foundation through awards KSEF-2913-RDE-016 and KSEF-3266-RDE-018 (both to DWW and SJ Price, UK Forestry). We also thank the many societies that have provided graduate student research grants, including the Society of Systematic Biologists, the Society for the Study of Evolution, the Society for the Study of Amphibians and Reptiles, the Society for Freshwater Science, the Kentucky Society of Natural History, and the Kentucky Academy of Sciences. Our research would not have been possible without the exceptional computational resources provided by the University of Kentucky Center for Computational Sciences and Lipscomb High Performance Computing Cluster. We also wish to thank the many people who have been instrumental in laying down the groundwork for much of research and providing invaluable consulting and assistance with lab work and computation, including S. Randal Voss, Jeramiah Smith, Catherine Linnen, Mary Foley, Alan Lemmon, and Emily Moriarty Lemmon.

References

Ali OA, O'Rourke SM, Amish SJ, Meek MH, Luikart G, Jeffres C, et al. RAD capture (Rapture): flexible and efficient sequence-based genotyping. Genetics. 2016;202:389–400.

Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH. Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. Genetics. 2011;188:799–808.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet. 2016;17:81–92.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One. 2008;3:e3376.

Beçak W, Beçak ML, Schreiber G, Lavalle D, Amorim FO. Interspecific variability of DNA content in Amphibia. Cell Mol Life Sci. 1970;26:204–6.

Bi K, Vanderpool D, Singhal S, Linderoth T, Moritz C, Good JM. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics. 2012;13:403.

Bryson RW Jr, Zarza E, Grummer JA, Parra-Olea G, Flores-Villela O, Klicka J, et al. Phylogenomic insights into the diversification of salamanders in the *Isthmura bellii* group across the Mexican highlands. Mol Phylogenet Evol. 2018;125:78–84.

Campbell NR, Harmon SA, Narum SR. Genotyping-in-thousands by sequencing (GT-seq): a cost effective SNP genotyping method based on custom amplicon sequencing. Mol Ecol Res. 2015;15:855–67.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. Stacks: building and genotyping loci de novo from short-read sequences. G3. 2011;1:171–82.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: an analysis tool set for population genomics. Mol Ecol. 2013;22:3124–40.

Czypionka T, Krugman T, Altmüller J, Blaustein L, Steinfartz S, Templeton AR. Ecological transcriptomics – a non-lethal sampling approach for endangered fire salamanders. Methods Ecol Evol. 2015;6:1417–25.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. 2011;12:499–510.

Eaton DA. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics. 2014;30:1844–9.

Elewa A, Wang H, Talavera-Lopez C, Joven A, Brito G, Kumar A, et al. Reading and editing the *Pleurodeles waltl* genome reveals novel features of tetrapod regeneration. Nat Commun. 2017;8:2286.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6: e19379.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst Biol. 2012;61:717–26.

Fan JB, Chee MS, Gunderson KL. Highly parallel genomic assays. Nat Rev Genet. 2006;7:632–44.

Feng YJ, Liu QF, Chen MY, Liang D, Zhang P. Parallel tagged amplicon sequencing of relatively long PCR products using the Illumina HiSeq platform and transcriptome assembly. Mol Ecol Res. 2016;16:91–102.

Fitzpatrick BM, Johnson JR, Kump DK, Smith JJ, Voss SR, Shaffer HB. Rapid spread of invasive genes into a threatened native species. Proc Natl Acad Sci U S A. 2010;107:3606–10.

Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhue C, Pudlo P, et al. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Mol Ecol. 2013;22:3165–78.

Gibbs HL, Denton RD. Cryptic sex? Estimates of genome exchange in unisexual mole salamanders (*Ambystoma* sp.). Mol Ecol. 2016;25:2805–15.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009;27:182–9.

Graham CF, Glenn TC, McArthur AG, Boreham DR, Kieran T, Lance S, et al. Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). Mol Ecol Res. 2015;15:1304–15.

Gregory TR. Animal genome size database. 2018. http://www.genomesize.com. Accessed 14 Apr 2018.

Guo Y, Long J, He J, Li CI, Cai Q, Shu XO, et al. Exome sequencing generates high quality data in non-target regions. BMC Genomics. 2012;13:194.

Habermann B, Bebin AG, Herklotz S, Volkmer M, Eckelt K, Pehlke K, et al. An *Ambystoma mexicanum* EST sequencing project: analysis of 17,352 expressed sequence tags from embryonic and regenerating blastema cDNA libraries. Genome Biol. 2004;5:R67.

Hara Y, Tatsumi K, Yoshida M, Kajikawa E, Kiyonari H, Kuraku S. Optimizing and benchmarking de novo transcriptome sequencing: from library preparation to assembly evaluation. BMC Genomics. 2015;16:977.

Hedtke SM, Morgan MJ, Cannatella DC, Hillis DM. Targeted enrichment: maximizing orthologous gene comparisons across deep evolutionary time. PLoS One. 2013;8:e67908.

Hime PM, Hotaling S, Grewelle RE, O'Neill EM, Voss SR, Shaffer HB, et al. The influence of locus number and information content on species delimitation: an empirical test case in an endangered Mexican salamander. Mol Ecol. 2016;25:5959–74.

Hoffberg SL, Kieran TJ, Catchen JM, Devault A, Faircloth BC, Mauricio R, et al. RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. Mol Ecol Res. 2016;16:1264–78.

Irisarri I, Baurain D, Brinkmann H, Delsuc F, Sire J-Y, Kupfer A, et al. Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nat Ecol Evol. 2017;1:1370–8.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science. 1992;258:818–21.

Kapusta A, Suh A, Feschotte C. Dynamics of genome size evolution in birds and mammals. Proc Natl Acad Sci U S A. 2017;114:E1460–9.

Keinath MC, Timoshevskiy VA, Timoshevskaya NY, Tsonis PA, Voss SR, Smith JJ. Initial characterization of the large genome of the salamander *Ambystoma mexicanum* using shotgun and laser capture chromosome sequencing. Sci Rep. 2015;5:16413.

Keinath MC, Voss SR, Tsonis PA, Smith JJ. A linkage map for the newt *Notophthalmus viridescens*: insights in vertebrate genome and chromosome evolution. Dev Biol. 2017;426:211–8.

Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: a resource for timelines, timetrees, and divergence times. Mol Biol Evol. 2017;34:1812–9.

Lemmon AR, Lemmon EM. High-throughput identification of informative nuclear loci for shallow-scale phylogenetics and phylogeography. Syst Biol. 2012;61:745–61.

Lemmon EM, Lemmon AR. High-throughput genomic data in systematics and phylogenetics. Annu Rev Ecol Evol Syst. 2013;44:99–121.

Lemmon AR, Emme SA, Lemmon EM. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst Biol. 2012;61:727–44.

Lepais O, Weir JT. SimRAD: an R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. Mol Ecol Res. 2014;14:1314–21.

Licht LE, Lowcock LA. Genome size and metabolic rate in salamanders. Comp Biochem Physiol B Biochem Mol Biol. 1991;100:83–92.

Linnen CR, Poh YP, Peterson BK, Barrett RD, Larson JG, Jensen JD, et al. Adaptive evolution of multiple traits through multiple mutations at a single gene. Science. 2013;339:1312–6.

Looso M, Preussner J, Sousounis K, Bruckskotten M, Michel CS, Lignelli E, et al. A de novo assembly of the newt transcriptome combined with proteomic validation identifies new protein families expressed during tissue regeneration. Genome Biol. 2013;14:R16.

Lucas LK, Gompert Z, Gibson JR, Bell KL, Buerkle CA, Nice CC. Pervasive gene flow across critical habitat for four narrowly endemic, sympatric taxa. Freshw Biol. 2016;61:933–46.

Madison-Villar MJ, Sun C, Lau NC, Settles ML, Mueller RL. Small RNAs from a big genome: the piRNA pathway and transposable elements in the salamander species *Desmognathus fuscus*. J Mol Evol. 2016;83:126–36.

Matsunami M, Kitano J, Kishida O, Michimae H, Miura T, Nishimura K. Transcriptome analysis of predator- and prey-induced phenotypic plasticity in the Hokkaido salamander (*Hynobius retardatus*). Mol Ecol. 2015;24:3064–76.

McCartney-Melstad E, Mount GG, Shaffer HB. Exon capture optimization in amphibians with large genomes. Mol Ecol Res. 2016;16:1084–94.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 2007;17:240–8.

Mohlhenrich ER, Mueller RL. Genetic drift and mutational hazard in the evolution of salamander genomic gigantism. Evolution. 2016;70:2865–78.

Murphy MO, Jones KS, Price SJ, Weisrock DW. A genomic assessment of population structure and gene flow in an aquatic salamander identifies the roles of spatial scale, barriers, and river architecture. Freshw Biol. 2018;63:407–19.

Newman CE, Austin CC. Sequence capture and next-generation sequencing of ultraconserved elements in a large-genome salamander. Mol Ecol. 2016;25:6162–74.

Nowoshilow S, Schloissnig S, Fei JF, Dahl A, Pang AW, Pippel M, et al. The axolotl genome and the evolution of key tissue formation regulators. Nature. 2018;554:50–5.

Nunziata SO, Lance SL, Scott DE, Lemmon EM, Weisrock DW. Genomic data detect corresponding signatures of population size change on an ecological time scale in two salamander species. Mol Ecol. 2017;26:1060–74.

Olmo E. Quantitative variations in the nuclear DNA and phylogenesis of the Amphibia. Caryologia. 1973;26:43–68.

O’Neill EM, Schwartz R, Bullock CT, Williams JS, Shaffer HB, Aguilar-Miguel X, et al. Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. Mol Ecol. 2013;22:111–29.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One. 2012;7:e37135.

Portik DM, Smith LL, Bi K. An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). Mol Ecol Res. 2016;16:1069–83.

Putta S, Smith JJ, Walker JA, Rondet M, Weisrock DW, Monaghan J, et al. From biomedicine to natural history research: EST resources for ambystomatid salamanders. BMC Genomics. 2004;5:54.

Qi Z, Zhang Q, Wang Z, Ma T, Zhou J, Holland JW, et al. Transcriptome analysis of the endangered Chinese giant salamander (*Andrias davidianus*): immune modulation in response to *Aeromonas hydrophila* infection. Vet Immunol Immunopathol. 2016;169:85–95.

Rodríguez A, Burgon JD, Lyra M, Irisarri I, Baurain D, Blaustein L, et al. Inferring the shallow phylogeny of true salamanders (*Salamandra*) by multiple phylogenomic approaches. Mol Phylogenet Evol. 2017;115:16–26.

Sessions SK. Evolutionary cytogenetics in salamanders. Chromosom Res. 2008;16:183–201.

Shen XX, Liang D, Feng YJ, Chen MY, Zhang P. A versatile and highly efficient toolkit including 102 nuclear markers for vertebrate phylogenomics, tested by resolving the higher level relationships of the Caudata. Mol Biol Evol. 2013;30:2235–48.

Shen XX, Liang D, Chen MY, Mao RL, Wake DB, Zhang P. Enlarged multilocus data set provides surprisingly younger time of origin for the Plethodontidae, the largest family of salamanders. Syst Biol. 2016;65:66–81.

Smith JJ, Kump DK, Walker JA, Parichy DM, Voss SR. A comprehensive expressed sequence tag linkage map for tiger salamander and Mexican axolotl: enabling gene mapping and comparative genomics in *Ambystoma*. Genetics. 2005;171:1161–71.

Smith JJ, Putta S, Zhu W, Pao GM, Verma IM, Hunter T, et al. Genic regions of a large salamander genome contain long introns and novel genes. BMC Genomics. 2009;10:19.

Sun C, Mueller RL. Hellbender genome sequences shed light on genomic expansion at the base of crown salamanders. Genome Biol Evol. 2014;6:1818–29.

Sun C, Shepard DB, Chong RA, Lopez Arriaza J, Hall K, Castoe TA, et al. LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. Genome Biol Evol. 2012;4:168–83.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013;31:46–53.

Voss SR, Smith JJ. Evolution of salamander life cycles: a major-effect quantitative trait locus contributes to discrete and continuous variation for metamorphic timing. Genetics. 2005;170:275–81.

Weisrock DW, Shaffer HB, Storz BL, Storz SR, Voss SR. Multiple nuclear gene sequences identify phylogenetic species boundaries in the rapidly radiating clade of Mexican ambystomatid salamanders. Mol Ecol. 2006;15:2489–503.

Wielstra B, Arntzen JW. Kicking *Triturus arntzeni* when it’s down: large-scale nuclear genetic data confirm that newts from the type locality are genetically admixed. Zootaxa. 2014;3802:381–8.

Wielstra B, Duijm E, Lagler P, Lammers Y, Meilink WR, Ziermann JM, et al. Parallel tagged amplicon sequencing of transcriptome-based genetic markers for *Triturus* newts with the ion torrent next-generation sequencing platform. Mol Ecol Res. 2014a;14:1080–9.

Wielstra B, Arntzen JW, van der Gaag KJ, Pabijan M, Babik W. Data concatenation, Bayesian concordance and coalescent-based analyses of the species tree for the rapid radiation of *Triturus* newts. PLoS One. 2014b;9:e111011.

Zieliński P, Stuglik MT, Dudek K, Konczal M, Babik W. Development, validation and high-throughput analysis of sequence markers in nonmodel species. Mol Ecol Res. 2014a;14:352–60.

Zieliński P, Dudek K, Stuglik MT, Liana M, Babik W. Single nucleotide polymorphisms reveal genetic structuring of the carpathian newt and provide evidence of interspecific gene flow in the nuclear genome. PLoS One. 2014b;9:e97431.

Zieliński P, Nadachowska-Brzyska K, Dudek K, Babik W. Divergence history of the Carpathian and smooth newts modelled in space and time. Mol Ecol. 2016;25:3912–28.