THESIS

THE EFFECTS OF GENOME EXPANSION ON TRANSPOSABLE ELEMENT DIVERSITY IN SALAMANDERS

Submitted by

Ava Haley

Department of Biology

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2021

Master's Committee:

Advisor: Rachel Mueller

Rachel Mueller
Daniel Sloan
Mark Stenglein

ABSTRACT

THE EFFECTS OF GENOME EXPANSION ON TRANSPOSABLE ELEMENT DIVERSITY IN SALAMANDERS

Transposable elements (TEs) are repetitive sequences of DNA that replicate and proliferate throughout genomes. Taken together, all the TEs in a genome form a diverse community of sequences, which can be studied to draw conclusions about genome evolution. TE diversity can be measured using ecological models for species distribution that consider richness and evenness of communities. It is currently not well studied how genome expansion impacts the diversity of transposable elements. However, there are a few models that predict TE diversity decreasing as genomes expand due to varying mechanisms such as selection against ectopic recombination and competition between TEs and silencing machinery. Salamanders are known to have some of the largest vertebrate genomes. Salamanders of the genus *Plethodon* in particular have very large genomes consisting of high levels of TEs, with sizes ranging from 30 to 70 Gigabases (Gb). Here, I use Oxford Nanopore sequencing to generate low-coverage genomic sequences for four species of *Plethodon* that encompass two independent genome expansion events, one in the eastern clade and one in the western clade: *Plethodon glutinosus* (41.4 Gb), *P. cinereus* (30.5 Gb), *P. idahoensis* (71.7 Gb), and *P. vehiculum* (50.5 Gb). I classified the TEs in these datasets using RepeatMasker and DnaPipeTE and found ~51 superfamilies which accounted for 27-32% of the genomes. For each genome I calculated the Simpson's and Shannon's diversity indices to quantify diversity, taking into account both TE richness and evenness. In all cases, the values for Simpson's index were within 0.75 and 0.79, and for Shannon's index all species were within 1.88 and 1.99. We conclude that once genomes

reach large sizes, they maintain high levels of TE diversity at the superfamily level, in contrast to observations made by previous studies done on smaller genomes.

ACKNOWLEDGEMENTS

This study could not have been completed without the guidance and expertise of my thesis advisor and mentor, Dr. Rachel Mueller. I want to express my thanks and gratitude for her dedication, help, and constant enthusiasm towards my research and thesis writing. I could not have imagined a better advisor and mentor for my master's project, and she made my experience at Colorado State University unforgettable.

I would also like to thank my committee member Dr. Daniel Sloan for his constant support and help with many of the computing tasks required to complete my project. I could not have completed my research in a timely manner without his guidance. His introduction to genetics course was the reason I became interested in genomics in the first place, and to have the opportunity to have him as a committee member was incredibly meaningful to me.

I'd like to also thank my other committee member, Dr. Mark Stenglein, for his support, guidance, and advice for efficiently completing my research project. I also want to thank my friend and lab mate, Michael Itgen, for teaching me many of the skills needed to complete this project, and for sparking my interest in bioinformatics. He has been a constant support throughout my entire graduate school experience.

TABLE OF CONTENTS

INTRODUCTION

Genome sizes vary greatly among eukaryotes, encompassing sizes as small as 0.002 Gb in the eukaryotic fungi *Encephalitozoon cuniculi* and as large as 670 Gb in *Amoeba dubia*, a difference of 330,000-fold (Gregory 2021). Among animals, the differences span 6,650-fold; the smallest animal genome is 0.02 Gb in the nematode *Pratylenchus coffeae* and the marbled lungfish *Protopterus aethiopicus* has the largest animal genome at ~133 Gb (Gregory 2021). Human genomes are only 3.4 Gb (Gregory 2021). Salamanders include many of the largest animal genomes, reaching as large as 120 Gb in *Necturus lewisi*. The salamander clade also includes substantial genome size variation. Across species, salamander genomes range between 10 Gb and 120 Gb (Herrick and Sclavi 2018). This 12-fold range is wide in comparison to other vertebrate clades. For example, in mammals, genome sizes range from 1.6 to 6.3 Gb, and in birds, the range is only ~1 to 2.2 Gb (Kapusta et al. 2017).

The main proximate cause for such large genomes in salamanders is the proliferation of transposable elements (Sun and Mueller 2014), which underlie differences in genome size across diverse taxa (Muñoz-López and García-Pérez 2009). Transposable elements (TEs) are DNA sequences that replicate and insert themselves throughout the genome. As transposable elements are constantly replicating, copy number tends to increase. The fraction of the genome made up of TE varies greatly across the tree of life, ranging from just 0.1% in the fungi *Pseudozyma antarctica* (Castanera et al. 2017), all the way up to ~90% in a species of lily, *Fritillaria imperialis* (Ambrožová et al. 2011).

In salamander genomes, up to 47% of the total DNA is classified as recognizable repeats (Sun et al. 2012). This percentage does not include older TEs that have accumulated mutations

over time and become unrecognizable by annotation software, so the genome is likely to be made up of significantly more transposable element-derived sequences than are detected. Because the majority of TEs serve no function in the genome as far as coding for proteins or regulating gene expression, they eventually accumulate mutations leading to decay, causing them to go unnoticed during TE annotation (Venner et al. 2009). These mutated TEs are effectively neutral, as they do not typically serve any immediate functions and aren't deleterious to the organism (Bennetzen and Park 2018). The large amount of variation in salamander genome sizes reflects different species containing different amounts of transposable elements, with some genomes only being made up of 25% detectable transposable elements (Sun et al. 2011).

Transposable elements are broken down into two broad classes. The first class of transposable elements are retrotransposons, which replicate by utilizing the host's transcriptional machinery to create an RNA intermediate. The RNA intermediate is then reverse transcribed into a cDNA copy and inserted back into the genome using the TE enzymatic machinery, a mechanism referred to as "copy and paste" (Bourque et al. 2018). The second class of transposable elements are DNA transposons, which do not have an RNA intermediate and instead move as the direct, excised DNA sequence itself, reinserting into a different location in the genome. This mechanism is referred to as "cut and paste" (Muñoz-López and García-Pérez 2009). Each class contains both autonomous and non-autonomous TEs. Autonomous TEs contain open reading frames coding for transposition proteins, while non-autonomous TEs do not contain these ORFs, but transpose using transposase enzymes that are "borrowed" from other nearby transposable elements. LINE retrotransposons are an example of autonomous TEs, and SINE retrotransposon elements are an example of nonautonomous TEs (Wessler 2006).

Transposable elements can be further broken down into 9 orders and 39 superfamilies, which are commonly classified using a unified system designed by Thomas Wicker based on structure and length, sequence similarity, and replication strategy (Wicker et al. 2007). However, classification systems are often updated as new discoveries are made, and more diverse classifications also exist that can be utilized alongside Wicker's classification, such as classifications that incorporate more evolutionary or phylogenetic approaches (Jurka et al. 2005, Arkhipova 2017, Goerner-Potvin and Bourque 2018). Some superfamilies can be found in almost all eukaryotes, such as *Gypsy* and *Copia* of the LTR order (Borque et al. 2018). However, most of these classes and their superfamilies are highly variable within different genomes, with some superfamilies existing at significantly higher or lower frequencies than others depending on the species. For example, in the caecilian *I. bannanicus*, the retrotransposon DIRS makes up ~30% of the genome and the retrotransposon LTR/Gypsy makes up ~1% (Wang et al. 2020), while in salamanders, LTR/Gypsy is the most abundant (Wang et al. 2020). In contrast, class II DNA transposons make up the majority of transposable elements in actinopterygian (ray-finned fish) genomes at 35-60%, while retrotransposons exist at a significantly lower amount (Sotero-Caio et al. 2017). Superfamilies can be further classified into families and subfamilies.

Transposable elements replicating and spreading in genomes can cause unwanted deleterious effects, such as interrupting functioning genes (Parhad and Theurkauf 2019), expending host energy by using up replication, transcription, and translation factors (Cavalier-Smith 2005), impacting regulation of gene expression (Meagher and Vassiliadis 2005), and causing ectopic recombination-mediated deletions and duplications (Lim and Simmons 1994). Due to these negative impacts on the genome and its host, silencing mechanisms such as the PIWI-interacting RNA pathway, or piRNA pathway, have evolved to restrict transposable

element proliferation (Lisch and Bennetzen 2011). The pathway relies on proteins complexed with small RNAs to silence TE loci transcriptionally and post-transcriptionally (Aravan 2007). TEs can also be deleted through recombination (Devos et al. 2002). However, the variation in genome size and transposable element content across eukaryotes suggests that these silencing and deletion mechanisms differ in activity across organisms (Mueller 2018).

As genomes expand, the number of transposable elements increases (Kidwell 2002). However, how exactly the diversity of the transposable element community in a genome changes with expansion is not yet well understood (Elliot and Gregory 2015). The diversity of TE communities within genomes can be measured in a similar manner to species diversity in ecological communities (Venner et al. 2009, Linquist et al. 2015). Measuring diversity in ecological communities considers the number of species, or richness, and the abundance of each species, or evenness. Though diversity in TEs hasn't been as well-studied, it can be viewed in an analogous way as the richness and evenness of active TE classes, orders, or superfamilies in a genome (Wang et al. 2020). One way this has been done is by using the Simpson and Shannon diversity indices (Nunes et al. 2010, Shannon 1948).

The genomes of different species can differ in TE diversity, independent of differences in overall TE amounts or superfamilies present (Abrusan and Krambeck 2006). For example, although different genomes may be populated by the same superfamilies, their diversity can differ if one genome contains one TE superfamily that makes up the majority of the TE content, while another genome may contain a relatively even amount of each superfamily (Abrusan and Krambeck 2006). Similarly, two genomes with very different transposable element makeups may end up with the same diversity indices, due to the same richness and same evenness, regardless of having different superfamilies (Wang et al. 2021).

Although the relationship between genome size and transposable element diversity has not been widely studied, several analyses suggest that transposable element diversity will decrease with increasing genome size. Petrov et al. (2003) suggested that, because ectopic recombination is more likely to cause harmful effects in smaller genomes due to the higher chance of deletion or duplication interrupting a functioning gene, smaller genomes will select for more diverse TE communities. With more diverse TE communities, ectopic recombination is less likely to occur because there are fewer identical off-target sites to drive errors in crossing-over. In large genomes, the chances of interrupting a functioning gene during ectopic recombination-mediated deletion or duplication are lower. In addition, recombination rates per base pair are lower, which decreases the likelihood of ectopic recombination overall. Thus, larger genomes can be more permissive to low-diversity TE communities (Petrov 2003).

Furano et al. (2004) suggested that, because ectopic recombination can cause harmful deletions, genomes with lower recombination rates (as well as lower ectopic recombination rates) are more permissive to TE activity. TEs therefore increase, which causes genome size to increase. As genome size increases, one type of TE may more successfully exploit host enzymes, outcompeting the other TEs. This would result in decreased TE diversity (Furano 2004). Finally, Boissinot predicts that genomes with lower recombination rates have higher insertion levels of active TEs, which leads to an arms race between the host silencing mechanisms and the TE superfamily proliferation mechanisms, leading to decreased TE diversity (Boissinot 2016).

It has been previously observed that in small genomes (<1000Mb), the diversity of the TE landscape tends to be maximized, with the most TE superfamilies being present. For example, in *Branchiostoma floridae* (lancelet) and *Bombyx mori* (silkworm) which both have <600Mb genomes, researchers found 39 active TE superfamilies (Elliot and Gregory 2015).

However, as genomes start to expand, it appears that there is no monotonic relationship between genome size and TE superfamily richness and instead, richness increases and then decreases with increasing genome size. In reality, there is still an unclear relationship between genome size and TE diversity (Elliot and Gregory 2015).

The current study was done with the goal of testing the hypothesis that TE diversity decreases with genome expansion. In order to achieve this, we chose the salamander genus *Plethodon* as a study system due to a wide range of genome sizes, but high similarity in physical traits and life history. This genus of salamander contains at least 55 species (Highton 2012) with a huge amount of diversity in their genome sizes, but not a lot of diversity in their physical appearances (Figure 1). Two species' genomes from the two main *Plethodon* clades were sequenced — *Plethodon cinereus* and *Plethodon glutinosus* from eastern North America, and *Plethodon vehiculum* and *Plethodon idahoensis* from the Pacific Northwest. Within each of these clades, there exists a wide range in genome size, with the eastern species ranging from 30 to 41 Gb, and the western species ranging from 50 to 72 Gb (Itgen et al. 2021). With such a large range in genome sizes, we can observe how the transposable element diversity changes as the genomes drastically grow in size within two independent genome expansion events.

The genus *Plethodon* is at least 52 million years old (Kumar et al. 2017). Woodland salamanders are part of the only family of lungless salamanders, Plethodontidae. They are only found in North America, more commonly along cool and moist regions in the East and West. These salamanders live in a moist environment, and due to the lack of lungs, they respire through their skin (AmphibiaWeb 2021). *Plethodon* salamanders do not have an aquatic larval stage, and instead lay their eggs on land and go through direct development (Highton 1962). The eastern clade is concentrated along the east coast and Appalachian Mountains, and the western species

are only found in the Pacific Northwest. *Plethodon* are carnivorous animals (Howard 2003), and typically are found underneath rotting vegetation and debris (Mizuno and Macgregor 1974). These salamanders are a good indicator of a healthy habitat, as they are highly sensitive to pH and other environmental changes (Welsh and Hodgson 2013).

Thus far, the majority of projects examining transposable element biology have been done using annotations from fully assembled genomes, typically from short read sequencing, or a combination of short and long sequencing (Shahid and Slotkin 2020). Studying and annotating larger genomes has been challenging because of the large number of TEs that make up many big genomes, as genomes with a large number of repeats often result in misassembled TEs (Shahid and Slotkin 2020); human genome assemblies are 16.2% shorter than expected due to the loss of TEs during assembly (McCoy 2014). In addition, TEs can occasionally be mis-annotated as genes (Bennetzen and Park 2018).

Sequencing technology has advanced significantly over the last two decades. Short read sequencing, such as Illumina, has been widely used since 2005 (Berglund et al. 2011). Long read sequencing, such as Oxford Nanopore and Pacific Biosciences — also referred to as third generation sequencing — became available in 2011 (Amarasinghe 2020). Since then, long read sequencing has increased in popularity due to its many advantages. Long read sequencing can produce reads $\geq$ 10 Kbp in length (Amarasinghe 2020), versus the typical short read sequencing length of up to 300 bp. Longer reads are also able to increase classification accuracy, as well as improving de novo assembly accuracy (Amarasinghe 2020). Specific to studies of TE biology, long reads are much more likely to contain entire TE sequences on a single read, versus short reads that first need to be assembled to reconstruct possible full-length TEs (Shahid and Slotkin 2020).

While the increased error rate of long read sequencing is considered significant in some studies, it has also been shown in previous studies that TE detection is more successful with longer reads, even with low coverage (Shahid and Slotkin 2020). Additionally, long read sequencing has been previously proven to be effective at classification despite the error rate; when using BLAST against reference genomes to classify reads of both Illumina and Oxford Nanopore data separately, 97% of animal reads from 20 different genomes were correctly classified down to the genus level (Pearman et al. 2020). While it is not possible to do sequence-level comparisons among individual TE insertions with Oxford Nanopore data because of the error rate, these previous studies suggest that TEs are able to be classified down to the superfamily level accurately.

In this study, we test the hypothesis that the diversity of the TE landscape decreases with genome expansion by comparing TE diversity indices for four species of *Plethodon* salamanders that capture two independent instances of genome expansion. We rely exclusively on Oxford Nanopore long read sequencing data with no existing genome assembly to reference, demonstrating the power of this method for quantifying TE community diversity. Using both Simpson and Shannon's diversity indices, we find that there is no support for a decrease in TE diversity at the superfamily level as genome size expands. Although diversity indices in the larger genomes are slightly lower, the difference between the species in each clade is negligible.

Figure 1 from top to bottom: Plethodon cinereus, Plethodon glutinosus, Plethodon vehiculum, Plethodon idahoensis (Amphibiaweb, Todd Pierson, William Flaxington, Gary Nafis)

MATERIALS AND METHODS

Tissue Collection

Plethodon cinereus and *P. glutinosus* were collected from South Cherry Valley and Oneonta, Otsego County, New York, under the New York State Department of Environmental Conservation scientific collection permit #2303. *Plethodon vehiculum* was collected from Pacific County, Washington, under the scientific collection permit # ITGEN 17-309 issued by the Washington Department of Fish and Wildlife. *Plethodon idahoensis* was collected in Shoshone County, Idaho, under the wildlife collection permit #180226 issued by the Idaho Department of Fish and Game.

Animals were euthanized via submersion in 10% buffered MS222. Tissues were collected and stored in RNALater at -20ºC. All work was completed according to the Mueller Lab's IUCAC protocols (17-7189A).

DNA Extraction

DNA extraction was performed using a Qiagen DNeasy Blood and Tissue kit. 0.2g of trunk skin and muscle tissue was used for each species. The manufacturer's protocol was followed, except for the following: vortexing the samples was foregone in order to retain the longest fragments possible. Every few hours during incubation, the tubes were flicked to ensure efficient lysing. Each centrifuge time was doubled to ensure all solution passed through the spin column. 30 µl of elution buffer was used instead of the suggested 200 µl in the final step, in order to increase DNA concentration.

DNA was quantified using 1µl per sample in a Qubit Fluorometer.

Library Preparation

Library preparation was done using a Ligation Sequencing Kit (SQK-LSK109), a Flow Cell Priming Kit (EXP-FLP002), and a Native Barcoding Expansion 13-24 (EXP-NBD114) from Oxford Nanopore. New England Biolabs consumables included a NEB Blunt/TA Ligase Master Mix (M0367), NEBNext® Quick Ligation Reaction Buffer (NEB B6058), and NEBNext® Companion Module for Oxford Nanopore Technologies® Ligation Sequencing (cat # E7180S).

DNA Repair and End Prep

2 µg of genomic DNA was used instead of the suggested 1 µg, due to prior less successful runs prompting the use of more DNA. 1 µl of each end-prepped sample was quantified using a Qubit fluorometer.

Native Barcode Ligation

Four distinct barcodes were used in this experiment, one per sample. 1000 ng of each end-prepped sample was used for barcode ligation. Note that this is twice the amount of suggested sample per the protocol, but earlier less successful runs prompted the use of more DNA.

Two samples at a time were pooled together into a clean 1.5 ml Eppendorf tube. *P. glutinosus* and *P. cinereus* were pooled together, and *P. vehiculum* and *P. idahoensis* were pooled together to equal about 850 ng of DNA per tube, slightly more than the 700 ng suggested by the protocol. Pooled samples were quantified with 1 µl on the Qubit fluorometer.

Adapter ligation and clean-up

We only wanted to enrich for sequences greater than 3 kb, so the Long Fragment Buffer was used. Adapter-ligated DNA libraries were quantified with 1 µl on the Qubit Fluorometer.

Priming and loading the SpotON flow cell

This part of the procedure was performed two separate times, with only two species occupying one flow cell.

DNA sequencing

Sequencing was done on the Oxford Nanopore MinION sequencer with the MinKnow software. The sequencer was run for 72 hours with the base calling setting of extremely fast.

Transposable Element Annotation

Porechop was used to trim adapters and barcodes (Wick et al. 2017).

There are a variety of transposable element annotation programs available, but most of them are geared towards complete genome assemblies or short-read shotgun sequencing data. Finding a pipeline that would efficiently work on low-coverage MinION data was a challenge. Our goals were 1) to find the most effective annotation tools possible, enabling accurate calculation of the diversity indices for each genome, and 2) to achieve consistent annotation levels across species, allowing them to be compared without the introduction of bias. In a previous study annotating TEs in the caecilian amphibian *Ichthyophis bannanicus*, RepeatMasker and DnaPipeTE together were responsible for annotating 94.1% of the TE sequences (Wang et al. 2020). Additionally, a TE annotation study done on a beetle *Dichotomius (Luederwaldtinia) schiffleriso* also found that RepeatMasker and DnaPipeTE together annotated 95% of all of the detected TEs in the genome (Amorim 2020). Therefore, these two programs together were chosen based on their previous successful application to characterizing TE landscapes. DnaPipeTE is a program that detects TE sequences based on repetitiveness by using the program Trinity to assemble repeats using low-coverage data. RepeatMasker uses a user-specified TE library to locate interspersed repeats that are likely to be transposable elements, as well as low-complexity DNA. Typically, RepeatMasker is used to mask detected TEs from the

genome of interest in order to allow analysis of the non-repetitive portions, but for studies focused on TE biology such as this one, the sequences identified by RepeatMasker become the subject of downstream analysis.

Using those two programs, my pipeline was completed as follows: 1) Raw trimmed reads were queried using RepeatMasker against both RepBase and a custom repeat library, which contained known TEs from six other salamanders from the family Plethodontidae (*Aneides flavipunctatus, Desmognathus ochrophaeus*, *Batrachoseps nigriventris, Bolitoglossa occidentalis, Bolitoglossa rostrata, and Eurycea tynerensis)* as well as the hellbender salamander, *Cryptobranchus alleganiensis* (Sun et al. 2011, Sun and Mueller 2014). Raw trimmed reads were also run through DnaPipeTE. 2) Repetitive sequences identified using DnaPipeTE were queried using RepeatMasker against the salamander library for annotation, as they are only annotated down to TE order level by DnaPipeTE itself. 3) A custom Perl script was used to parse out each RepeatMasker TE based on its base pair location within each read, as many reads contained multiple TEs. 4) Finally, the repetitive sequences detected by DnaPipeTE and TEs detected by RepeatMasker were combined for each species to characterize the total TE landscape for each species. We are assuming that the sequence data is a random subsample of the total genome sequence.

Measuring Diversity

TE diversity was measured for each species using both the Simpson's and Shannon diversity indices in two different ways. In both methods, TE superfamilies were considered as species. In the first method, the total numbers of detected TE sequences annotated to each superfamily were considered as the number of individuals per "species." In the second method, the total numbers of base pairs for each annotated superfamily were used for total presence of

individuals per "species." The second method differs from the first in that using base pair measurements takes into account the different sizes of TEs, as some can be significantly longer than others and therefore take up more space in the genome. Unknown repeats (i.e. sequences identified by DnaPipeTE as repetitive, but not classified as known TEs) were excluded from the analysis, as were TEs that could only be annotated down to the level of Class (i.e. LTR). Simpson's diversity index considers both the number of species present in a population, as well as the relative abundance of each of those species. Diversity increases as both richness and evenness increase. Simpson's diversity index is expressed as the variable D, calculated by: $D = \frac{\sum n(n-1)}{N(N-1)}$, with n being the total number of individuals per species and N being total number of all individuals. D is the probability that two individuals at random pulled from a community will be from the same species. Since diversity decreases as D increases, this number is often expressed as 1 – D, or the Gini-Simpson's index instead, which is more intuitive. The greater the value of 1 – D, the greater the diversity (Nunes et al. 2010). The Shannon's diversity index is similar to Simpson's index in that it takes into account both richness and evenness of an ecological community and is represented by the variable H, which is calculated by: $H = -\sum_{i=1}^{s} p_i \, ln \, p_i$, with p_i being the proportion of species (*i)* relative to the total number of species. H increases as evenness and richness increase, so the higher the value of H, the greater the diversity (Shannon 1948). Shannon's diversity index is more sensitive to sample size and rarer species than Simpson's index is (Mouillot and Leprêtre 1999), so the Shannon index may be a more accurate representation of genome diversity because of the presence of many low frequency repeats. However, it's also possible that with low coverage data rare repeats may go undetected, so it's important to use both indices with this type of dataset.

RESULTS

For *Plethodon vehiculum* and *P. idahoensis,* the MinION generated 2.11 Gb of data and 512,830 reads, with an N50 of 7.49 kb (Figure 2). For *Plethodon glutinosus* and *P. cinereus,* the MinION generated 4.15 Gb of data and 1.22 million reads, with an N50 of 6.59 kb (Figure 2).

The combined outputs for RepeatMasker and DnaPipeTE resulted in the following amounts of repeats for each species: 1,476,209 for *P. glutinosus*, 2,153,518 for *Plethodon cinereus,* 898,214 for *P. idahoensis,* and 807,344 for *P. vehiculum.* In all four species, ~99% of the repeats were detected by RepeatMasker, and the remainder by DnaPipeTE.

All four species contained at least 52 superfamilies. Using both methods of calculating diversity – the total number of detected repeats and the total number of base pairs for each TE superfamily – the superfamily *Gypsy* of the LTR order made up the majority of the repeats of all four genomes, taking up 23% and 33% of *Plethodon glutinosus* (total number of individual repeats and total base pairs, respectively), 20% and 27% in *P. cinereus,* 31% and 36% in *P. idahoensis,* and 24% and 30% in *P. vehiculum.* The second most abundant TE superfamily in all four species was *L2* from the order LINE, making up 17% and 15% of the repeats in *P. glutinosus,* 19% and 18% of *P. cinereus,* 16% and 19% of *P. idahoensis,* and 18% and 18% of *P. vehiculum.* The third most abundant superfamily was *DIRS* of the order DIRS, making up 8% and 11% of the total repeats in *P. glutinosus,* 9% and 13% in *P. cinereus,* 9% and 11% in *P. idahoensis,* and 8% and 11% in *P. vehiculum.* The least abundant superfamily across all four species was the DNA transposon Sola, in the TIR order. This TE only made up 0.0002% and 0.0006% of *P. glutinosus* and *P. cinereus* TEs and base pairs, 0.0002% and 0.0007% of *P. idahoensis,* and 0.0001% and 0.0005% of *P. vehiculum.* Many repeats were unable to be

classified, and we refer to those as "unknown." 31% and 22% of the total repeats were classified as "unknown" in *P. glutinosus*, 30% and 21% in *P. cinereus,* 22% and 15% in *P. idahoensis,* and 26% and 20% in *P. vehiculum* (Table 1).

Using the total number of base pairs classified as repeats compared to the total number of base pairs sequenced, 28% of the *P. glutinosus* genome was classified as known TEs, 27% of the *P. cinereus* genome, 32% of the *P. idahoensis* genome, and 30% of the *P. vehiculum* genome. In contrast, 8% of the *P. glutinosus* genome was classified as unknown repeats, 7% of the *P. cinereus genome,* 6% of the *P. idahoensis* genome, and 7% of the *P. vehiculum* genome (Table 1).

The eastern clade (*P. glutinosus* and *P. cinereus*) contained two superfamilies that were not detected in the western clade: DNA transposon Zator, and Retrotransposon Proto2. *Plethodon idahoensis* contained a superfamily that was not detected in the other three species, Tad1 of the LINE order. *Plethodon vehiculum* was the only species that lacked the superfamily R1, a LINE element. The total makeup of TEs of each genome can be seen in Table 1.

Using the total numbers of classified repeats, the Simpson diversity index ranged from 0.80 to 0.82 in *Plethodon glutinosus* and *P. cinereus* and 0.78 to 0.81 in *Plethodon idahoensis* and *P. vehiculum* (Table 2). The Shannon's diversity index ranged from 2.05 to 2.16 in *P. glutinosus* and *P. cinereus* and 2.03 to 2.12 in *P. idahoensis* and *P. vehiculum* (Table 2). Using the total numbers of base pairs occupied by each TE, the Simpson diversity index ranged from 0.76 to 0.79 in *Plethodon glutinosus* and *P. cinereus* and 0.75 to 0.78 in *Plethodon idahoensis* and *P. vehiculum* (Table 2). The Shannon's diversity index ranged from 1.88 to 1.99 in *P. glutinosus* and *P. cinereus* and 1.85 to 1.96 in *P. idahoensis* and *P. vehiculum* (Table 2). The diversity indices are slightly lower for the larger genomes within each pairwise comparison.

However, this pattern isn't seen when we rank all four species by genome size and compare indices; diversity does not decrease monotonically as genome size increases.

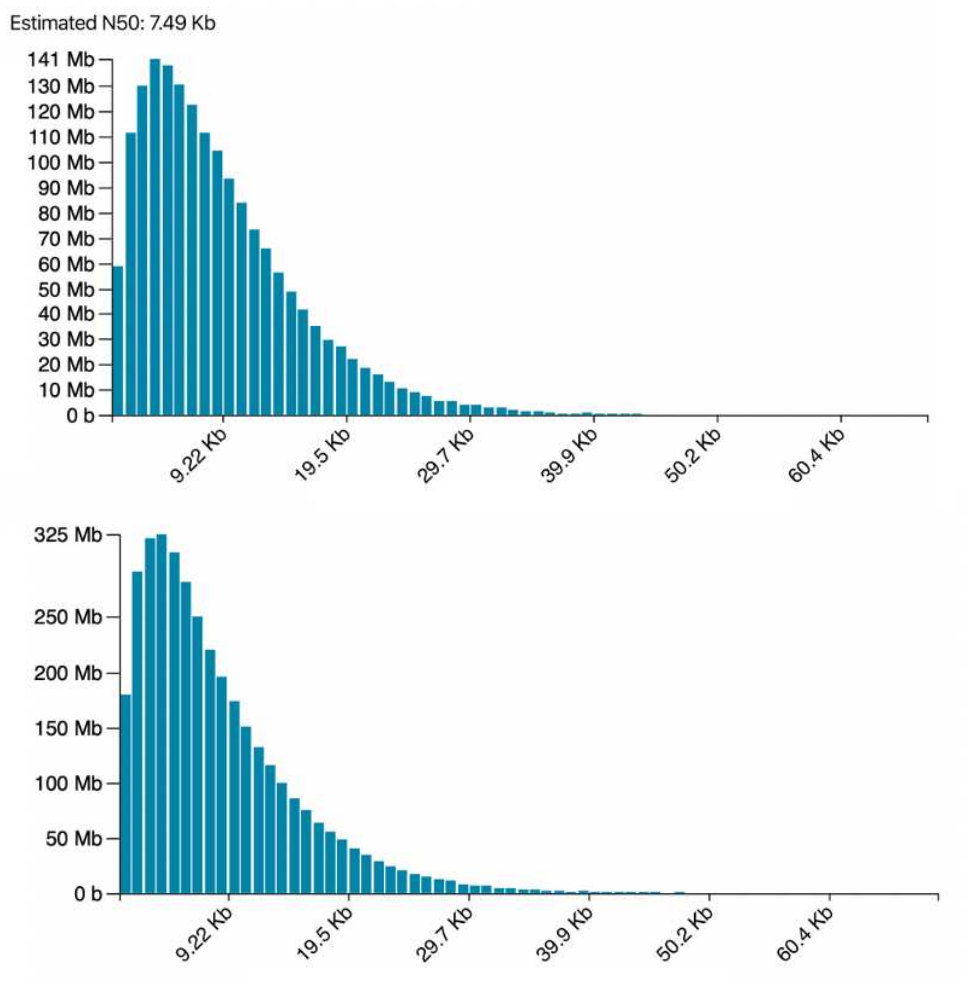

Figure 2 Top: Read lengths for Plethodon vehiculum and idahoensis Bottom: Read lengths for Plethodon glutinosus and cinereus. Note that the axes are different, reflecting that more reads were sequenced for Plethodon glutinosus and cinereus.

Table 1 Summary of repeats in the four genomes. Asterisks (*) indicate that the superfamily was detected at <0.001%

Order	**Superfamily**	**% of total repeats (individual repeats)**				**% of total repeats (base pairs occupied by repeats)**			
		P. glutinosus	*P. cinereus*	*P. idahoensis*	*P. vehiculum*	*P. glutinosus*	*P. cinereus*	*P. idahoensis*	*P. vehiculum*
Class I - Retrotransposon - Autonomous									
LTR	*ERV*	4.609	4.331	3.249	3.801	6.802	6.642	4.422	5.713
	Gypsy	23.411	20.004	30.806	24.438	33.025	27.487	36.072	29.924
	Bel-Pao	0.006	0.008	0.008	0.006	0.002	0.002	0.003	0.002
	Copia	0.122	0.108	0.077	0.205	0.103	0.102	0.081	0.252
	Bhikari	0.108	0.116	0.054	0.085	0.093	0.092	0.039	0.057
	Unknown LTR	0.024	0.012	0.030	0.008	0.013	0.006	0.040	0.004
DIRS	*DIRS*	7.818	8.523	8.573	7.974	10.610	12.531	10.766	11.459
LINE	*Ngaro*	0.335	2.188	0.069	0.385	0.396	2.552	1.431	0.497
	Penelope	1.965	1.242	0.710	0.970	0.981	0.849	0.484	0.628
	Jockey	0.022	0.030	0.036	0.062	0.020	0.025	0.028	0.051
	L1	2.870	3.287	4.435	5.473	1.575	1.905	2.607	3.137
	L2	17.284	18.804	16.297	17.622	15.257	17.697	18.921	17.506
	RTE	1.130	1.244	0.961	1.113	0.805	0.961	0.737	0.864
	R1	0.002	0.001	0.015	0.000	0.001	0.001	0.013	0.000
	R2	0.001	0.002	0.003	0.002	0.000*	0.001	0.002	0.001
	I	0.019	0.024	0.028	0.013	0.010	0.014	0.016	0.007
	CR1	0.986	1.320	0.743	0.900	0.649	0.934	0.517	0.681
	Proto2	0.000*	0.000*	0.000	0.000	0.000*	0.000*	0.000	0.000
	Tad1	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.000
	Unknown LINE	0.077	0.085	0.069	0.067	0.056	0.071	0.053	0.054
Class I - Retrotransposon - Non-autonomous									
SINE	*7SL*	0.002	0.003	0.004	0.000	0.001	0.002	0.002	0.000
	5S	0.027	0.027	0.125	0.131	0.009	0.010	0.046	0.054
	tRNA	0.143	0.178	0.140	0.123	0.041	0.054	0.055	0.038
	B4	0.007	0.012	0.005	0.006	0.002	0.004	0.001	0.002
	Deu	0.897	1.040	1.151	1.829	0.478	0.580	0.624	1.003
	MIR	0.368	0.428	0.120	0.271	0.176	0.217	0.060	0.145
	Unknown SINE	0.001	0.001	0.002	0.001	0.001	0.001	0.002	0.001
Class II - DNA Transposon - Subclass 1									
TIR	*hAT*	1.588	1.258	0.384	0.421	1.325	0.877	0.209	0.235
	Tc1-Mariner	0.629	0.909	1.305	0.688	0.382	0.625	0.828	0.426
	PIF-Harbinger	1.511	1.582	2.379	2.302	1.261	1.287	1.452	1.868
	PiggyBac	0.186	0.209	2.287	1.779	0.108	0.127	1.581	1.014
	Sola	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*
	MuDR	0.014	0.016	0.024	0.042	0.008	0.009	0.013	0.048
	P	0.006	0.004	0.002	0.004	0.002	0.002	0.001	0.002
	Zisupton	0.058	0.075	0.125	0.058	0.029	0.046	0.051	0.030
	Kolobok	0.019	0.029	0.012	0.006	0.011	0.019	0.005	0.003
	Academ	0.015	0.020	0.004	0.009	0.011	0.015	0.003	0.007
	Dada	0.006	0.011	0.005	0.004	0.003	0.006	0.002	0.002
	Ginger	0.016	0.038	0.017	0.040	0.007	0.016	0.008	0.017
	IS3EU	0.003	0.004	0.004	0.003	0.002	0.003	0.001	0.002

	MULE	0.008	0.011	0.008	0.005	0.003	0.005	0.003	0.002
	Merlin	0.004	0.004	0.006	0.003	0.001	0.001	0.002	0.001
	CMC/En-Spm	0.099	0.159	0.219	0.171	0.046	0.076	0.097	0.014
	Novosib	0.000*	0.001	0.001	0.000*	0.000*	0.000*	0.000*	0.000*
	Zator	0.011	0.015	0.000	0.000	0.014	0.023	0.000	0.000
Crypton	*Crypton*	0.018	0.017	0.011	0.007	0.006	0.008	0.004	0.004
Maverick	*Maverick*	0.480	0.610	0.584	0.934	0.690	1.064	0.792	1.225
Helitron	*Helitron*	1.546	1.330	1.755	1.662	2.739	2.355	3.236	3.317
Unknown Superfamilies		0.103	0.108	0.117	0.122	0.071	0.084	0.077	0.097
Unable to be classified		31.337	30.385	21.898	26.200	22.142	20.545	14.574	19.586

Table 2 Simpson and Shannon's diversity indices for all four species

	Gini-Simpson's Diversity Index (1-D) ***Using total TE copy number***	*Shannon's Diversity Index (H)* ***Using total TE copy number***	*Gini-Simpson's Diversity Index (1-D)* ***Using total base pair number***	*Shannon's Diversity Index (H)* ***Using total base pair number***
Plethodon glutinosus – 41.4 Gb	0.80	2.05	0.76	1.88
Plethodon cinereus – 30.5 Gb	0.82	2.16	0.79	1.99
Plethodon idahoensis – 71.7 Gb	0.78	2.03	0.75	1.85
Plethodon vehiculum – 50.5 Gb	0.81	2.12	0.78	1.96

DISCUSSION

TE superfamily diversity does not decrease with genome expansion

Using both the Simpson and Shannon's diversity indices, our results do not support the hypothesis that TE superfamily diversity decreases as genome size increases in the 30 – 70 Gb size-range. The small differences in diversity indices are unlikely to reflect meaningful differences in TE diversity. In previous studies done on transposable element diversity using several vertebrate models, we see a two-fold difference in diversity indices in genomes only 0.9 Gb different from one another – the pufferfish genome (0.4 Gb) has diversity indices of 1.0 (Simpson) and 2.1 (Shannon), indicating high diversity, while the chicken genome (1.3 Gb) has diversity indices of 0.50 and 0.90, indicating very low diversity (Wang et al. 2021). With such a large difference in indices at a small absolute increase in genome size (0.9 Gb), it would be expected that the salamander indices would differ more if they were in fact different from one another in diversity, as the salamander genomes are different from each other by 10 to 20 Gb.

The diversity indices calculated here for *Plethodon* (Table 1) are similar to other salamander diversity indices. Wang et al. (2021) calculated these values for five different salamander species, and the values were between 0.71 and 0.79 for Gini-Simpson's index, and 1.61 and 2.26 for Shannon's index (Wang et al. 2020) With these salamanders, there was also no pattern observed between diversity and genome size, but it was also an imperfect study system as the species studied are highly phylogenetically diverged, meaning they will have more differences in their genome biology overall that could obscure differences in TE diversity stemming from genome size. Because our study system consisted of four closely related species,

which will have much more similar genomes overall, we can be more confident that diversity does not decrease as genome size increases.

Our results indicate that the models that predict a decrease in diversity may fit across a range of smaller genome sizes, but as genomes reach massive sizes, they no longer apply. Not only do we see high diversity in the detected TE superfamilies, but large genomes also contain inactive and degraded TEs (Novák et al 2020), which are diverse in sequence and therefore unlikely to mediate ectopic recombination. Thus, large genomes do not appear to be characterized by a low-diversity sequence community overall.

A possible explanation for the lack of decreased TE diversity with increased genome size in our dataset is that the models that predict a decrease in diversity as genomes expand do not accurately capture the dynamics of TEs and their hosts in all cases. Although ectopic recombination is less harmful in larger genomes (Petrov 2003), this may not translate directly into less diverse TE communities; it is possible (though unlikely) that ectopic recombination is still sufficiently harmful in large genomes to exert selective pressure to maintain diverse TE communities.

There is also the possibility that the number of TE superfamilies becomes maxed out after the genome reaches a certain size (Elliot and Gregory 2015), and there may be mechanisms that keep these superfamilies at the same evenness, which would keep the diversity the same. I propose several such mechanisms here: though Furano (2004) suggests that one TE superfamily may outcompete the others by exploiting host enzymes, it's possible that in large genomes with a lot of repeats, several superfamilies are able to coexist long-term, perhaps trading who is competitively dominant. It's also possible that host enzymes are not rate limited and no competition occurs at all, so TEs continue to proliferate as they would in a small genome. In turn

diversity is not affected. Boissinot (2016) predicts an arms race between silencing machinery and TE superfamilies, but perhaps as genomes expand in size and TEs continue to proliferate, silencing mechanisms are compromised for several superfamilies of TEs, resulting in no effect on, or even an increase in, diversity. Or it's possible that there is less competition overall between the silencing machinery and TE superfamilies leading to less suppression of TEs, removing the main cause of decreased diversity.

Finally, it is also possible that annotating only down to the superfamily level — considering every superfamily member as the same "species" — is not sensitive enough. In reality, each family within a superfamily has its own unique classification based on sequence divergence, so to treat them all as one may not be enough for picking up a difference in diversity. Measuring diversity at a more specific scale such as family could show a more accurate difference in diversity. For example, Boissinot (2016) shows that in mammals, one family in a TE superfamily can vastly outcompete the other families and increase in abundance, which leads to a reduction in evenness, and therefore a reduction in diversity of active families. In contrast, in non-mammalian vertebrates, most families within a superfamily exist and are active at relatively similar abundance (Boissinot 2016), which produces high evenness and high diversity. This indicates that diversity can differ at the family level even if it is not different at the superfamily level, so diversity may be different at the family level for salamanders as well.

Challenges adapting short-read TE annotation pipelines to long-read data

In studies aimed towards sequencing the whole genome, the first step for annotation is creating an assembly. Genome assemblies involve piecing together a large number of smaller pieces of DNA to recreate a draft of what the genome would look like, resulting in longer

sequences of DNA called contigs (Pop et al. 2004). Assembling TE sequences is inherently challenging because they are repetitive, which means that reads representing TE loci can assemble equally well to multiple other reads representing different TE loci (Miller et al. 2010). Thus, transposable elements can result in mis-assembly of the genome, due to tandem repeats being collapsed together, and sometimes they are even left out entirely (McCoy 2014). Most genome assembly programs (SPAdes, for example) are equipped for short read sequences, occasionally supplemented by longer ones, and higher coverage (at least 1x) (Bankevich et al. 2012). Low-coverage short-read shotgun sequencing datasets can also be readily assembled, where the goal is to assemble contigs that represent high copy number and recently proliferating TEs, rather than assembling an entire genome. However, with low-coverage MinION data, assembly is difficult due to the high error rate. When the sequence reads of identical repeat sequences can be as much as 15% different from one another, they are unlikely to assemble together (Miller et al. 2010).

Because of the large genome sizes in my focal taxa, my sequence dataset was low coverage (between 0.01x and 0.07x for all four species). I attempted various assembly programs made for long read data, such as Canu (Koren et al. 2017), but continued to run into computational issues due to the low coverage. Because our analyses required only the annotation of TEs to the superfamily level, and we found 1 to 94 entire TE sequences within single reads, we were able to move forward without an assembled genome. Based on these results, we conclude that, although assembly of low-coverage MinION data is not feasible, the unassembled reads contain information content that is sufficient for some applications including overall TE landscape characterization.

Challenges applying long-read sequencing technology to amphibian samples

I ended up running both groups three times on a new flow cell each time, for a total of six MinION runs. My first few runs with the MinION sequencer were not as successful as I had hoped for, with my first and second runs only generating ~20,000 reads and less than 1 Gb of data. My third and fourth runs were better, generating around 100,000 reads, but only 1 Gb of data. With my final two runs, I had about 2 Gb of data for one group and 4 Gb for the other. While these were significantly more than my earlier unsuccessful runs, I was initially discouraged as Oxford Nanopore claims that one could generate up to 30 Gb of sequence data. However, very few studies have been done in the laboratory using the MinION sequencer on amphibian DNA. A study done on the ornate burrowing frog, *Platyplectrum ornatum,* also attempted to do Oxford Nanopore long read sequencing, and came out with similar results. They were only able to generate ~1.5 Gb of sequence data from each flow cell. They redid the run with a similar species of frog, and came out with similar results, leading them to suggest the possibility of limitations with nanopore flow cells with frog DNA (Lamichhaney et al. 2021). Another study done on the frog *Leptopelis vermiculatus* only resulted in 109,047 and 181,123 reads (Menegon et al. 2017). Finally, a study done directly in the field on the toad *Atelopus ignescens* only resulted in 503 reads (Pomerantz et al. 2017). However, these last two studies are not entirely comparable because they were working with PCR products rather than genomic DNA. These leads me to suspect that amphibian DNA overall has its limitations with Oxford Nanopore sequencing. It is known that high G+C content can negatively impact MinION sequencing (Laver et al. 2015), but we know that that isn't the case here, as the G+C content was below 50% for all four *Plethodon* species. Therefore, differences could be at DNA, cell, or tissue level. We know that genome size is not the issue, as in Lamichhaney's (2012) study, their frog

study system had a small genome size (~1.06 Gb), so something beyond genome and chromosome size may be affecting the MinION output, such as lipids or cellular membranes.

REFERENCES

1. Abrusán G, Krambeck H-J. 2006. Competition may determine the diversity of transposable elements. *Theoretical Population Biology* 70:364–375.
2. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* 21:30.
3. Ambrožová K, Mandáková T, Bureš P, Neumann P, Leitch IJ, Koblížková A, Macas J, Lysak MA. 2011. Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of Fritillaria lilies. *Annals of Botany* 107:255–268.
4. Amorim IC, Melo ES, Moura RC, Wallau GL. 2020. Diverse mobilome of Dichotomius (Luederwaldtinia) schiffleri (Coleoptera: Scarabaeidae) reveals long-range horizontal transfer events of DNA transposons. *Mol Genet Genomics* 295:1339–1353.
5. AmphibiaWeb Database Search. Available from: https://amphibiaweb.org/
6. Animal Genome Size Database: Home. Available from: http://www.genomesize.com/index.php
7. Arkhipova IR. 2017. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mobile DNA* 8:19.
8. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* 19:455–477.
9. Bennetzen JL, Kellogg EA. 1997. Do Plants Have a One-Way Ticket to Genomic Obesity? *Plant Cell* 9:1509–1514.
10. Bennetzen JL, Park M. 2018. Distinguishing friends, foes, and freeloaders in giant genomes. *Current Opinion in Genetics & Development* 49:49–55.
11. Berglund EC, Kiialainen A, Syvänen A-C. 2011. Next-generation sequencing technologies and applications for human genetic history and forensics. *Investigative Genetics* 2:23.
12. Boissinot S, Sookdeo A. 2016. The Evolution of LINE-1 in Vertebrates. *Genome Biol Evol* 8:3485–3507.
13. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, et al. 2018. Ten things you should know about transposable elements. *Genome Biology* 19:199.
14. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Available from: https://genome.cshlp.org/content/27/5/722
15. Castanera R, Borgognone A, Pisabarro AG, Ramírez L. 2017. Biology, dynamics, and applications of transposable elements in basidiomycete fungi. *Appl Microbiol Biotechnol* 101:1337–1350.
16. Devos KM, Brown JKM, Bennetzen JL. 2002. Genome Size Reduction through Illegitimate Recombination Counteracts Genome Expansion in Arabidopsis. *Genome Res.* 12:1075–1079.
17. Elliott TA, Gregory TR. 2015. Do larger genomes contain more diverse transposable elements? *BMC Evolutionary Biology* 15:69.
18. Goerner-Potvin P, Bourque G. 2018. Computational tools to unmask transposable elements. *Nature Reviews Genetics* 19:688–704.

19. Goubert C. 2021. clemgoub/dnaPipeTE. Available from: https://github.com/clemgoub/dnaPipeTE
20. Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M. 2015. De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes aegypti). *Genome Biology and Evolution* 7:1192–1205.
21. Gregory TR. Applying ecological models to communities of genetic elements: the case of neutral theory. Available from: https://core.ac.uk/reader/189689016?utm_source=linkout
22. Hubley R. 2021. rmhubley/RepeatMasker. Available from: https://github.com/rmhubley/RepeatMasker
23. Itgen et al. 2021, unpublished.
24. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110:462–467.
25. Kapusta A, Suh A, Feschotte C. 2017. Dynamics of genome size evolution in birds and mammals. *Proc Natl Acad Sci USA* 114:E1460–E1469.
26. Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* 115:49-63
27. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27:722–736.
28. Kumar, G. Stecher, M. Suleski, and S.B. Hedges, 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution.* 34:1812-1819
29. Lamichhaney S, Catullo R, Keogh JS, Clulow S, Edwards SV, Ezaz T. 2021. A bird-like genome from a frog: Mechanisms of genome size reduction in the ornate burrowing frog, *Platyplectrum ornatum*. *Proc Natl Acad Sci USA* 118:e2011649118.
30. Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ. 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification* 3:1–8.
31. Lim JK, Simmons MJ. 1994. Gross chromosome rearrangements mediated by transposable elements in Drosophila melanogaster. *BioEssays* 16:269–275.
32. Linquist S, Cottenie K, Elliott TA, Saylor B, Kremer SC, Gregory TR. 2015. Applying ecological models to communities of genetic elements: the case of neutral theory. *Mol Ecol* 24:3232–3242.
33. Lisch D, Bennetzen JL. 2011. Transposable element origins of epigenetic gene regulation. *Current Opinion in Plant Biology* 14:156–161.
34. McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, Petrov DA, Fiston-Lavier A-S. 2014. Illumina TruSeq Synthetic Long-Reads Empower De Novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements. *PLOS ONE* 9:e106689.
35. Meagher TR, Vassiliadis C. 2005. Phenotypic impacts of repetitive DNA in flowering plants. *New Phytologist* 168:71–80.
36. Menegon M, Cantaloni C, Rodriguez-Prieto A, Centomo C, Abdelfattah A, Rossato M, Bernardi M, Xumerle L, Loader S, Delledonne M. 2017. On site DNA barcoding by nanopore sequencing. *PLOS ONE* 12:e0184741.
37. Miller JR, Koren S, Sutton G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics* 95:315–327.

38. Mouillot D, Leprêtre A. 1999. A comparison of species diversity estimators. *Res Popul Ecol* 41:203–215.
39. Mueller RL. 2017. piRNAs and Evolutionary Trajectories in Genome Size and Content. *J Mol Evol* 85:169–171.
40. Muñoz-López M, García-Pérez JL. 2010. DNA Transposons: Nature and Applications in Genomics. *Curr Genomics* 11:115–128.
41. Novák P, Guignard MS, Neumann P, Kelly LJ, Mlinarec J, Koblížková A, Dodsworth S, Kovařík A, Pellicer J, Wang W, et al. 2020. Repeat-sequence turnover shifts fundamentally in species with large genomes. *Nat. Plants* 6:1325–1329.
42. Nunes AP, Silva AC, Paiva ACD. 2010. Detection of masses in mammographic images using geometry, Simpson's Diversity Index and SVM. *International Journal of Signal and Imaging Systems Engineering* 3:40–51.
43. Parhad SS, Theurkauf WE. Rapid evolution and conserved function of the piRNA pathway. *Open Biology* 9:180-181.
44. Pearman WS, Freed NE, Silander OK. 2020. Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads. *BMC Bioinformatics* 21:220.
45. Pomerantz A, Peñafiel N, Arteaga A, Bustamante L, Pichardo F, Coloma LA, Barrio-Amorós CL, Salazar-Valenzuela D, Prost S. 2017. Real-time DNA barcoding in a remote rainforest using nanopore sequencing. *Gigascience* 7:4
46. Pop M, Phillippy A, Delcher AL, Salzberg SL. 2004. Comparative genome assembly. *Briefings in Bioinformatics* 5:12.
47. Shahid S, Slotkin RK. 2020. The current revolution in transposable element biology enabled by long reads. *Current Opinion in Plant Biology* 54:49–56.
48. Shannon CE. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal* 27:379-423
49. Sun C, Mueller RL. 2014. Hellbender genome sequences shed light on genomic expansion at the base of crown salamanders. *Genome Biol Evol* 6:1818–1829.
50. Sun C, Shepard DB, Chong RA, López Arriaza J, Hall K, Castoe TA, Feschotte C, Pollock DD, Mueller RL. 2012. LTR Retrotransposons Contribute to Genomic Gigantism in Plethodontid Salamanders. *Genome Biol Evol* 4:168–183.
51. Venner S, Feschotte C, Biémont C. 2009. Transposable elements dynamics: toward a community ecology of the genome. *Trends Genet* 25:317–323.
52. Wang J, Itgen MW, Wang H, Gong Y, Jiang J, Li J, Sun C, Sessions SK, Mueller RL. 2020. Gigantic Genomes Can Provide Empirical Tests of TE Dynamics Models — An Example from Amphibians. *bioRxiv*.
53. Welsh HH, Hodgson GR. 2013. Woodland salamanders as metrics of forest ecosystem recovery: a case study from California's redwoods. *Ecosphere* 4:59.
54. Wessler SR. 2006. Transposable elements and the evolution of eukaryotic genomes. *PNAS* 103:17600–17601.
55. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom* 3
56. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* 8:973–982.

57. Wierzbicki F, Schwarz F, Cannalonga O, Kofler R. 2020. Generating high quality assemblies for genomic analysis of transposable elements. *bioRxiv*