



Transposable Element Diversity Remains High in Gigantic Genomes

Ava Louise Haley¹ · Rachel Lockridge Mueller¹

Received: 28 July 2021 / Accepted: 8 June 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

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 models for ecological community diversity that consider species richness and evenness. Several models predict TE diversity decreasing as genomes expand because of selection against ectopic recombination and/or competition among TEs to garner host replicative machinery and evade host silencing mechanisms. Salamanders have some of the largest vertebrate genomes and highest TE loads. Salamanders of the genus *Plethodon*, in particular, have genomes that range in size from 20 to 70 Gb. Here, we 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 (*Plethodon cinereus*, 29.3 Gb vs. *Plethodon glutinosus*, 38.9 Gb) and one in the western clade (*Plethodon vehiculum*, 46.4 Gb vs *Plethodon idahoensis*, 67.0 Gb). We classified the TEs in these genomes and found > 40 TE superfamilies, accounting for 22–27% of the genomes. We calculated Simpson's and Shannon's diversity indices to quantify overall TE diversity. In both pairwise comparisons, the diversity index values for the smaller and larger genome were almost identical. This result indicates that, when genomes reach extremely large sizes, they maintain high levels of TE diversity at the superfamily level, in contrast to predictions made by previous studies on smaller genomes.

Keywords Genome size · *Plethodon* · Transposon ecology · Diversity index

Introduction

Genome sizes vary ~ 75,000-fold among eukaryotes, from ~ 0.002 Gb (e.g., in the eukaryotic fungus *Encephalitozoon cuniculi*) to ~ 150 Gb (e.g., in the monocot *Paris japonica*) (Pellicer et al. 2010). Across animals, the differences span 6,650-fold (Gregory 2022). Salamanders, one of the three clades of amphibians comprising 773 extant species (AmphibiaWeb 2022), include many of the largest animal genomes, ranging from ~ 9 Gb in *Thorius spilogaster* to 120 Gb in *Necturus lewisi* (Decena-Segarra et al. 2020; Gregory 2022). The main proximate cause for their large and variably sized genomes is the proliferation of transposable elements (Sun et al. 2012a, b), which contribute to

differences in genome size across diverse taxa (Wells and Feschotte 2020). Transposable elements (TEs) are DNA sequences that replicate and insert themselves throughout the genome. The percentage of the genome made up of TEs varies greatly across the tree of life, from ~ 0.1% (e.g., in the fungus *Pseudozyma antarctica*) to ~ 90% (e.g., in the lily *Fritillaria imperialis*) (Ambrozova et al. 2011; Castanera et al. 2017). In salamander genomes, ~ 25% to ~ 50% of the total DNA has been classified as recognizable TEs depending on the species (Sun et al. 2012a, b; Sun and Mueller 2014; Nowoshilow et al. 2018). Because the majority of TEs serve no initial protein-coding or regulatory function in the genome, they accumulate mutations, which eventually cause them to be undetectable during TE annotation (Venner et al. 2009). In plant and animal lineages with gigantic genomes, low rates of DNA removal through deletion contribute to the accumulation of TE sequences over time, producing genome expansion and high levels of TE-derived, but ultimately unrecognizable, sequence (Nystedt et al. 2013; Frahy et al. 2015; Kelly et al. 2015; Novák et al. 2020; Niu et al. 2022). Thus, the ~ 25% to ~ 50% estimates reported in salamanders

Handling editor: John Bracht.

✉ Rachel Lockridge Mueller
rachel.mueller@colostate.edu

¹ Department of Biology, Colorado State University, Fort Collins, CO 80523-1878, USA

do not include older TE insertions that have likely accumulated mutations and become unrecognizable (Keinath et al. 2015).

Transposable elements are categorized into two classes. The first is the 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 TE enzymatic machinery (Bourque et al. 2018). The second is the 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 (Muñoz-López and García-Pérez 2010). Within these classes, TEs are further categorized into 9 orders and > 39 superfamilies, commonly classified using the Wicker unified system (Wicker et al. 2007), although other classifications also exist (Jurka et al. 2005; Arkhipova 2017). Many superfamilies can be found in almost all eukaryotes, such as *Gypsy/mdg-4* and *Copia* of the LTR order (Bourque et al. 2018). Most superfamilies are variable across different genomes, existing at higher or lower proportions depending on the species. For example, in the caecilian *I. bannanicus*, the Class I retrotransposon DIRS makes up ~30% of the genome and the retrotransposon *Gypsy/mdg-4* makes up ~1%, while in salamanders, *Gypsy/mdg-4* is the most abundant (Wang et al. 2021). In contrast, class II DNA transposons make up 39%–60% of some teleost genomes, while retrotransposons exist at lower levels (Sotero-Caio et al. 2017).

Taken together, all of the TEs in a genome form a community of sequences, which can be studied to draw conclusions about genome evolution. As genomes expand, the number of TEs typically increases (Kidwell 2002; Elliott and Gregory 2015b). However, how the diversity of the overall TE community changes with expansion is not yet well understood (Elliott and Gregory 2015a). TE diversity within genomes can be measured in an analogous way to species diversity in ecological communities (Abrusán and Krambeck 2006; Venner et al. 2009; Linquist et al. 2015). Analyses of ecological diversity quantify the number of species, or richness, and the abundance of each species, or evenness using the Simpson and Shannon diversity indices (Shannon 1948; Simpson 1949). TE diversity can be approached in a similar way using richness and evenness of TE types (e.g., superfamilies) in a genome (Wang et al. 2021).

Several analyses have suggested that TE diversity will be highest in smaller genomes. TEs can have negative effects on the fitness of their “hosts” by causing recombination at ectopic, or non-homologous, sites, which can lead to deletions and duplications (Langley et al. 1988; Petrov et al. 2003). Because ectopic recombination is more likely to delete or duplicate a functional sequence in smaller genomes, small genome size should select for

more diverse TE communities, lowering the number of 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 can be lower, depending on chromosome number, which decreases the likelihood of ectopic recombination overall. Thus, larger genomes can be more permissive to low-diversity TE communities. For these same reasons, larger genomes can be more permissive to TE activity overall, producing a genomic environment in which competition to exploit host replicative machinery, and/or evade host silencing machinery, can lead to a decrease in diversity (Furano et al. 2004; Abrusán and Krambeck 2006; Boissinot and Sookdeo 2016).

In this study, we test the hypothesis that TE diversity is lower in larger genomes. We chose the salamander genus *Plethodon* (family Plethodontidae) as a study system due to the wide range of genome sizes, but high similarity in physical traits and life history, that exists across the 58 species (Petranka 1998; Gregory 2022). We analyzed two species' genomes from each of the two main *Plethodon* clades—*P. cinereus* (29.3 Gb genome) and *P. glutinosus* (38.9 Gb) from the eastern clade and *P. vehiculum* (46.4 Gb) and *P. idahoensis* (67.0 Gb) from the western clade. The median divergence time between the eastern and western clades, based on 11 published studies, is 45 million years (Kumar et al. 2017). *Plethodon cinereus* and *P. glutinosus* span the basal split within the eastern plethodon clade (median divergence 11.1 mya, adjusted divergence 15.6 mya, 3 published studies (Kumar et al. 2017)). Similarly, *P. vehiculum* and *P. idahoensis* span the basal split within the western plethodon clade (median divergence 31 mya, 7 published studies (Kumar et al. 2017)). Phylogenetic reconstructions of ancestral genome sizes for *Plethodon* based on different taxonomic sampling schemes confirm that the size difference between *P. vehiculum* and *P. idahoensis* reflects an increase along the *P. idahoensis* lineage since their point of common ancestry, and that the size difference between *P. cinereus* and *P. glutinosus* reflects an increase along the *P. glutinosus* lineage since their point of common ancestry that may have been accompanied by a decrease along the *P. cinereus* lineage (Newman et al. 2016; Itgen et al. 2022). Thus, our sampling encompasses two independent genome expansion events. We rely exclusively on Oxford Nanopore long read sequencing data—with no existing reference genome assembly—to quantify TE community diversity, validating our method using both Oxford Nanopore data and the full genome assembly of the model salamander *Ambystoma mexicanum*. Using both Simpson and Shannon's diversity indices, we find that TE diversity at the superfamily level is similar across our focal taxa, despite evolutionary changes in genome size. We discuss our findings in light of

hypotheses for TE proliferation and silencing dynamics in large genomes.

Materials and Methods

Tissue Collection

Plethodon cinereus and *Plethodon 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 Washington Department of Fish and Wildlife scientific collection permit # ITGEN 17-309. *Plethodon idahoensis* was collected in Shoshone County, Idaho, under the Idaho Department of Fish and Game wildlife collection permit #180226. Published genome sizes exist for all four species of *Plethodon* and vary across studies (Gregory 2022), but we use our own lab's measurements because they were performed on individuals collected at the same time and from the same locality as those sequenced here. Genome sizes for the species are *P. cinereus* (29.3 Gb), *P. glutinosus* (38.9 Gb), *P. vehiculum* (46.4 Gb), and *P. idahoensis* (67.0 Gb) (Itgen et al. 2022). Animals were euthanized via submersion in 10% buffered MS-222. Tissues were collected and stored in RNALater at -20°C. All work was completed according to the Colorado State University IUCAC protocol (17-7189A).

DNA Extraction, Library Preparation, and DNA Sequencing

DNA extraction was performed from 0.2 g of trunk skin and muscle tissue using a Qiagen DNeasy Blood and Tissue kit for each species. The manufacturer's protocol was followed except that (1) samples were flicked instead of vortexed to retain the longest DNA fragments possible, (2) centrifuge times were doubled to ensure all solution passed through the spin column, and (3) 30 µl of elution buffer was used to increase final DNA concentration.

Library preparation was done using a Ligation Sequencing Kit (SQK-LSK109), a Flow Cell Priming Kit (EXP-FLP002), and a Native Barcoding Expansion Kit 13–24 (EXP-NBD114) from Oxford Nanopore. New England Biolabs consumables used were an NEB Blunt/TA Ligase Master Mix (M0367), NEBNext® Quick Ligation Reaction Buffer (NEB B6058), and NEBNext® Companion Module for Oxford Nanopore Technologies® Ligation Sequencing (E7180S). For DNA repair and end prep, the amount of input genomic DNA was increased to 2 µg from the suggested 1 µg. For native barcode ligation, 1000 ng of end-prepped sample was used, twice the amount of suggested sample per

the manufacturer's protocol. A distinct barcode was used for each species. Following barcoding, *P. glutinosus* and *P. cinereus* were pooled together, and *P. vehiculum* and *P. idahoensis* were pooled together to equal about 850 ng of total DNA per pooled sample pair, slightly more than the 700 ng suggested by the protocol. The Long Fragment Buffer was used during adapter ligation. Throughout the protocol, samples were quantified with 1 µl on the Qubit fluorometer. Priming and loading the SpotON flow cells (R9.4.1) were performed two separate times, with two species occupying one flow cell. Sequencing was performed on the Oxford Nanopore MinION sequencer with the MinKnow software (v. 3.6.5). The sequencer was run for 72 h with the base calling setting of extremely fast. Porechop was used to trim adapters and barcodes (Wick et al. 2017).

Transposable Element Annotation

Our goals were (1) to find the most effective TE annotation tools for low-coverage MinION data 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 *Ichthyophis bannanicus*, RepeatMasker and DnaPipeTE together annotated 94.1% of the TE sequences (Wang et al. 2021). Additionally, in a TE annotation study on the beetle *Dichotomius (Luederwaldtinia) schiffleriso*, RepeatMasker and DnaPipeTE together annotated 95% of all of the detected TEs in the genome (Amorim et al. 2020). Although neither study relied on low-coverage MinION data, we initially chose these two programs together based on these previous successful applications. RepeatMasker uses a user-specified library to identify TEs based on sequence similarity, while DnaPipeTE detects TE sequences based on repetitiveness by using Trinity to assemble repeats from low-coverage data. 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. Initial exploratory analyses demonstrated that RepeatMasker annotated > 99% of the TEs in our *Plethodon* dataset and DnaPipeTE annotated < 1%. Thus, for our final analysis, we relied exclusively on RepeatMasker.

Our 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*, *Batrachoseps nigriventris*, *Bolitoglossa occidentalis*, *Bolitoglossa rostrata*, *Desmognathus ochrophaeus*, and *Eurycea tynerensis*) as well as the hellbender salamander (*Cryptobranchus alleganiensis*, family

Cryptobranchidae) and the axolotl (*Ambystoma mexicanum*, family Ambystomatidae) (Sun et al. 2012a, b; Nowoshilow et al. 2018). (2) 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. (3) Finally, the TEs detected by RepeatMasker were summarized for each species to characterize the total TE landscape for each species. Sequences that were identified as being repetitive, but not able to be classified, were referred to as “unknown repeats.” We estimated the percentage of each genome occupied by each TE superfamily, as well as by unknown repeats, by dividing the base pairs annotated to each superfamily by the total base pairs sequenced for each genome. We are assuming that the sequence data are a random subsample of the total genome sequence. Although we did not screen samples for possible contaminants, we treated all samples identically, and do not anticipate that this introduced bias into our comparative results.

Measuring Diversity of the Genomic TE Communities

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 are considered as species. In the first method, the total numbers of detected TE sequences annotated to each superfamily (either full-length or fragmented) 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 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 is expressed as the variable D , calculated by $D = \frac{\sum n(n-1)}{N(N-1)}$ (Simpson 1949). 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, which is more intuitive. The Shannon's diversity index is represented by the variable H , which is calculated by $H = -\sum_{i=1}^s p_i \ln p_i$ (Shannon 1948). The higher the value of H , the greater the diversity. Shannon's diversity index is more sensitive to sample size and rarer species than is Simpson's index (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, with low-coverage data, rare repeats may go undetected, so we used both indices.

Validation of Low-Coverage MinION Datasets for TE Diversity Index Calculation

We validated our overall low-coverage, MinION-based approach to TE community diversity measurement using the axolotl, *Ambystoma mexicanum*, which is the only salamander with a completely sequenced genome (genome size ~32 Gb) (Nowoshilow et al. 2018). We obtained a small, low-coverage MinION dataset for *A. mexicanum* from GenBank (ERX713866; 0.0222 Gb, or 0.0007× coverage; N50=4 kb) and trimmed it using Prowler, with a PHRED quality score cutoff of 12 (Lee et al. 2021). Next, we ran the trimmed dataset through our TE annotation and diversity index pipeline. We then compared the diversity index results to those we obtained in a previous study using the full *A. mexicanum* genome sequence (Wang et al. 2021) and found them to be within ~5% of one another (Shannon index based on the whole genome = 2.26; Shannon index based on low-coverage MinION dataset = 2.18; Gini–Simpson index based on the whole genome = 0.89; Gini–Simpson index based on low-coverage MinION dataset = 0.86). Because the *A. mexicanum* MinION dataset is at least one order of magnitude smaller than the datasets for *Plethodon* that we analyze here, we can conclude that our approach yields robust estimates of TE community diversity. This is in line with previous work demonstrating that aspects of the overall TE community in salamanders can be revealed reliably with low-coverage data, making these large genomes accessible to genomic study (Sun and Mueller 2014; Frahy et al. 2015).

Results and Discussion

Transposable Element Levels Are Similar Across Genome Sizes

For *Plethodon cinereus* and *P. glutinosus*, the MinION generated 4.15 Gb of total data and 1.22 million reads, with an N50 of 6.59 kb. 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. After trimming, this translates into 0.07× coverage for *P. cinereus* (1.99 Gb, N50 = 7.8 kb), 0.03× for *P. glutinosus* (1.34 Gb, N50 = 5.3), 0.01× for *P. vehiculum* (0.7 Gb, N50 = 8.3 kb), and 0.01× for *P. idahoensis* (0.7 Gb, N50 = 6.8 kb). These values are lower than expected based on MinION technology specs, although low data yield in applications of MinION sequencing to amphibian samples has also been reported in other studies (Menegeon et al. 2017; Pomerantz et al. 2018; Lamichhaney et al. 2021).

The RepeatMasker pipeline identified the following numbers of repeats for each species: 1,862,970 for *P. cinereus*; 1,338,850 for *P. glutinosus*; 700,019 for *P. vehiculum*; and 726,561 for *P. idahoensis*. Between one and 94 individual

TE sequences were annotated within single reads. Overall, the percentage of the genome composed of TEs (calculated as bp TE/ total dataset bp) ranged from 22% in *P. cinereus* to 27% in *P. idahoensis*, with an additional 5–10% composed of unknown repeats (Table 1). For each of the two genome expansion events encompassed by the pairwise comparisons—the lineage leading to *P. glutinosus* in the eastern clade and the lineage leading to *P. idahoensis* in the western clade—the percentage of the genome composed of recognizable TEs does not increase nearly as much as the genome size itself. The *P. glutinosus* genome is ~33% larger than that of *P. cinereus*, but the percentage of recognizable TEs is only 2% higher. Similarly, the *P. idahoensis* genome is ~45% larger than the *P. vehiculum* genome, but the percentage of recognizable TEs is only 1% higher. This result suggests that the increase in genome size is attributable to the accumulation of TEs that have persisted long enough to accumulate mutations and become unrecognizable, which in turn suggests decreased rates of TE deletion rather than recent bursts of TE proliferation. Interestingly, earlier DNA reassociation kinetic studies (i.e., Cot-curve comparisons) suggested that the percentage of repetitive DNA was much higher in the larger genome of *P. vehiculum* (80%) than in the smaller genome of *P. cinereus* (60%), a pattern that our results do not corroborate (Mizuno and Macgregor 1974).

Transposable Element Landscapes Are Similar Across Genome Sizes

All four species contained at least 40 TE superfamilies, which varied in relative abundance by 5 orders of magnitude within each genome (Table 2). Using both methods of calculating relative abundance—the total number of individual TE sequences and the total number of base pairs occupied by the TE superfamily—*Gypsy/mdg-4* (order LTR) was the most abundant in all four genomes, followed by L2 (order LINE) and DIRS (order DIRS). *Gypsy/mdg-4* accounted for 17–28% of the total repeats, and 25–33% of the total repeat base pairs; L2 accounted for 17–19% of the total repeats (16–21% base pairs); and DIRS accounted for 6–7% of the total repeats (10–12% base pairs) (Table 2). Overall, the most abundant TE superfamilies were dominated by retrotransposons; PIF-Harbinger and Helitron were the only DNA

transposons that exceeded 1% of the repeats in all four species. Unknown repeats accounted for 26–41% of the total repeats (16%–29% base pairs).

The percentage of the total genome occupied by each of the top 15 most abundant TE superfamilies is summarized in Figs. 1 and 2. *Gypsy/mdg-4* accounted for 7–11% of the total genomic sequence in each genome. All four species had the same six most abundant TE superfamilies, in the same rank order: *Gypsy/mdg-4*, L2, DIRS, ERV, Helitron, and L1. Thus, we infer that, in both cases of genome expansion—on the lineage leading to *P. idahoensis* in the western clade, and on the lineage leading to *P. glutinosus* in the eastern clade—the most abundant superfamilies all contributed to genome expansion through an increase in copy number, reflecting increased proliferation and/or decreased deletion. There are more differences in rank abundance among the lower-frequency superfamilies, but with our low-coverage dataset, there is more error associated with those estimates. Overall, the four species contained nearly identical detected TE superfamilies.

Transposable Element Superfamily Diversity Remains Unchanged as Genome Size Increases in Salamanders

For both pairwise comparisons—*P. cinereus* and *P. glutinosus* in the eastern clade, and *P. vehiculum* and *P. idahoensis* in the western clade—the diversity indices were similar between the smaller and larger genomes, demonstrating that a 10–20 Gb difference in genome size was not associated with a substantial change in TE community diversity measured at the superfamily level (Table 3). When the abundance of each TE superfamily was measured using TE copy number, the differences in Gini–Simpson's index were 0.01 for the eastern clade ($1-D=0.77$ and 0.76 for *P. cinereus* and *P. glutinosus*, respectively) and 0.03 for the western clade ($1-D=0.78$ and 0.75 for *P. vehiculum* and *P. idahoensis*, respectively). Using total base pairs occupied by each TE superfamily, the differences for both pairwise comparisons were 0.03. Using TE copy number, the differences in Shannon index were 0.03 for the eastern clade ($H=1.94$ and 1.91 for *P. cinereus* and *P. glutinosus*, respectively) and 0.11 for the western clade ($H=1.99$ and 1.88 for *P. vehiculum* and *P. idahoensis*, respectively). Using total base pairs occupied by

Table 1 Total number of repeat sequences detected in each genome, and the percentage of the overall sequence data occupied by classifiable and unclassifiable repeats

	Genome size (Gb)	Clade	Total # detected repeats	% sequence data occupied by classified TEs	% sequence data occupied by unknown repeats
<i>P. cinereus</i>	29.3	Eastern	1,862,970	22	8
<i>P. glutinosus</i>	38.9	Eastern	1,338,850	24	10
<i>P. vehiculum</i>	46.4	Western	700,019	26	8
<i>P. idahoensis</i>	67.0	Western	726,561	27	5

Table 2 Percentage of total repetitive sequence in each genome that is composed of each TE superfamily as well as unknown repeats

Order	Superfamily	% of total repeats (individual repeats)				% of total repeats (base pairs occupied by repeats)			
		<i>P. glutinosus</i>	<i>P. cinereus</i>	<i>P. idahoensis</i>	<i>P. vehiculum</i>	<i>P. glutinosus</i>	<i>P. cinereus</i>	<i>P. idahoensis</i>	<i>P. vehiculum</i>
Class I—Retrotransposon—Autonomous									
LTR	<i>ERV</i>	3.167%	2.990%	2.195%	2.458%	5.197%	5.206%	3.854%	4.642%
	<i>Gypsy/mdg-4</i>	19.003%	17.220%	27.984%	21.261%	28.560%	24.900%	33.113%	27.227%
	<i>Bel-Pao</i>	0.003%	0.002%	0.000%	0.002%	0.001%	0.001%	0.000%	0.001%
	<i>Copia</i>	0.141%	0.129%	0.075%	0.211%	0.150%	0.155%	0.093%	0.278%
	<i>Bhikari</i>	0.089%	0.095%	0.052%	0.063%	0.065%	0.070%	0.036%	0.040%
	<i>Foamy</i>	0.074%	0.073%	0.096%	0.059%	0.099%	0.129%	0.160%	0.088%
	<i>Unknown LTR</i>	0.015%	0.005%	0.005%	0.004%	0.008%	0.003%	0.003%	0.002%
DIRS	<i>DIRS</i>	5.900%	6.671%	7.081%	5.805%	9.654%	11.884%	10.116%	10.297%
LINE	<i>Ngaro</i>	0.264%	0.321%	0.896%	0.299%	0.363%	0.483%	1.495%	0.479%
	<i>Penelope</i>	1.948%	1.343%	0.872%	1.026%	0.944%	0.884%	0.519%	0.604%
	<i>Jockey</i>	0.029%	0.040%	0.043%	0.080%	0.028%	0.038%	0.034%	0.076%
	<i>L1</i>	2.644%	3.260%	4.746%	5.767%	1.494%	1.942%	2.571%	3.196%
	<i>L2</i>	16.493%	18.834%	16.709%	17.487%	15.539%	18.759%	20.702%	18.355%
	<i>RTE</i>	0.987%	1.144%	0.919%	1.026%	0.695%	0.882%	0.677%	0.784%
	<i>R1</i>	0.022%	0.029%	0.032%	0.025%	0.015%	0.023%	0.035%	0.023%
	<i>R2</i>	0.000*	0.001%	0.009%	0.001%	0.000*	0.001%	0.006%	0.001%
	<i>I</i>	0.015%	0.018%	0.021%	0.008%	0.010%	0.013%	0.012%	0.005%
	<i>CR1</i>	0.881%	1.198%	0.702%	0.918%	0.590%	0.873%	0.462%	0.721%
	<i>Tad1</i>	0.002%	0.001%	0.002%	0.001%	0.001%	0.000%	0.001%	0.001%
	<i>Unknown LINE</i>	0.070%	0.081%	0.068%	0.057%	0.050%	0.063%	0.049%	0.042%
Class I—Retrotransposon—Non-autonomous									
SINE	<i>7SL</i>	0.000*	0.000*	0.000*	-	0.000*	0.000*	0.000*	-
	<i>5S</i>	0.011%	0.006%	0.107%	0.079%	0.004%	0.002%	0.037%	0.031%
	<i>tRNA</i>	0.180%	0.251%	0.166%	0.224%	0.089%	0.093%	0.069%	0.106%
	<i>B4</i>	0.007%	0.013%	0.004%	0.007%	0.002%	0.004%	0.001%	0.002%
	<i>Deu</i>	1.403%	1.667%	1.309%	2.142%	0.740%	0.926%	0.673%	1.237%
	<i>MIR</i>	0.162%	0.196%	0.143%	0.186%	0.077%	0.097%	0.068%	0.097%
Class II—DNA Transposon—Subclass I									
TIR	<i>hAT</i>	1.358%	1.188%	0.311%	0.385%	1.334%	0.905%	0.189%	0.217%
	<i>Tc1-Mariner</i>	0.387%	0.613%	1.378%	0.589%	0.299%	0.527%	0.853%	0.381%
	<i>PIF-Harbinger</i>	1.513%	1.679%	2.759%	2.572%	1.378%	1.498%	1.571%	2.005%
	<i>PiggyBac</i>	0.205%	0.243%	3.025%	2.150%	0.115%	0.141%	1.626%	1.076%
	<i>Sola</i>	0.001%	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*
	<i>MuDR</i>	0.012%	0.015%	0.025%	0.043%	0.007%	0.009%	0.014%	0.051%
	<i>P</i>	0.003%	0.005%	0.002%	0.004%	0.002%	0.004%	0.001%	0.002%
	<i>Kolobok</i>	0.001%	0.001%	0.000*	0.013%	0.001%	0.001%	0.000*	0.010%
	<i>Academ</i>	0.021%	0.037%	0.012%	0.015%	0.030%	0.057%	0.017%	0.023%
	<i>MULE</i>	0.000*	0.000*	-	-	0.000*	0.000*	-	-
	<i>CMC/En-Spm</i>	0.043%	0.085%	0.098%	0.000%	0.025%	0.046%	0.053%	0.000%
	<i>Novosib</i>	0.000*	0.001	0.001	0.002%	0.000*	0.000*	0.000*	0.000*
Crypton	<i>Crypton</i>	0.000*	0.001	0.000%	0.000*	0.000*	0.001	0.000%	0.000*
Maverick	<i>Maverick</i>	0.277%	0.343%	0.329%	0.593%	0.562%	0.880%	0.660%	0.968%
Helitron	<i>Helitron</i>	1.532%	1.292%	1.620%	1.681%	2.895%	2.408%	3.334%	3.397%
Unable to be classified		40.847%	38.560%	25.938%	32.472%	28.794%	25.868%	16.579%	23.365%

*Indicates that the superfamily was detected at <0.001%

Fig. 1 Percentage of the sequence dataset composed of the 15 most abundant TE superfamilies in *Plethodon vehiculum* and *Plethodon idahoensis* (western subclade). *P. idahoensis* has a larger genome than *P. vehiculum* (67.0 Gb vs. 46.4 Gb). Superfamilies are ordered from highest to lowest abundance in *P. vehiculum*

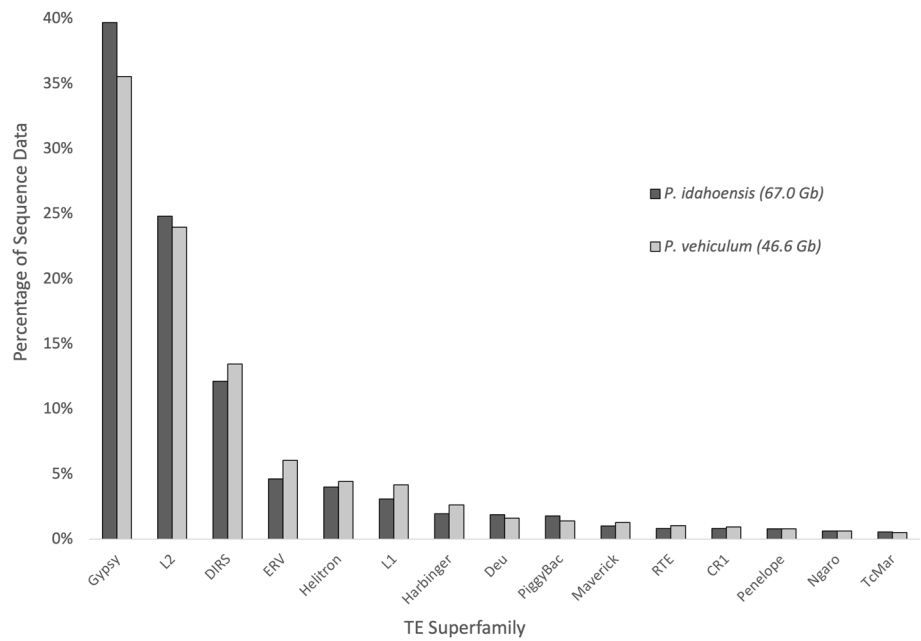


Fig. 2 Percentage of the sequence dataset composed of the 15 most abundant TE superfamilies in *Plethodon cinereus* and *Plethodon glutinosus* (eastern subclade). *P. glutinosus* has a larger genome than *P. cinereus* (38.9 Gb vs. 29.3 Gb). Superfamilies are ordered from highest to lowest abundance in *P. cinereus*

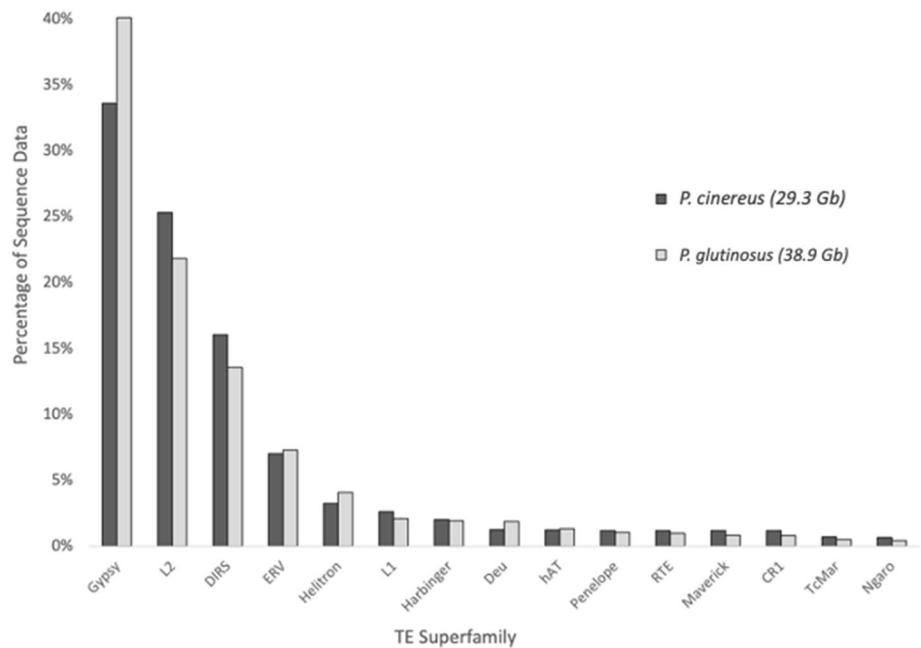


Table 3 Simpson and Shannon's diversity indices for TE superfamily diversity

	Gini-Simpson Index (1-D) Using total TE copy number	Shannon's Index (H) Using total TE copy number	Gini-Simpson Index (1-D) Using total base pair number	Shannon's Index (H) Using total base pair number
<i>Plethodon cinereus</i> —29.3 Gb	0.77	1.94	0.79	1.95
<i>Plethodon glutinosus</i> —38.9 Gb	0.76	1.91	0.76	1.87
<i>Plethodon vehiculum</i> —46.4 Gb	0.78	1.99	0.79	1.98
<i>Plethodon idahoensis</i> —67.0 Gb	0.75	1.88	0.76	1.86

each TE superfamily, the differences for the pairwise comparisons were 0.12. For context, a comparative study of TE superfamily diversity across vertebrates that encompassed species with smaller genomes found that the pufferfish *Takifugu rubripes* (0.4 Gb genome) had a Gini–Simpson index of 1.0 and Shannon index of 2.1, whereas the chicken *Gallus gallus* (1.3 Gb genome) had a Gini–Simpson index of 0.5 and Shannon index of 0.9, differing by 0.5 and 1.2, respectively (Wang et al. 2021). These differences are an order of magnitude greater than the differences we report in salamanders. Other pairwise comparisons of TE superfamily diversity in vertebrate genomes that differ in relative size by about the same amounts as the salamanders we study here reveal both similar and different levels of diversity; for example, *G. gallus* (1.3 Gb) versus the frog *Xenopus tropicalis* (1.7 Gb) differ by 0.4 in Gini–Simpson index and 1.34 in Shannon index, whereas *X. tropicalis* versus the lizard *Anolis carolinensis* (2.2 Gb) differ by 0.01 in Gini–Simpson index and 0.17 in Shannon index (Wang et al. 2021).

The diversity index values we report for *Plethodon* fall within the range reported for five species of salamanders that represent three families (Ambystomatidae, Cryptobranchidae, and Plethodontidae), two different types of datasets (whole-genome assembly and low-coverage 454 genome skimming), and a range of genome sizes (15–55 Gb) (Wang et al. 2021). In that study, there was no correlation between genome size and TE superfamily diversity in salamanders. However, the species analyzed (*Desmognathus ochrophaeus*, *Batrachoseps nigriventris*, *Ambystoma mexicanum*, *Aneides flavipunctatus*, and *Cryptobranchus alleganiensis*) were phylogenetically quite divergent, including spanning the basal-most split in the salamander clade, and these large evolutionary distances could be associated with overall differences in genome biology that would obscure changes in TE diversity stemming from genome size. In addition, the deep evolutionary history encompassed by those five species captured increases and decreases along the lineages leading to the focal taxa (Sessions 2008). In contrast, our study system consisted of four more closely related species within the genus *Plethodon*, which are expected to have much more similar genomes overall. In addition, our taxon sampling yielded two pairwise comparisons in which the larger of the two genomes resulted from an increase in genome size since the pairs' points of common ancestry (Newman et al. 2016; Itgen et al. 2022). Thus, the current study is a more powerful system for detecting decreases in TE diversity with increases in genome size. The fact that we do not see this pattern suggests that TE superfamily diversity remains high in enormous genomes. In addition, large genomes contain high levels of inactive and degraded TEs (Novák et al. 2020), which are diverse in sequence. Thus, large genomes do not appear to be characterized by a low-diversity sequence community overall.

Our results suggest 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. The richness of TE superfamilies may reach a maximum after the genome reaches a certain size (Elliott and Gregory 2015a)—we see ~40 superfamilies represented in each *Plethodon* genome—and TE dynamics in large genomes may keep these superfamilies at the same evenness. Some of the suggested mechanisms predicting decreased diversity include competition among TEs to exploit host enzymes (Furano et al. 2004) or evade host silencing machinery (Boissinot and Sookdeo 2016); our results suggest that these competitive interactions may not be relevant among TE superfamilies in large genomes. 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 to detect relevant changes in TE diversity because each superfamily consists of multiple divergent families. For example, in mammals, one L1 family evades host silencing to be active at a time, whereas in lizards and other non-mammalian vertebrates, multiple active L1 families coexist, demonstrating differences in active TE diversity within the same superfamily (Boissinot and Sookdeo 2016). Overall, our results demonstrate that substantial increases in genome size occur without associated changes in TE diversity at the superfamily level.

Acknowledgements We acknowledge members of A. Haley's Master's Degree committee, D. Sloan and M. Stenglein, for valuable assistance. We thank members of the Mueller and Sloan labs for discussion. We thank M. Itgen for *Plethodon* tissues. We thank E. Anderson for assistance running Prowler. Funding was provided by the National Science Foundation (1911585 to RLM) and Colorado State University.

Funding Funding was provided by the National Science Foundation (1911585 to RLM) and Colorado State University.

Data Availability Sequence data are deposited in the NCBI Sequence Read Archive (SRA) under accession number PRJNA749318.

Code Availability Not applicable.

Declarations

Conflict of interest Not applicable.

Consent to Participate NA.

Consent for Publication NA.

Ethical Approval Animal use was done in accordance with Colorado State University's IACUC protocol number 17-7189A.

References

- Abrusán G, Krambeck H-J (2006) Competition may determine the diversity of transposable elements. *Theor Pop Biol* 70:364–375
- Ambrozova K, Mandakova T, Bures P, Neumann P, Leitch IJ, Kobizkova A, Macas J, Lysak MA (2011) Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of *Fritillaria* lilies. *Ann Bot* 107:255
- Amorim I, Melo E, Moura R, Wallau G (2020) Diverse mobilome of *Dichotomius (Luederwaldtinia) schiffleri* (Coleoptera: Scarabaeidae) reveals long-range horizontal transfer events of DNA transposons. *Molec Genet Genom* 295:1339–1353
- AmphibiaWeb: information on amphibian biology and conservation [Internet]. 2022. Berkeley, California. <http://amphibiaweb.org/>.
- Arkhipova IR (2017) Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mob DNA* 8:1–14
- Boissinot S, Sookdeo A (2016) The evolution of LINE-1 in vertebrates. *Genome Biol Evol* 8:3485–3507
- 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 Biol* 19:199
- Castanera R, Borgognone A, Pisabarro AG, Ramírez L (2017) Biology, dynamics, and applications of transposable elements in basidiomycete fungi. *App Microbiol Biotech* 101:1337–1350
- Decena-Segarra LP, Bizjak-Mali L, Kladnik A, Sessions SK, Rovito SM (2020) Miniaturization, genome size, and biological size in a diverse clade of salamanders. *Am Nat* 196:634–648
- Elliott TA, Gregory TR (2015a) Do larger genomes contain more diverse transposable elements? *BMC Evol Biol* 15:1–10
- Elliott TA, Gregory TR (2015b) What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philos Trans R Soc B* 370:20140331
- Frahry MB, Sun C, Chong R, Mueller RL (2015) Low levels of LTR retrotransposon deletion by ectopic recombination in the gigantic genomes of salamanders. *J Mol Evol* 80:120–129
- Furano AV, Duvernell DD, Boissinot S (2004) L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. *Trends Genet* 20:9–14
- Gregory TR (2022) Animal genome size database. <http://www.genomesize.com>
- Itgen MW, Siegel DS, Sessions SK, Mueller RL (2022) Genome size drives morphological evolution in organ-specific ways. *Evol. Press*.
- Jurka J, Vapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Gen Res* 110:462–467
- Keinath MC, Timoshevskiy VA, Timoshevskaya NY, Tsonis PA, Voss SR, Smith JJ (2015) Initial characterization of the large genome of the salamander *Ambystoma mexicanum* using shotgun and laser capture chromosome sequencing. *Sci Rep* 5:16413
- Kelly LJ, Renny-Byfield S, Pellicer J, Macas J, Novák P, Neumann P, Lysak MA, Day PD, Berger M, Fay MF (2015) Analysis of the giant genomes of *Fritillaria (Liliaceae)* indicates that a lack of DNA removal characterizes extreme expansions in genome size. *New Phytol* 208:596–607
- Kidwell MG (2002) Transposable elements and the evolution of genome size in eukaryotes. *Genetica* 115:49–63
- Kumar S, Stecher G, Suleski M, Hedges SB (2017) TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol* 34:1812–1819
- 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 Nat Acad Sci* 118.
- Langley CH, Montgomery E, Hudson R, Kaplan N, Charlesworth B (1988) On the role of unequal exchange in the containment of transposable element copy number. *Genet Res* 52:223–235
- Lee S, Nguyen LT, Hayes BJ, Ross EM (2021) Prowler: a novel trimming algorithm for Oxford Nanopore sequence data. *Bioinformatics* 37:3936–3937
- 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. *Molec Ecol* 24:3232–3242
- 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
- Mizuno S, Macgregor HC (1974) Chromosomes, DNA sequences, and evolution in salamanders of the genus *Plethodon*. *Chromosoma* 48:239–296
- Mouillot D, Leprêtre A (1999) A comparison of species diversity estimators. *Res Pop Ecol* 41:203–215
- Muñoz-López M, García-Pérez JL (2010) DNA transposons: nature and applications in genomics. *Curr Genom* 11:115–128
- Newman CE, Gregory R, Austin CC (2016) The dynamic evolutionary history of genome size in North American woodland salamanders. *Genome* 60:285–292
- Niu S, Li J, Bo W, Yang W, Zuccolo A, Giacomello S, Chen X, Han F, Yang J, Song Y et al (2022) The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell* 185:204–217.e214
- Novák P, Guignard MS, Neumann P, Kelly LJ, Mlinarec J, Kobližková A, Dodsworth S, Kovařík A, Pellicer J, Wang W (2020) Repeat-sequence turnover shifts fundamentally in species with large genomes. *Nat Plants* 6:1325–1329
- Nowoshilow S, Schloissnig S, Fei J-F, Dahl A, Pang AWC, Pippel M, Winkler S, Hastie AR, Young G, Roscito JG et al (2018) The axolotl genome and the evolution of key tissue formation regulators. *Nature* 554:50–55
- Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A et al (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature* 497:579–584
- Pellicer J, Fay MF, Leitch IJ (2010) The largest eukaryotic genome of them all? *Bot J Linn Soc* 164:10–15
- Petranks JW (1998) Salamanders of the United States and Canada. Smithsonian Institution Press, Washington
- Petrov DA, Aminetzach YT, Davis J, Bensasson D, Hirsch AE (2003) Size matters: non-LTR retrotransposable elements and ectopic recombination in *Drosophila*. *Mol Biol Evol* 20:880–892
- Pomerantz A, Peñafiel N, Arteaga A, Bustamante L, Pichardo F, Coloma L, Barrio-Amorós C, Salazar-Valenzuela D, Prost S (2018) Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *Gigasci* 7:4
- Sessions SK (2008) Evolutionary cytogenetics in salamanders. *Chromosome Res* 16:183–201
- Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423
- Simpson EH (1949) Measurement of diversity. *Nature* 163:688–688
- Sotero-Caio CG, Platt RN II, Suh A, Ray DA (2017) Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biol Evol* 9:161–177
- Sun C, Mueller RL (2014) Hellbender genome sequences shed light on genome expansion at the base of crown salamanders. *Genome Biol Evol* 6:1818–1829
- Sun C, Arriaza JRL, Mueller RL (2012a) Slow DNA loss in the gigantic genomes of salamanders. *Genome Biol Evol* 4:1340–1348

- Sun C, Shepard DB, Chong RA, Arriaza JL, Hall K, Castoe TA, Feschotte C, Pollock DD, Mueller RL (2012b) LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. *Genome Biol Evol* 4:168–183
- Venner S, Feschotte C, Biemont C (2009) Dynamics of transposable elements: towards a community ecology of the genome. *Trends Genet* 25:317–323
- Wang J, Itgen MW, Wang H, Gong Y, Jiang J, Li J, Sun C, Sessions SK, Mueller RL (2021) Gigantic genomes provide empirical tests of transposable element dynamics models. *Genom Proteom Bioinform* 19:123–139
- Wells JN, Feschotte C (2020) A field guide to eukaryotic transposable elements. *Ann Rev Genet* 54:539–561
- Wick R, Volkening J, Loman N (2017) Porechop. Github <https://github.com/rrwick/Porechop>.
- Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capi P, Chalhou B, Flavell A, Leroy P, Morgante M, Panaud O (2007) A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* 8:973–982