Received: 25 June 2019 | Revised: 27 August 2019 | Accepted: 27 August 2019

DOI: 10.1111/1755-0998.13080

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

TERAD: Extraction of transposable element composition from RADseq data

Solomon T. C. Chak | **Dustin R. Rubenstein**

Department of Ecology, Evolution and Environmental Biology, Columbia University, New York, NY, USA

Correspondence
Solomon T. C. Chak, Department of Ecology, Evolution and Environmental Biology, Columbia University, New York, NY, USA.
Email: tc2856@columbia.edu

Funding information
Simons Foundation; Life Sciences Research Foundation; National Science Foundation, Grant/Award Number: IOS-1121435, IOS-1257530 and IOS-1439985; Columbia University; College of William and Mary

Abstract

Transposable elements (TEs) – selfish DNA sequences that can move within the genome – comprise a large proportion of the genomes of many organisms. Although low-coverage whole-genome sequencing can be used to survey TE composition, it is noneconomical for species with large quantities of DNA. Here, we utilize restriction-site associated DNA sequencing (RADSeq) as an alternative method to survey TE composition. First, we demonstrate in silico that double digest restriction-site associated DNA sequencing (ddRADseq) markers contain the same TE compositions as whole genome assemblies across arthropods. Next, we show empirically using eight *Synalpheus* snapping shrimp species with large genomes that TE compositions from ddRADseq and low-coverage whole-genome sequencing are comparable within and across species. Finally, we develop a new bioinformatic pipeline, TERAD, to extract TE compositions from RADseq data. Our study expands the utility of RADseq to study the repeatome, making comparative studies of genome structure for species with large genomes more tractable and affordable.

KEYWORDS
bioinformatics, ddRADseq, genome size, mobile genetic element, *Synalpheus*

1 | INTRODUCTION

Transposable elements (TEs) are selfish genetic elements that can move within a genome via a copy-and-paste or cut-and-paste mechanism (Feschotte, Jiang, & Wessler, 2002). Ubiquitous among all organisms, TEs are the main contributor of genome size differences across eukaryotes (Gregory, 2005b), and they play a critical role in evolution by introducing mutation and facilitating genomic rearrangement (Bourque et al., 2018). Cross-species comparative studies have been instrumental in understanding how ecology affects TE abundance (Kalendar, Tanskanen, Immonen, Nevo, & Schulman, 2000) and how TEs drive evolution (Feiner, 2016; Staton & Burke, 2015).

Low-coverage whole-genome sequencing (LC-WGS) has been shown to be an economical method to survey TEs (Goubert et al., 2015; Rasmussen & Noor, 2009; Talla et al., 2017; Wicker et al., 2008). This approach makes use of the repetitive nature of TEs so that at low sequencing depths, the sequence library will be more likely to be dominated by TEs than by nonrepetitive regions. LC-WGS has enabled comparative analysis of TE composition across species in a phylogenetic framework, including examinations of TE dynamics and genome size evolution among *Drosophila* species (Sessegolo, Burlet, & Haudry, 2016) and *Leptidea* butterfly populations (Talla et al., 2017). Furthermore, Lefébure et al. (2017) combined LC-WGS and transcriptomic data to test the relationship between rates of selection and TEs expansion among 22 asellid isopods.

Despite the potential of LC-WGS for comparative studies exploring the role of TEs in genome evolution, in species with large genomes, LC-WGS can still be cost-prohibitive. Apart from Angiosperms that have a~ 1000-fold difference in genome size (Bennett & Leitch, 2005) (although not all of this variation is attributable to TEs), at least several animal lineages are known to vary considerably in genome size and also contain species with exceptionally large genomes. These include salamanders (14–120 Gb)

(Nowoshilow et al., 2018), Orthoptera (1.52–16.56 Gb) and crustaceans (0.14–63.20 Gb) (Hultgren, Jeffery, Moran, & Gregory, 2018). However, the underlying mechanism of how these large genomes formed remains unclear because most of our knowledge about genome size change is based on lineages with small genomes (Dufresne & Jeffery, 2011). Understanding when and how these large genomes originated requires multi-species comparisons in a phylogenetic context (Dufresne & Jeffery, 2011), yet progress has been limited by a lack of cost-effective tools to survey TE compositions in lineages with large and variable genomes.

Here, we develop a new bioinformatic pipeline (TERAD) to estimate TE composition across species using double digest restriction-site associated DNA sequencing (ddRADseq) data. Restriction-site associated DNA sequencing (RADSeq) is an economical and widely used reduced representational approach to develop genome-wide SNPs for studies in ecology and evolution (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Peterson, Weber, Kay, Fisher, & Hoekstra, 2012). Yet, its application to studying repetitive elements like TEs has not been evaluated previously, although it has been attempted by Trucchi, Frajman, Haverkamp, Schönswetter, and Paun (2017). We show in silico and empirically that TE composition can be extracted from ddRADseq markers. First, we show that TE abundance is positively associated with whole genome assemblies and simulated ddRADseq markers across arthropods. Next, we perform ddRADseq and LC-WGS using eight species of *Synalpheus* snapping shrimps and show that ddRADseq markers can estimate interspecific differences in TE abundance in a way comparable to LC-WGS data. We use *Synalpheus* (Bate, 1888) shrimps because genome sizes are large and vary across species (Jeffery, Hultgren, Chak, Gregory, & Rubenstein, 2016). Ultimately, this study expands the utility of RADseq to study the repeatome, making comparative studies of genome structure for species with large genomes more tractable and affordable.

2 | METHODS

2.1 | Comparison of TEs from ddRADseq and whole genome assemblies across arthropods

We compared the proportions of TEs between simulated ddRADseq markers (see below for explanation) and whole genome assemblies using 16 arthropod species from three major lineages of arthropods (Hexapoda, Chelicerata and Crustacea; Table S1). We used all of the crustacean whole genome assemblies available at the time of analysis (February 2018), as well as other arthropod genomes in which both whole genomes and genome size data were available. Using a custom script (see simddRAD.sh), we simulated ddRADseq markers from whole genome assemblies using the five combinations of dual restriction enzymes (SbfI-EcoRI, SphI-EcoRI, EcoRI-MspI, SphI-MluCI and NlaIII-MluCI) with increasing genome coverage and a wide size selection criterion (300 ± 36 bp). These enzyme combinations were reported by Peterson et al. (2012) to generate ddRADseq markers across two orders of magnitude of genome coverage in most species. We used RepeatMasker (Smit, Hubley, & Green, 2015) and the arthropod repeat database in Repbase v 20,181,026 (Bao, Kojima, & Kohany, 2015) to identify TEs in both simulated ddRADseq markers and whole genome assemblies. For genome assemblies, we calculated the proportion of base pairs of the genome that contained TEs. For ddRADseq markers, we calculated the proportional count of ddRADseq markers that contained TEs because ddRADseq only samples a small fraction of the genome and therefore the proportion of TE base pairs is unlikely to be accurate. We examined the correlations between the proportions of TEs from genome assemblies and ddRADseq using phylogenetic generalized least-squares (PGLS) regressions with a maximum-likelihood tree based on cytochrome oxidase subunit 1 (COI) sequences (Table S1). PGLS regression models the correlation between species traits while controlling for statistical independence between species due to shared phylogenetic history (Blomberg, Lefevre, Wells, & Waterhouse, 2012). We also performed PGLS regressions separately for each TE subclass (LINE, SINE, LTR, DNA transposon and unclassified TEs).

2.2 | Restriction enzymes bias in TE database

The choice of restriction enzymes in a RADseq protocol is based on their cut frequencies and the target library size (Peterson et al., 2012). However, when using RADseq to survey TEs, the cut frequencies of the restriction enzymes in the TE should also be considered. Therefore, we tested whether the number of restriction sites of TEs in the Repbase arthropod repeat database (Bao et al., 2015) differs among enzymes. Using custom scripts (see tallyRE.sh), we tallied and compared the number of restriction sites in known TEs for EcoRI, MluCI, MspI, NlaIII, SbfI and SphI, as well as for the five enzyme combinations detailed above (from Peterson et al., 2012).

2.3 | Empirical comparison of ddRADseq and LC-WGS

We compared TE composition between data from ddRADseq and LC-WGS in eight snapping shrimp species in the genus *Synalpheus* (Alpheidae) (Table S2). We used one sample per species for both ddRADseq and LC-WGS and then added three additional samples (four total) per species for ddRADseq. We extracted genomic DNA using several walking legs from alcohol-preserved specimens using Qiagen DNeasy Tissue Kits (Qiagen). Extracted DNA was quantified using a Qubit 3.0 Fluorometer with the dsDNA HS assay (ThermoFisher Scientific) and visualized on 2% agarose gels.

For ddRADseq, we followed the protocol in Peterson et al. (2012) using EcoRI and MspI and a wide size selection criteria (338–414 bp). Briefly, we digested 1,000 ng of genomic DNA with EcoRI and MspI (New England Biolabs) and then cleaned up the digested DNA using Agencourt AMPure XP beads (Beckman Coulter Life Sciences). We ligated the double-digested DNA with barcoded adaptors that were fivefold in excess to prevent the formation of chimeras. We pooled and bead-cleaned up barcoded samples before size selection using a Pippin Prep and a dye-free 2% Agarose gel cassette with internal

standards (CDF2010; Sage Science). We performed a 10-cycle PCR using a Phusion PCR kit according to manufacturer protocols (New England Biolabs) with multiplexed primers and adjusted PCR products to 10 μM for sequencing on an Illumina HiSeq2500 (125 bp pair-end; New York Genome Center).

For LC-WGS, we sent 1,500 ng of genomic DNA to Novogene (Chula Vista, CA) for TruSeq PCR-free library preparation (Illumina) and 150 bp pair-end sequencing on an Illumina NovaSeq to obtain at least 1X coverage according to the respective genome size of each species (Jeffery et al., 2016). We used different sequencing platforms for ddRADseq and LC-WGS for economic reasons. The accuracy in assigning TEs (in subsequent analyses) to both LC-WGS and ddRADseq reads should improve when sequencing platforms with longer read lengths are used.

2.4 | TE analysis for ddRADseq data

We developed a pipeline, TERAD, to extract TE composition from ddRADseq data (Figure 1). Our analysis showed that while MspI cut sites are present in 86% of arthropod TEs, EcoRI cut sites occur less frequently (56%). To remove the bias due to the rarity of EcoRI cut sites, our analysis included only TEs that did not have an EcoRI restriction site (Figure 1). Therefore, we analysed only the EcoRI-ends of the paired-end reads. We used a custom assembly scheme to analyse TEs that began after the EcoRI restriction site. These TEs may or may not have MspI cut sites because each TE may or may not extend to the end of the other pair-end read.

We used process_radtags from Stacks (v 1.37) (Catchen, Amores, Hohenlohe, Cresko, & Postlethwait, 2011) to demultiplex raw reads and discarded reads with low quality scores (phred score <10) with the -r option. We combined single-end reads that were in-phase with the paired-end reads and ones that were out-of-phase and trimmed all reads to 120 bp. Reads were aligned using cd-hit-est (Fu, Niu, Zhu, Wu, & Li, 2012; Li & Godzik, 2006) with minimal alignment coverage of 40 bp for both reads and a sequence identity threshold of 0.95 (i.e. maximum of two mismatches in every 40 bp). These settings allowed reads that were identical in one section (at least 40 bp) due to the presence of TEs to be aligned and clustered to a representative read. We then used RepeatMasker and a custom repeat database to identify TEs from the representative reads. The custom repeat database included the Repbase arthropod database as well as TEs generated using RepeatModeler (Smit et al., 2015) from three decapod whole genome assemblies: *Eriocheir sinensis* (Song et al., 2016), *Neocaridina denticulata* (Kenny et al., 2014) and *Procambarus virginalis* (Gutekunst et al., 2018). We also used RepeatProteinMask to identify TEs based on a database of TE-encoded proteins. The protein databases were extracted from the nr GenBank databases and translations of interspersed repeat consensus sequences. For each representative read (from cd-hit) that contained a TE, we compared the cd-hit's alignment starting position and the TE starting position to determine whether the EcoRI restriction site was internal or external to the TE (Figure 1). We used an R script to summarize results from cd-hit, RepeatMasker and RepeatProteinMask and then calculated the proportional count of reads that contained major TE groups (i.e. DNA transposon, LTR, LINE, SINE, Helitron and unclassified TEs) and other forms of repetitive DNA (i.e. low-complexity repeats, microsatellites, minisatellites and RNA). The TERAD pipeline (https://github.com/solomonchak/TERAD) takes a fasta file with single-end reads from one sample and outputs a csv file with the TE composition summary and several intermediate summary files from cd-hit, RepeatMasker and RepeatProteinMask that are useful for other analyses. Detailed installation instructions, example, and explanations of output files from each step are available at the GitHub site.

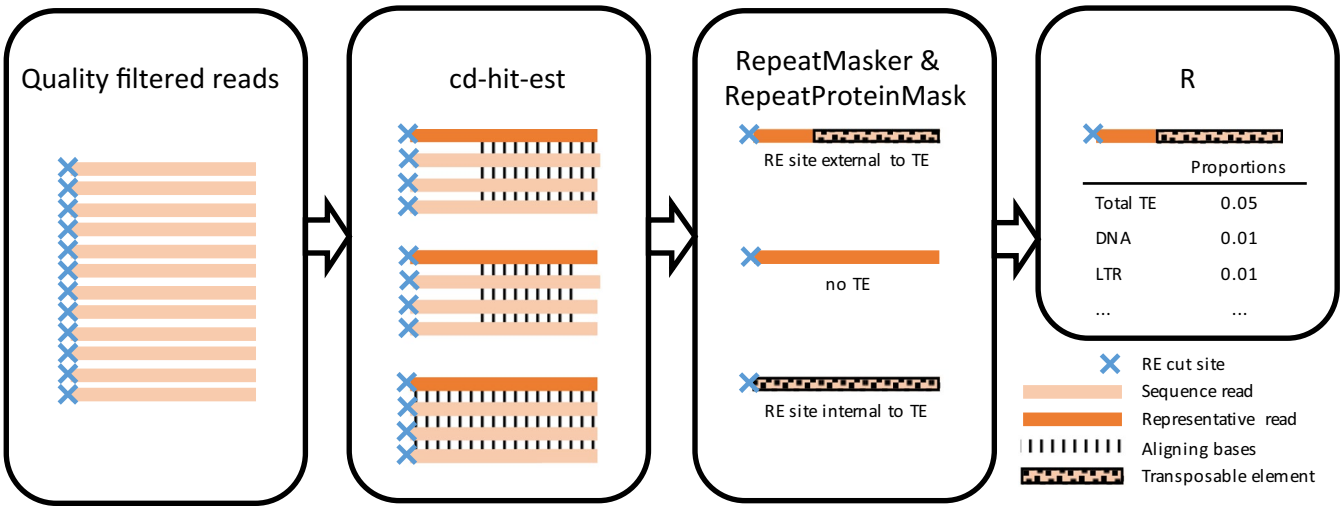

FIGURE 1 Schematic diagram of the TERAD workflow. First, quality-filtered reads were assembled using cd-hit-est, allowing for partially aligned reads to be clustered to a representative read. Then, transposable elements (TEs) were identified using RepeatMasker and RepeatProteinMask. Finally, we used an R script to extract representative reads where the restriction enzyme (RE) site is external to the TE and calculated the proportional count of reads that contained major TE groups [Colour figure can be viewed at wileyonlinelibrary.com]

2.5 | TE analysis for LC-WGS data

We performed unsupervised quality trimming on raw reads using UrQt (Modolo & Lerat, 2015) and removed mitochondrial DNA using Norgal (Al-Nakeeb, Petersen, & Sicheritz-Pontén, 2017). We used dnaPipeTE (Goubert et al., 2015) to estimate TEs in the LC-WGS data. This pipeline uses single-end raw low-coverage genomic reads to produce a precise estimate of TE composition and has been used across multiple arthropod taxa (Goubert et al., 2015; Kapun et al., 2018; Talla et al., 2017). Preliminary tuning showed that two iterations on oneword with 0.05X coverage gave the best assembly N50 metric in the Trinity step of dnaPipeTE. For each sample, we tallied the genome percentages of major TE groups corresponding to the ddRADseq analysis.

2.6 | Data analysis

We compared the results from the ddRADseq and LC-WGS data in three ways. First, we compared the identities of TE superfamilies between ddRADseq and LC-WGS within each of the eight shrimp species. We extracted TE classifications and counts from the LC-WGS data based on the output 'reads_per_component_and_annotation' from dnaPipeTE and those from the ddRADseq data based on the output 'TE.summary1' from TERAD. We then matched the two datasets by TE superfamilies and calculated how many superfamilies were shared between the two datasets for each species. We ignored TEs with unclassified classes and superfamilies (12% and 17% of the ddRADseq and LC-WGS data, respectively) in this analysis because their identities were uncertain. Most of these unknown TEs were identified de-novo from RepeatModeler from the three decapod genomes. Second, we tested whether more abundant TE superfamilies in the ddRADseq data were also more abundant in LC-WGS data. We ranked TE superfamilies by counts for each dataset and ran a linear regression between ranks from the ddRADseq and LC-WGS data for each species. Finally, we tested whether species with more TEs based on LC-WGS data also had more TEs based on ddRADseq data. We compared the proportional count of TEs in ddRADseq data and the proportion of TE base pairs in LC-WGS data across species using PGLS regression based on a *Synalpheus* phylogeny (Chak, Duffy, Hultgren, & Rubenstein, 2017). We ran one regression using one sample per species that was used for both ddRADseq and LC-WGS and a second regression using four samples per species for ddRADseq by taking the median TE proportion for each species. For all analyses, we ran supplementary analyses within each major TE subclass.

3 | RESULTS

We found that genome assemblies with a higher proportion of TE base pairs had a higher proportional count of TEs in their simulated ddRADseq markers, regardless of the combinations of restriction enzymes used (PGLS: all $p < .003$, Table S3, Figure 2). In terms of specific TE subclass, the enzyme combinations SphI-EcoRI and

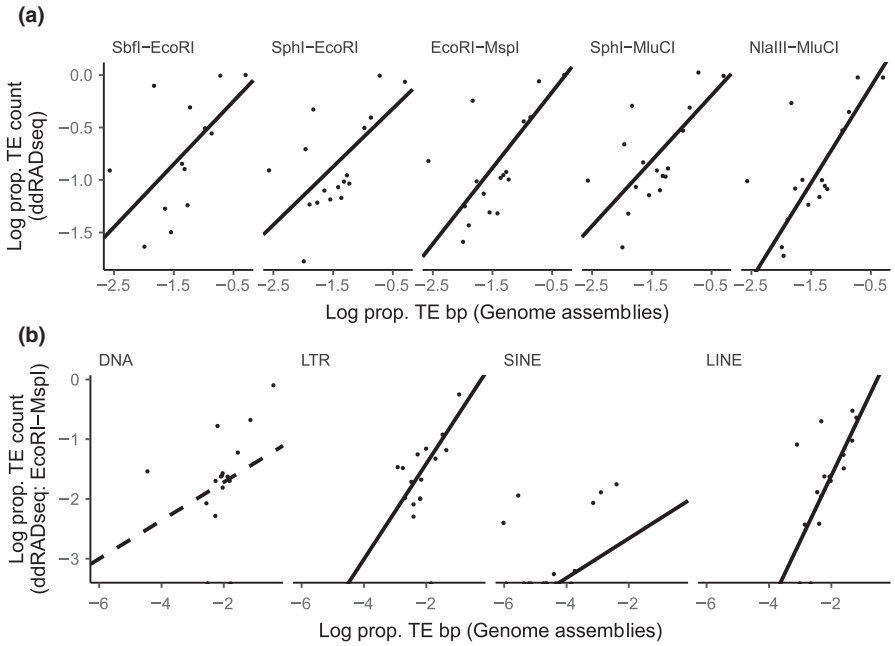

FIGURE 2 Concordance of transposable elements (TE) composition between simulated double digest restriction-site associated DNA sequencing (ddRADseq) markers and genome assemblies across arthropods. (a) Linear correlations between the proportional count of TEs from five sets of simulated ddRADseq markers (i.e. proportions of ddRADseq marker that contained TEs) and the proportion of TE base pairs (bp) from whole genome assemblies. (b) Linear correlations between each TE subclass between genome assemblies and simulated ddRADseq markers generated by the restriction enzymes EcoRI and MspI. Regression lines are based on phylogenetic generalized least-squares regressions that controlled for phylogenetic independence (solid lines: $p < .05$, dashed lines: $p > .05$)

EcoRI-MspI generated the best correlations between simulated ddRADseq and whole genome assemblies (Table S3, Figure 2). Therefore, ddRADseq markers, particularly using enzymes that generate medium levels of genome coverage, appear to more accurately reflect the TE composition of whole genome assemblies.

The arthropod TE database contains 10,392 sequences. The percentage of TE sequences that contain at least one restriction site varies drastically depending on the restriction enzyme used (SbfI: 6%, SphI: 41%, EcoRI: 56%, MspI: 85%, NlaIII: 92% and MluCI: 97%). The percentage of TE sequences that contain two cut sites in the case of ddRADseq is slightly lower (SbfI-EcoRI: 4%, SphI-EcoRI: 32%, EcoRI-MspI: 54%, SphI-MluCI: 41% and NlaIII-MluCI: 90%). Therefore, the choice of restriction enzymes may slightly bias the number of TEs recovered from ddRADseq. However, the bias is likely to be smaller when the cut sites lay outside of a TE. Therefore, our TERAD pipeline attempted to reduce such bias by analyzing ddRAD markers where the relatively rarer restriction cut sites for EcoRI lay outside TEs.

We compared TE compositions between data from ddRADseq and LC-WGS across eight *Synalpheus* species using three analyses. First, comparing the 801 unique TE superfamilies across two datasets, 74.45% (median; range across species: 65.35%–91.23%) of the TE superfamilies from the ddRADseq data were found in the LC-WGS data, and 89.87% (median; range: 66.67%–94.37%) of TE superfamilies from the LC-WGS data were found in the ddRADseq data. Results were similar within TE subclasses (see Table S4), but SINEs showed the weakest relationship, probably due to their smaller sizes. Second, more abundant TE families (i.e. those with higher ranks) from that ddRADseq data were consistently more abundant in the LC-WGS data in all species (all $p < .0001$, Figure 3). Results were qualitatively similar within TE subclasses, especially with the more abundant DNA transposons and LINEs (see Table S5). Finally, species with a higher proportion of TE base pairs in the LC-WGS data had higher proportional counts in the ddRADseq data when using either the same eight samples across species (PGLS, $p = .024$, Figure 4) or four samples per species (PGLS, $p = .005$, Figure 4). Analyses within major TE subclasses showed qualitatively similar trends (see Table S6, Figure S1).

4 | DISCUSSION

Here, we show both in silico and empirically that ddRADseq markers can be used to estimate and compare TE composition across species. First, we show that among arthropod genome assemblies, interspecific differences in the genome proportion of total TEs and major TE subclasses can be recovered from simulated ddRADseq markers, despite the fact that known TE sequences are biased in the number of restriction enzyme cut sites. Furthermore, using our approach of sampling ddRADseq markers that have the rare cutter sites external to the TEs, we show that interspecific differences in TE abundance and composition estimated from ddRADseq are similar

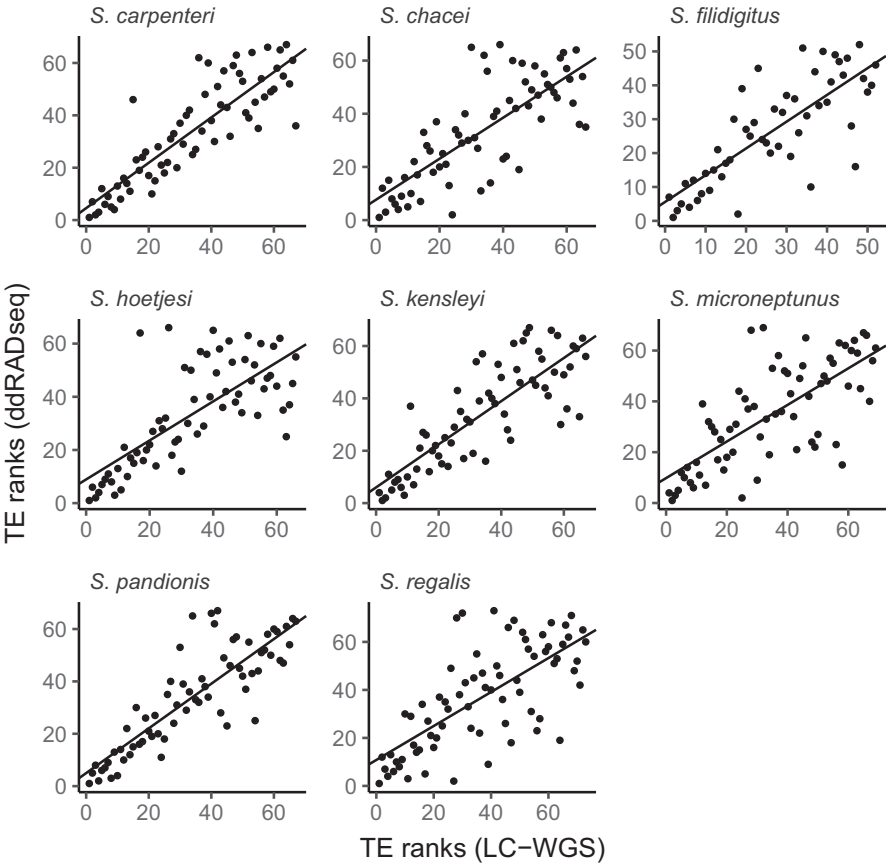

FIGURE 3 Ranks of transposable elements (TE) superfamilies based on double digest restriction-site associated DNA sequencing (ddRADseq) and low-coverage whole-genome sequencing (LC-WGS) data in eight *Synalpheus* species. Solid lines are linear regression slopes (all $p < .0001$)

to those estimated from LC-WGS. Therefore, our results suggest that ddRADseq is a reliable method for comparative studies of TE composition across species and can be a practical and economical alternative to the more expensive LC-WGS and long-read sequencing. Perhaps most importantly, since ddRADseq data are already available in many species, our method provides a means for re-analysing these data to extract TE composition for comparative analysis without the need for additional sequencing.

The use of RADseq to study TEs across species is especially cost-effective for lineages with large genome sizes in which LC-WGS can be cost-prohibitive. A comparative approach in these lineages will be especially useful for helping to clarify the evolutionary relationships between TE composition and other organismal, ecological and cellular traits (Gregory, 2005a). Since ddRADseq is often used to develop SNP markers that are used in population genomics (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013), the ability to extract TE composition from ddRADseq may further expand the utility of this convenient sequencing protocol. For example, SNPs markers from ddRADseq can be used to infer demographic history and effective population size (N_e) (Excoffier, Dupanloup, Huerta-Sanchez, Sousa, & Foll, 2013; Kamm, Terhorst, Durbin, & Song, 2019), which is hypothesized to be an important driving force of genome size and TE evolution (Ågren & Wright, 2011; Lefébure et al., 2017; Lynch & Conery, 2003). Thus, linking population genetics approaches with analyses of genome structure using the same sequence data may open up a wealth of new types of comparative evolutionary studies (Rubenstein et al., 2019).

Since ddRADseq is sampling only a small fraction of the genome (e.g. 0.14%–0.48% among 17 species using EcoRI and MspI) (Peterson et al., 2012), TE distribution will probably affect the accuracy of our method. The distribution of TEs varies in different lineages such that some lineages (e.g. mammals) have genomes where LINEs and SINE non-LTR retrotransposons dominate (Platt, Vandewege, & Ray, 2018), whereas other lineages (e.g. teleost fishes) have highly diverse TE superfamilies (Chalopin, Naville, Plard, Galiana, & Volff, 2015). Low-coverage sampling of the genome is ideal for species with TE distributions that have a few dominant TE families, each with high copy numbers. In contrast, when species have many TE families, each with low copy numbers, low-coverage sampling may be less accurate. While TE diversity does not show a simple relationship with genome size (Elliott & Gregory, 2015) (i.e. plants with smaller genomes tend to have greater TE diversity, but fungi with smaller genomes tend to have lower TE diversity), smaller genomes tend to have lower copy numbers per TE family than larger genomes (Kidwell, 2002). Therefore, ddRADseq may be less well suited to survey TEs in species with very small genome sizes, but LC-WGS in these species is also more cost-effective. Finally, the limitation of low-coverage sampling of the genome is not unique to ddRADseq and would similarly affect LC-WGS (Rasmussen & Noor, 2009). Furthermore, older and inactive TEs tend to accumulate more mutations (Bourque et al., 2018), which could create new restriction sites. Thus, TEs that are sampled from RADseq may be more sensitive towards older elements than newly active elements. Thus, the choice of method to sample TE composition will ultimately depend upon the TE landscape of a specific lineage.

The frequency of restriction enzyme cut sites in TEs can affect the TE estimates from RAD data. We tried to reduce this bias bioinformatically by analysing only TEs that contain the enzyme that cut more frequently among known TEs. We show that despite this bias, TE composition estimated from ddRADseq data is comparable to those from LC-WGS data. This may be due to the fact that the TE database often does not contain full-length TEs, and thus, the estimated number of restriction sites from a TE database may not actually reflect the true cut frequency. Parts of the TE that are not in the database may be less biased in restriction site and are identified by RepeatProteinMask as sequences with known TE proteins. Nonetheless, researchers interested in extracting information about TE composition from new or preexisting RADseq data should first explore the frequency of restriction sites in their

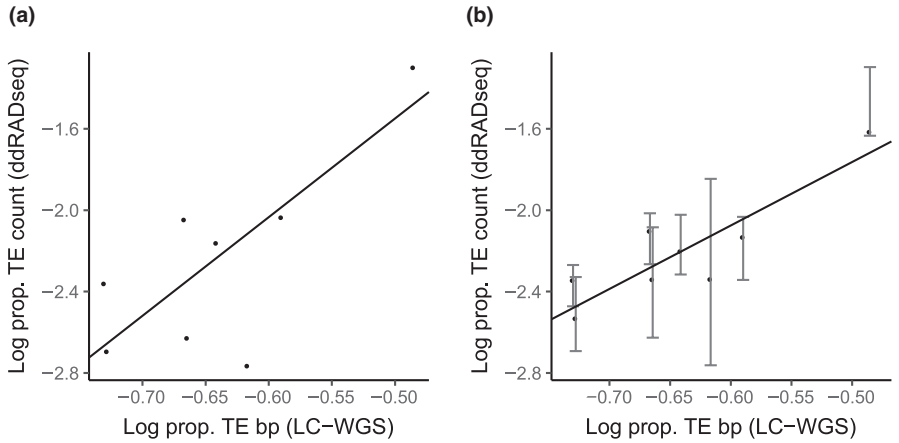

FIGURE 4 Relationships between proportional counts of transposable elements (TEs) from double digest restriction-site associated DNA sequencing (ddRADseq) and the proportions of TE base pairs (bp) from low-coverage whole-genome sequencing (LC-WGS) across eight *Synalpheus* species. (a) Results based on ddRADseq and LC-WGS on the same eight samples. (b) Results based on the median of four samples per species for ddRADseq. Regression lines are based on phylogenetic generalized least-squares regressions (all $p < .05$). Error bars in (b) indicate ranges of proportional counts of TEs

respective lineage-specific TE databases to estimate potential bias. For preexisting ddRADseq data, our TERAD pipeline provides a way to reduce such bias by sampling TEs with a selected restriction site. For new RADseq projects, depending on the study aim and organism, a single-enzyme RADseq protocol should theoretically be less biased in TE sampling, but it may not be as economical as ddRADseq (Peterson et al., 2012), though this needs to be examined empirically. Although we used ddRADseq here, our TERAD pipeline can also be used for single-enzyme RADseq. Furthermore, since the cut frequencies of restriction enzymes are variable across broad eukaryotic groups (Herrera, Reyes-Herrera, & Shank, 2015), our method should be used with caution when comparing across very different lineages.

In conclusion, we developed a bioinformatic pipeline to extract TE composition from RAD data and showed both theoretically and empirically that this method can be used for comparative analysis of TE abundance across species. Our work expands the utility of RADseq to studying the repeatome, especially for lineages that have large genome sizes in which LC-WGS is not yet economical.

ACKNOWLEDGEMENTS

S.T.C.C. was supported by the Simons Foundation via the Life Sciences Research Foundation. D.R.R. was supported by the US National Science Foundation (IOS-1121435, IOS-1257530, IOS-1439985). This work made use of the HPC resources from Columbia University and the College of William and Mary.

AUTHOR CONTRIBUTIONS

The study was designed and coordinated by S.T.C.C. and D.R.R.; statistical analyses were conducted and the manuscript was drafted by S.T.C.C; the manuscript was revised by both authors.

DATA ACCESSIBILITY

Supporting data are available from the Supporting Information and Dryad (doi:10.5061/dryad.2375033). Sequence reads were deposited in NCBI's Sequence Read Archive (BioProject accession PRJNA560035). TERAD is available at GitHub (https://github.com/solomonchak/TERAD) and in the Supporting information.

ORCID

Solomon T. C. Chak https://orcid.org/0000-0002-6362-4414
Dustin R. Rubenstein https://orcid.org/0000-0002-4999-3723

REFERENCES

Ågren, J. A., & Wright, S. (2011). Co-evolution between transposable elements and their hosts: A major factor in genome size evolution? *Chromosome Research*, *19*(6), 777–786. https://doi.org/10.1007/s10577-011-9229-0

Al-Nakeeb, K., Petersen, T. N., & Sicheritz-Pontén, T. (2017). Norgal: Extraction and de novo assembly of mitochondrial DNA from whole-genome sequencing data. *BMC Bioinformatics*, *18*, 510. https://doi.org/10.1186/s12859-017-1927-y

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). *Harnessing the power of RADseq for ecological and evolutionary genomics*. Nature Reviews: Genetics, advance online publication. https://doi.org/10.1038/nrg.2015.28

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*, *6*, 11. https://doi.org/10.1186/s13100-015-0041-9

Bate, C. S. (1888). Report on the Crustacea Macrura dredged by HMS Challenger during the years 1873–76. *Representative Voyage of Challenger Zoology*, *24*, 1–942.

Bennett, M. D., & Leitch, I. J. (2005). Genome size evolution in plants. In T. R. Gregory (Ed.), *The Evolution of the Genome* (pp. 89–162). Burlington, MA: Academic Press.

Blomberg, S. P., Lefevre, J. G., Wells, J. A., & Waterhouse, M. (2012). Independent contrasts and PGLS regression estimators are equivalent. *Systematic Biology*, *61*(3), 382–391. https://doi.org/10.1093/sysbio/syr118

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., … Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biology*, *19*(1), 199. https://doi.org/10.1186/s13059-018-1577-z

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: Building and genotyping loci de novo from short-read sequences. *G3: Genes*, *Genomes*. *Genetics*, *1*(3), 171–182.

Catchen, J. M., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. https://doi.org/10.1111/mec.12354

Chak, S. T. C., Duffy, J. E., Hultgren, K. M., & Rubenstein, D. R. (2017). Evolutionary transitions towards eusociality in snapping shrimps. *Nature Ecology & Evolution*, *1*, 0096. https://doi.org/10.1038/s41559-017-0096

Chalopin, D., Naville, M., Plard, F., Galiana, D., & Volff, J.-N. (2015). Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biology and Evolution*, *7*(2), 567–580. https://doi.org/10.1093/gbe/evv005

Dufresne, F., & Jeffery, N. (2011). A guided tour of large genome size in animals: What we know and where we are heading. *Chromosome Research*, *19*(7), 925–938. https://doi.org/10.1007/s10577-011-9248-x

Elliott, T. A., & Gregory, T. R. (2015). Do larger genomes contain more diverse transposable elements? *BMC Evolutionary Biology*, *15*(1), 69. https://doi.org/10.1186/s12862-015-0339-8

Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, *9*(10), e1003905. https://doi.org/10.1371/journal.pgen.1003905

Feiner, N. (2016). Accumulation of transposable elements inHoxgene clusters during adaptive radiation of *Anolis* lizards. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1840), https://doi.org/10.1098/rspb.2016.1555

Feschotte, C., Jiang, N., & Wessler, S. R. (2002). Plant transposable elements: Where genetics meets genomics. *Nature Reviews: Genetics*, *3*(5), 329–341. https://doi.org/10.1038/nrg793

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

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*(4), 1192–1205. https://doi.org/10.1093/gbe/evv050

Gregory, T. R. (2005a). Genome size evolution in animals. In T. R. Gregory (Ed.), *The Evolution of the Genome* (pp. 4–87). Burlington, MA: Academic Press.

Gregory, T. R. (2005b). Synergy between sequence and size in Large-scale genomics. *Nature Reviews: Genetics*, *6*(9), 699–708. https://doi.org/10.1038/nrg1674

Gutekunst, J., Andriantsoa, R., Falckenhayn, C., Hanna, K., Stein, W., Rasamy, J., & Lyko, F. (2018). Clonal genome evolution and rapid invasive spread of the marbled crayfish. *Nature Ecology & Evolution*, https://doi.org/10.1038/s41559-018-0467-9

Herrera, S., Reyes-Herrera, P. H., & Shank, T. M. (2015). Predicting RAD-seq marker numbers across the eukaryotic tree of life. *Genome Biology and Evolution*, *7*(12), 3207–3225. https://doi.org/10.1093/gbe/evv210

Hultgren, K. M., Jeffery, N. W., Moran, A., & Gregory, T. R. (2018). Latitudinal variation in genome size in crustaceans. *Biological Journal of the Linnean Society*, *123*(2), 348–359. https://doi.org/10.1093/biolinnean/blx153

Jeffery, N. W., Hultgren, K., Chak, S. T. C., Gregory, T. R., & Rubenstein, D. R. (2016). Patterns of genome size variation in snapping shrimp. *Genome*, *59*(6), 393–402. https://doi.org/10.1139/gen-2015-0206

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., & Schulman, A. H. (2000). Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proceedings of the National Academy of Sciences, USA*, *97*(12), 6603–6607. https://doi.org/10.1073/pnas.110587497

Kamm, J. A., Terhorst, J., Durbin, R., & Song, Y. S. (2019) Efficiently inferring the demographic history of many populations with allele count data. *Journal of the American Statistical Association*, *00*(0), 1–16. https://doi.org/10.1101/287268

Kapun, M., Aduriz, M. G. B., Staubach, F., Vieira, J., Obbard, D., Goubert, C., … Wiberg, R. A. W. (2018). Genomic analysis of european drosophila populations reveals longitudinal structure and continent-wide selection, *bioRxiv*, 313759.

Kenny, N., Sin, Y., Shen, X., Zhe, Q. U., Wang, W., Chan, T., … Hui, J. (2014). Genomic sequence and experimental tractability of a new decapod shrimp model, *Neocaridina Denticulata*. *Marine Drugs*, *12*(3), 1419. https://doi.org/10.3390/md12031419

Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, *115*(1), 49–63.

Lefébure, T., Morvan, C., Malard, F., François, C., Konecny-Dupré, L., Guéguen, L., … Douady, C. J. (2017). Less effective selection leads to larger genomes. *Genome Research*, *27*, 1016–1028. https://doi.org/10.1101/gr.212589.116

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Lynch, M., & Conery, J. S. (2003). The Origins of Genome Complexity. *Science*, *302*(5649), 1401–1404. https://doi.org/10.1126/science.1089370

Modolo, L., & Lerat, E. (2015). UrQt: An efficient software for the Unsupervised Quality trimming of NGS data. *BMC Bioinformatics*, *16*(1), 137. https://doi.org/10.1186/s12859-015-0546-8

Nowoshilow, S., Schloissnig, S., Fei, J.-F., Dahl, A., Pang, A. W. C., Pippel, M., … Myers, E. W. (2018). The axolotl genome and the evolution of key tissue formation regulators. *Nature*, *554*(7690), 50–55. https://doi.org/10.1038/nature25458

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, *7*(5), e37135. https://doi.org/10.1371/journal.pone.0037135

Platt, R. N., Vandewege, M. W., & Ray, D. A. (2018). Mammalian transposable elements and their impacts on genome evolution. *Chromosome Research*, *26*(1-2), 25–43. https://doi.org/10.1007/s10577-017-9570-z

Rasmussen, D., & Noor, M. (2009). What can you do with 0.1x genome coverage? A case study based on a genome survey of the scuttle fly *Megaselia scalaris* (Phoridae). *BMC Genomics*, *10*(1), 382.

Rubenstein, D. R., Ågren, J. A., Carbone, L., Elde, N. C., Hoekstra, H. E., Kapheim, K. M., … Hofmann, H. A. (2019). Coevolution of genome architecture and social behavior. *Trends in Ecology & Evolution*, *34*, 844–855. https://doi.org/10.1016/j.tree.2019.04.011

Sessegolo, C., Burlet, N., & Haudry, A. (2016). Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. *Biology Letters*, *12*(8), https://doi.org/10.1098/rsbl.2016.0407

Smit, A. F. A., Hubley, R., & Green, P. (2015). *RepeatMasker Open-4.0*., http://www.repeatmasker.org.

Song, L., Bian, C., Luo, Y., Wang, L., You, X., Li, J., … Xu, P. (2016). Draft genome of the Chinese mitten crab, *Eriocheir Sinensis*. *Gigascience*, *5*(1), 5. https://doi.org/10.1186/s13742-016-0112-y

Staton, S. E., & Burke, J. M. (2015). Evolutionary transitions in the Asteraceae coincide with marked shifts in transposable element abundance. *BMC Genomics*, *16*, 623. https://doi.org/10.1186/s12864-015-1830-8

Talla, V., Suh, A., Kalsoom, F., Dincă, V., Vila, R., Friberg, M., … Backström, N. (2017). Rapid increase in genome size as a consequence of transposable element hyperactivity in wood-white (*Leptidea*) butterflies. *Genome Biology and Evolution*, *9*(10), 2491–2505. https://doi.org/10.1093/gbe/evx163

Trucchi, E., Frajman, B., Haverkamp, T. H. A., Schönswetter, P., & Paun, O. (2017). Genomic analyses suggest parallel ecological divergence in Heliosperma pusillum (Caryophyllaceae). *New Phytologist*, *216*(1), 267–278. https://doi.org/10.1111/nph.14722.

Wicker, T., Narechania, A., Sabot, F., Stein, J., Vu, G. T. H., Graner, A., … Stein, N. (2008). Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. *BMC Genomics*, *9*(1), 1–15. https://doi.org/10.1186/1471-2164-9-518

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Chak STC, Rubenstein DR. TERAD: Extraction of transposable element composition from RADseq data. *Mol Ecol Resour*. 2019;19:1681–1688. https://doi.org/10.1111/1755-0998.13080