

1 Teaching transposon classification as a means to crowd source the

2 curation of repeat annotation – a tardigrade perspective

3 Valentina Peona^{1,2,3,*}& Jacopo Martelossi^{4,*}, Dareen Almojil⁵, Julia Bocharkina⁶, Ioana
4 Brännström^{7,8}, Max Brown⁹, Alice Cang¹⁰, Tomas Carrasco Valenzuela^{11,12}, Jon DeVries¹³,
5 Meredith Doellman^{14,15}, Daniel Elsner¹⁶, Pamela Espindola Hernandez¹⁷, Guillermo Friis
6 Montoya¹⁸, Bence Gaspar¹⁹, Danijela Zagorski²⁰, Paweł Hałakuc²¹, Beti Ivanovska²²,
7 Christopher Laumer²³, Robert Lehmann²⁴, Ljudevit Luka Boštančić²⁵, Rahia Mashhoodi²⁶,
8 Sofia Mazzoleni²⁷, Alice Mouton²⁸, Maria Nilsson Janke²⁵, Yifan Pei^{1,29}, Giacomo Potente³⁰,
9 Panagiotis Provataris³¹, José Ramón Pardos³², Ravindra Raut³³, Tomasa Sbaffi³⁴, Florian
10 Schwarz³⁵, Jessica Stapley³⁶, Lewis Stevens³⁷, Nusrat Sultana³⁸, Radka Symonova³⁹,
11 Mohadeseh Tahami⁴⁰, Alice Urzi⁴¹, Heidi Yang⁴², Abdullah Yusuf⁴³, Carlo Pecoraro⁴⁴,
12 Alexander Suh^{1,45*}

13

14 ¹ Department of Organismal Biology – Systematic Biology, Evolutionary Biology Centre,
15 Uppsala University, SE-752 36 Uppsala, Sweden

16 ² Swiss Ornithological Institute Vogelwarte, Sempach, CH-6204, Switzerland

17 ³ Department of Bioinformatics and Genetics, Swedish Natural History Museum, Stockholm,
18 Sweden

19 ⁴ Department of Biological Geological and Environmental Science, University of Bologna,
20 Via Selmi 3, 40126 Bologna, Italy

21 ⁵ New York University Abu Dhabi, Saadiyat Island, United Arab Emirates

22 ⁶ Skolkovo Institute of Science and Technology, Moscow, Russia

23 ⁷ Oslo University, Natural History Museum, Oslo, Norway

24 ⁸ Uppsala University, Department of Ecology and Genetics, Uppsala, Sweden

25 ⁹ Anglia Ruskin University

26 ¹⁰ University of Arizona, Tucson, AZ, USA

27 ¹¹ Evolutionary Genetics Department, Leibniz Institute for Zoo and Wildlife Research, Berlin

28 10315, Germany

29 ¹² Berlin Center for Genomics in Biodiversity Research, Berlin 14195, Germany

30 ¹³ Reed College, Portland, OR, United States of America

31 ¹⁴ Department of Ecology and Evolution, The University of Chicago, Chicago, IL 60637

32 ¹⁵ Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556

33 ¹⁶ Evolutionary Biology & Ecology, University of Freiburg, Germany

34 ¹⁷ Helmholtz Zentrum München, Research Unit Comparative Microbiome Analysis (COMI).

35 Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany

36 ¹⁸ Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AE, UK

37 ¹⁹ Institute of Evolution and Ecology, University of Tuebingen, Germany

38 ²⁰ Institute of Botany, Czech Academy of Sciences, Průhonice, Czechia

39 ²¹ Institute of Evolutionary Biology, Faculty of Biology, Biological and Chemical Research

40 Centre, University of Warsaw, Warsaw, Poland

41 ²² Institute of Genetics and Biotechnology, Hungarian University of Agriculture and Life

42 Sciences, Budapest, Hungary

43 ²³ The Natural History Museum, Cromwell Road, London SW6 7SJ

44 ²⁴ Biological and Environmental Science and Engineering Division, King Abdullah University

45 of Science and Technology (KAUST), Thuwal, Saudi Arabia

46 ²⁵ LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG),

47 Senckenberganlage 25, 60325 Frankfurt, Germany

48 ²⁶ Department of Genetics, Environment & Evolution, Centre for Biodiversity & Environment

49 Research, University College London

50 ²⁷ Department of Ecology, Faculty of Science, Charles University, Prague

51 ²⁸ INBIOS-Conservation Genetic Lab, University of Liege, Belgium

52 ²⁹ Centre for Molecular Biodiversity Research, Leibniz Institute for the Analysis of
53 Biodiversity Change, Adenauerallee 127, 53113 Bonn, Germany

54 ³⁰ Department of Systematic and Evolutionary Botany, University of Zurich

55 ³¹ NGS Core Facility, DKFZ-ZMBH Alliance, German Cancer Research Center, 69120
56 Heidelberg, Germany

57 ³² Departamento de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias
58 Naturales (MNCN-CSIC), José Gutiérrez Abascal 2, 28006 Madrid, Spain.

59 ³³ National Institute of Technology Durgapur

60 ³⁴ Molecular Ecology Group (MEG), National Research Council of Italy – Water Research
61 Institute (CNR-IRSA), Verbania, Italy

62 ³⁵ Eurofins Genomics Europe Pharma and Diagnostics Products & Services Sales GmbH

63 ³⁶ Plant Pathology Group, Institute of Integrative Biology, ETH Zurich

64 ³⁷ Tree of Life, Wellcome Sanger Institute, Cambridge CB10 1SA, UK

65 ³⁸ Department of Botany, Jagannath University, Dhaka-1100, Bangladesh

66 ³⁹ Institute of Hydrobiology, Biology Centre of the Czech Academy of Sciences, České
67 Budějovice, Czech Republic

68 ⁴⁰ Department of Biological and Environmental Science, University of Jyväskylä, P.O. Box
69 35, 40014 Jyväskylä, Finland

70 ⁴¹ Centogene GmbH, Am Strande 7, 18055 Rostock, Germany

71 ⁴² Department of Ecology & Evolutionary Biology, University of California, Los Angeles,
72 Los Angeles, California, United States of America

73 ⁴³ Zell- und Molekularbiologie der Pflanzen, Technische Universität Dresden

74 ⁴⁴ Physalia-courses, Berlin, 10249 Germany

75 ⁴⁵ School of Biological Sciences, University of East Anglia, Norwich Research Park, NR4
76 7TU Norwich, United Kingdom

77 *Corresponding authors: VP: valentina.peona@vogelwarte.ch, JM:

78 jacopo.martelossi2@unibo.it, AS: a.suh@leibniz-lib.de

79

80 Key words: transposable elements, manual curation, library, annotation, non-model organism,

81 genome assembly

82

83 **Abstract**

84 The advancement of sequencing technologies results in the rapid release of hundreds of new
85 genome assemblies a year providing unprecedented resources for the study of genome
86 evolution. Within this context, the significance of in-depth analyses of repetitive elements,
87 transposable elements (TEs) in particular, is increasingly recognized in understanding genome
88 evolution. Despite the plethora of available bioinformatic tools for identifying and annotating
89 TEs, the phylogenetic distance of the target species from a curated and classified database of
90 repetitive element sequences constrains any automated annotation effort. Manual curation of
91 raw repeat libraries is deemed essential due to the frequent incompleteness of automatically
92 generated consensus sequences. However, manual curation and classification are time-
93 consuming processes that offer limited short-term academic rewards and are typically
94 confined to a few research groups where methods are taught through hands-on experience.
95 Crowd sourcing efforts could offer a significant opportunity to bridge the gap between
96 learning the methods of curation effectively and empowering the scientific community with
97 high-quality, reusable repeat libraries. Here, we present an example of such crowd sourcing
98 effort developed through both in-person and online courses built around a collaborative peer-
99 reviewed teaching process that can be used as teaching reference guide for similar projects.
100 The collaborative manual curation of TEs from two tardigrade species, for which there were
101 no TE libraries available, resulted in the successful characterization of hundreds of new and
102 diverse TEs: A hidden treasure awaits discovery within non-model organisms.

103 **Background**

104 The importance of in-depth analyses of repetitive elements, particularly transposable elements
105 (TEs), is becoming more and more fundamental to understand genome evolution and the
106 genetic basis of adaptation [1]. While there is a wealth of bioinformatic tools available for the

107 identification and annotation of TEs (https://tehub.org/en/resources/repeat_tools), any
108 automated annotation effort is limited by the phylogenetic distance of the target species to a
109 database of curated and classified repetitive element sequences [2]. For example, in birds
110 where zebra finch and chicken have well-characterized repetitive elements because their
111 genomes were first sequenced in large consortia during the pre-genomics era [3,4], automated
112 annotation of other bird genomes will render most repeats as correctly classified [5,6]. On the
113 other hand, in taxa as diverse and divergent as insects, up to 85% of repetitive sequences can
114 remain of “unknown” classification in non-*Drosophila* species [7]. This is problematic.
115 Inferences about the mobility and accumulation of TEs, as well as their potential effects on
116 the host, are not feasible for unclassified repeats, as well as for incorrectly classified repeats if
117 the automated classification is based on short, spurious nucleotide sequence similarity [8,9].
118 The reference bias in TE classification reflects the history of the TE field in the genomics era:
119 In the 1990s and 2000s, there were usually multiple people tasked with TE identification,
120 classification, and annotation for each genome project, yielding manually curated consensus
121 sequences (namely representative sequences which quality was controlled and improved) and
122 fully classified TE libraries deposited in databases such as Repbase [2]. Over the last ten years,
123 however, the number of genome projects both of individual labs as well as large consortia has
124 increased exponentially and so have speed and number of automated TE annotation efforts
125 [10–12], while time and personnel have remained limited for curated TE annotation efforts.
126 Similar to taxonomic expertise required for identifying and classifying organisms, TE
127 identification and classification need hands-on experience with manual curation for months or
128 even years per genome [1] which is usually taught through knowledge passed within genome
129 projects and research groups. Recent efforts [13–15] have started to make manual curation
130 accessible to a broader scientific audience, with the aim to increase reproducibility and
131 comparability. However, what cannot be changed is that there are hundreds if not thousands

132 of genomes per TE-interested researcher with more or less pressing priority for time-
133 consuming manual curation.

134 Low scalability and people power are major obstacles that need to be overcome by the many
135 facets of computational biology where curation is essential. Annotation efforts of other
136 genomic features have shown that crowd sourcing through teaching [16–22], or “course
137 sourcing” as we call it, has the benefit of providing participants with hands-on skills for
138 curation and experience on how to reconcile biology with technical limitations, while
139 simultaneously sharing the workload of time-consuming curation across multiple people
140 working on different parts at the same time. Thus, we argue that a TE curation effort that
141 would take months or years for a single person may fit into a few days or weeks of teaching,
142 of course as long as reproducibility and comparability are ensured throughout course duration.

143 Here, we present our “course sourcing” experience from two iterations of a Physalia Course
144 on TE identification, classification, and annotation. We focused on two species of tardigrades
145 as a case study to motivate student-centered learning through direct contribution to scientific
146 knowledge: Tardigrades are, to our knowledge, the most high-ranking animal phylum without
147 curated TE annotation, very clearly illustrated by the fact that in previous genome analyses,
148 almost all repeats remained of “unknown” classification [23]. Tardigrades are a diverse group
149 of aquatic and terrestrial animals which show extraordinary ability to survive extreme
150 environments by entering the state of cryptobiosis [24]. This animal clade comprises almost
151 1,200 described species belonging to Panarthropoda [25] and the two species used in the
152 courses are closely related and belong to the Hypsibiidae family [23].

153 The first course took place in person in June 2018 in Berlin across five full-time work days:
154 The first three days familiarized the 13 participants with the biology of TEs, concepts for
155 classification, and methods for annotation using the tardigrade *Hypsibius dujardini*, while the
156 last two days had a student-centered learning format where each participant was able to

157 deepen knowledge where needed and curate as many TEs as possible from the target species.
158 The second course took place virtually in June 2021 due to the Covid-19 pandemic and
159 comprised five afternoons in the Berlin time zone to minimize Zoom fatigue. The overall
160 format was similar to the prior in-person course but with 24 participants and focusing on
161 another tardigrade, *Ramazzottius varieornatus*, which the participants identified to have not a
162 single shared TE family with the tardigrade *H. dujardini* curated in the 2018 course. Between
163 the two courses, the participants were able to uncover a vast diversity of TEs and successfully
164 curate almost 500 consensus sequences. We demonstrate therefore that a collaborative
165 approach is a valuable means to achieve significant results for the scientific community and
166 we hope to share with the community a teaching reference for future similar efforts, because:
167 A hidden treasure always awaits discovery in non-model organisms.

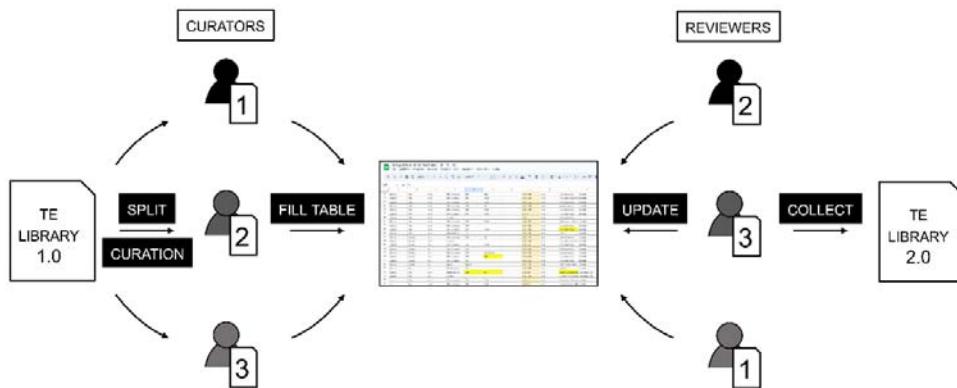
168 **Results and Discussion**

169 Incorporating crowd sourcing efforts within a classroom setting (“course sourcing”) can
170 represent an invaluable opportunity for teaching, while simultaneously contributing to the
171 scientific community. However, course sourcing also presents its own unique challenges,
172 particularly in terms of minimizing errors, maximizing reproducibility and student
173 engagement. Drawing from our experience in both in-person and virtual settings, we
174 identified several crucial factors in teaching TE manual curation that must be considered
175 during the organization and supervision of such course, like: a) establishing a standardized
176 approach for curation and classification of TE consensus sequences; b) implementing a peer-
177 review process between participants to check on the quality of the curated consensus sequence;
178 c) maintaining meticulous version control of the libraries. Here, we describe how we
179 addressed these points. First, to establish a standard approach to manual curation, we
180 implemented methods widely used in the TE community that have been recently reviewed in
181 detail [13,14]. The approach, briefly, consists in producing and inspecting multi-sequence

182 alignments for each of the consensus sequences automatically generated by RepeatModeler
183 [10]. Each nucleotide position of the “alignable part” of the alignment is carefully inspected to
184 identify the correct termini of the TE while correcting for any ambiguous base or gap. To
185 correct for ambiguous bases, we applied a majority rule and assigned the most representative
186 IUPAC nucleotide character for each position in the alignment (see **Methods**). To correct the
187 consensus sequences where gaps of different lengths are present, we considered each
188 insertion/deletion length as independent events so that a majority rule was applicable to these
189 regions as well. When very complex regions could not be unambiguously solved, stretches of
190 10 N nucleotides were inserted as placeholder (gap) in the consensus sequence. The TE
191 classification followed the nomenclature used by RepeatMasker to ensure direct compatibility
192 with the tool and its suite of scripts for downstream analysis. Second, when participants
193 completed the curation of their consensus sequences, then their results would go through a
194 peer-review process where both the quality of the sequence and its classification were revised
195 by other participants (or course faculty). During the in-person edition, a random set of
196 consensus sequences curated by one participant was assigned to another participant, while in
197 the second online edition, all sequences were reviewed by the two instructors and one
198 participant (**Figure 1**). The review of the TE sequences continued after the official conclusion
199 of the course. To ensure reproducibility and the documentation of the entire decision-making
200 process for classification, all steps and details of classification were recorded in a shared
201 Google Sheet. The tables would include the changes in consensus sequence names, names of
202 the curators and reviewers and additional comment (**Figure 1, Table S1**). Whenever a change
203 was introduced in a consensus sequence (either in the nucleotide sequence itself or in the
204 classification), the new version was directly added to the multi-sequence alignment file used
205 for curation together with the original one. Keeping all the versions of a consensus in the
206 same alignment file and respective notes in the tables allows the implementation of a basic
207 version control useful to check on the steps leading to a particular decision. From the re-

208 iteration of the course, we noticed three particularly challenging points for beginners that need
209 an extra supervision effort. The most challenging points are the identification of the correct
210 termini, target site duplications (a hallmark of transposition for the vast majority of TEs) if
211 any, and the correct spelling of the TE categories for classification in accordance with the
212 RepeatMasker nomenclature rules. The last point is of particular importance especially if the
213 repeat annotation is visualized as a landscape using the RepeatMasker scripts (e.g.,
214 calcDivergence.pl and createRepeatLandscape.pl) to not cause computing errors and
215 downstream misinterpretations.

216 Finally, all the tutorials to obtain and curate a TE library are available on the GitHub
217 repository linked to this paper: <https://github.com/ValentinaPeona/Tardigrate>.



218

219 **Figure 1.** Schematic representation of the peer-reviewed process of TE curation.

220 Improvement of the transposable element libraries

221 To generate the TE libraries, we first ran RepeatModeler and RepeatModeler2 on both species
222 and obtained 489 and 900 consensus sequences for *H. dujardini* and *R. varieornatus*
223 respectively (**Table 1**). Then the course participants manually curated as many consensus
224 sequences as possible. In about three course days plus voluntary efforts by some participants
225 after each course, the participants were able to curate 286 consensus sequences (58%) of the

226 *H. dujardini* library and 145 consensus sequences (16%) of the *R. varieornatus* library (**Table**
227 **S1-3**). Given the lack of previously curated libraries from closely related species, most of the
228 consensus sequences were automatically classified as “unknown” by RepeatModeler, but the
229 thorough process of manual curation successfully reclassified 305 unknown consensus
230 sequences (out of a total of 431 curated sequences, 71%) into known categories of elements.
231 After manual curation, we found that most of the two species’ libraries are comprised of DNA
232 transposons and a minority of retrotransposons (**Table 1**). Since many consensus sequences
233 remained uncurated and unclassified, it is possible that the relative percentages of the
234 categories change in the future, but we expect, especially from the composition of the *H.*
235 *dujardini* library, to mostly find additional (non-autonomous) DNA transposons among the
236 unclassified.

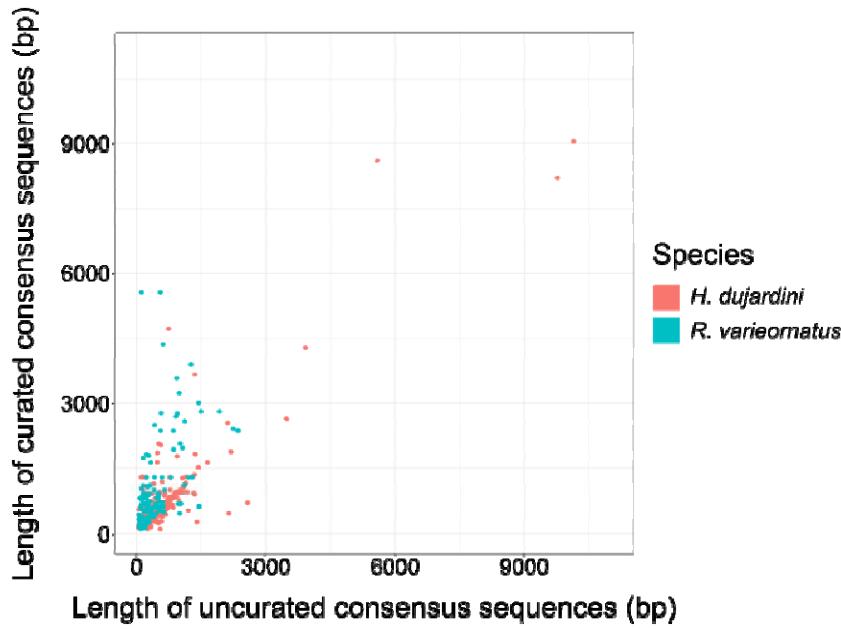
237

238 **Table 1:** Overview of classification of tardigrade repeats in the curated libraries. The libraries
239 here described contain both curated and uncurated consensus sequences.

Species	DNA	LINE	LTR	SINE	Unknown
<i>Hypsibius dujardini</i>	247	12	29	2	199
<i>Ramazzottius varieornatus</i>	203	35	11	-	651

240

241 The process of manual curation improved the overall level of TE classification of the libraries
242 but also the quality of the individual consensus sequences by correctly identifying their
243 termini and in general by extending their sequence. Indeed, by comparing the lengths of the
244 consensus sequences for the same element, we can notice a marked increase in length after
245 curation (**Figure 2**).



246

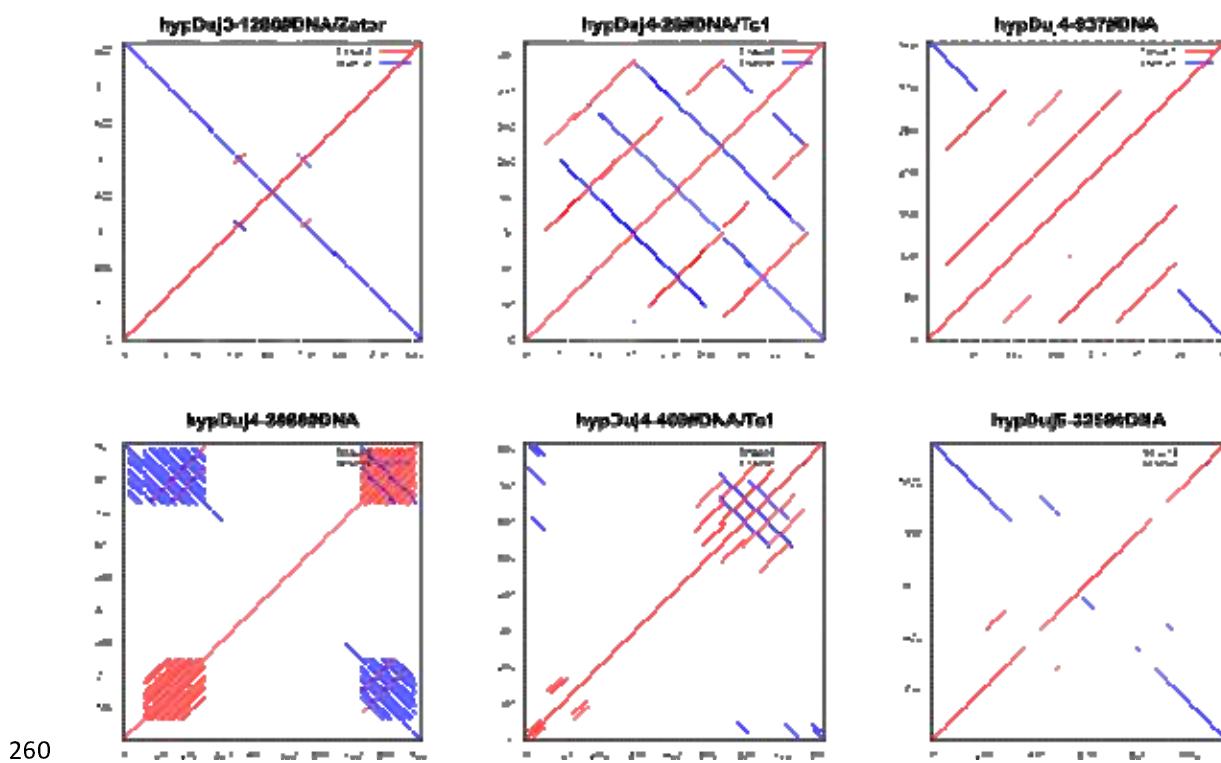
247 **Figure 2.** Comparison of the length of the consensus sequences before and after manual
248 curation.

249

250 **Diversity of transposable elements**

251 When looking at the diversity of repeats in the partially curated libraries (libraries comprising
252 both curated and uncurated consensus sequences), we identified a total of 419 Class II DNA
253 consensus sequences belonging to the superfamilies/clades CMC, MULE, TcMar, Sola,
254 PiggyBac, Tc4, PIF-Harbinger, Zator, hAT, Maverick, and P. Many of these elements are
255 non-autonomous and show a remarkable diversity of internal structures (**Figure 3**). For Class
256 I retrotransposons, we found 40 LINEs belonging to the superfamilies/clades CR1, CRE, R2,
257 R2-NesL, L2, RTE-X and RTE-BovB and other 35 LTRs belonging to the
258 superfamilies/clades DIRS, Gypsy, Ngaro and Pao.

259

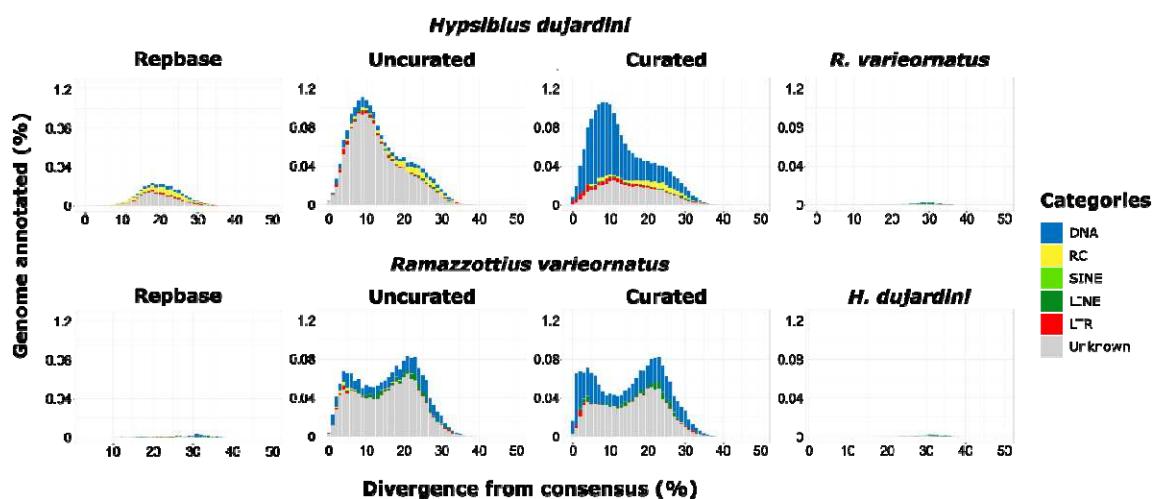


261 **Figure 3.** Dotplots of six DNA transposons from the library of *Hypsibius dujardini* produced
262 with the MAFFT online server. These elements were selected by course participants for
263 aesthetic reasons.

264
265 To highlight the importance of generating and using custom repeat libraries for the organisms
266 of interest as well as their curation, we masked the two tardigrade genomes and compared
267 how the annotation and accumulation patterns change when using general repeat libraries (in
268 this case the Repbase library for Arthropoda) and species-specific ones before and after
269 curation (**Figure 4, Table 2**). The use of the known repeats for Arthropoda available on
270 Repbase provided a poor and insufficient annotation for both species (all the following
271 percentages are given for *H. dujardini* and then for *R. varieornatus*) where only 1.95% and
272 0.26% of the assemblies were annotated as interspersed repeats and the accumulation patterns
273 were characterized only by likely old insertions. Then the use of species-specific, albeit
274 uncurated, libraries completely changed the percentage of TEs annotated (16.38% and
275 15.66%) and their accumulation patterns that showed many recently accumulated insertions.

276 While the shape and percentages of the repeat landscapes did not drastically change after the
277 manual curation of the libraries, the curated libraries clearly highlighted a large accumulation
278 of DNA transposons in recent and ancient times alike that were either not present in the other
279 landscapes or were hidden among the “unknown” repeats. Especially for *R. varieornatus*, the
280 curation highlighted a higher accumulation of repeats in the very recent times (1-5% of
281 divergence). This higher accumulation of DNA transposons in recent times is also in line with
282 the finding of multiple putatively active transposable element subfamilies (**Table 3**). Finally,
283 the use of the repeat library of one species to annotate the other species (reciprocal masking)
284 resulted to be almost as insufficient as the use of the Repbase library for Arthropoda stressing
285 once again how important it is to have a capillary knowledge of the repeatome for correct
286 biological interpretations.

287



288

289 **Figure 4.** Repeat landscapes of the genomes of *H. dujardini* and *R. varieornatus* annotated
290 with the Repbase (Arthropoda clade), uncurated and curated of both tardigrades combined
291 libraries, and with libraries of the reciprocal species (only species-specific repeats). The
292 divergence from consensus calculated with the Kimura 2-parameter distance model is shown
293 on the x-axis. The percentage of genome annotated is shown on the y-axis.

294

295 **Table 2.** Number of base pairs annotated and percentages of the main TE categories.

Species	Library	DNA (bp)	DNA (%)	LINE (bp)	LINE (%)	SINE (bp)	SINE (%)	LTR (bp)	LTR (%)	Unknown (bp)	Unknown (%)	Total (bp)	Total (%)
<i>Hypsibius dujardini</i>	Repbase Arthropoda	347033	0.34	75334	0.07	264	0	200462	0.2	1370894	1.34	1993987	1.95
	Uncurated	1681052	1.65	310239	0.3	5366	0.01	514564	0.5	14199202	13.92	16710223	16.38
	Curated	11149552	10.93	290632	0.28	2424	0	868156	0.85	4658887	4.57	16869651	16.63
<i>Ramazzottius varieornatus</i>	<i>R. varieornatus</i>	62676	0.06	60480	0.06	0	0	8437	0.01	60917	0.06	192510	0.19
	Repbase Arthropoda	68902	0.12	33938	0.06	266	0	16972	0.03	23959	0.04	144037	0.26
	Uncurated	1753754	3.16	413647	0.75	4486	0.01	134451	0.24	6375274	11.5	8681612	15.66
296	Curated	3385077	6.11	454742	0.82	1320	0	145257	0.26	4880857	8.81	8867253	16
	<i>H. dujardini</i>	45939	0.08	40575	0.07	1320	0	6334	0.01	49444	0.09	143612	0.26

297

298 **Table 3:** List of repeat subfamilies with putatively ongoing activity, i.e., at least 10 copies
299 with 0% distance to consensus.

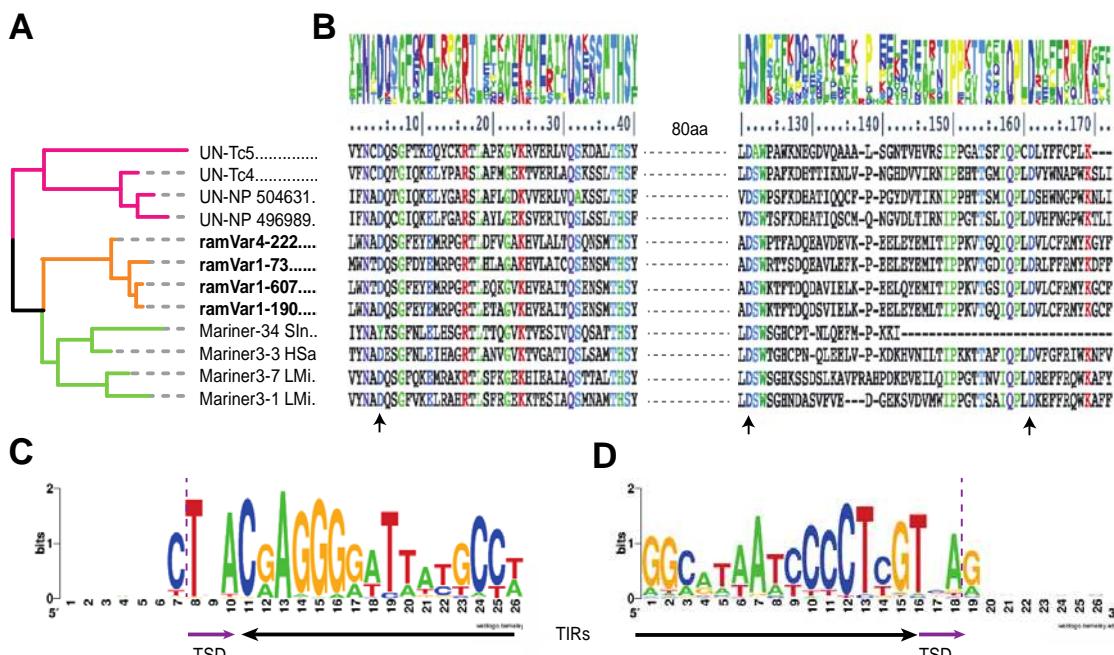
TE category	<i>Hypsibius dujardini</i>	<i>Ramazzottius varieornatus</i>
DNA transposon	7	3
LTR retrotransposon	3	0
Unknown	0	2

300

301 As a demonstrative example of the contribution of the collaborative curation process in
302 providing novel insights into TEs diversity, taxonomic distribution and biology, we decided to
303 deeply characterize consensus sequences that we classified as Tc4. These elements have a
304 rather limited taxonomic distribution, few references in bibliography exist, and they
305 incompletely duplicate the target site upon transposition [26] which can impose challenges for
306 their classification. The Tc4 transposons are DDD elements firstly discovered in
307 *Caenorhabditis elegans* [26] where they recognize the interrupted palindrome CTNAG as
308 target site for insertion, and cause duplication of only the central TNA trinucleotide.
309 Regarding their taxonomic distribution, consensus sequences for Tc4 elements are known and
310 deposited only for nematodes and arthropods in RepeatPeps, Repbase and DFAM.
311 Phylogenetic analyses based on DDD segments confidently placed the four tardigrade Tc4

312 consensus sequences identified in *R. varieornatus* within the Tc4 clade in a sister relationship
313 with arthropod elements and with a branching pattern that reassemble the Panarthropoda
314 group (tardigrades + onychophorans + arthropods) within Ecdysozoa [27] (**Figure 5A**). The
315 DDD catalytic domain resulted to be highly conserved between different phyla (**Figure 5B**)
316 and the target site of tardigrades mirror what was previously observed in nematodes (i.e.,
317 C|TNA|G where “|” marks the transposase cut site **Figure 5C-D**). We could therefore
318 hypothesize that these elements first originated during the diversification of Ecdysozoa.
319 However, broader comparative analyses involving more early-diverging Metazoa clades are
320 necessary to confirm this lineage-specific origin.

321



322

323 **Figure 5.** Characterization and phylogenetic analyses of Tc4 elements. (A) Phylogenetic tree
324 of Tc4 consensus sequences based on DDD catalytic domains identified in the *R. varieornatus*
325 consensus sequences, highlighted in bold and orange, together with representative sequences
326 extracted from the RepeatPeps library from nematodes (pink) and insects (green). All nodes
327 received maximal support value. (B) Alignment of DDD catalytic domains of sequences

328 included in phylogenetic analyses. Residues conserved in more than 80% of the sequences are
329 colored. Arrows highlight catalytic DDD residues. Sequence logos of 5' (C) and 3' (D) ends
330 of Tc4 elements used to curate the *R. varieornatus* consensus sequences. Black and purple
331 arrows denote terminal inverted repeats (TIRs) and target site duplications (TSDs),
332 respectively. The purple dotted line marks the transposase cut on the CTNAG target site.

333

334 **Contributions from the course participants**

335 During both editions of the course, participants were free to explore their favorite topics
336 within the scope of the syllabus and we here share two contributions developed by the
337 participants that can be useful for the entire community. First, an additional repeat library of
338 130 consensus sequences (119 of which are DNA transposons) was produced with the use of
339 REPET for *R. varieornatus* (**Table S4**). Second, a guide for the classification of TEs from
340 multisequence alignments (**File S1**) that can be a useful starting point for beginners and
341 complementary to more extensive guides [13,14].

342 **Conclusion**

343 As shown here and in many other studies, repeat annotation is key to correctly identify and
344 interpret patterns of genome evolution and proper annotation is based on a thorough curation
345 of the repeat libraries [8,9,28]. However, it is hard for curation efforts to keep up with the
346 sheer amount of genomes released every year as curation done by single laboratories may
347 require months or even years for a single genome. Recent developments of machine learning-
348 based tools to automatize the curation and classification processes are promising [29–32] and
349 there are additional tools to facilitate the curation process like TEAid [13] and EarlGrey [33].
350 Until fully automatized, reliable tools are developed and there are manual curation training
351 sets for understudied taxa, we emphasize the need to implement manual curation for repeat
352 libraries as well as to find alternative ways to deal with the curation of hundreds of new

353 libraries. Here we presented one such alternative approach, namely a peer-reviewed course
354 sourcing effort designed to be as reproducible and comparable as possible and where the
355 hands-on tutorials were designed to be meaningful for the participants because they dealt with
356 real unexplored data and directly contributed to the scientific community. The two iterations
357 of this course sourcing effort resulted in the successful curation of hundreds of new and
358 diverse TEs and we hope that this experience and teaching framework can be of use for the
359 genomic and TE communities and to be applicable to other types of data/analysis that need
360 manual curation (e.g., genome assemblies [21,22] and gene annotations).

361 **Materials and Methods**

362 **Genome assemblies**

363 For this study, we used the genome assemblies of the two tardigrade species: *Hypsibius*
364 *dujardini* (GCA_002082055.1) and *Ramazzottius varieornatus* (GCA_001949185.1)
365 produced by sequencing a pool of male and female individuals by Yoshida et al. [34]. The
366 *Hypsibius dujardini* genome was assembled using long PacBio and short Illumina reads
367 whereas the *Ramazzottius varieornatus* genome was assembled using a combination of Sanger
368 and Illumina reads [34].

369 **Raw repetitive element library**

370 To start the *de novo* characterization of transposable elements, we ran RepeatModeler on *H.*
371 *dujardini* and RepeatModeler2 on *R. varieornatus* [35] using the option -LTR_struct and
372 obtained a library of raw consensus sequences for each of the genomes. RepeatModeler and
373 RepeatModeler2 automatically named the consensus sequences with the prefix “rnd” that we
374 replaced with the abbreviations of the species names: “hypDuj” for *H. dujardini* and “ramVar”
375 for *R. varieornatus*.

376 The two libraries were then compared to find similar sequences belonging either to the same
377 family or subfamily by using, respectively, the 80-80-80 rule [36] and the 95-80-98 rule [37].
378 The comparison was done by masking the library of *R. varieornatus* with the library of *H.*
379 *dujardini* using RepeatMasker [38].

380 **Manual curation of the consensus sequences**

381 After the generation of the libraries of raw consensus sequences, we proceeded with the
382 collaborative peer-reviewed manual curation step. The participants of the course were split
383 into ten groups and each group received about 80 consensus sequences to curate.

384 The first step of the curation consisted in the alignment of the raw consensus sequences to the
385 genome of origin using BLAST [39]. The best 20 BLAST hits were selected aligned with
386 their raw consensus sequence with MAFFT [40] which produced a multisequence alignment
387 for each consensus sequence ready to be manually curated (script RepeatModelerPipeline.pl).

388 Each of the multisequence alignment was then inspected to: 1) find the actual boundaries of
389 the repetitive elements; 2) build a new consensus sequence with Advanced Consensus Maker
390 (<https://hcv.lanl.gov/content/sequence/CONSENSUS/AdvConExplain.html>); 3) fix
391 ambiguous base and gap calls in the new consensus sequence following a majority rule; 4)
392 find sequence hallmarks to define the repetitive elements as transposable elements (e.g., target
393 site duplication, long terminal repeats, terminal inverted repeats or other motifs). Every new
394 consensus sequence was reported in a common Excel table (**Table S1**). To quantitatively
395 measure the improvement of the repeat libraries after manual curation, we compared the
396 length of consensus sequences before and after curation.

397 In all the figures and tables, the term “curated” indicates that the library mentioned contains
398 manually curated consensus sequences as well as all the consensus sequences that remained
399 uncurated. Finally, we consider each consensus sequence as a proxy for a transposable
400 element subfamily. However, the consensus sequences were not checked for redundancy and

401 not clustered into families and subfamilies using the 80-80-80 or 95-80-98 rules for
402 nomenclature because the focus of the study was on classifying the consensus sequences into
403 superfamilies and orders of transposable elements.

404 The code used to produce the consensus sequences and their alignments is provided as tutorial
405 on the GitHub repository <https://github.com/ValentinaPeona/TardigradeTE>.

406 **Classification**

407 The new consensus sequences were classified using sequence characteristics retrieved by the
408 alignments (e.g., target site duplications, terminal repeats) and homology information
409 retrieved through masking the sequences with Censor [41,42] following the recommendations
410 from [36] and [43]. When the information retrieved by the alignments and Censor were not
411 enough to provide a reliable classification of the elements, the sequences were further
412 analyzed for the presence of informative protein domains using Conserved Domain Database
413 [44–46].

414 Since the course participants in general had never curated transposable element alignments
415 before, we decided to implement a peer-review process. For the first course (*H. dujardini*), the
416 results of each participant were sent to another participant to check the curated alignments and
417 independently retrieve key information for the classification. The independent sequences and
418 classifications would be compared and fixed if necessary. In the second course (*R.*
419 *varieornatus*), all sequences were inspected by the same 3 reviewers only who applied the
420 same process as previously described.

421 **Comparative analysis of the repetitive content**

422 The genome assemblies of both tardigrade species were masked with RepeatMasker 4.1.10
423 using four different types of TE libraries: 1) known Arthropoda consensus sequences from
424 Repbase; 2) raw uncurated consensus sequences from the respective species; 3) curated
425 consensus sequences together with the consensus sequences that were not curated from the

426 respective species; 4) curated consensus sequences together with the consensus sequences that
427 were curated from the other species. The RepeatMasker output files were then used to get the
428 percentages of the genomes annotated as TEs and to visualize the landscapes of the
429 accumulation of repeats.

430 Finally, we estimated the number of putative active transposable elements in the two genomes
431 by filtering the RepeatMasker annotation for elements that show at least 10 copies with a 0%
432 divergence from their consensus sequences.

433 **Characterization of Tc4 elements**

434 During the manual curation process, participants found types of DNA transposons that are
435 currently considered to have a rather restricted phylogenetic distribution like Tc4 Mariner
436 elements, therefore more in-depth analyses were run on these elements. The protein domains
437 of known Tc elements were compared to the Tc4 consensus sequences from the tardigrade
438 species and phylogenetic relationships were established.

439 Protein homologies of the partially curated repeat libraries were collected using BlastX (e-
440 value 1e-05) [47] against a database of TE-related protein (RepeatPeps library) provided with
441 the RepeatMasker installation. We extracted the amino acid translation of each hit on Tc4
442 elements based on the coordinates reported in the BlastX output. Resulting protein sequences
443 were aligned together with all members of the TcMar superfamily present in RepeatPeps
444 library using MAFFT (*L-INS-i* mode) [48] and the alignment was manually inspected to
445 identify and isolate the catalytic DDD domain. The resulting trimmed alignment was used for
446 phylogenetic inference with IQ-TREE-2 [49], identifying the best-fit evolutionary model with
447 ModelFinder2 and assessing nodal support with 1000 UltraFastBootstrap replicates [50]. The
448 resulting maximum likelihood tree was mid-point rooted and the Tc4 subtree extracted for
449 visualization purposes. The DDD segments of Tc4 elements were re-aligned using T-Coffee
450 in *expresso* mode [51] to produce conservation scores. A sequence logo of 5' and 3'

451 boundaries of identified Tc4 elements was produced extracting all sequences used to curate
452 the four *R. varieornatus* Tc4 elements and keeping the first 15 bp and 11 bp before and after
453 the terminal inverted repeats (TIRs), respectively.

454 **Additional transposable element library**

455 Participants ran REPET tool V3.0 [52] to produce a de novo transposable element library for
456 *R. varieornatus* in parallel to the one generated by RepeatModeler2. A custom TE library
457 composed by repeats from Repbase and from *H. dujardini* was used to aid REPET in the
458 classification process. Only consensus sequences that showed two or more full-length copies
459 in the *R. varieornatus* genome were retained in the new library. Furthermore, the consensus
460 sequences were scanned for protein domains and presence of TIRs or long terminal repeats
461 (LTRs).

462 **Abbreviations**

463 LTR: Long Terminal Repeats

464 TE: transposable element

465 TIR: Terminal Inverted Repeats

466 **Declarations**

467 *Ethics approval and consent to participate*

468 Not applicable.

469 *Consent for publication*

470 Not applicable.

471 *Availability of data and materials*

472 All data generated or analyzed during this study are included in this published article and its
473 supplementary information files. All newly curated repeat consensus sequences were
474 deposited in Dfam. The code for the tutorials used in the course as well as for the analysis of
475 the manuscript can be found on GitHub: <https://github.com/ValentinaPeona/Tardigrate>.

476 *Competing interests*

477 Carlo Pecoraro is founder of Physalia Courses (<http://www.physalia-courses.org/>) but had no
478 role in the design of the study.

479 *Authors' contributions*

480 AS conceived the project and VP contributed to its development. VP and JM analyzed the
481 data. AS, VP, JM wrote the manuscript, and all authors revised the manuscript. MT, AM, DA,
482 JS, GP provided additional contributions to the teaching material. All authors except CP
483 contributed to the curation of the repeat library. CP provided and maintained the
484 computational infrastructure during the courses. Authors are listed in alphabetical order.

485 *Acknowledgements*

486 Part of the analysis were performed on resources provided by the Swedish National
487 Infrastructure for Computing (SNIC) through Uppsala Multidisciplinary Center for Advanced
488 Computational Science (UPPMAX) and CSC-IT Finland.

489

490 **References**

491 1. Osmanski AB, Paulat NS, Korstian J, Grimshaw JR, Halsey M, Sullivan KAM, et al. Insights into
492 mammalian TE diversity through the curation of 248 genome assemblies. *Science* (1979) [Internet].
493 2023;380:eabn1430. Available from: <https://doi.org/10.1126/science.abn1430>

494 2. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic
495 genomes. *Mob DNA* [Internet]. 2015;6:11. Available from: <https://doi.org/10.1186/s13100-015-0041-9>

497 3. Wicker T. The repetitive landscape of the chicken genome. *Genome Res* [Internet]. 2004;15:126–
498 36. Available from: <http://genome.cshlp.org/content/15/1/126.abstract>

499 4. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, et al. Sequence and
500 comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.
501 *Nature* [Internet]. 2004;432:695–716. Available from: <https://doi.org/10.1038/nature03154>

502 5. Boman J, Frankl-Vilches C, da Silva dos Santos M, de Oliveira EHC, Gahr M, Suh A. The Genome of
503 Blue-Capped Cordon-Bleu Uncovers Hidden Diversity of LTR Retrotransposons in Zebra Finch. *Genes*
504 (Basel) [Internet]. 2019;10:301. Available from: <https://www.mdpi.com/2073-4425/10/4/301>

505 6. Kapusta A, Suh A, Feschotte C. Dynamics of genome size evolution in birds and mammals. *Proc
506 Natl Acad Sci U S A* [Internet]. 2017;114:E1460–9. Available from:
507 <http://www.pnas.org/content/114/8/E1460.abstract>

508 7. Sproul J, Hotaling S, Heckenhauer J, Powell A, Marshall D, Larracuente AM, et al. 600+ insect
509 genomes reveal repetitive element dynamics and highlight biodiversity-scale repeat annotation
510 challenges. *Genome Res* [Internet]. 2023; Available from:
511 <http://genome.cshlp.org/content/early/2023/09/22/gr.277387.122.abstract>

512 8. Platt RN, Blanco-Berdugo L, Ray DA. Accurate transposable element annotation is vital when
513 analyzing new genome assemblies. *Genome Biol Evol* [Internet]. 2016;8:403–10. Available from:
514 <https://doi.org/10.1093/gbe/evw009>

515 9. Peona V, Blom MPK, Xu L, Burri R, Sullivan S, Bunikis I, et al. Identifying the causes and
516 consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Mol
517 Ecol Resour* [Internet]. 2021;21:263–86. Available from:
518 <https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13252>

519 10. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for
520 automated genomic discovery of transposable element families. *Proceedings of the National
521 Academy of Sciences* [Internet]. 2020;117:9451–7. Available from:
522 <https://doi.org/10.1073/pnas.1921046117>

523 11. Zeng L, Kortschak RD, Raison JM, Bertozi T, Adelson DL. Superior ab initio identification,
524 annotation and characterisation of TEs and segmental duplications from genome assemblies. *PLoS
525 One* [Internet]. 2018;13:e0193588-. Available from: <https://doi.org/10.1371/journal.pone.0193588>

526 12. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, et al. Combined
527 Evidence Annotation of Transposable Elements in Genome Sequences. *PLoS Comput Biol* [Internet].
528 2005;1:e22-. Available from: <https://doi.org/10.1371/journal.pcbi.0010022>

529 13. Goubert C, Craig RJ, Bilat AF, Peona V, Vogan AA, Protasio A V. A beginner’s guide to manual
530 curation of transposable elements. *Mob DNA* [Internet]. 2022;13:7. Available from:
531 <https://doi.org/10.1186/s13100-021-00259-7>

532 14. Storer JM, Hubley R, Rosen J, Smit AFA. Curation Guidelines for de novo Generated Transposable
533 Element Families. *Curr Protoc* [Internet]. 2021;1:e154. Available from:
534 <https://doi.org/10.1002/cpz1.154>

535 15. Elliott TA, Heitkam T, Hubley R, Quesneville H, Suh A, Wheeler TJ, et al. TE Hub: A community-
536 oriented space for sharing and connecting tools, data, resources, and methods for transposable
537 element annotation. *Mob DNA* [Internet]. 2021;12:16. Available from:
538 <https://doi.org/10.1186/s13100-021-00244-0>

539 16. Leung W, Shaffer CD, Chen EJ, Quisenberry TJ, Ko K, Braverman JM, et al. Retrotransposons Are
540 the Major Contributors to the Expansion of the *Drosophila ananassae* Muller F Element. *G3*
541 *Genes|Genomes|Genetics* [Internet]. 2017;7:2439–60. Available from:
542 <https://doi.org/10.1534/g3.117.040907>

543 17. Moya ND, Stevens L, Miller IR, Sokol CE, Galindo JL, Bardas AD, et al. Novel and improved
544 *Caenorhabditis briggsae* gene models generated by community curation. *BMC Genomics*. 2023;24.

545 18. Chang WH, Mashouri P, Lozano AX, Johnstone B, Husić M, Olry A, et al. Phenotate: crowdsourcing
546 phenotype annotations as exercises in undergraduate classes. *Genetics in Medicine* [Internet].
547 2020;22:1391–400. Available from: <https://doi.org/10.1038/s41436-020-0812-7>

548 19. Zhou N, Siegel ZD, Zarecor S, Lee N, Campbell DA, Andorf CM, et al. Crowdsourcing image analysis
549 for plant phenomics to generate ground truth data for machine learning. *PLoS Comput Biol* [Internet].
550 2018;14:e1006337-. Available from: <https://doi.org/10.1371/journal.pcbi.1006337>

551 20. Singh M, Bhartiya D, Maini J, Sharma M, Singh AR, Kadarkaraisamy S, et al. The Zebrafish
552 GenomeWiki: a crowdsourcing approach to connect the long tail for zebrafish gene annotation.
553 Database [Internet]. 2014;2014:bau011. Available from: <https://doi.org/10.1093/database/bau011>

554 21. Prost S, Winter S, De Raad J, Coimbra RTF, Wolf M, Nilsson MA, et al. Education in the genomics
555 era: Generating high-quality genome assemblies in university courses. *Gigascience* [Internet].
556 2020;9:giaa058. Available from: <https://doi.org/10.1093/gigascience/giaa058>

557 22. Prost S, Petersen M, Grethlein M, Hahn SJ, Kuschik-Macollek N, Olesiuk ME, et al. Improving the
558 Chromosome-Level Genome Assembly of the Siamese Fighting Fish (*Betta splendens*) in a University
559 Master's Course. *G3 Genes|Genomes|Genetics* [Internet]. 2020;10:2179–83. Available from:
560 <https://doi.org/10.1534/g3.120.401205>

561 23. Yoshida Y, Koutsovoulos G, Laetsch DR, Stevens L, Kumar S, Horikawa DD, et al. Comparative
562 genomics of the tardigrades *Hypsibius dujardini* and *Ramazzottius varieornatus*. Tyler-Smith C, editor.
563 *PLoS Biol* [Internet]. 2017;15:e2002266. Available from:
564 <https://doi.org/10.1371/journal.pbio.2002266>

565 24. Møbjerg N, Halberg KA, Jørgensen A, Persson D, Bjørn M, Ramløv H, et al. Survival in extreme
566 environments – on the current knowledge of adaptations in tardigrades. *Acta Physiologica* [Internet].
567 2011;202:409–20. Available from: <https://doi.org/10.1111/j.1748-1716.2011.02252.x>

568 25. Peter D, Bertolani R, Guidetti R. Actual checklist of Tardigrada species. 2019;

569 26. Yuan JY, Finney M, Tsung N, Horvitz HR. Tc4, a *Caenorhabditis elegans* transposable element with
570 an unusual fold-back structure. *Proceedings of the National Academy of Sciences*. 1991;88:3334–8.

571 27. Giribet G, Edgecombe GD. Current Understanding of Ecdysozoa and its Internal Phylogenetic
572 Relationships. *Integr Comp Biol* [Internet]. 2017;57:455–66. Available from:
573 <https://doi.org/10.1093/icb/icx072>

574 28. Peona V, Kutschera VE, Blom MPK, Irestedt M, Suh A. Satellite DNA evolution in Corvoidea
575 inferred from short and long reads. *Mol Ecol* [Internet]. 2022;0–64. Available from:
576 <https://onlinelibrary.wiley.com/doi/10.1111/mec.16484>

577 29. Panta M, Mishra A, Hoque MT, Atallah J. ClassifyTE: a stacking-based prediction of hierarchical
578 classification of transposable elements. *Bioinformatics* [Internet]. 2021;37:2529–36. Available from:
579 <https://doi.org/10.1093/bioinformatics/btab146>

580 30. Orozco-Arias S, Lopez-Murillo LH, Piña JS, Valencia-Castrillon E, Tabares-Soto R, Castillo-Ossa L, et
581 al. Genomic object detection: An improved approach for transposable elements detection and
582 classification using convolutional neural networks. *PLoS One* [Internet]. 2023;18:e0291925-. Available
583 from: <https://doi.org/10.1371/journal.pone.0291925>

584 31. Bickmann L, Rodriguez M, Jiang X, Makalowski W. TEclass2: Classification of transposable
585 elements using Transformers. *bioRxiv* [Internet]. 2023;2023.10.13.562246. Available from:
586 <http://biorxiv.org/content/early/2023/10/16/2023.10.13.562246.abstract>

587 32. Orozco-Arias S, Isaza G, Guyot R, Tabares-Soto R. A systematic review of the application of
588 machine learning in the detection and classification of transposable elements. Nakai K, editor. *PeerJ*
589 [Internet]. 2019;7:e8311. Available from: <https://doi.org/10.7717/peerj.8311>

590 33. Baril T, Imrie RM, Hayward A. Earl Grey: a fully automated user-friendly transposable element
591 annotation and analysis pipeline. *bioRxiv* [Internet]. 2022;2022.06.30.498289. Available from:
592 <http://biorxiv.org/content/early/2022/07/02/2022.06.30.498289.abstract>

593 34. Yoshida Y, Koutsopoulos G, Laetsch DR, Stevens L, Kumar S, Horikawa DD, et al. Comparative
594 genomics of the tardigrades *Hypsibius dujardini* and *Ramazzottius varieornatus*. Tyler-Smith C, editor.
595 *PLoS Biol.* 2017;15:e2002266.

596 35. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for
597 automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A*.
598 2020;117:9451–7.

599 36. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification
600 system for eukaryotic transposable elements. *Nat Rev Genet.* 2007;8:973–82.

601 37. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering Transposable Element Diversification in
602 De Novo Annotation Approaches. *PLoS One.* 2011;6:e16526.

603 38. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0 [Internet]. 2015. Available from:
604 <http://www.repeatmasker.org>

605 39. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture
606 and applications. *BMC Bioinformatics.* 2009;10:421.

607 40. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: Multiple sequence alignment, interactive
608 sequence choice and visualization. *Brief Bioinform.* 2018;20:1160–6.

609 41. Kapitonov V V., Jurka J. A universal classification of eukaryotic transposable elements
610 implemented in Repbase. *Nat Rev Genet.* 2008;9:411–2.

611 42. Kohany O, Gentles AJ, Hankus L, Jurka J. Annotation, submission and screening of repetitive
612 elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics.* 2006;7:474.

613 43. Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev
614 Genet.* 2007;41:331–68.

615 44. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, et al. CDD: a
616 Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.*
617 2011;39:D225–9.

618 45. Marchler-Bauer A, Bryant SH. CD-Search: protein domain annotations on the fly. *Nucleic Acids
619 Res.* 2004;32:W327–31.

620 46. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. CDD/SPARCLE: the conserved
621 domain database in 2020. *Nucleic Acids Res.* 2020;48:D265–8.

622 47. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture
623 and applications. *BMC Bioinformatics* [Internet]. 2009;10:421. Available from:
624 <https://doi.org/10.1186/1471-2105-10-421>

625 48. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: Multiple sequence alignment, interactive
626 sequence choice and visualization. *Brief Bioinform* [Internet]. 2018;20:1160–6. Available from:
627 <https://doi.org/10.1093/bib/bbx108>

628 49. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, et al. IQ-TREE
629 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol.*
630 2020;37:1530–4.

631 50. Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast
632 bootstrap approximation. *Mol Biol Evol.* 2018;35:518–22.

633 51. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple
634 sequence alignment. *J Mol Biol.* 2000;302:205–17.

635 52. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering Transposable Element Diversification in
636 De Novo Annotation Approaches. *PLoS One* [Internet]. 2011;6:e16526. Available from:
637 <https://doi.org/10.1371/journal.pone.0016526>

638