Computational discovery of hidden breaks in 28S rRNA across eukaryotes and consequences for RNA Integrity Numbers.

Paschalis Natsidis[1], Philipp H. Schiffer[1], Irepan Salvador-Martínez[1], Maximilian J. Telford[1*]

1. Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Ecology, University College London, Gower Street, London WC1E 6BT

*Corresponding Author: m.telford@ucl.ac.uk

Abstract.

In some eukaryotes, a 'hidden break' has been described in which the 28S ribosomal RNA molecule is cleaved into two subparts. The break is common in protostome animals but has also been reported in some vertebrates and non-metazoan eukaryotes. We present a new computational approach to determine the presence of the hidden break in 28S rRNAs using mapping of RNA-Seq data. We find an homologous break is present across protostomes although has been lost in some taxa. We show that rare breaks in vertebrates are not homologous to the protostome break. A break is found in just 3 out of 307 species of non-animal eukaryotes studied but these are located in the same position as the protostome break suggesting a striking instance of convergent evolution. RNA Integrity Numbers (RIN) rely on intact 28s rRNA and will be consistently underestimated in the great majority of animal species with a break.

21 **Introduction**.

22 Ribosomes are made up of up to 80 ribosomal proteins and three (in prokaryotes) or four (in eukaryotes)
23 structural ribosomal RNAs named according to their sizes: in eukaryotes these are the 5S (~120
24 nucleotides), the 5.8S (~150 nucleotides), the 18S (~1800 nucleotides) and the 28S (~4000 to 5000
25 nucleotides). The 5.8S, 18S and 28S rRNAs are initially transcribed as a single RNA operon (the 5S is at a
26 separate locus in eukaryotes). The 18S and 5.8S are separated by the Internal Transcribed Spacer 1 (ITS1)
27 and 5.8S and 28S are separated by ITS2. The initial transcript is cleaved into three functional RNAs by
28 removing ITS1 and ITS2 (Fig. 1A).
29
30 In some species this picture is complicated by observations of a 'hidden break' in the 28S ribosomal RNA.
31 In organisms with a hidden break, the 28S rRNA molecule itself is cleaved into two approximately equal
32 sized molecules of ~2000 nucleotides each (Fig. 1A) (Winnebeck et al., 2010). These two RNAs (the 5'
33 28Sα and the 3' 28Sβ) are nevertheless intimately linked by inter-molecular hydrogen bonding within the
34 large subunit of the ribosome, just as the 28S and 5.8S rRNAs as well as different regions of the intact
35 28S rRNA are in species without a break.
36
37 The hidden break, first described in pupae of the silkmoth *Hyalophora cecropia* (Applebaum et al., 1966)
38 manifests itself experimentally when total RNA is extracted and separated according to size using gel
39 electrophoresis. If the RNA is first denatured by heating in order to separate hydrogen bonded stretches
40 of RNAs before electrophoresis, the hydrogen bonded 28Sα and 28Sβ subunits separate. These two
41 molecules, being approximately the same size, then migrate on a gel at the same rate as each other and,
42 coincidentally, at the same rate as the almost identically sized 18S rRNA molecule. The effect is that, rather
43 than observing two distinct ribosomal RNA bands on the gel of ~2000 and ~4000 nucleotides, a single
44 (and more intense) band composed of three different molecules each approximately 2kb long (18S,
45 28Sα and 28Sβ) is seen (McCarthy et al., 2015).
46
47 This behaviour of rRNA molecules in species with a hidden break can have interesting and potentially
48 detrimental practical consequences. The integrity of 18S and 28S rRNA molecules, as evidenced by the
49 presence of 2kb and 4-5kb bands/peaks following electrophoresis, is widely used as an indication that a
50 total RNA sample is not degraded. This is an important quality control step for experiments that will use
51 the total RNA sample for subsequent experiments such as transcriptome sequencing, Northern blotting
52 or rtPCR. The requirement to observe two rRNA peaks in an intact RNA sample has been formalised as
53 an important part of the RIN (RNA Integrity Number, Schroeder et al. 2006). For RNA samples extracted
54 from species possessing the hidden break, if conditions allow the separation of the two 28S subunits, a
55 single rRNA band/peak will be observed. The absence of two distinct rRNA peaks will give the impression
56 of a degraded sample with a low RIN, even when the RNA is not degraded.

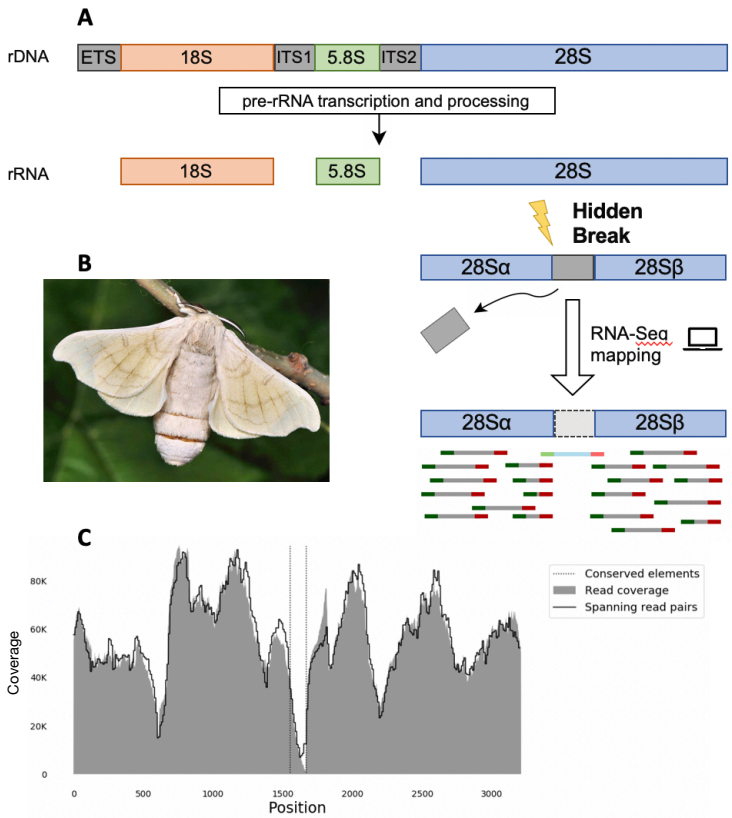

Figure 1. The hidden break and how to diagnose it. A) After post-transcriptional processing, the eukaryotic pre-rRNA molecule is cleaved to produce the 18S, 5.8S and 28S subunits. In some species, the 28S subunit gets cleaved into two subunits, a phenomenon known as the hidden break. B) The domestic silkmoth (*Bombyx mori*, Lepidoptera) is one of the few species with a characterised hidden break region (Fujiwara & Ishikawa (1986), photo by Paolo Mazzei, www.leps.it) C) RNA-Seq reads of *B. mori* mapped onto its 28S rRNA sequence. The hidden break is revealed as a drop in the read coverage and the number of spanning read pairs in the middle of the 28S sequence. The break is located between two highly conserved flanking regions (vertical lines).

Since its description in the silkmoth, the 28S hidden break has been found in numerous other species, although it is by no means universal. The most significant systematic attempts to catalogue the hidden break using gel electrophoresis date to the 1970s when Ishikawa published a series of papers (summarised in ref. Ishikawa, 1977a). Ishikawa observed the break in several arthropods and one or two representatives of other protostome phyla (Annelida, Mollusca, Rotifera, Platyhelminthes, Phoronida/Brachiopoda and Ectoprocta). The hidden break was not found in members of Deuterostomia (Chordata and Echinodermata) (Ishikawa, 1973) but was also absent from two species of nematodes (Ishikawa, 1977b) and from a species of chaetognath (Ishikawa, 1975b) (then thought to be deuterostome relatives but now known to be protostomes). The hidden break was recorded as ambiguous in non-bilaterian Cnidaria and absent in Porifera (Ishikawa 1975a, Ishikawa 1975b). The essence of Ishikawa's work was to show that the hidden break was a characteristic of the Protostomia although, surprisingly, a hidden break was also described in several unicellular eukaryotes: *Euglena*, *Acanthamoeba* and *Tetrahymena* (Ishikawa, 1975b) as well as

in plant chloroplasts (Bieri et al., 2017) The precise location of the hidden break is not revealed by electrophoresis.

Despite Ishikawa's work and other sporadic publications describing the presence of the hidden break in different taxa (Mertz et al. 1991, Karlstedt et al. 1992, Winnebeck et al. 2010, Haçarız & Sayers 2013, McCarthy 2015, DeLeo 2018), it is still currently unclear which groups of organisms possess the hidden break, whether the feature is homologous in those species that have a split 28S and has been lost in those that lack it, or whether a hidden break has evolved more than once. Finally, it is not known what the functional consequences of the hidden break might be.

Here we describe a new computational method to diagnose the presence of a hidden break in the many taxa for which large-scale RNA sequence data are available. Rather than requiring fresh material to be available for extracting total RNA to be run on an electrophoresis gel, our method is predicated on the expectation that, in taxa with a break, few if any RNA-Seq reads will map to the region that becomes excised from the initial 28S rRNA when this is broken to form the 28Sα and 28Sβ (Fig. 1A). Similarly, when paired end sequences are available, we expect there to be few if any pairs of reads spanning the region of the break, as the two sides of the break are on separate molecules.

We have tested our method by detecting known instances of the hidden break and have expanded our investigation to members of all metazoan phyla for which suitable RNA-Seq data currently exist. We also use our method to examine reported instances of the hidden break outside of the protostomes to establish whether there is evidence that these patchily distributed observations reveal hidden breaks that are homologous to the break commonly observed in the protostome animals. Finally, we have expanded our search to a diversity of eukaryotes: to verify the described instances of a break; to determine where in the molecule this exists; and to search for novel instances of a hidden break.

Results and Discussion.

Testing the method using known examples with and without the break.

To date, the presence of the hidden break has been established experimentally using electrophoresis of total RNA (McCarthy et al., 2015). A comparison of untreated and heat denatured total RNA shows the conversion of the large 28S band into two smaller bands running coincident with the 18S band: in effect the 28S band disappears (McCarthy et al., 2015). Our computational method relies, instead, on the expectation that RNA sequencing reads derived from RNA extracted from organisms possessing a 28S rRNA molecule that has been split by a deletion will show two characteristics. First, there will be an obvious absence of reads mapped to the genomic rRNA location precisely at the position of the split; second, if

112 paired end sequencing data are available, there should be no read pairs spanning from one side of the
113 split to the other as these are separate molecules.
114
115 The position of the split has been mapped experimentally in very few species, including the silkworm
116 *Bombyx mori* (Fujiwara & Ishikawa, 1986) (Fig 1B). Our method makes the accurate mapping of the position
117 of the split a simple process which is of importance: if we are to be confident that this character is
118 homologous between species it is necessary to show that the break occurs in an homologous region of
119 the molecule. The procedure for mapping reads and counting spanning reads are described in Materials
120 and Methods. Figure 1C shows an example of the pattern of read depth and numbers of spanning reads
121 that we observe in an organism in which the split is known to exist (*Bombyx*). We indicate in addition the
122 position of the break, which is known experimentally in *Bombyx mori*. This experimentally mapped break
123 is surrounded by highly conserved stretches of nucleotides allowing us to identify the homologous region
124 of the gene across eukaryotes.
125
126 *Expanding to most metazoan phyla.*
127 The hidden break has been characterised in a number of animal species and has largely been considered
128 to be specific to the protostomes. We have used our new method to expand the search for a potential
129 hidden break to members of all animal phyla for which we have found suitable RNA-Seq data. In total we
130 have examined 331 metazoan species including members of all but 4 animal phyla (Onychophora,
131 Loricifera, Micrognathozoa and Gnathostomulida). The majority of the species analysed came from the
132 Arthropoda (169) and Mollusca (31). We have also analysed 31 chordates, 28 echinoderms, as well as 2
133 hemichordates and 7 xenacoelomorphs. We have also examined data from 20 non-bilaterian species (14
134 cnidarians species, 3 ctenophores, 1 placozoan and 2 poriferans)
135
136 For the most part our results confirm those of Ishikawa; in almost all protostomes providing unambiguous
137 results (17 out of the 19 protostomian phyla) we observe the existence of the hidden break. Some
138 representative examples are shown in Fig. 2A. In each observation of the break we find it is bounded by
139 the same conserved sequence regions of the 28S molecule. Our ability to map the break means we have
140 been able to show that, across the protostomes, it is present in the same position in the molecule and we
141 conclude that this is an homologous character throughout this major group of animals. This is the first time
142 that this finding is supported by such a large taxon sampling and across such a breadth of taxa.
143
144 Unlike previous findings (Ishikawa, 1975b), we have found evidence for the hidden break in a chaetognath;
145 at the time of Ishikawa's work members of this phylum were thought to be relatives of the deuterostomes
146 but are now known to be protostomes (Papillon et al., 2004).
147

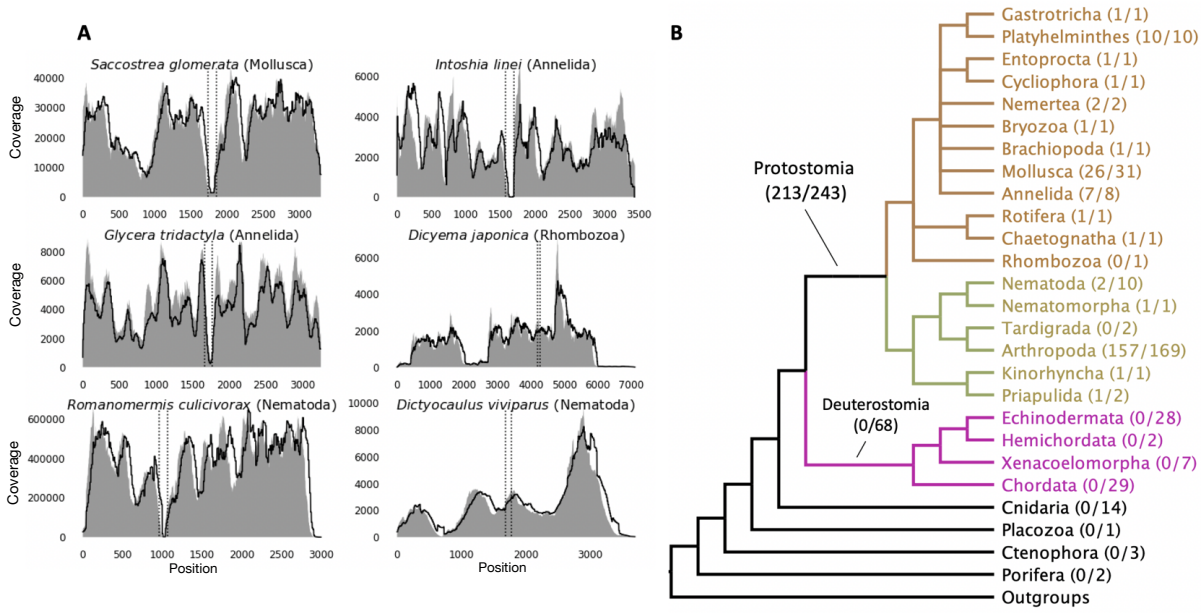

Figure 2. The hidden break is found in the protostome animals. A) Six examples of applying our method in different metazoan species. Evidence for a hidden break was found in most molluscs and annelids, including the orthonectid *I. linei*. The very fast evolving rhombozoan *D. japonica* is one of the few examples where the break was secondarily lost. The Nematoda appear to be a special case, with the break present in the enoplean species, e.g. *R. culicivorax*, but not in the Chromadorea, e.g. *D. viviparus* or the model *C. elegans* (not shown). B) The phylogenetic distribution of hidden break across metazoan phyla shows that it is a protostome synapomorphy, with occasional secondary losses.

Also noteworthy is the presence of this protostomian character in the orthonectid *Intoshia linei* although not in the dicyemid *Dicyema japonica*. While these two highly simplified animals were initially classified together within the phylum Mesozoa as a group intermediate between Protozoa and Metazoa, the dicyemids and orthonectids are now known to be separate lineages and both to be taxa within the protostomian Lophotrochozoa (Schiffer et al., 2018). The presence of the hidden break in *Intoshia* fits with this protostomian affiliation and the break has presumably been lost in *Dicyema*.

We also find evidence for the absence of the hidden break in a number of other protostomian clades (Fig 2). We confirm, for example, Ishikawa's work showing that there is no break in certain nematodes (Ishikawa, 1977b), or in the in the aphid *Acyrthosiphon pisum* (Ogino et al., 1990). We are also able to show for the first time that there is no break in the 28S rRNA of the priapulid *Priapulus caudatus* or in 2 species of tardigrades. Interestingly, in most of these cases, expanding our sampling to relatives of species without the break uncovers taxa that possess a hidden break. We find the break in the enoplean nematodes *Romanomermis* and *Trichinella* and also in the sister group of the nematodes - the Nematomorpha. The same is true of priapulids as we find the break in a second priapulid - *Halicryptus spinulosus*. While the pea aphid lacks a break (Ogino et al., 1990), we have shown the break is present in

157 other species of arthropod. The break is also known to be absent in *Thrips tabaci* (Macharia et al., 2015) yet we find evidence for a hidden break in four other closely related members of the thrips.

We find that almost every protostome species where we demonstrate the absence of the hidden break has relatives possessing it, showing that the lack of a break is a derived character - the break existed in an ancestor and has been lost. The exceptions to this rule are the tardigrades and dicyemids for which we currently have limited data. While absence of a break is infrequent in our sample of protostomes, losses have nevertheless occurred repeatedly suggesting that loss is easily achieved and relatively easily tolerated.

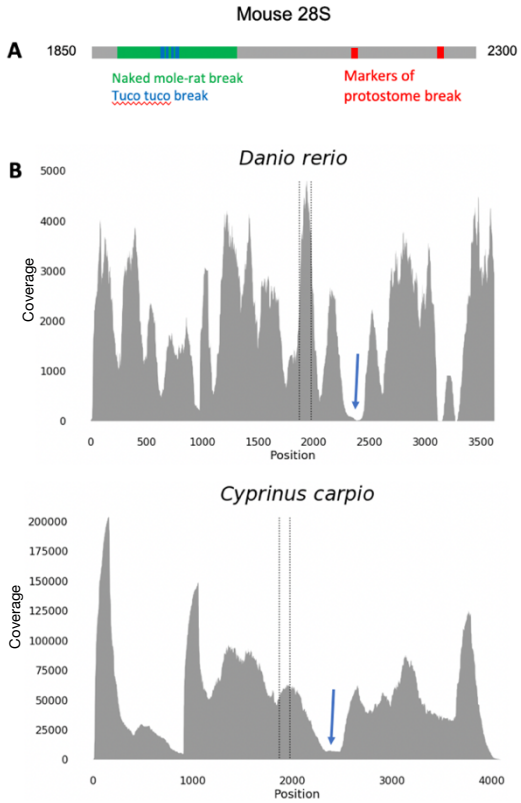

Figure 3. Vertebrate hidden breaks are not homologous to protostome breaks. A) Using the mouse 28S rRNA sequence as a reference, we show that the break regions recorded in rRNA extracted from naked mole-rat and Tuco-tuco testes are not homologous to the protostome hidden break. In both species the break is 5' of the two conserved markers we found surrounding the break in protostomes. **B)** The break recorded in cyprinid fish is not homologous to the protostome break; here we found the break is (blue arrows) 3' of the two conserved markers we found surrounding the break in protostomes.

Non Protostomian Metazoans.

We next applied our method to the two clades of deuterostomes: the Xenambulacraria (Philippe et al., 2019) (29 echinoderms, 2 hemichordates and 7 xenacoelomorphs) and the Chordata (2 urochordates and 29 vertebrates). We show that the break found in protostomes does not exist in any of the sampled species. There are, nevertheless, two special case previously described in chordates: hystricomorph

204 rodents, where species of the genus *Ctenomys* (the south american Tuctuc) and the naked mole-rat
205 (*Heterocephalus glaber*) are found to possess the hidden break (Melen 1999, Azpurua 2013). We show,
206 however, that the hidden break found in these two rodents is not homologous to the protostome hidden
207 break. We were unable to retrieve 28S sequences for these two species, however by using the mouse 28S
208 sequence as reference, we managed to locate these two breaks, as well as the two conserved markers
209 that enclose the protostome break. As it can be seen in fig. 3A, the two recorded rodent breaks do not fall
210 between these markers.
211
212 In more detail, the *Ctenomys* hidden break corresponds to nucleotides 1930-1950 in mouse 28S, while
213 the naked mole-rat hidden break falls between nucleotides 1880-2020 of the mouse 28S. The two
214 conserved satellites of the protostome hidden break lie in the positions 2149 and 2264 of the same mouse
215 28S sequence. The rodent hidden break site is located nearly 300bp upstream of the conserved
216 protostome site. The hidden break has also been reported in cyprinid fish (Leipoldt & Engel, 1983) and we
217 have analysed two members of this clade. Our results show that the break in cyprinids is also in a different
218 location that of protostomes and does not match the break described in rodents (Fig 3B).
219
220 We next looked at non-bilaterian animals to determine whether they possess the hidden break. Ishikawa's
221 work had been inconclusive in this regard - at least in the Cnidaria (Ishikawa, 1975a) - leaving open the
222 possibility that the break found across Protostomia is a primitive character that had been lost in the
223 deuterostomes. In a total of 20 non-bilaterian species (2 sponges, 14 cnidarians, 3 ctenophores, and the
224 placozoan *Trichoplax adhaerens*) we find no evidence for the existence of a hidden break (Fig 2B). The
225 distribution of the break in the animal kingdom points to the hidden break having appeared in the lineage
226 leading to the common ancestor of the protostomes.
227
228 *Convergent evolution of the hidden break outside Metazoa?*
229 Alongside the animals, there have been reports of a hidden 28S break in a small number of distantly related
230 species of eukaryotes, the ciliate *Tetrahymena* (Bostock et al., 1971), the amoebozoan *Acanthamoeba*
231 (Stevens & Pachler, 1972) and the desert truffle *Tirmania nivea* (Navarro-Ródenas et al., 2018). We have
232 investigated these taxa using our pipeline (Fig. 4A) to define the position of their breaks. We have further
233 extended our analyses of non-animal eukaryotes to over 300 non-metazoan eukaryotes, including 119
234 plants, 64 fungi and 124 other taxa (Fig. 4B). The hidden break described in plant chloroplast 23S rRNAs
235 (as opposed to nuclear 28S rRNA) does not appear to be located in the same region of the molecule
236 although the evolutionary distance makes alignment difficult (Bieri et al., 2017).
237
238 In *Acanthamoeba*, gel electrophoresis of heat denatured LSU rRNA shows that it separates into fragments
239 corresponding to molecular weights of 0.88 × 10^6 and 0.60 × 10^6 daltons (Stevens & Pachler, 1972).

Our results provide some support for this, as we find a break in the molecule that would divide it into fragments of 2320 and 1890 nucleotides (a ratio of 0.88:0.71). As suggested by the unequal sizes, the break is in a different position to the protostome 28S break. Surprisingly, the same is not true of the other two described instances of non-animal hidden breaks. We find a break as predicted in both *Tetrahymena* (Bostock et al., 1971) and the desert truffle *Tirmania* (Navarro-Ródenas et al., 2018) and in both cases the break is positioned in the same region of the 28S rRNA as the protostome hidden break. Our analysis of over 300 other eukaryotes (albeit heavily biased towards plants and fungi) found only one other clear instance of a break; this was in the unicellular opisthokont *Capsaspora owczarzaki* and, as we observed in *Tetrahymena* and *Tirmania*, the break was in the same position as the protostomian break.

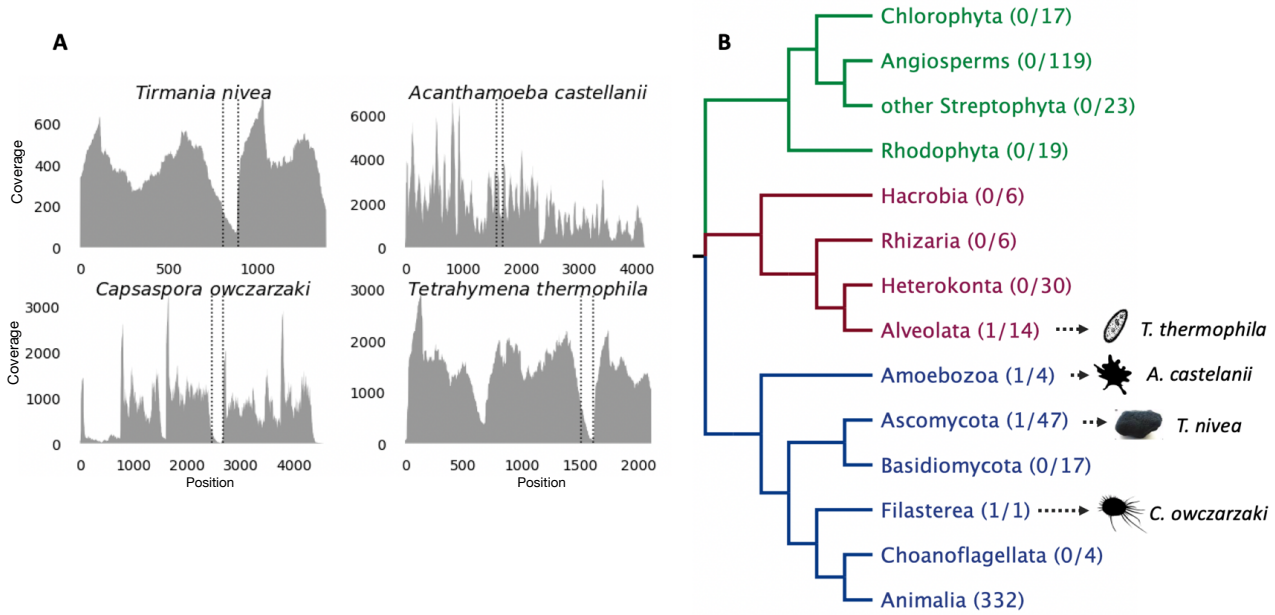

Figure 4. Distribution of the hidden break across non-metazoan eukaryotes. A) Four non-metazoan eukaryotes show evidence of a break. In three of these, (*T. nivea*, *C. owczarzaki*, *T. thermophila*) our analysis shows a hidden break that in a homologous position to the protostome one. In *A. castellanii*, the drop in read coverage is observed 3' of the two conserved markers we found surrounding the break in protostomes. This occurrence of hidden break has previously been characterised experimentally (see main text). B) The phylogenetic distribution of hidden break across Eukaryota suggests that the trait evolved convergently in the protostome ancestor and in a few taxa outside the Metazoa.

Conclusions

We present a computational approach that enabled us to perform a thorough and taxonomically broad examination of the 28S rRNA molecules of the animals. Our results strongly support Ishikawa's observation of a hidden break that evolved in the common ancestor of the protostomes. We have searched for the hidden break in members of almost all protostomian phyla and have demonstrated its existence in a homologous region of the molecule in almost all cases - the exceptions being the tardigrades and

264 dicyemids/rhombozoans. By expanding beyond the published observations of occasional absences of the
265 break in protostomes we show that in almost all cases of a species lacking the break, we find its existence
266 in sister taxa, implying that the absence is a derived state rather than a primitive absence. We interpret
267 the observed lack of the hidden break in tardigrades and dicyemids as due to loss of the character in
268 these lineages (at least in those few we have been able to sample).
269
270 We have also examined other instances of a break previously recorded in non-protostome animals (in two
271 groups of rodents and cyprinid fish) and have been able to show in each case that, while experiments
272 have suggested that they may share a mechanism with the introduction of protostomian break, they occur
273 in non-homologous regions of the molecule and this, together with their phylogenetic distribution, shows
274 that they are convergently evolved instances of a 28S break.
275
276 The large evolutionary distance between the non-metazoan eukaryotes in which we have confirmed or
277 discovered LSU hidden breaks as well as the lack of a break in most sampled eukaryotic taxa suggests
278 these too are cases of convergent evolution. This conclusion makes the fact that the three, presumably
279 convergently evolved, breaks known in *Tetrahymena*, *Tirmania* and *Capsaspora* all fall within the same
280 region of the molecule as the break in protostomes particularly notable and strongly suggestive both of a
281 common mechanism and, probably, a common functional reasons for the evolution of these breaks.
282 However, what the function of the break might be is still unknown.
283
284 Our findings have important ramifications for the use of RNA Integrity Numbers or RIN (Schroeder et al.
285 2006). The RIN, relying as it does on evidence for the integrity of the distinct 28S rRNA molecule, will tend
286 to produce artificially low values for the great majority of protostomes in cases where denaturation of the
287 RNA is possible. This source of error means that experimenters need to be careful in interpreting RIN when
288 evaluating RNA samples from the more than 95% of animal species that are protostomes (Chapman,
289 2009). The scale of this potential problem suggests that an alternative to the standard RIN that takes into
290 account the protostomian hidden break would be a valuable development for many researchers.
291
292 The emergence of a break between two regions of the large subunit ribosomal RNA is already known to
293 have occurred previously in eukaryotic evolution. The 5.8S rRNA, in addition to the 28S rRNA, forms part
294 of the large subunit of the ribosome (Henras et al., 2015). As discussed, the eukaryotic 5.8S and 28S
295 rRNAs are initially transcribed as a single molecule and are subsequently separated by the excision of the
296 intervening Internal Transcribed Spacer (ITS2). A separate 5.8S rRNA does not, however, exist in bacteria
297 where a sequence homologous to the eukaryotic 5.8S forms an uninterrupted part of the 23S rRNA (the
298 homolog of the eukaryote 28S) (Henras et al., 2015). The separation of 5.8S and 28S and the evolution of
299 the ITS2 seem to be close counterparts of the protostomian separation of 28S α and 28S β. We propose

300 that the rapidly-evolving, excised spacer sequence that lies between 28S α and 28S β. should be
301 considered as a third, protostome-specific, Internal Transcribed Spacer - the ITS3.

302
303 **Materials and Methods**
304
305 *General method*
306 Our method to identify the hidden break computationally rests on the assumption that, in species with the
307 hidden break, RNA-Seq datasets will contain very few reads, in particular paired-end reads, in the hidden
308 break region but will have normal levels of coverage on its flanking regions (28Sα and 28Sβ). We
309 established a computational pipeline to determine the existence of the hidden break in representative
310 species of as many phyla as possible in a semi-automated fashion.
311
312 *Identifying the hidden break region*
313 In a first step we tried to identify conserved sequences that could be used as flanking markers for the
314 hidden break region described in protostomes (Ishikawa, 1976), to help us determine potential homology
315 with breaks in other organisms. We aligned 28S sequences from *B. mori*, which has a characterised break
316 region, and other species, with and without a described hidden break, using mafft (Katoh et al. 2002). We
317 identified two conserved 20-mers (5'-AGUGGAGAAGGGUUCCAUGU-3' and 5'-
318 CGAAAGGGAATCGGGTTTAA-3') closely flanking the hidden break region. These two 20-mers can be
319 used as markers to identify the hidden break region characterised in protostomes when looking at other
320 species.
321
322 *Establishing a computational pipeline to search for the hidden break*
323 We used Python v3.7.0 to establish a semi-automated pipeline that proceeds from read mapping to results
324 visualization (as figures of read coverage). Firstly, our pipeline employs kallisto v0.44 (Bray et al., 2016) to
325 map paired RNA-Seq reads against the 28S sequence from the respective species. To ensure that kallisto
326 pseudoalignments are not influencing the analysis (Wu et al., 2018) we also tested the slower STAR (Dobin
327 et al., 2013) and bwa (Li & Durbin, 2009) aligners and found no notable differences in the results (data not
328 shown). Next, we use the mapped reads to calculate read coverage for each position in the 28S sequence
329 using BEDtools v2.27.1 (Quinlan & Hall, 2010). Finally, the pipeline uses Python's 'matplotlib' library to
330 produce plots of the depth of read mapping to along the 28S rRNA molecule. We then visually inspected
331 these plots. A species was considered to have the hidden break if there was an obvious drop in the read
332 coverage between the two conserved flanking 20-mers. The method was tested in data from *Bombyx*
333 *mori*, which has an experimentally identified and characterised hidden break.
334
335 *Data collection and large-scale application of the pipeline*
336 We retrieved all available entries (as of May 2019) from the SILVA database (Quast et al., 2013) in a single
337 FASTA file comprising 633,348 eukaryotic 28S sequences. We filtered these data to obtain a set of

338 sequences that could be analysed with our method by removing all duplicates, (i.e. entries coming from
339 the same species) and discarding all sequences shorter than 2,000bp. After these filtering steps we
340 retained 28S sequences from 12,460 species.
341
342 We next searched for RNA-Seq data for each of these species on the SRA database using eutils. We set
343 a threshold requirement of at least 1 Gigabase of RNA-Seq data and imposed an upper limit of 4
344 Gigabases for our analysis. We retrieved data within these size boundaries from 1,024 species.
345 A list of all retrieved species and the corresponding data sets and can be found under:
346 https://github.com/pnatsi/hiddenbreak/tree/master/data/suitable_species.tsv.
347
348 To achieve maximum representation of animal phyla in our results, we manually added 40 species
349 (highlighted in bold, Supp. Table 1) to those emerging from the semi-automated pipeline. The 28S
350 sequences for those species were retrieved from NCBI and RNAcentral databases, and the paired RNA-
351 Seq data from the SRA database. For these species only we also calculated the proportion of read pairs
352 spanning each residue of 28S using 'samtools' v1.9 (option 'view -F 12', Li et al. 2009). This metric was
353 highly correlated to the read coverage and was not applied to the rest of the species.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 764840. MT and PS were supported by European Research Council advanced award (ERC-2012-AdG 322790)

Competing Interests

The authors declare no competing interests.

361 **References**

362 Applebaum SW, Ebstein RP, Wyatt GR (1966) "Dissociation of ribosomal ribonucleic acid from silkmoth
363 pupae by heat and dimethylsulfoxide: Evidence for specific cleavage points" *Journal of Molecular*
364 *Biology*, **21**(1):29-41

365 Azpurua J, Ke Z, Chen IX, Zhang Q, Ermolenko DN, Zhang ZD, Gorbunova V, Seluanov A. (2013). "Naked
366 mole-rat has increased translational fidelity compared with the mouse, as well as a unique 28S
367 ribosomal RNA cleavage" *PNAS,* **110**(43):17350-5

368 Bieri P, Leibundgut M, Saurer M, Boehringer D, Ban N (2017) "The complete structure of the chloroplast
369 70S ribosome in complex with translation factor pY" *EMBO Journal* **36**(4):475-86

370 Bostock CJ, Prescott DM, Lauth M (1971) "Lability of 26 S ribosomal RNA in *Tetrahymena pyriformis*"
371 *Experimental Cell Research,* **66**(1):260–2

372 Bray NL, Pimentel H, Melsted P, Pachter L (2016) "Near-optimal probabilistic RNA-seq quantification"
373 *Nature Biotechnology*, **34**(5):525–7

374 Chapman AD (2009) "Numbers of living species in Australia and the world" *Toowomba, Australia: 2nd*
375 *edition* Australian Biodiversity Information Services

376 DeLeo DM, Pe JL, Bracken-Grissom HD (2018). "RNA profile diversity across Arthropoda: guidelines,
377 methodological artifacts, and expected outcomes" *Biology Methods and Protocols,* **3**(1):bpy012

378 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013)
379 "STAR: ultrafast universal RNA-seq aligner" *Bioinformatics* **29**(1):15-21

380 Fujiwara H, Ishikawa H (1986) "Molecular mechanism of introduction of the hidden break into the 28S
381 rRNA of insects: implication based on structural studies" *Nucleic Acids Research,* **14**(16):6393–6401

382 Haçariz O, Sayers G (2013) "*Fasciola hepatica* - Where is 28S ribosomal RNA?" *Experimental Parasitology*,
383 **135**(2):426–9

384 Henras AK, Plisson-Chastang C, O'Donohue MF, Chakrabarty A, Gleizes PE (2015) "An overview of pre-
385 ribosomal RNA processing in eukaryotes" *Wiley Interdisciplinary Reviews RNA* **6**(2):225-42

386 Ishikawa H (1973) "Comparative studies on the thermal stability of animal ribosomal RNA's"
387 Comparative *Biochemistriy and Physiology Part B: Comparative Biochemistry* **46**(2):217-27
388
389 Ishikawa H (1975a) "Comparative studies on the thermal stability of animal ribosomal RNA's—II.
390 Sea-anemones (Coelentarata)" Comparative *Biochemistriy and Physiology Part B: Comparative*
391 *Biochemistry* **50**(1):1-4
392
393 Ishikawa H (1975b) "Comparative studies on the thermal stability of animal ribosomal RNA's—III.
394 Sponge (Porifera) and the other species (*Tetrahymena* and *Lachnus*)" Comparative *Biochemistriy*
395 *and Physiology Part B: Comparative Biochemistry* **51**(1):81-5

396 Ishikawa H (1977a) "Evolution of ribosomal RNA" *Comparative Biochemistry and Physiology,* **58**(1):1-7

Ishikawa H (1977b) “Comparative studies on the thermal stability of animal ribosomal RNA's—IV. Thermal stability and molecular integrity of ribosomal RNA's from several protostomes (rotifers, round-worms, liver-flukes and brine-shrimps)” *Comparative Biochemistriy and Physiology Part B: Comparative Biochemistry* **56**(3):229-34

Karlstedt K, Paatero G, Mäkelä JH, Wikgren BJ (1992) “A hidden break in the 28.0S rRNA from *Diphyllobothrium dendriticum*” *Journal of Helminthology*, **66**(3):193–7

Katoh K, Misawa K, Kuma K, Miyata T (2002) “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform” *Nucleic Acids Research* **30**(14):3059-66

Leipoldt M, Engel W (1983) “Hidden breaks in ribosomal RNA of phylogenetically tetraploid fish and their possible role in the diploidization process” *Biochemical Genetics*, **21**(7–8):819–841

Li H, Durbin R (2009) “Fast and accurate short read alignment with Burrows-Wheeler transform” *Bioinformatics* **25**(14):1754-60

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. (2009) “The Sequence Alignment / Map format and SAMtools” *Bioinformatics,* **25**(16):2078–9

Macharia RW, Ombura FL, Aroko EO (2015) “Insects’ RNA profiling reveals absence of “hidden break” in 28S ribosomal RNA molecule of onion thrips, *Thrips tabaci*” *Journal of Nucleic Acids*, **2015**:965294

McCarthy SD, Dugon MM, Power AM (2015). “‘Degraded’ RNA profiles in Arthropoda and beyond” *PeerJ*, **3**:e1436

Melen GJ, Pesce CG, Rossi MS, Kornblihtt AR (1999) “Novel processing in a mammalian nuclear 28S pre-rRNA: Tissue-specific elimination of an ‘intron’ bearing a hidden break site” *EMBO Journal*, **18**(11):3107–18.

Mertz PM, Bobek LA, Rekosh DM, LoVerde PT (1991) “*Schistosoma haematobium* and *S. japonicum*: Analysis of the ribosomal RNA genes and determination of the “gap” boundaries and sequences” *Experimental Parasitology*, **73**(2):137–49

Navarro-Ródenas A, Carra A, Morte A (2018) “Identification of an alternative rRNA post-transcriptional maturation of 26S rRNA in the kingdom Fungi” *Frontiers in Microbiology*, **9**:994

Ogino K, Eda-Fujiwara H, Fujiwara H, Ishikawa H (1990) “What causes the aphid 28S rRNA to lack the hidden break?” *Journal of Molecular Evolution* **30**(6):509–13

Papillion D, Perez Y, Caubit X, Le Parco Y (2004) “Identification of Chaetognaths as Protostomes is supported by the analysis of their mitochondrial genome” *Molecular Biology and Evolution* **21**(11):2122-9

Philippe H, Poustka AJ, Chiodin M, Hoff KJ, Dessimoz C, Tomiczek B, Schiffer PH, Müller S, Domman D, Horn M, Kuhl H, Timmermann B, Satoh N, Hikosaka-Katayama T, Nakano H, Rowe ML, Elphick MR, Thomas-Chollier M, Hankeln T, Mertes F, Wallberg A, Rast JP, Copley RR, Martinez P, Telford MJ

434 (2019) “Mitigating anticipated effects of systematic errors supports sister-group relationship between
435 Xenacoelomorpha and Ambulacraria” *Current Biology* **29**(11):1818-26

436 Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Glo FO, Yarza P (2013) “The SILVA ribosomal RNA
437 gene database project: improved data processing and web-based tools” *Nucleic Acids Research*
438 **41**:D590-6

439 Quinlan AR, Hall IM (2010) “BEDTools: a flexible suite of utilities for comparing genomic features”
440 *Bioinformatics*, **26**(6):841–2

441 Schiffer PH, Robertson HE, Telford MJ (2018) “Orthonectids are highly degenerate annelid worms” *Current*
442 *Biology*, **28**(12):1970-4

443 Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow
444 M, Ragg T (2006) “The RIN: an RNA integrity number for assigning integrity values to RNA
445 measurements” **7**:3

446 Stevens A, &Pachler P (1972) “Discontinuity of 26 s rRNA in *Acanthamoeba castellani*” *Journal of Molecular*
447 *Biology*, **66**(2):225–37

448 Wu DC, Yao J, Ho KS, Lambowitz AM, Wilke CO (2018) “Limitations of alignment-free tools in total RNA-
449 seq quantifications” *BMC Genomics,* **19**(1):510