www.nature.com/scientificreports

SCIENTIFIC REPORTS
natureresearch

OPEN

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

Paschalis Natsidis, Philipp H. Schiffer, Irepan Salvador-Martínez & Maximilian J. Telford*

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 (arthropods, molluscs, annelids etc.), but a break 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 a homologous break is present across protostomes although it has been lost in a small number of taxa. We show that rare breaks in vertebrate 28S rRNAs are not homologous to the protostome break. A break is found in just 4 out of 331 species of non-animal eukaryotes studied and, in three of these, the break is 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.

Ribosomes are made up of up to 80 ribosomal proteins and three (in prokaryotes) or four (in eukaryote cytoplasmic ribosomes) structural ribosomal RNAs named according to their sizes: in eukaryotes the cytoplasmic ribosomal RNAs (rRNAs) are the 5S (~120 nucleotides), the 5.8S (~150 nucleotides), the 18S (~1800 nucleotides) and the 28S (~4000 to 5000 nucleotides). The 5.8S, 18S and 28S rRNAs are initially transcribed as a single RNA operon (the 5S is at a separate locus in eukaryotes). The 18S and 5.8S are separated by the Internal Transcribed Spacer 1 (ITS1) and 5.8S and 28S are separated by ITS2. The initial transcript is cleaved into three functional RNAs by removing ITS1 and ITS2 (Fig. 1A).

In some species this picture is complicated by observations of a 'hidden break' in the 28S rRNA. In organisms with a hidden break, the 28S rRNA molecule itself is cleaved into two approximately equal sized molecules of ~2000 nucleotides each (Fig. 1B)[1]. These two RNAs (the 5′ 28Sα and the 3′ 28Sβ) are nevertheless intimately linked by inter-molecular hydrogen bonding within the large subunit of the ribosome, just as the 28S and 5.8S rRNAs, as well as different regions of the intact 28S rRNA are in species without a break.

The hidden break, first described in pupae of the silkmoth *Hyalophora cecropia*[2], manifests itself experimentally when total RNA is extracted and separated according to size using gel electrophoresis. If the RNA is denatured by heating in order to separate hydrogen bonded stretches of RNAs before electrophoresis, the hydrogen bonded 28Sα and 28Sβ subunits separate. These two molecules, being approximately the same size, migrate on a gel at the same rate as each other and, coincidentally, at the same rate as the almost identically sized 18S rRNA molecule. The effect is that, rather than observing two distinct ribosomal RNA bands on the gel of ~2000 and ~4000 nucleotides, a single (and more intense) band composed of three, different molecules, each approximately 2 kb long (18S, 28Sα and 28Sβ) is seen[3]. This also applies to electropherograms where we normally expect two distinct peaks for 18S and 28S molecules (Fig. 1A); in species that possess the hidden break only one peak is observed, corresponding to the equally sized 18S, 28Sα and 28Sβ (Fig. 1B).

The processing of the polycistronic ribosomal RNAs into individual rRNA molecules has been studied in depth, mostly in yeast and to a lesser extent in plants and vertebrates[4]. It has been shown to be a complex process

Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, Gower Street, London, WC1E 6BT, UK. *email: m.telford@ucl.ac.uk

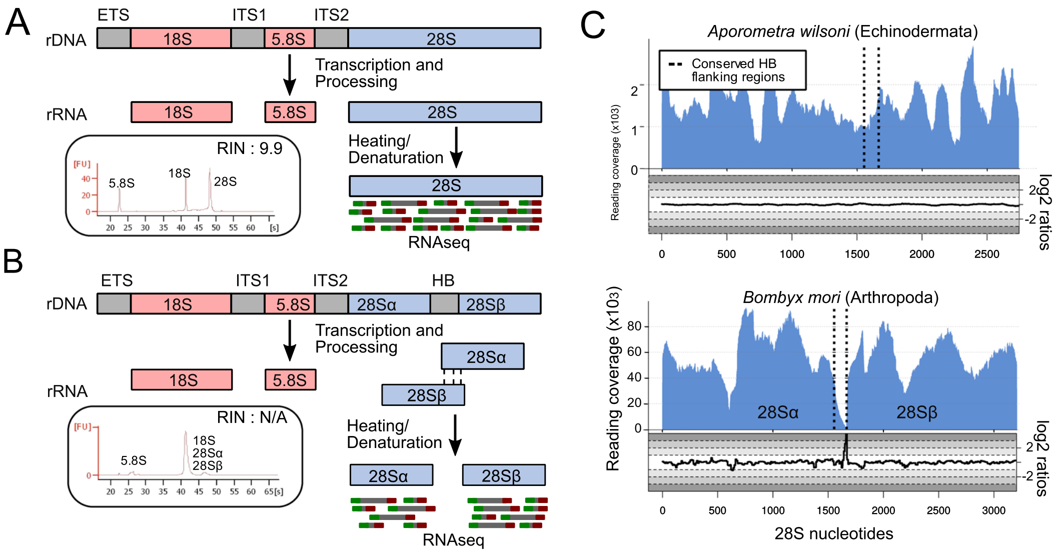

Figure 1. The hidden break and how to diagnose it. (**A**) After post-transcriptional processing, the eukaryotic pre-rRNA molecule is cleaved to produce the 18S, 5.8S and 28S subunits. The RIN works normally in species that follow this rule. (**B**) In some species, the 28S subunit gets cleaved into 28Sα and 28Sβ, a phenomenon known as the hidden break. This phenomenon can be detected via electrophorograms where, in species with a break, only one peak, instead of two, is observed. (**C**) Our method allows the computational discovery of hidden breaks, by mapping RNA-Seq reads onto the 28S rRNA sequence and measuring (i) the read coverage and (ii) the $\log_2$ of the ratio of forward/reverse reads mapped in each position of the 28S. In species with a break, e.g. *Bombyx mori*, a drop in coverage is observed in the middle of the break region. Additionally, the $\log_2$ of the forward/reverse reads ratio drops before the hidden break site and increases immediately after it. These features are not observed in *Aporometra wilsoni*, an animal that does not possess the break. The vertical lines represent the conserved regions flanking the hidden break site.

involving many proteins. The additional step of dividing the single 28S rRNA into two subunits is peculiar to protostomes and (as we show) a few other taxa, which are unlikely to be homologous instances. The processing of the hidden break does not occur in model organisms typically used to study ribosome processing and the mechanism is far from clear. There are three observations of particular interest, however. First is the observation that the excised ribonucleotides (which have been studied in a handful of species, see[5] and[6]) are located at the end of a stem structure and centred on the tetranucleotide UAAU, suggesting these features might be recognised by whatever machinery is involved in the excision. Second, it has been observed in the fly *Musca carnaria* that the precise cut can be achieved *in vitro* simply by the addition of RNase to intact ribosomes suggesting that the excision may occur following ribosome assembly and not necessarily requiring a complex set of proteins to achieve it[7]. In yeast, the D7a loop where the hidden break is located is found on the surface of the ribosome, perhaps making it particularly accessible to enzymatic cleavage[8]. Finally, it has been shown that if the 28S rRNA from the fly *Sciara coprophila*, which has a hidden break, is heterologously expressed in *Xenopus laevis* oocytes, the rRNA is processed into the typical two fragments[9]. This suggests that the processing of the hidden break depends on features inherent in the 28S rRNA itself, which can be recognised by cellular machinery present even in species that do not normally divide their 28S rRNA. What these features are is presently not yet clear.

This behaviour of rRNA molecules in species with a hidden break can have interesting and potentially detrimental practical consequences. The integrity of 18S and 28S rRNA molecules, as evidenced by the presence of ~2 kb and ~4 kb bands/peaks following electrophoresis, is widely used as an indication that a total RNA sample is not degraded. This is an important quality control step for experiments that will use the total RNA sample for subsequent experiments such as transcriptome sequencing, Northern blotting or rtPCR.

The requirement to observe two rRNA peaks in an intact RNA sample has been formalised as an important part of the RNA Integrity Number (RIN[10]). The RIN consists of a number, from 10 (intact) to 1 (totally degraded), that describes the quality of an RNA sample. The calculation of the RIN is performed by an automated algorithm that selects features from an electropherogram (e.g. 28S peak height) and uses these features to calculate the RIN. The RIN algorithm is based on a Bayesian learning approach, that used hundreds of RNA samples for training each with an expert-assigned quality rank from 1 to 10. After several models were trained using artificial neural networks, the best was chosen for prediction of previously unseen test data[10]. Importantly, the samples used for the RIN algorithm training were obtained from species without a hidden break, namely human, rat, and mouse[10]. For RNA samples extracted from species possessing the hidden break, if conditions allow the separation of the two 28S subunits, a single rRNA band/peak will be observed. The absence of two distinct rRNA peaks will give the impression of a degraded sample with a low RIN, even when the RNA is not degraded.

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[11]). 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[12]) but was also absent from two species of nematodes[13] and from a species of chaetognath[14] (then thought to be deuterostome relatives but now known to be protostomes[15]). The hidden break was recorded as ambiguous in non-bilaterian Cnidaria and absent in Porifera[14,16]. 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*[8] as well as in plant chloroplasts[17]. The precise location of the hidden break is not revealed by these studies using electrophoresis.

Despite Ishikawa's work and other sporadic publications describing the presence of the hidden break in different taxa[1,3,18–21], 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.

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. An example of this analysis for a species without (*Aporometra wilsoni)* and a species with (*Bombyx mori*) the hidden break are shown in Fig. 1C. Similarly, in the region immediately before the break we expect a large proportion of the mapped reads to be reverse reads (red in Fig. 1B), while the majority of the reads that mapped immediately after the break will be forward reads (green in Fig. 1B). The ratios of forward to reverse reads mapped in each position can also be seen in Fig. 1C. Lastly, 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 selected 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[3]. 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[3]. 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 three characteristics. First, there will be an obvious absence of reads mapped to the genomic rRNA location precisely at the position of the split; second, we expect that reads mapped right before the break will mostly be reverse reads, in contrast with the region after the break where most mapped reads will be forward reads; third, if paired-end sequencing data are available, there should be no read pairs spanning from one side of the split to the other as these are separate molecules.

The position of the split has been mapped experimentally in very few species, including the silkworm *Bombyx mori*[5]. Our method makes the accurate mapping of the position of the split a simple process which is of importance: if we are to be confident that this character is homologous between species it is necessary to show that the break occurs in a homologous region of the molecule. The procedure for mapping reads and counting forward and reverse reads, as well as spanning read pairs are described in Materials and Methods. Figure 1C shows an example of the pattern of read depth and forward/reverse reads ratios that we observe in an organism without a split (*Aporometra wilsoni*) and in an organism in which the split is known to exist (*Bombyx mori*)[5]. We also indicate the position of the break, which is known experimentally in *Bombyx mori*. This experimentally mapped break is surrounded by highly conserved stretches of nucleotides allowing us to identify the homologous region of the gene across eukaryotes.

We used three different metrics (read coverage, forward/reverse reads ratio, number of spanning read pairs) to investigate the existence of a hidden break in 28S sequences. We noticed that the read coverage is more easily interpreted than the other two, in the cases where all three metrics were all applied. Our results showed that the forward/reverse reads ratio metric contained more noise than the read coverage, and that the number of spanning read pairs across the break region is highly correlated with the read coverage.

Expanding to most protostomian phyla. The hidden break has been characterised in a number of animal species and has largely been considered to be specific to the protostomes. We have used our method to expand the search for a potential hidden break to members of all animal phyla for which we have found suitable RNA-Seq data. In total, we have examined 347 metazoan species including members of all but 4 animal phyla (Onychophora, Loricifera, Micrognathozoa and Gnathostomulida). The majority of the species analysed came from the Arthropoda (169) and Mollusca (31). We have also analysed 31 chordates, 28 echinoderms, as well as 2 hemichordates and 7 xenacoelomorphs. We have also examined data from 36 non-bilaterian species (29 cnidarians species, 3 ctenophores, 1 placozoan and 3 poriferans).

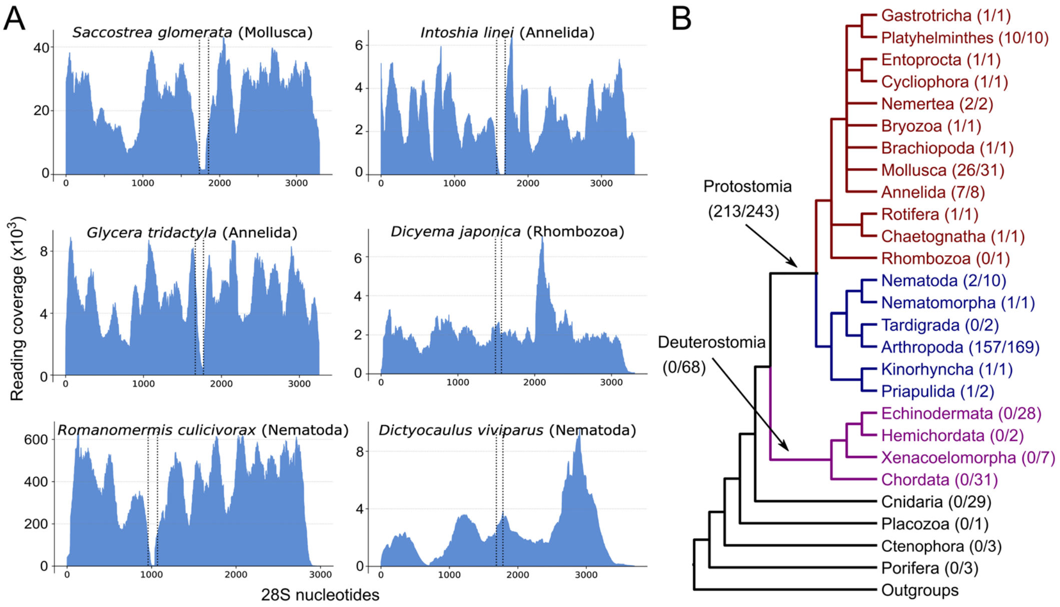

Figure 2. The hidden break is found in most protostome animals. (**A**) Six examples of the application our method to different metazoan species. Evidence for a hidden break was found in most molluscs and annelids, including the orthonectid *Intoshia linei*. The fast-evolving dicyemid *Dicyema japonica* is one of the few examples where the break has been secondarily lost. The Nematoda also appear to be a special case, with the break present in the enoplean species, e.g. *Romanomermis culicivorax*, but not in the Chromadorea, e.g. *Dictyocaulus viviparus* or the model *Caenorhabditis elegans* (not shown). (**B**) The phylogenetic distribution of the hidden break across metazoan phyla shows that it is a protostome synapomorphy, with occasional secondary losses. The numbers in parentheses show the proportion of the species having the hidden break among the total of analysed species for this group.

For the most part, our results confirm those of Ishikawa; in almost all protostomes providing unambiguous results (17 out of the 19 protostomian phyla we were able to test) we observe the existence of the hidden break. Representative examples are shown in Fig. 2A. In each observation of the break we find it to be bounded by the same conserved sequence regions of the 28S molecule. Our ability to map the break means we have been able to show that, across the protostomes, it is present in the same position in the molecule and we conclude that this is a homologous character throughout this major group of animals. This is the first time that this finding is supported by such a large taxon sampling and across such a breadth of taxa.

In contrast to previous findings[14], we have found evidence for the hidden break in a chaetognath; at the time of Ishikawa's work members of this phylum were thought to be relatives of the deuterostomes but are now known to be protostomes[15]. 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[22]. 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 that the hidden break is absent in a number of other protostomian clades (Fig. 2B). We confirm, for example, Ishikawa's work showing that there is no break in certain rhabditid nematodes[13], or in the aphid *Acyrthosiphon pisum*[23]. 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 culicivorax* and *Trichinella spiralis* 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[23], we have shown the break is present in 157 other species of arthropod. The break is also known to be absent in *Thrips tabaci*[24] 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.

Non protostomian metazoans. We next applied our method to the two clades of deuterostomes: the Xenambulacraria[25] (29 echinoderms, 2 hemichordates and 7 xenacoelomorphs) and the Chordata (2 urochordates and 29 vertebrates). We find that the break found in protostomes does not exist in any of the sampled species. There are,

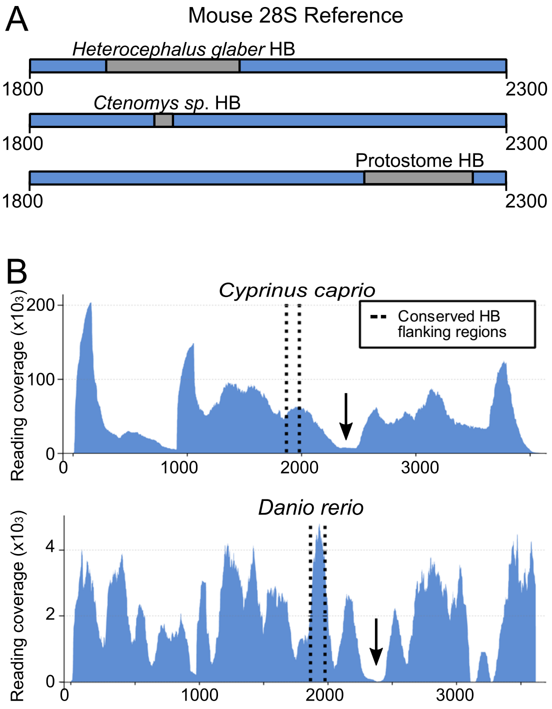

Figure 3. Vertebrate hidden breaks are not homologous to protostome breaks. (**A**) Using the mouse 28S rRNA sequence as a reference, we show that the break regions recorded in rRNA extracted from naked mole-rat and Tuco-tuco testes are not homologous to the protostome hidden break. In both species the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

nevertheless, two special cases previously described in vertebrates: hystricomorph rodents, where species of the genus *Ctenomys* (the south american Tuctuc) and the naked mole-rat (*Heterocephalus glaber*) are found to possess the hidden break[26,27]. We show, however, that the break found in these two rodents is not homologous to the protostome hidden break. We were unable to retrieve 28S sequences for these two species, however, by using the mouse 28S rRNA sequence as reference, we were able to locate the approximate region of these two breaks and compare these to the two conserved markers that enclose the protostome break. The two recorded rodent breaks do not fall between the protostome break markers (Fig. 3A). In more detail, the *Ctenomys* hidden break corresponds to nucleotides 1930–1950 in mouse 28S, while the naked mole-rat hidden break falls between nucleotides 1880–2020 of the mouse 28S. The two conserved satellites of the protostome hidden break lie in the positions 2149 and 2264 of the same mouse 28S sequence. Thus, the rodent hidden break site is located nearly 300 bp upstream of the conserved protostome site.

The hidden break has also been reported in cyprinid fish[28] and we have analysed two members of this clade. Our results show that the break in cyprinids is also in a different location from that of protostomes and does not match the break described in rodents (Fig. 3B). The size ratio of the two 28S fragments we have predicted (~2300:1300 = 1.77:1) corresponds to the ratio of the two 28S fragments that are described for cyprinid fish[33], (a ratio of ~1.75:1). The sites of the protostome, rodent and cyprinid hidden breaks were mapped onto the homologlous region of the secondary structure of the human 28S ribosomal RNA to visualise their distinct relative positions (Supplementary Fig. 1)

We next looked at non-bilaterian animals to determine whether they possess the hidden break. Ishikawa's work had been inconclusive in this regard - at least in the Cnidaria[11] - leaving open the possibility that the break found across Protostomia is a primitive characteristic of animals that has been lost in the deuterostomes. In a total of 36 non-bilaterian species (3 sponges, 29 cnidarians, 3 ctenophores, and the placozoan *Trichoplax adhaerens*) we find no evidence for the existence of a hidden break (Fig. 2B). The distribution of the break in the animal kingdom points to the hidden break having appeared in the lineage leading to the common ancestor of the protostomes.

Convergent evolution of the hidden break outside Metazoa?

Alongside the animals, there have been reports of a hidden 28S rRNA break in a small number of distantly related species of eukaryotes, the ciliate *Tetrahymena pyriformis*[29], the amoebozoan *Acanthamoeba castelanii*[30] and the desert truffle *Tirmania nivea*[31]. We have investigated these taxa using our pipeline (Fig. 4A) to define the position of their breaks. We have greatly extended our analyses of non-animal eukaryotes to over 300 non-metazoans, including 119 plants, 64 fungi and 148 other taxa (Fig. 4B). The hidden break previously described in plant chloroplast 23 S rRNAs (as opposed to nuclear 28S rRNA) is not located in the same region of the molecule as the protostomian break[17].

In *Acanthamoeba*, gel electrophoresis of heat denatured LSU rRNA showed that it separates into fragments corresponding to molecular weights of 0.88×10^6 and 0.60×10^6 daltons[30]. Our computational results provide support for this, as we find a break in the molecule that would divide it into fragments of approximately 2320 and

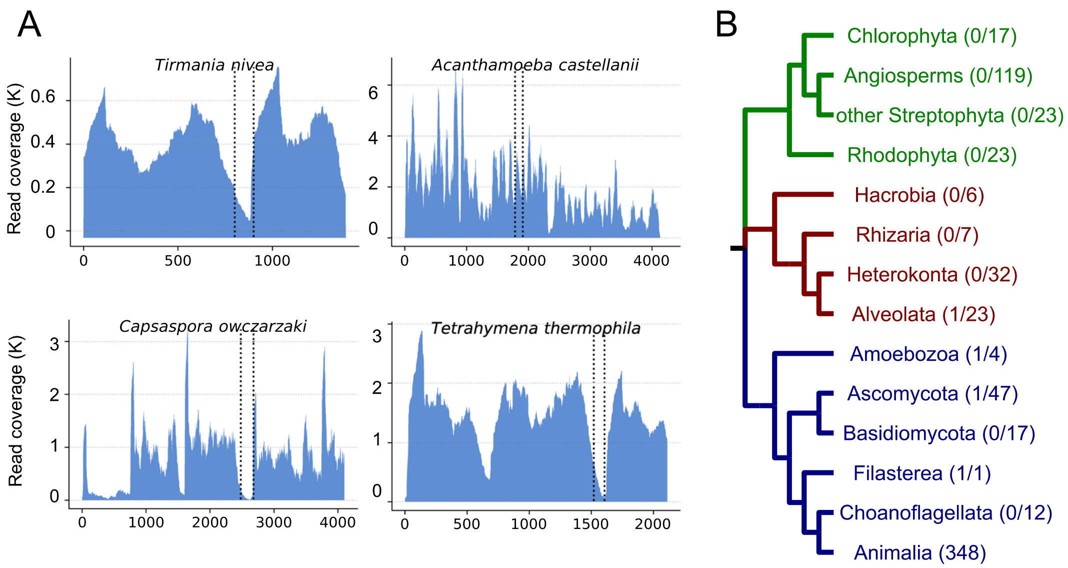

Figure 4. Distribution of the hidden break across non-metazoan eukaryotes. (**A**) Four non-metazoan eukaryotes show evidence of a break. In three of these, (*Tirmania nivea*, *Capsaspora owczarzaki*, *Tetrahymena thermophila*) our analysis shows a hidden break that is in a homologous position to the protostome one. In *Acanthamoeba castellanii*, the drop in read coverage is observed 3' of the two conserved markers we found surrounding the break in protostomes. This occurrence of a 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. The numbers in parentheses show the proportion of the species having the hidden break among the total of analysed species for this group.

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 confirm the finding of a break in both *Tetrahymena*[29] and the desert truffle *Tirmania*[31] 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.

Variability of the hidden break region among bilaterians. In order to identify potential signals of the hidden break in the sequence of the 28S rRNA, we took a closer look at the region of the molecule where the hidden break occurs. We analysed 5,562 bilaterian 28S sequences from species of phyla possessing the hidden break and 626 sequences from species of phyla where the break was not reported (Table 1). In every sequence we located the two conserved protostome markers and considered the nucleotides that separated them. The results showed first, that the average distance between the markers in species with the break was approximately 120 nucleotides, while in species without the break the average distance was found to be 97 nucleotides (Fig. 5); second, in species without the break the distribution of the marker distances was much narrower than in the species with the break (Fig. 5). These results suggest that the region that is spliced out during the hidden break process is more prone to length variation, while in species without the hidden break the same region has a more conserved length. Even though the two distributions were found to differ significantly (Welch's *t*-test, $p = 2.713\text{e}^{-06}$), it is likely that the phylogenetic structure of the dataset influences this observation.

We also calculated the AU content of the region near the hidden break site in species with and without the hidden break (Table 1). The results show that in species with the break the AU% in the spliced region is higher than it is in the complete 28S molecule, while in species without the break the AU% of the homologous region is the same, if not lower than in the whole 28S sequence.

This agrees with previous suggestion of an AU-rich (and thus less stable) region of nucleotides being responsible for the dissociation of the 28S into two subunits[3]. We also counted the occurrence and the frequencies of the tetranucleotide UAAU, that has been proposed to be part of the cleavage mechanism. This showed that, with the exception of Gastrotricha, UAAU is present in less than half of the examined species with a break, suggesting that the hidden break is not dependent on the presence of this motif.

Conclusions

We present a computational approach that enabled us to perform a thorough and taxonomically broad examination of the 28S rRNA molecules of the eukaryotes for the presence of the hidden break. 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

	No. of seqs	Mean	Median	mean (AT) - whole 28S	mean(AT) - break region	% of species with UAAU
Protostomes with hidden break						
Arthropoda	4,175	128.92	106	45.95%	51%	30.5%
Mollusca	555	107.43	105	42.79%	55.27%	7.2%
Platyhelminthes	147	117.30	116	49.72%	52.42%	23.12%
Nematoda (Enoplea)	18	141.22	118	50.50%	60.02%	22.2%
Nemertea	94	105.52	103	44.24%	53.86%	12.7%
Gastrotricha	74	101.32	102	48.23%	60.74%	75%
Halicryptus	1	95	95	53.66%	53.09%	N/A
Protostomes without hidden break						
Nematoda (Chromadorea)	275	92.30	92	50.75%	53.56%	51.6%
Tardigrada	7	139	132	45.71%	45.64%	28.5%
Priapulus	1	89	89	55.10%	54.16%	N/A
Dicyema	1	78	78	43.35%	40.74%	N/A
Deuterostomes						
Chordata	136	92.72	93	38.82%	33.98%	0%
Echinodermata	71	99.38	100	41.67%	29.94%	0%
Xenacoelomorpha	115	86.95	86	51.72%	52.29%	11.3%

Table 1. Mean and median distances of the two conserved markers of the protostomian break, AU contents of the whole 28S sequence and the region near the hidden break and proportion of sequences with the UAAU 4-mer per phylum. In protostome phyla that possess the break, the distances between the two markers are in general larger than in the protostomes that have lost the break (with the exception of 2 species of tardigrades) and deuterostomes. The region that containing the break site has higher AU content compared to the average of the whole 28S molecule in break-possessing phyla, which is not the case in the species without the break. The UAAU tetranucleotide is not found in most protostome phyla with a hidden break and was not found in deuterostome phyla (except Xenacoelomorpha).

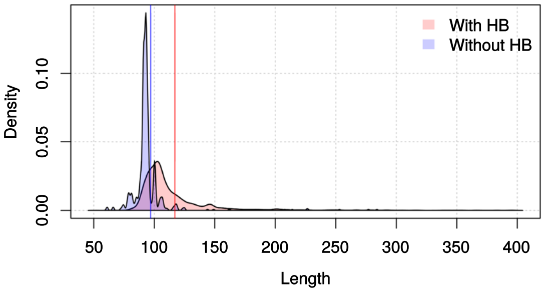

Figure 5. Lengths of the region that contains the hidden break in protostomes, as defined by the distance between the two conserved markers that surround it. In species with a break, the average length of the region is larger (approximately 118 vs 97 nucleotides) and the length distribution is broader than in the species that do not have it, possibly because the deleted section is under weak selection. The two vertical lines depict the means of the two distributions. Distances larger than 300 bp are not shown (5 not shown for no break, 116 not shown with break). Supplementary Fig. 1. Homologous positions of different hidden breaks indicated on secondary structure of human 28S rRNA. The orange arrow points to the site of the protostome hidden break, the red arrow indicates the rodent hidden break, and the pink arrow shows the cyprinid fish hidden break. The image was created using RiboVision server[43] 597 (http://apollo.chemistry.gatech.edu/RiboVision/).

existence in a homologous region of the molecule in almost all cases - the exceptions being the poorly sampled tardigrades and dicyemids. By expanding beyond the published observations of occasional absences of the break in protostomes we show that in almost all cases of a species lacking the break, we find its existence in sister taxa, implying that the absence is a derived state rather than a primitive absence. We interpret the observed lack of the hidden break in tardigrades and dicyemids as due to loss of the character in these lineages (at least in those few we have been able to sample).

We have also examined other instances of a break previously recorded in non-protostome animals (in two groups of rodents and cyprinid fish) and have been able to show in each case that, while they may share a mechanism with the introduction of protostomian break, they occur in non-homologous regions of the molecule and this, together with their phylogenetic distribution, shows that they are convergently evolved instances of a 28S break.

The large evolutionary distance between the non-metazoan eukaryotes in which we have confirmed or discovered 28S rRNA hidden breaks as well as the lack of a break in most sampled eukaryotic taxa suggests these, too are rare cases of convergent evolution. This conclusion makes the fact that the three, presumably convergently evolved breaks known in *Tetrahymena*, *Tirmania* and *Capsaspora* all fall within the same region of the molecule as the break in protostomes particularly notable and strongly suggestive both of a common mechanism and, probably, a common functional reasons for the evolution of these breaks. However, what the function of the break might be is still unknown.

We observe drops in read coverage at other locations of the 28S rRNA in many species and do not draw conclusions regarding whether these may be additional breaks in the 28S molecule or whether they represent other factors affecting the sequencing of these regions. For the protostomes the sheer reproducibility of the conserved drop in the same location across thousands of species meant that we could draw solid conclusions regarding its existence across clades. For protostomes with no hidden break the consistent coverage across the break region was also reliable (could not occur if there was a break) - and no additional evidence was needed. For species outside of the protostomes we required a good level of coverage across the gene and putative instances of a break were then confirmed with other data, e.g. existing electropherogram data or the ratio of forward and reverse reads as we have described. As we focussed on the cases described above, drops in coverage elsewhere in different species do not impact our conclusions.

Our findings have important ramifications for the use of RNA Integrity Numbers or RIN[10]. The RIN, relying as it does on evidence for the integrity of the distinct 28S rRNA molecule, will tend to produce artificially low values for the great majority of protostomes in cases where denaturation of the RNA is possible. This source of error means that experimenters need to be careful in interpreting RIN when evaluating RNA samples from the more than 95% of animal species that are protostomes[32]. The scale of this potential problem suggests that an alternative to the standard RIN that takes into account the protostomian hidden break would be a valuable development for many researchers.

The emergence of a break between two regions of the large subunit rRNA is already known to have occurred previously in eukaryotic evolution. The 5.8S rRNA, in addition to the 28S rRNA, forms part of the large subunit of the ribosome[4]. As discussed, the eukaryotic 5.8S, 18S and 28S rRNAs are initially transcribed as a single molecule and are subsequently separated by the excision of the intervening Internal Transcribed Spacer (ITS2). A separate 5.8S rRNA does not, however, exist in bacteria where a sequence homologous to the eukaryotic 5.8S forms an uninterrupted part of the 23 S rRNA (the homolog of the eukaryote 28S)[4]. The separation of 5.8S and 28S and the evolution of the ITS2 seem to be close counterparts of the protostomian separation of 28Sα and 28Sβ. We propose that the rapidly evolving, excised spacer sequence that lies between 28Sα and 28Sβ should be considered as a third, protostome-specific, Internal Transcribed Spacer - the ITS3.

Methods

General method.

Our method to identify the hidden break rests on the assumption that, in species with the hidden break, RNA-Seq datasets will contain very few reads, in particular paired-end reads, in the hidden break region but will have normal levels of coverage on its flanking regions (28Sα and 28Sβ). We established a computational pipeline to determine the existence of the hidden break in representative species of as many phyla as possible in a semi-automated fashion.

Identifying the hidden break region.

In a first step we identified conserved sequences that could be used as flanking markers for the hidden break region described in protostomes[13], to help us determine potential homology of breaks in other organisms with the protostome break. We aligned 28S rRNA sequences from *Bombyx mori*, which has a characterised break region[5] and from other species, with and without a described hidden break, using mafft[33]. We identified two highly conserved 20-mers (5′-AGUGGAGAAGGGUUCCAUGU-3′ and 5′-CGAAAGGGAATCGGGTTTAA-3′) flanking the hidden break region. These two 20-mers can be used as markers to identify the hidden break region characterised in protostomes when looking at other species.

Establishing a pipeline to search for the hidden break.

We used Python v3.7.0 to establish a semi-automated pipeline that proceeds from read mapping to results visualization (as graphs of read coverage). Firstly, our pipeline employs kallisto v0.44[34] to map paired-end RNA-Seq reads against the respective 28S rRNA sequence from the species of interest. To ensure that kallisto pseudoalignments are not influencing the analysis[35], we also tested the slower STAR[36] and bwa[37] aligners and found no notable differences in the results (data not shown). Next, we used these mapped reads to calculate read coverage for each position in the 28S sequence using BEDtools v2.27.1[38]. Finally, the pipeline uses Python's 'matplotlib' library to produce plots of the depth of read mapping along the 28S rRNA molecule. We then inspected these plots. A species was considered to have the hidden break if there was an obvious drop in the read coverage between the two conserved flanking 20-mers. The method was tested using data from *Bombyx mori*, which has an experimentally identified and characterised hidden break.

Data collection and large-scale application of the pipeline.

We retrieved all available entries (as of May 2019) from the SILVA database[39] in a single FASTA file comprising 633,348 eukaryotic 28S sequences. We filtered these data to obtain a set of sequences that could be analysed with our method by removing all duplicates (i.e. entries coming from the same species) and discarding all sequences shorter than 2,000 bp. After these filtering steps we retained 28S RNA sequences from 12,460 species.

We next searched for RNA-Seq data for each of these species on the SRA database using eutils. We set a threshold requirement of at least 1 Gigabase of RNA-Seq data and imposed an upper limit of 4 Gigabases for our analysis. We retrieved data within these size boundaries from 1,024 species. A list of the species that produced an interpretable plot can be found in Supplementary Table 1, and the corresponding coverage plots can be seen in Supplementary File 1.

To achieve maximum representation of animal phyla in our results, we manually added 80 species (highlighted in bold, Supplementary Table 1) to those emerging from the semi-automated pipeline. The 28S rRNA sequences for these species were retrieved from NCBI and RNAcentral databases[40] and the paired-end RNA-Seq data from the SRA database[41]. For 40 species only we also calculated the proportion of read pairs spanning each residue of 28S using SAMtools v1.9 (option 'view -F 12'[42]). This metric was highly correlated to the read coverage and was not applied to the rest of the species. We also calculated the number of forward and reverse reads mapped in the interval between 300 nucleotides before the break and 300 nucleotides after using SAMtools (options 'view -f 67' for forward and 'view -f 131' for reverse reads). We then calculated the ratios of forward/reverse and reverse/forward reads for each position and visualised the result, using logarithms (base 2) of ratios to make the fluctuations symmetric. This analysis was run for 17 species (10 with and 7 without the hidden break) and the results can be seen in Supplementary File 2.

Variability of the hidden break region among bilaterians. We downloaded all 9,713 bilaterian 28S sequences available in SILVA database[39] as of March 2019. We removed duplicates to be left with 28S sequences from 6,188 bilaterians. We located the two conserved markers of the protostomian break in each sequence and counted the number of nucleotides that separate them. The detailed results of this analysis can be found in Supplementary File 3. We also measured the AU content of these intervening sequences and compared this to the overall 28S rRNA gene AU content. We counted instance of the UAAU tetranucleotide (and other AU rich tetranucleotides) within the intervening sequences as these are potential signals for the hidden break.

Data availability

A list of all species from which data were retrieved and the corresponding data sets and be found under: https://github.com/pnatsi/hiddenbreak/tree/master/data/suitable_species.tsv. The code that performs the analysis and plots the result can be found under: https://github.com/pnatsi/HBinspector.

Received: 14 October 2019; Accepted: 27 November 2019;
Published online: 20 December 2019

References

1. Winnebeck, E. C., Millar, C. D. & Warman, G. R. Why does insect RNA look degraded? *J. Insect. Sci.* **10**, 159 (2010).
2. Applebaum, S. W., Ebstein, R. P. & Wyatt, G. R. Dissociation of ribosomal ribonucleic acid from silkmoth pupae by heat and dimethylsulfoxide: Evidence for specific cleavage points. *J. Mol. Biol.* **21**, 29–41 (1966).
3. McCarthy, S. D., Dugon, M. M. & Power, A. M. 'Degraded' RNA profiles in Arthropoda and beyond. *PeerJ.* **3**, e1436 (2015).
4. Henras, A. K., Plisson-Chastang, C., O'Donohue, M. F., Chakrabarty, A. & Gleizes, P. E. An overview of pre-ribosomal RNA processing in eukaryotes. *Wiley Interdiscip. Rev. RNA.* **6**, 225–242 (2015).
5. Fujiwara, H. & Ishikawa, H. Molecular mechanism of introduction of the hidden break into the 28S rRNA of insects: implication based on structural studies. *Nucleic Acids Res.* **14**, 6393–6401 (1986).
6. Sun, S., Xie, H., Sun, Y., Song, J. & Li, Z. Molecular characterization of gap region in 28S rRNA molecules in brine shrimp *Artemia parthenogenetica* and planarian *Dugesia japonica*. *Biochemistry (Mosc).* **77**, 411–417 (2012).
7. Lava-Sanches, P. A. & Puppo, S. Occurrence *in vivo* of "hidden breaks" at specific sites of the 26 S ribosomal RNA of *Musca carnaria*. *J. Mol. Biol.* **95**, 9–20 (1975).
8. Ben-Shem, A. *et al.* The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science.* **334**, 1524–1529 (2011).
9. Basile-Borgia, A. E., Dunbar, D. A. & Ware, V. C. Heterologous rRNA expression: internal fragmentation of *Sciara coprophila* 28S rRNA within microinjected *Xenopus laevis* oocytes. *Insect Mol. Biol.* **14**, 523–536 (2005).
10. Schroeder, A. *et al.* The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* **7**, 3 (2006).
11. Ishikawa, H. Evolution of ribosomal. *RNA. Comp. Biochem. and Physiol. B.* **58**, 1–7 (1977).
12. Ishikawa, H. Comparative studies on the thermal stability of animal ribosomal. *RNA's. Comp. Biochem. and Physiol. B.* **46**, 217–227 (1973).
13. Ishikawa, H. 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). *Comp. Biochem. and Physiol. B.* **56**, 229–234 (1977).
14. Ishikawa, H. Comparative studies on the thermal stability of animal ribosomal RNA's—III. Sponge (Porifera) and the other species (*Tetrahymena* and *Lachnus*). *Comp. Biochem. and Physiol. B.* **51**, 81–85 (1975).
15. Marlétaz, F., Peijnenburg, K. T. C. A., Goto, T., Satoh, N. & Rokhsar, D. S. A new spiralian phylogeny places the enigmatic arrow worms among gnathiferans. *Curr. Biol.* **29**, 312–8 (2019).
16. Ishikawa, H. Comparative studies on the thermal stability of animal ribosomal RNA's—II. Sea-anemones (Coelenterata). *Comp. Biochem. and Physiol. B.* **50**, 1–4 (1975).
17. Bieri, P., Leibundgut, M., Saurer, M., Boehringer, D. & Ban, N. The complete structure of the chloroplast 70S ribosome in complex with translation factor pY. *EMBO Journal.* **36**, 475–486 (2017).
18. Mertz, P. M., Bobek, L. A., Rekosh, D. M. & LoVerde, P. T. *Schistosoma haematobium* and *S. japonicum*: Analysis of the ribosomal RNA genes and determination of the "gap" boundaries and sequences. *Exp. Parasitol.* **73**, 137–149 (1991).
19. Karlstedt, K., Paatero, G., Mäkelä, J. H. & Wikgren, B. J. A hidden break in the 28.0S rRNA from *Diphyllobothrium dendriticum*. *J. Helminthol.* **66**, 193–197 (1992).
20. Haçariz, O. & Sayers, G. *Fasciola hepatica* - Where is 28S ribosomal RNA? *Exp. Parasitol.* **135**, 426–429 (2013).
21. DeLeo, D. M., Pe, J. L. & Bracken-Grissom, H. D. RNA profile diversity across Arthropoda: guidelines, methodologi,cal artifacts, and expected outcomes. *Biology Methods and Protocols.* **3**, bpy012 (2018).
22. Schiffer, P. H., Robertson, H. E. & Telford, M. J. Orthonectids are highly degenerate annelid worms. *Curr. Biol.* **28**, 1970–1974 (2018).
23. Ogino, K., Eda-Fujiwara, H., Fujiwara, H. & Ishikawa, H. What causes the aphid 28S rRNA to lack the hidden break?. *J. Mol. Evol.* **30**, 509–513 (1990).
24. Macharia, R. W., Ombura, F. L. & Aroko, E. O. Insects' RNA profiling reveals absence of "hidden break" in 28S ribosomal RNA molecule of onion thrips, *Thrips tabaci*. *J. Nucleic Acids.* **2015**, 965294 (2015).
25. Philippe, H. *et al.* Mitigating anticipated effects of systematic errors supports sister-group relationship between Xenacoelomorpha and Ambulacraria. *Curr. Biol.* **29**, 1818–1826 (2019).
26. Melen, G. J., Pesce, C. G., Rossi, M. S. & Kornblihtt, A. R. Novel processing in a mammalian nuclear 28S pre- rRNA: Tissue-specific elimination of an 'intron' bearing a hidden break site. *EMBO Journal.* **18**, 3107–3118 (1999).

27. Azpurua, J. *et al*. Naked mole-rat has increased translational fidelity compared with the mouse, as well as a unique 28S ribosomal RNA cleavage. *PNAS.* **110**, 17350–17355 (2013).
28. Leipoldt, M. & Engel, W. Hidden breaks in ribosomal RNA of phylogenetically tetraploid fish and their possible role in the diploidization process. *Biochem. Genet.* **21**, 819–841 (1983).
29. Bostock, C. J., Prescott, D. M. & Lauth, M. Lability of 26S ribosomal RNA in *Tetrahymena pyriformis*. *Exp. Cell. Res.* **66**, 260–262 (1971).
30. Stevens, A. & Pachler, P. Discontinuity of 26 s rRNA in *Acanthamoeba castellani*. *J. Mol. Biol.* **66**, 225–237 (1972).
31. Navarro-Ródenas, A., Carra, A. & Morte, A. Identification of an alternative rRNA post-transcriptional maturation of 26S rRNA in the kingdom Fungi. *Front. Microbiol.* **9**, 994 (2018).
32. Chapman, A. D. Numbers of living species in Australia and the world. Toowomba, Australia: 2nd edition *Australian Biodiversity Information Services*, http://www.environment.gov.au/system/files/pages/2ee3f4a1-f130-465b-9c7a-79373680a067/files/nlsaw-2nd-complete.pdf (2009).
33. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
34. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
35. Wu, D. C., Yao, J., Ho, K. S., Lambowitz, A. M. & Wilke, C. O. Limitations of alignment-free tools in total RNA-seq quantifications. *BMC Genomics.* **19**, 510 (2018).
36. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* **29**, 15–21 (2013).
37. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* **25**, 1754–1760 (2009).
38. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* **26**, 841–842 (2010).
39. Quast, C. *et al*. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–596 (2013).
40. The RNAcentral Consortium. RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Res.* **47**, D1250–1251 (2019).
41. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **47**, D8–D13 (2019).
42. Li, H. *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics.* **25**, 2078–2079 (2009).
43. Bernier, C. *et al*. RiboVision: Visualization and analysis of ribosomes. *Faraday Discuss.* **169**, 1–12 (2014).

Acknowledgements

We thank Kate Rawlinson for the electropherograms shown in Fig. 1. We thank members of the Telford lab for useful feedback. This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 764840 to M.T.; the European Research Council (ERC-2012-AdG 322790) to M.T. and Human Frontier Science Program (HFSP RGP0002/2016) to M.T.

Author contributions

M.J.T. proposed original idea. M.J.T., I.S.M., P.H.S. developed idea and guided research. P.N. conducted research with help from I.S.M., P.H.S. M.J.T. wrote initial draft. M.J.T., P.N., P.H.S., I.S.M. contributed to final draft. P.N. and I.S.M. prepared figures and tables. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-55573-1.

Correspondence and requests for materials should be addressed to M.J.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.