7970–7976 *Nucleic Acids Research, 2018, Vol. 46, No. 15* *Published online 18 May 2018*
doi: 10.1093/nar/gky414

Group I introns are widespread in archaea

Eric P. Nawrocki[1,*], Thomas A. Jones[2,3] and Sean R. Eddy[2,3,4,*]

[1]National Center for Biotechnology Information, U.S. National Library of Medicine, Bethesda, MD 20894, USA, [2]Howard Hughes Medical Institute, Harvard University, Cambridge, USA, [3]Department of Molecular and Cellular Biology, Harvard University, Cambridge, USA and [4]School of Engineering and Applied Sciences, Harvard University, Cambridge, USA

Received March 25, 2018; Revised April 27, 2018; Editorial Decision May 02, 2018; Accepted May 04, 2018

ABSTRACT

Group I catalytic introns have been found in bacterial, viral, organellar, and some eukaryotic genomes, but not in archaea. All known archaeal introns are bulge-helix-bulge (BHB) introns, with the exception of a few group II introns. It has been proposed that BHB introns arose from extinct group I intron ancestors, much like eukaryotic spliceosomal introns are thought to have descended from group II introns. However, group I introns have little sequence conservation, making them difficult to detect with standard sequence similarity searches. Taking advantage of recent improvements in a computational homology search method that accounts for both conserved sequence and RNA secondary structure, we have identified 39 group I introns in a wide range of archaeal phyla, including examples of group I introns and BHB introns in the same host gene.

INTRODUCTION

Group I introns are canonical examples of catalytic RNAs (1,2). They have been found in bacteria, organelles, bacteriophages, and a few eukaryotic nuclear genomes, but none have yet been found in archaea (3–5). Except for a few group II catalytic introns (6), all known archaeal introns are so-called bulge-helix-bulge (BHB) introns (7,8), named for a consensus RNA structure motif processed by the archaeal transfer RNA (tRNA) splicing endoribonuclease (9–11).

Tocchini-Valentini *et al.* (12) suggested that group I introns went extinct in archaea by evolving into simpler BHB introns, by co-opting a tRNA intron endonuclease for splicing and then decaying to lose RNA catalysis. This idea parallels the widely accepted hypothesis that in eukaryotes, spliceosomal introns and the *trans*-acting spliceosomal machinery evolved from catalytic group II introns (13–15). The evolutionary history and phylogenetic distribution of the known types of introns—eukaryotic spliceosomal introns, eukaryotic tRNA introns and archaeal BHB introns, and sporadically distributed group I and group II introns—have long been subjects of intense interest (16–18).

Group I introns might exist in archaea yet have been missed, because they are difficult to identify by sequence analysis (19). They conserve a distinctive catalytic RNA secondary and tertiary structure, but often show little primary sequence conservation, so sequence similarity search methods such as BLAST (20) can fail to detect them. They occur most often in ribosomal RNA (rRNA) genes or tRNA genes, where the presence an intervening sequence is easily recognized in sequence alignments, but distinguishing an unexpected group I from an expected BHB intron in an archaeal host gene requires RNA structure prediction. Many group I introns encode a homing endonuclease gene (HEG) in their RNA sequence, which is responsible for a DNA mobility event that propagates the intron into intronless host genes (5). HEGs can be identified by sequence similarity, but homologous HEGs are also found in many other classes of mobile DNA elements, including archaeal BHB introns, group II introns, inteins, and freestanding mobile HEGs (21,22). Many HEG homologs are known in archaea, and to date have been attributed to one of these other classes of mobile elements.

One way to identify group I introns is to perform a computational search for their conserved RNA structural features. Various computational methods have been developed for searching genome sequences for a combination of RNA structure and sequence similarity (23,24), including some programs specifically designed for group I intron detection (19,25). Perhaps the most sensitive general-purpose method for searching for homologs of a given RNA multiple sequence alignment and consensus RNA secondary structure uses probability models called covariance models (CMs) (26–28). Infernal, a CM software package for RNA homology search and alignment (29), is the basis for the Rfam RNA database, which contains CMs for 2500+ RNA families (30).

CM methods are computationally demanding. Until recently, database-wide Infernal searches for large RNA consensus structures required an infeasible amount of time. Over the past several years, a series of advances have greatly

*To whom correspondence should be addressed. Tel: +1 617 496 6757; Fax: +1 617 496 3894; Email: seaneddy@fas.harvard.edu
Correspondence may also be addressed to Eric P. Nawrocki. Email: nawrocke@ncbi.nlm.nih.gov

accelerated Infernal. It is now possible to do sensitive large-scale homology searches for consensus RNA structures as large as group I introns (29,31).

Here, we describe using Infernal to identify 39 group I introns in a wide variety of archaeal phyla.

MATERIALS AND METHODS

Subtype-specific group I intron alignments were downloaded from the GISSD database (46) in October 2014. The IB4 alignment was edited to correct two mistakes in secondary structure annotation, and the IA3 alignment was edited to correct a format issue.

All archaeal sequences in GenBank were retrieved in September 2017 using the Entrez query 'txid2157[orgn]' (the NCBI Taxonomy ID number for 'Archaea' is 2157) which returned 591 443 sequences comprised of 6 710 959 751 total nucleotides.

Infernal v1.1.2 (29) was used to build and calibrate CMs from group I intron alignments, using the cmbuild and cmcalibrate programs with default parameters. The cmsearch program was used with two different parameter settings to search each of the 15 CMs against the archaeal sequence data, once with default parameters, and once with the command line option `--anytrunc` which can improve performance on truncated or interrupted sequences, such as group I introns interrupted by embedded HEGs. Hits with an *E*-value $\leq$0.01 were kept, and overlapping hits were removed, keeping the hit with the lowest *E*-value. Highly similar hits were removed such that no two remaining hits were more than 97% identical given their alignment to the CM.

Homing endonuclease gene (HEG) homologs were detected by translating each of the 39 introns in all six frames and searching all ORFS $\geq$20aa against Pfam 31.0 (47) using the HMMER 3.1b2 hmmscan program (48), with an *E*-value threshold of $\leq 10^{-3}$.

Large subunit (LSU) and small subunit (SSU) rRNA sequences were identified using cmsearch in all sequences with an IB4 or IA3 hit using the Rfam 13.0 (31) archaeal LSU and SSU models (RF01959 and RF02540) with the command line option `--rfam`, which accelerates searches by enforcing strict sequence-based filters.

To check that host LSU and SSU rRNA sequences were archaeal in origin, we scored each rRNA sequence with the domain-specific Rfam covariance models LSU_rRNA_archaea, LSU_rRNA_bacteria, and LSU_rRNA_eukarya (RF02540, RF02541, RF02543) or the corresponding SSU models (RF01959, RF00177, RF01960) and checked that the highest-scoring alignment was to the archaeal profile. This is only suitable as a coarse-grained check; similarity scores to multiple alignment profiles are a crude substitute for phylogenetic classification on trees.

Taxonomic classification of 25 SSU rRNAs identified on the same contigs as the introns was done on the RDPClassifier web server (49) using training set 16 and default parameters, at a threshold of $\geq$80% confidence.

RESULTS AND DISCUSSION

The input to an Infernal search is a multiple RNA sequence alignment with consensus RNA secondary structure annotation. Group I introns have been classified into five types (IA, IB, IC, ID and IE) and further subdivided into fourteen subtypes (IA1, etc.) based on variation in conserved consensus secondary structure (50,51). No single alignment, such as the Rfam database group I intron alignment (Intron_gp1, RF00028), seemed likely to capture this structural diversity well, so we followed the approach of Lang *et al.* (19) and used subtype-specific models. We obtained fourteen subtype-specific alignments of group I introns from the Group I Intron Sequence and Structure Database (GISSD) (46), lightly edited to correct some errors in consensus structure annotation, and built CMs from each of them with Infernal. The IA3 subtype structural alignment, for example, contains 56 sequences, 205–374nt long, with an average pairwise sequence identity of 45%, and an annotated consensus structure of 81 base pairs. The IB4 subtype alignment contains 89 sequences, 203–392nt long, with 44% average pairwise sequence identity, and 71 annotated consensus base pairs.

The quantity and diversity of known archaeal genome sequence has grown rapidly recently, due to metagenomic studies of uncultivated archaea (34,36–44), including representatives of several newly proposed archaeal phyla and superphyla (33,35,52). We obtained a 6.7Gb file of all 591 443 archaeal sequences in GenBank using a taxonomic query of 'txid2157[orgn]' at NCBI.

We searched this archaeal sequence data with each of the 14 group I subtype CMs and with the Rfam Intron_gp1 model. Most of the 15 searches take less than 30 minutes on four CPU cores. The IA1 and IC3 subtype alignments include large HEG insertions, and these larger queries take 1–4 h. The Infernal search program, cmsearch, assigns log-odds probability scores to RNA secondary structure alignments and ranks putative hits by a measure of the statistical significance, the *E*-value (expectation value), the estimated number of false positives at that score threshold. After removing overlapping hits and keeping the hit with the lowest *E*-value, there were no significant hits at a threshold of $E < 0.01$ for ten subtypes and the RF00028 model. IA2 and IC3 searches found three and two significant hits, respectively, but upon examination, we were unsure whether these were truly in archaeal sequences, as opposed to misannotated or contaminating bacterial or phage DNA, and we did not consider these further.

The IB4 and IA3 CM searches found a total of 95 significant hits with an *E*-value of 0.01 or less, after we removed overlapping, lower scoring hits. We examined these hits and removed 25 that corresponded to redundant identical sequences in the archaeal database; seven where the sequence record does not contain a complete intron; three where we were not confident that the sequence folded into a complete group I consensus structure; and four where we were not confident that the host gene is an archaeal sequence. Because introns may contain nonconserved insertions including HEGs, a single intron may be identified by Infernal in more than one local alignment piece. After assigning the remaining hits to single introns, we identified a total of 39 nonredundant, full length putative group I introns, 27 IB4s and 12 IA3s. Most were identified by at least one strongly significant hit; 30 of the 39 have at least one hit of $E < 10^{-10}$.

Table 1. Summary of 39 identified archaeal group I introns showing the name we use to refer to each intron; GenBank accession; phylum assignment (R: inferred by RDPClassifier; G: annotated in GenBank; –: unclassified by either); coordinates of the intron in the source sequence; *E*-value of the most significant Infernal hit to this intron; insertion position in the host ribosomal RNA gene in canonical *E. coli* coordinate numbering (32); and citation for the source sequence. For four phylum assignments, noted by ?, Genbank annotation and RDPClassifier inference were in conflict, and we chose one (see Materials and Methods). For 14 assignments, noted by *, an SSU rRNA was present but RDPClassifier was unable to resolve a phylum-level assignment. IB4.16 is intentionally missing (see text). An expanded table is provided as a parseable .csv file in supplementary data.

Intron Name	Sequence Accession	Phylum	Intron Coordinates	Infernal *E*-value	Insertion Position	Reference
IA3.1	CP010426.1	Woesearchaeota (R)	664815–664111	2.2e–16	LSU/2593	(33)
IA3.2	LQMQ01000054.1	Euryarchaeota* (G)	4017–4815	4.6e–8	LSU/2500	(34)
IA3.3	LQMP01000030.1	Euryarchaeota (G+R)	66091–65273	3.3e–11	LSU/2500	(34)
IA3.4	JWKY01000014.1	–	16576–15805	4.4e–12	LSU/2593	(33)
IA3.5	ASMP01000002.1	Pacearchaeota? (R)	130592–130914	7.0e–33	LSU/2593	(35)
IA3.6	KP308561.1	Pacearchaeota (R)	16134–15526	2.9e–27	LSU/2593	(33)
IA3.7	KP308720.1	–*	23163–23468	3.5e–42	LSU/2593	(33)
IA3.8	KY476711.1	–*	66905–66301	1.1e–26	LSU/2593	(36)
IA3.9	MNVE01000016.1	Micrarchaeota* (G)	260342–261072	2.2e–15	LSU/2593	(37)
IA3.10	MNVF01000017.1	Micrarchaeota (G)	6638–7348	4.9e–16	LSU/2593	(37)
IA3.11	MNVH01000001.1	Micrarchaeota* (G)	9089–8693	1.6e–36	LSU/2593	(37)
IA3.12	MNVZ01000017.1	Pacearchaeota (G)	14519–14202	2.8e–30	LSU/2593	(37)
IB4.1	CBTY010000008.1	Thaumarchaeota (G+R)	399334–399054	1.0e–34	LSU/1931	(38)
IB4.2	KP308713.1	Woesearchaeota (R)	1026–667	2.3e–32	LSU/1931	(33)
IB4.3	KP308715.1	Woesearchaeota (R)	33361–33696	8.1e–39	LSU/1931	(33)
IB4.4	KP308717.1	Woesearchaeota (R)	8526–8928	3.0e–32	LSU/1931	(33)
IB4.5	KP308748.1	Woesearchaeota (R)	21438–21748	1.4e–25	LSU/1931	(33)
IB4.6	KP308561.1	Pacearchaeota (R)	17157–16694	9.1e–31	LSU/1931	(33)
IB4.7	KP308720.1	–*	21440–21763	1.4e–35	LSU/1923	(33)
IB4.8	AEIX01000219.1	Nanohaloarchaeota (G+R)	2242–1950	7.8e–21	SSU/1498	(39)
IB4.9	BA000048.1	Aigarchaeota? (G)	1314432–1313300	2.0e–7	LSU/1923	(40)
IB4.10	AEIX01000215.1	Nanohaloarchaeota (G)	10957–11697	4.5e–10	LSU/1931	(39)
IB4.11	AP011903.1	Thaumarchaeota? (R)	23185–23940	1.0e–6	LSU/1923	(40)
IB4.12	AQRL01000028.1	Woesearchaeota? (R)	30565–31289	9.2e–17	LSU/1931	(35)
IB4.13	ASPK01000003.1	Thaumarchaeota* (G)	12290–11538	7.5e–5	LSU/1923	(35)
IB4.14	AGCY01000458.1	Euryarchaeota (G+R)	2850–3533	1.1e–11	LSU/1931	(41)
IB4.15	KP308720.1	–*	21772–22514	3.3e–15	LSU/1931	(33)
IB4.17	MNVF01000017.1	Micrarchaeota (G)	5465–5748	2.4e–29	LSU/1923	(37)
IB4.18	MNVH01000001.1	Micrarchaeota* (G)	10442–9750	6.2e–8	LSU/1923	(37)
IB4.19	MNVH01000001.1	Micrarchaeota* (G)	13564–12830	2.7e–16	SSU/1498	(37)
IB4.20	MFRR01000014.1	Micrarchaeota (G)	65096–64410	5.0e–11	LSU/1923	(42)
IB4.21	MNVZ01000017.1	Pacearchaeota (G)	15452–15175	3.0e–27	LSU/1923	(37)
IB4.22	MFWP01000012.1	Pacearchaeota* (G)	2846–3187	5.2e–28	LSU/1931	(42)
IB4.23	MNFA01000040.1	Thaumarchaeota (G+R)	5119–5802	1.9e–10	LSU/1931	(43)
IB4.24	MNVG01000014.1	Micrarchaeota* (G)	9943–10190	1.7e–29	LSU/1931	(37)
IB4.25	MNVJ01000003.1	Micrarchaeota* (G)	25302–24603	6e–14	LSU/1931	(37)
IB4.26	MNVJ01000003.1	Micrarchaeota* (G)	28477–27735	4.3e–17	SSU/1498	(37)
IB4.27	MNWW01000056.1	Woesearchaeota (G+R)	8341–9021	1.3e–10	LSU/1931	(37)
IB4.28	NJDR01000009.1	Aenigmarchaeota (G)	12303–13059	3.5e–10	LSU/1923	(44)

Table 1 summarizes attributes of these introns, their source sequences, and their locations in their host genes.

We looked at three other lines of evidence, independent of Infernal similarity scores, that provide additional support for a conclusion that these are group I introns.

First, group I introns are most commonly found in rRNA and tRNA host genes (3), and they tend to be inserted at a small number of conserved homologous positions (32). We identified the host gene for each intron by similarity searches with flanking sequence. All 39 introns are in rRNA genes, 36 in large subunit (LSU) and 3 in small subunit (SSU). Figure 1 shows a schematic of the coordinates of six group I introns in LSU and SSU host genes, sometimes co-occurring with archaeal BHB introns in the same gene. We identified intron insertion positions in canonical *E. coli* coordinate numbering (Table 1). With the exception of the SSU/1498 insertion position, all insertion positions are previously observed insertion sites for group I introns (32).

Second, we looked for evidence of homing endonuclease genes in the putative introns. We translated each intron sequence in six frames and searched all ORFs of length $\geq$20 against all known Pfam protein domains (47), which include several distinct HEG families (53). Twenty-one of the 39 introns contain ORFs with significant similarity to a Pfam domain, all to LAGLI-DADG homing endonucleases (Pfam LAGLIDADG_1, LAGLIDADG_2, or LAGLIDADG_3; PF00961, PF03161, PF14528).

Third, the consensus secondary structure identified by an Infernal local alignment is typically just a subset of the base pairs in any individual RNA structure. Infernal is also unable to model pseudoknots, so for the P3/ P7 pseudoknot in the group I intron core, the Infernal models include the P3 helix (because its sequence is less conserved and has more covariation information) and P7 is aligned only as primary sequence. Identifying a P7 helix, other secondary structure elements typical of group I introns, and additional base

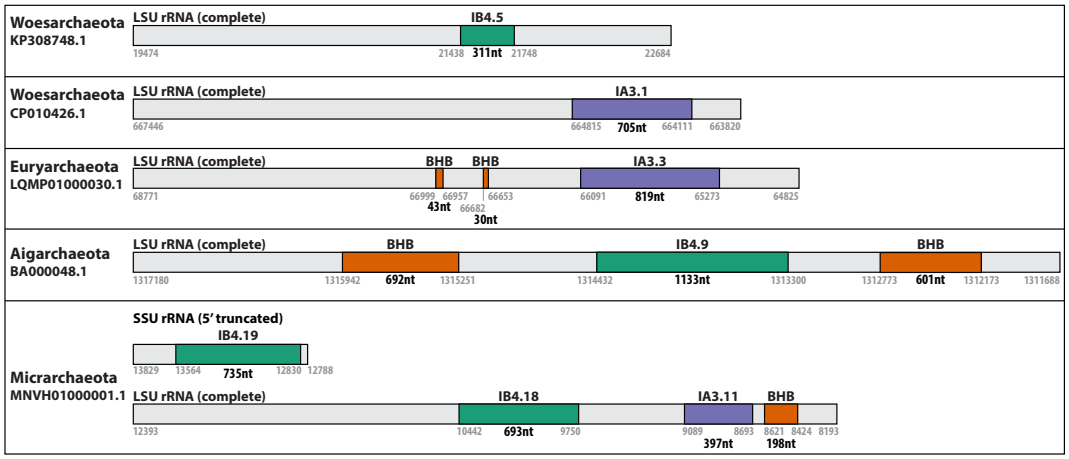

Figure 1. Coordinates of LSU and SSU rRNA host genes, group I introns and BHB introns in five archaeal sequences in GenBank.

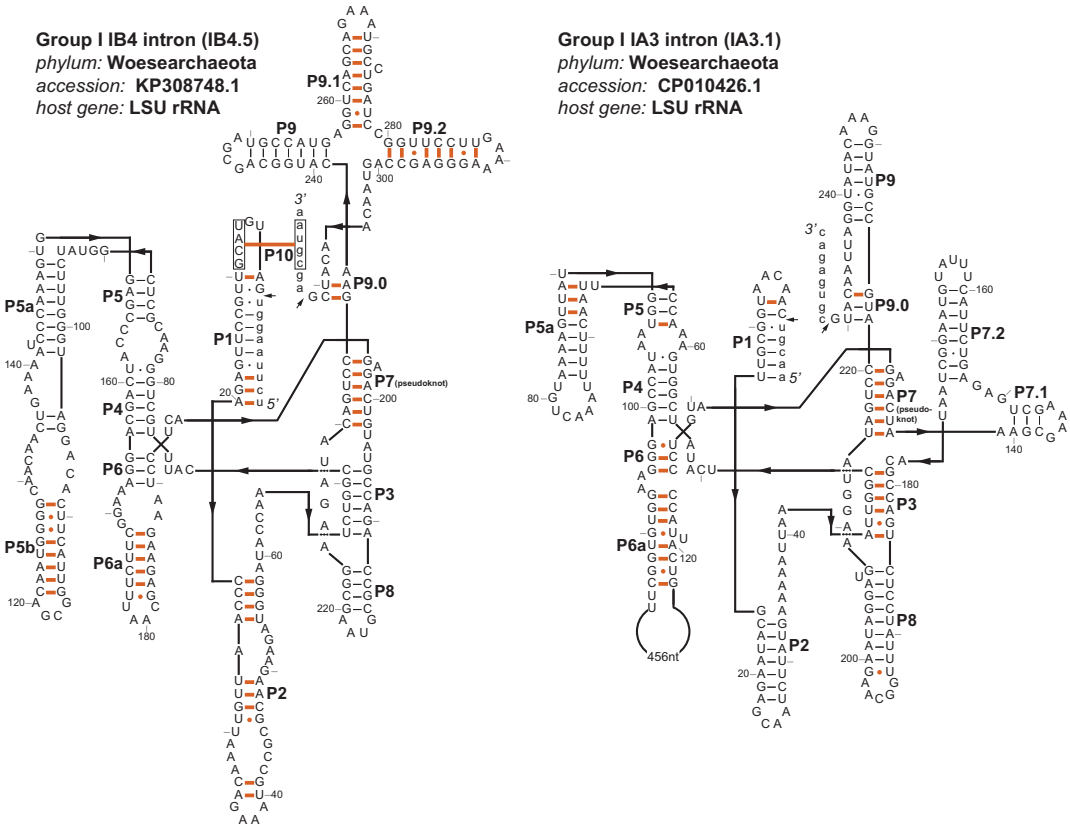

Figure 2. Predicted secondary structures of the IB4.5 and IA3.1 introns. Base pairs shown in thin black lines (or black dots, for GU pairs) are in the Infernal alignments. Those shown by thick red lines (or red dots) were manually predicted. Exonic residues are shown in lower case. Intron positions are numbered starting at 2; by convention, residue 1 is an exogenous added guanosine. Representation follows conventions of Cech *et al.* (45).

pairs that extend Infernal-predicted consensus helices provide more confidence in the predicted introns. We manually identified a P7 helix in all 39 predicted introns (Supplementary Table). We manually predicted complete group I intron structures for two introns, IB4.5 and IA3.1. For IB4.5, the cmsearch alignment predicted 58 consensus basepairs; our structure prediction includes 47 of these, and 56 other basepairs. IA3.1 contains a large insertion in P6 with homology to LAGLI-DADG homing endonucleases, so Infernal identified this intron in two local alignment pieces without identifying the P6 and P3 helices that span the insertion. Our structure prediction includes 54 of the 56 base pairs predicted by cmsearch, and 30 additional basepairs including P3 and P6 helices. Figure 2 shows the predicted complete IB4.5 and IA3.1 secondary structures, which are consistent with canonical group I intron structures (2).

The 39 introns are found in 31 different GenBank sequences annotated as archaeal, all of which were obtained

by metagenomic environmental sampling. Because GenBank annotation can be unreliable, especially for the phylogenetic source of sequences obtained in metagenomic samples, we also checked that each host rRNA gene sequence scored higher for archaeal SSU and LSU rRNA Rfam consensus models than to bacterial or eukaryotic models. An example of a sequence that failed this check is an LSU rRNA in GenBank MBAA01000200.1, a 4.7 kb contig that was binned into a Lokiarchaeota genome assembly, but appears to be bacterial. We had already named two introns from this contig (IA3.13 and IB4.16) before removing them from our analysis, which is why they are missing in Table 1.

We asked whether group I introns are broadly distributed across the archaeal phylogeny, or if they only occur in some particular clade. Specific phylum assignments are annotated in 25 of the 31 GenBank records. Both because of the unreliability of these annotations, and also because the taxonomy of archaeal phyla has been expanding (33,35,52), we sought an additional means of assigning sequences to archaeal phyla. SSU rRNA is a convenient and well-established phylogenetic marker for phylum-level classification, much more so than LSU rRNA. Although almost all the introns are in LSU rRNA genes, the genes for SSU and LSU rRNAs are typically adjacent, and we identified an SSU rRNA gene in 25 of the 31 source sequences. The SSU rRNA classification tool RDPClassifier (49) predicts a confident phylum assignment for 16 of these 25 SSU sequences. In four cases, RDP assignments conflicted with GenBank annotation. Three were cases where a new archaeal phylum was proposed after the GenBank sequence was deposited (for which we retained the RDP assignment), and one resulted from a mislabeled training sequence in RDPClassifier (for which we retained the Genbank annotation). Consensus phylum assignments are summarized in Table 1. Figure 3 shows that the introns are distributed widely across archaeal phyla and superphyla, even when counting only cases where GenBank and RDPClassifier phylum assignments strictly agree.

Various sources say that group I introns have not been identified in archaea before (3–5,12,55). However, we do note that in an analysis of an archaeal metagenomic sequencing survey, Nunoura *et al.* (40) referred in passing to an HEG-containing LSU rRNA intron, the same intron that we call IB4.11, as a group I intron. They apparently assumed that any HEG-containing intron is a group I. They did not show evidence to distinguish it from a BHB intron (which often contain HEGs), and they stated that archaeal group I introns were known, citing work on hyperthermophilic bacteria, not archaea (56).

Infernal searches are convenient but not strictly necessary to find these introns. BLAST searches can also identify significant similarities between some of these archaeal intron sequences and some known group I introns in bacteria and eukarya. One reason they have been missed in previous archaeal genome annotations may be that a few informative similarities would be hard to spot amidst a long list of BLAST hits to rRNA and HEG sequences.

The failure to identify group I introns in archaea has previously been attributed to their having evolved into BHB introns and gone extinct (12). It remains plausible that BHB introns derived from group I introns, but our work shows that group I introns are not extinct in archaea, and we identified cases where BHB and group I introns co-exist in the same host gene (Figure 1).

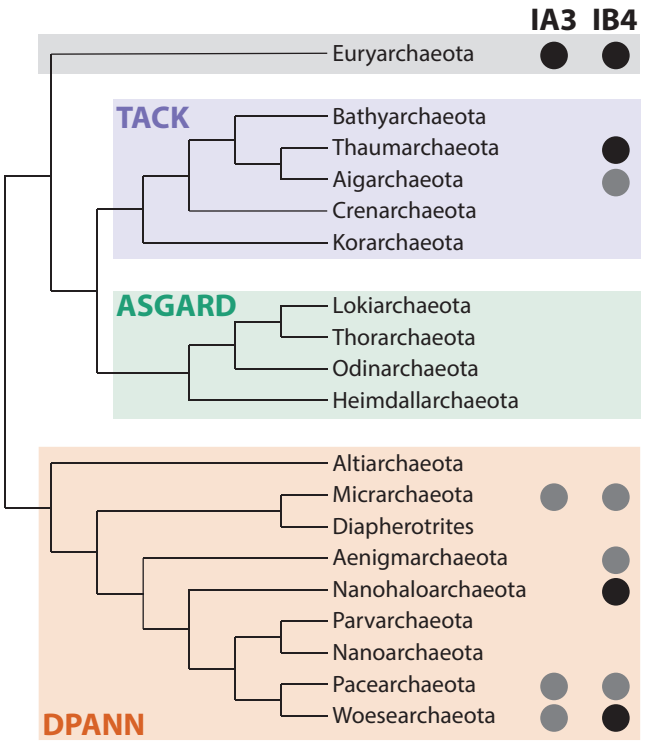

Figure 3. Distribution of identified group I introns in archaeal phyla. Taxonomic assignments of IA3 and IB4 introns from Table 1 are shown on a cladogram of archaeal phyla and superphyla (54). Dark circles indicate occurrence of one or more introns with higher confidence assignments, where GenBank annotation and RDPClassifier agreed.

Development of CM-based RNA similarity search methods was originally motivated by a desire to search for group I introns (57), but until recently, these methods were too computationally demanding to systematically search genome sequence data with consensus RNA structures this large (29,31). CM-based homology searches for essentially any structural RNA or RNA element are now feasible.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Alejandro Schäffer for his continued support of the project.

FUNDING

Intramural Research Program of the US National Institutes of Health National Library of Medicine; Howard Hughes Medical Institute; and the US National Institutes of Health [R01-HG009116 to S.R.E.]. Funding for open access charge: HHMI.
Conflict of interest statement. None declared.

REFERENCES

1. Kruger,K., Grabowski,P.J., Zaug,A.J., Sands,J., Gottschling,D.E. and Cech,T.R. (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell*, **31**, 147–157.
2. Vicens,Q. and Cech,T.R. (2006) Atomic level architecture of group I introns revealed. *Trends Biochem. Sci.*, **31**, 41–51.
3. Haugen,P., Simon,D.M. and Bhattacharya,D. (2005) The natural history of group I introns. *Trends Genet.*, **21**, 111–119.
4. Nielsen,H. and Johansen,S.D. (2009) Group I introns: moving in new directions. *RNA Biol.*, **6**, 375–383.
5. Hausner,G., Hafez,M. and Edgell,D.R. (2014) Bacterial group I introns: mobile RNA catalysts. *Mob. DNA*, **5**, 8.
6. McNeil,B.A., Semper,C. and Zimmerly,S. (2016) Group II introns: versatile ribozymes and retroelements. *Wiley Interdiscip. Rev. RNA*, **7**, 341–355.
7. Yoshihisa,T. (2014) Handling tRNA introns, archaeal way and eukaryotic way. *Front. Genet.*, **5**, 213.
8. Jay,Z.J. and Inskeep,W.P. (2015) The distribution, diversity, and importance of 16S rRNA gene introns in the order Thermoproteales. *Biol. Direct.*, **10**, 35.
9. Lykke-Andersen,J. and Garrett,R.A. (1994) Structural characteristics of the stable RNA introns of archaeal hyperthermophiles and their splicing junctions. *J. Mol. Biol.*, **243**, 846–855.
10. Diener,J.L. and Moore,P.B. (1998) Solution structure of a substrate for the archaeal pre-tRNA splicing endonucleases: the bulge-helix-bulge motif. *Mol. Cell.*, **1**, 883–894.
11. Berkemer,S.J., zu,Siederdissen, C.H., Amman,F., Wintsche,A., Will,S., Hofacker,I., Prohaska,S. and Stadler,P. (2015) Processed small RNAs in Archaea and BHB elements. *Genom. Comput. Biol.*, **1**, e18.
12. Tocchini-Valentini,G.D., Fruscoloni,P. and Tocchini-Valentini,G.P. (2011) Evolution of introns in the archaeal world. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 4782–4787.
13. Sharp,P.A. (1991) Five easy pieces. *Science*, **254**, 663.
14. Zimmerly,S. and Semper,C. (2015) Evolution of group II introns. *Mob. DNA*, **6**, 7.
15. Novikova,M. and Belfort,M. (2017) Mobile group II introns as ancestral eukaryotic elements. *Trends Genet.*, **33**, 773–783.
16. Gilbert,W. (1978) Why genes in pieces? *Nature*, **271**, 501.
17. Martin,W. and Koonin,E.V. (2006) Introns and the origin of nucleus-cytosol compartmentalization. *Nature*, **440**, 41–45.
18. Doolittle,W.F. (2014) The trouble with (group II) introns. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 6536–6537.
19. Lang,B.F., Laforest,M.-J. and Burger,G. (2007) Mitochondrial introns: a critical view. *Trends Genet.*, **23**, 119–125.
20. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
21. Belfort,M. and Roberts,R.J. (1997) Homing endonucleases: keeping the house in order. *Nucleic Acids Res.*, **25**, 3379–3388.
22. Hafez,M. and Hausner,G. (2012) Homing endonucleases: DNA scissors on a mission. *Genome*, **55**, 553–569.
23. Macke,T.J., Ecker,D.J., Gutell,R.R., Gautheret,D., Case,D.A. and Sampath,R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.*, **29**, 4724–4735.
24. Gautheret,D. and Lambert,A. (2001) Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J. Mol. Biol.*, **313**, 1003–1011.
25. Lisacek,F., Diaz,Y. and Michel,F. (1994) Automatic identification of group I intron cores in genomic DNA sequences. *J. Mol. Biol.*, **235**, 1206–1217.
26. Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
27. Sakakibara,Y., Brown,M., Hughey,R., Mian,I.S., Sjölander,K., Underwood,R.C. and Haussler,D. (1994) Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.*, **22**, 5112–5120.
28. Freyhult,E.K., Bollback,J.P. and Gardner,P.P. (2007) Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res.*, **17**, 117–125.
29. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
30. Kalvari,I., Argasinska,J., Quinones-Olvera,N., Nawrocki,E.P., Rivas,E., Eddy,S.R., Bateman,A., Finn,R.D. and Petrov,A.I. (2017) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.*, **46**, D335–D342 .
31. Nawrocki,E.P., Burge,S.W., Bateman,A., Daub,J., Eberhardt,R.Y., Eddy,S.R., Floden,E.W., Gardner,P.P., Jones,T.A., Tate,J. *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, **43**, D130–D137.
32. Jackson,S., Cannone,J., Lee,J., Gutell,R. and Woodson,S. (2002) Distribution of rRNA introns in the three-dimensional structure of the ribosome. *J. Mol. Biol.*, **323**, 35–52.
33. Castelle,C.J., Wrighton,K.C., Thomas,B.C., Hug,L., Brown,C.T., Wilkins,M.J., Frischkorn,K.R., Tringe,S.G., Singh,A., Markillie,L.M. *et al.* (2015) Genomic expansion of domain Archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.*, **25**, 690–701.
34. Baker,B.J., Saw,J.H., Lind,A.E., Lazar,C.S., Hinrichs,K.U., Teske,A.P. and Ettema,T.J. (2016) Genomic inference of the metabolism of cosmopolitan subsurface Archaea, Hadesarchaea. *Nat. Microbiol.*, **1**, 16002.
35. Rinke,C., Schwientek,P., Sczyrba,A., Ivanova,N.N., Anderson,I.J., Cheng,J.F., Darling,A., Malfatti,S., Swan,B.K., Gies,E.A. *et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, **499**, 431–437.
36. Paul,B.G., Burstein,D., Castelle,C.J., Handa,S., Arambula,D., Czornyj,E., Thomas,B.C., Ghosh,P., Miller,J.F., Banfield,J.F. *et al.* (2017) Retroelement-guided protein diversification abounds in vast lineages of Bacteria and Archaea. *Nat. Microbiol.*, **2**, 17045.
37. Probst,A.J., Castelle,C.J., Singh,A., Brown,C.T., Anantharaman,K., Sharon,I., Hug,L.A., Burstein,D., Emerson,J.B., Thomas,B.C. *et al.* (2017) Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO2 concentrations. *Environ. Microbiol.*, **19**, 459–474.
38. Lebedeva,E.V., Hatzenpichler,R., Pelletier,E., Schuster,N., Hauzmayer,S., Bulaev,A., Grigor'eva,N.V., Galushko,A., Schmid,M., Palatinszky,M. *et al.* (2013) Enrichment and genome sequence of the group I.1a ammonia-oxidizing Archaeon '*Ca.* Nitrosotenuis uzonensis' representing a clade globally distributed in thermal habitats. *PLOS ONE*, **8**, e80835.
39. Narasingarao,P., Podell,S., Ugalde,J.A., Brochier-Armanet,C., Emerson,J.B., Brocks,J.J., Heidelberg,K.B., Banfield,J.F. and Allen,E.E. (2012) De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *ISME J.*, **6**, 81–93.
40. Nunoura,T., Hirayama,H., Takami,H., Oida,H., Nishi,S., Shimamura,S., Suzuki,Y., Inagaki,F., Takai,K., Nealson,K.H. *et al.* (2005) Genetic and functional properties of uncultivated thermophilic crenarchaeotes from a subsurface gold mine as revealed by analysis of genome fragments. *Environ. Microbiol.*, **7**, 1967–1984.
41. Podell,S., Ugalde,J.A., Narasingarao,P., Banfield,J.F., Heidelberg,K.B. and Allen,E.E. (2013) Assembly-driven community genomics of a hypersaline microbial ecosystem. *PLOS ONE*, **8**, e61692.
42. Anantharaman,K., Brown,C.T., Hug,L.A., Sharon,I., Castelle,C.J., Probst,A.J., Thomas,B.C., Singh,A., Wilkins,M.J., Karaoz,U. *et al.* (2016) Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.*, **7**, 13219.
43. Butterfield,C.N., Li,Z., Andeer,P.F., Spaulding,S., Thomas,B.C., Singh,A., Hettich,R.L., Suttle,K.B., Probst,A.J., Tringe,S.G. *et al.* (2016) Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. *PeerJ*, **4**, e2687.
44. Dombrowski,N., Seitz,K.W., Teske,A.P. and Baker,B.J. (2017) Genomic insights into potential interdependencies in microbial hydrocarbon and nutrient cycling in hydrothermal sediments. *Microbiome*, **5**, 106.
45. Cech,T.R., Damberger,S.H. and Gutell,R.R. (1994) Representation of the secondary and tertiary structure of group I introns. *Nat. Struct. Biol.*, **1**, 273–280.

46. Zhou,Y., Lu,C., Wu,Q.J., Wang,Y., Sun,Z.T., Deng,J.C. and Zhang,Y. (2008) GISSD: group I intron sequence and structure database. *Nucleic Acids Res.*, **36**, D31–D37.
47. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
48. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLOS Comp. Biol.*, **7**, e1002195.
49. Wang,Q., Garrity,G.M., Tiedje,J.M. and Cole,J.R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, **73**, 5261–5267.
50. Michel,F. and Westhof,E. (1990) Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.*, **216**, 585–610.
51. Li,Z. and Zhang,Y. (2005) Predicting the secondary structures and tertiary interactions of 211 group I introns in IE subgroup. *Nucleic Acids Res.*, **33**, 2118–2128.
52. Zaremba-Niedzwiedzka,K., Caceres,E.F., Saw,J.H., Bäckström,D., Juzokaite,L., Vancaester,E., Seitz,K.W., Anantharaman,K., Starnawski,P., Kjeldsen,K.U. *et al.* (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, **541**, 353–358.
53. Stoddard,B.L. (2014) Homing endonucleases from mobile group I introns: discovery to genome engineering. *Mob. DNA.*, **5**, 7.
54. Spang,A., Caceres,E.F. and Ettema,T.J.G. (2017) Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science*, **357**, eaaf3883.
55. Lykke-Andersen,J., Aagaard,C., Semionenkov,M. and Garrett,R.A. (1997) Archaeal introns: splicing, intercellular mobility and evolution. *Trends Biochem. Sci.*, **22**, 326–331.
56. Nesbø,C.L. and Doolittle,W.F. (2003) Active self-splicing group I Introns in 23S rRNA genes of hyperthermophilic bacteria, derived from introns in eukaryotic organelles. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 10806–10811.
57. Eddy,S.R. (2015) Homology searches for structural RNAs: from proof of principle to practical use. *RNA*, **21**, 605–607.