PNAS RESEARCH ARTICLE | GENETICS

Massive intein content in *Anaeramoeba* reveals aspects of intein mobility in eukaryotes

Lucie Gallot-Lavallée[a,b,1], Jon Jerlström-Hultqvist[a,b,c,1], Paula Zegarra-Vidarte[c], Dayana E. Salas-Leiva[a,b,2], Courtney W. Stairs[d], Ivan Čepička[e], Andrew J. Roger[a,b], and John M. Archibald[a,b,1]

Edited by Marlene Belfort, University at Albany, State University of New York, Albany, NY; received April 19, 2023; accepted October 20, 2023

Inteins are self-splicing protein elements found in viruses and all three domains of life. How the DNA encoding these selfish elements spreads within and between genomes is poorly understood, particularly in eukaryotes where inteins are scarce. Here, we show that the nuclear genomes of three strains of *Anaeramoeba* encode between 45 and 103 inteins, in stark contrast to four found in the most intein-rich eukaryotic genome described previously. The *Anaeramoeba* inteins reside in a wide range of proteins, only some of which correspond to intein-containing proteins in other eukaryotes, prokaryotes, and viruses. Our data also suggest that viruses have contributed to the spread of inteins in *Anaeramoeba* and the colonization of new alleles. The persistence of *Anaeramoeba* inteins might be partly explained by intragenomic movement of intein-encoding regions from gene to gene. Our intein dataset greatly expands the spectrum of intein-containing proteins and provides insights into the evolution of inteins in eukaryotes.

genomics | microbiology | mobile genetic elements | inteins | evolution

Significance

The spread of mobile genetic elements within and between the genomes of diverse organisms has greatly impacted biology across the tree of life. Inteins are an enigmatic type of mobile genetic element capable of self-splicing at the protein level. While inteins are generally rare in eukaryotes, we show that three strains of the recently described eukaryotic microbe *Anaeramoeba* have many dozens of inteins encoded in their nuclear genomes. Genomic investigation shows that *Anaeramoeba*'s inteins were likely acquired by lateral transfer from viruses and bacteria and subsequently spread to many different genomic locations via allelic and, possibly, nonallelic homing. The promiscuous nature of inteins in *Anaeramoeba* makes them promising substrates for genetic engineering and biotechnological applications.

Inteins are internal protein elements capable of catalyzing their own excision (1, 2). Analogous to self-splicing mRNA introns, intein self-splicing results in peptide bond formation between the amino (N)- and carboxyl (C)-terminal exteins and reconstitution of the mature host protein. Many inteins contain a homing endonuclease (HE) domain (3) that facilitates their mobility at the DNA level. Homing begins with the creation of a double-strand break in an intein-free allele by the HE encoded by an intein-donor allele. Using the endogenous homologous recombination pathway, the cleaved allele then recombines with the intein-donor allele, resulting in intein spread (4, 5). This process also likely underlies how inteins move between the genomes of different organisms by lateral gene transfer (LGT) (6–8). HE-mediated allelic specificity means that intein coding regions found at the same location tend to share more sequence similarity to one another than to those in different locations (9).

Inteins were initially described in the Vacuolar ATPase (VMA) of *Saccharomyces cerevisiae* (10, 11), and the subsequent growth of sequence databases facilitated the systematic search for inteins in bacteria, archaea, eukaryotes, and viruses (12, 13). Inteins are more abundant in prokaryotes, with 25% of bacterial and 50% of archaeal genomes encoding at least one intein (13). While prokaryotic inteins are hosted by many different proteins (13), >60% of intein-containing proteins in prokaryotes are involved in DNA replication, recombination, and repair. One possible explanation for this bias is that the expression of inteins in concert with DNA repair enzymes reduces the damaging effects of nonspecific HE activity, thereby reducing selection against intein spread (3). It may also facilitate intein spread by ensuring that the recipient (intein-minus) allele is in an open chromatin state at the same time as the homing enzymes are active (3, 14). Another factor is that DNA replication, recombination, and repair genes are prevalent in phage genomes (13, 15–17), which could facilitate phage-mediated intein spread (13, 18). Inteins generally reside within evolutionarily conserved regions of proteins such as active sites, DNA-binding sites, and ligand-binding motifs (13, 14, 18). This nonrandom distribution is thought to hinder intein loss, as imprecise deletions of the intein sequence would negatively impact host protein function (18). Given that such sites are highly conserved across species, it could also facilitate HE-mediated homing and LGT of inteins between distantly related organisms (18). Interestingly, research suggests that some inteins have been exapted to the benefit of their host organisms as "environmental sensors," with conditional splicing playing a role in the regulation of host protein expression (19, 20).

The intein distribution landscape in eukaryotes differs greatly from prokaryotes. A recent analysis revealed 257 inteins residing in the nuclear or chloroplast genomes of 231 eukaryotes (12). These inteins were restricted primarily to unicellular species and to a much smaller array of proteins than in prokaryotes. Indeed, no nucleus-encoded DNA processing proteins have been reported to host inteins. Fungal *Pre-mRNA-processing factor 8* (Prp8)

Author contributions: L.G.-L., A.J.R., and J.M.A. designed research; L.G.-L., J.J.-H., P.Z.-V., D.E.S.-L., C.W.S., and I.Č. performed research; P.Z.-V. contributed new reagents/analytic tools; L.G.-L., J.J.-H., D.E.S.-L., C.W.S., I.Č., A.J.R., and J.M.A. analyzed data; and L.G.-L., J.J.-H., D.E.S.-L., A.J.R., and J.M.A. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2023 the Author(s). Published by PNAS. This article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]To whom correspondence may be addressed. Email: lucie.gallot-lavallee@dal.ca, jon.jerlstrom.hultqvist@icm.uu.se, or jmarchib@dal.ca.

[2]Present address: Department of Biochemistry, University of Cambridge, Cambridge CB2 1QW, United Kingdom.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2306381120/-/DCSupplemental.

Published November 29, 2023.

Downloaded from https://www.pnas.org by UPPSALA UNIVERSITETSBIBLIOTEK BMC-BIBLIOTEKET on January 2, 2024 from IP address 130.238.37.18.

Table 1. Inteins in *A. ignava* strain BMAN and *A. flamelloides* strains BUSSELTON2 and SCHOONER1

	BMAN	BUSSELTON2	SCHOONER1
No. of inteins	103	45	47
No. of exteins*	79	42	43
No. of distinct intein-containing KOGs*	49	17	17
Mean intein size (aa)	367.6	352.6	351.6
Minimum intein size	160	185	203
Maximum intein size	550	500	500
Anaeramoeba protein hosts with eukaryotic counterpart†	6	3	3
Anaeramoeba protein hosts with viral counterpart†	4	3	3
Anaeramoeba protein hosts with prokaryotic counterpart†	29	12	12
Novel protein hosts in *Anaeramoeba*	18	3	3

*Numbers differ because of recent duplications of several extein+intein coding regions in both *Anaeramoeba* species.
†Based on KOG: recent paralogs belonging to the same KOG were not counted multiple times.

and VMA are the most common hosts of nuclear inteins, followed by RNA polymerases (RNApol) in several fungi and some amoebozoans. Here we report the existence of a remarkably high number of inteins in the nuclear genomes of three strains of *Anaeramoeba*, a recently characterized genus of unicellular anaerobic eukaryotes belonging to the Metamonada, a eukaryotic "supergroup" noteworthy in having mitochondrion-related organelles with metabolisms adapted to low-oxygen environments (21–23). Analysis of their peculiar intein distribution profiles provides important insights into the origins and spread of inteins within and between the genomes of eukaryotes, prokaryotes, and viruses.

Results

Inteins Are Abundant in *Anaeramoeba*. Preliminary investigation of three *Anaeramoeba* nuclear genomes—*A. ignava* strain BMAN, *A. flamelloides* strain BUSSELTON2, and *A. flamelloides* strain SCHOONER1 (23)—revealed the presence of numerous putative intein-encoding regions. Using BLAST- and hidden Markov model–based approaches, we performed a systematic survey of the *Anaeramoeba* proteins using inteins from InBase and previous studies (12, 24). We predicted 103, 45, and 47 complete inteins in *A. ignava*, *A. flamelloides* (BUSSELTON2), and *A. flamelloides* (SCHOONER1), respectively (Table 1), far exceeding the highest number found in any other organism sequenced to date (16 in bacteria, 15 in archaea, and four nucleus-encoded inteins in the opisthokont *Capsaspora owczarzaki*) (13). We also found pseudogenized inteins (*SI Appendix*, section 1) but no "split" inteins (i.e., those undergoing *trans*-splicing). In *A. ignava* strain BMAN, 15 proteins were found to have two or more inteins; in BUSSELTON2 and SCHOONER1, three and four proteins have two or more inteins, respectively. Remarkably, the BMAN intron splicing protein Prp8 has five inteins, four at previously described sites [Prp8-a,-b,-d and -e (25)] and one at a novel site (Prp8-f) located in the RNAse H domain (residues AKWK/TAEE; Dataset S1). In addition to Prp8, one of the BMAN dynein paralogs hosts three inteins, as do the MCM6 and MCM8 proteins. A maximum of four inteins was previously described for a single protein (26), although we also found other instances of four inteins per protein in taxonomically diverse archaeal MCM proteins; i.e., in National Center for Biotechnology Information (NCBI) reference sequences: WP_083500658.1, WP_048152640, WP_014557079.1.

Splicing was confirmed for three *Anaeramoeba* inteins using a combination of shotgun proteomics and in vitro experiments. Approximately 4,000 proteins were detected in *A. flamelloides* BUSSELTON2 using nanoLC-MS/MS, although peptides diagnostic of splicing (i.e., those mapping to intein-extein boundaries) were rare. Two inteins in DNA-directed RNA polymerase 2 subunit 2 (RPB2) were nevertheless shown to be spliced out of their native extein host (*SI Appendix*, section 2). One of the three predicted inteins in the *A. ignava* BMAN activating signal cointegrator 1 complex 3 (g14419.t1) protein was also shown in vitro to splice out of a GFP-intein construct expressed in *Escherichia coli* (Fig. 1 and *SI Appendix*, section 2). This intein is noteworthy in that it is a singleton (it shows no obvious similarity to inteins described in other organisms; see below). Combined with these experimental data and the fact that many of the predicted *Anaeramoeba* inteins reside in presumed-to-be essential proteins (see below), examination of splicing domain sequences (*SI Appendix*, section 3) suggests that most inteins identified herein are capable of splicing.

Exceptional Functional Diversity of *Anaeramoeba* Intein-Containing Proteins. We used the Eukaryotic Orthologous Groups of proteins (KOG) database to assign a KOG and functional category to each *Anaeramoeba* intein-containing protein. *Anaeramoeba* inteins reside in 55 different KOGs (given that some proteins host several inteins and some inteins are found in species-specific paralogs, the number of unique KOGs is less than the total number of inteins). In contrast, other known eukaryotic nuclear inteins are restricted to 15 different KOGs, 13 of which were described by Green et al. (12) [we detected two additional intein-containing KOGs in the green alga *Chloropicon primus*: DNApol delta KOG0969 and zeta KOG0968 (protein accessions QDZ18123.1 and QDZ21998, respectively)]. *Anaeramoeba* intein-containing proteins corresponding to intein-bearing KOGs in other eukaryotes include (but are not limited to) Prp8 (BMAN, BUSSELTON2, and SCHOONER1), VMA (BMAN only), DNA-directed RNA polymerase 2 subunit 2 (RPB2) (BMAN, BUSSELTON2, and SCHOONER1), and Transcription initiation factor 5B (BMAN only). Numerous *Anaeramoeba* proteins with inteins have an intein-containing homolog in archaea, bacteria, or both (29, 12, and 12 for BMAN, BUSSELTON2, and SCHOONER1, respectively). Additionally, several intein-bearing proteins in *Anaeramoeba* have intein-containing homologs in *Nucleocytoviricota*, namely Lon proteases, DNApol alpha, delta, and epsilon, and the largest and second-largest subunit of RNA polymerase (the two latter proteins were previously the only intein-bearing proteins reported in both eukaryotes and viruses). The remaining KOGs (18 in BMAN and three each in BUSSELTON2 and SCHOONER1) represent novel host proteins (Table 1 and Dataset S1).

Downloaded from https://www.pnas.org by UPPSALA UNIVERSITETSBIBLIOTEK BMC-BIBLIOTEKET on January 2, 2024 from IP address 130.238.37.18.

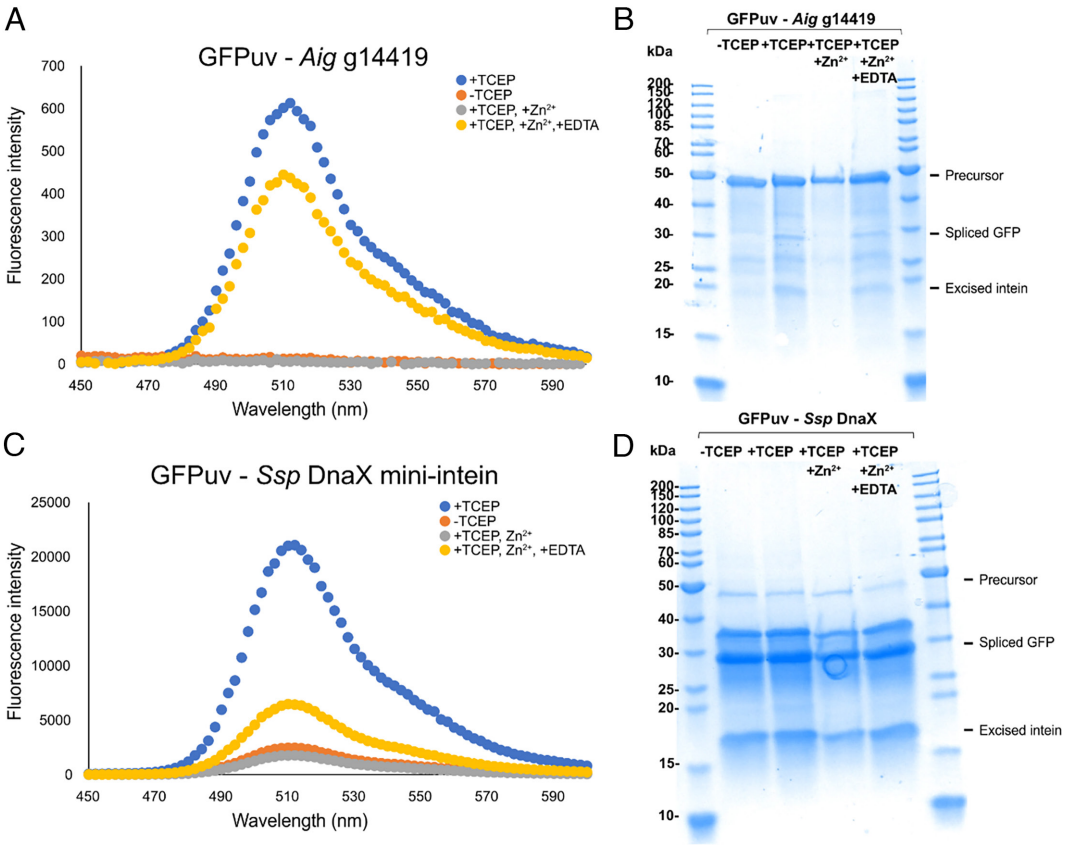

Fig. 1. Protein-splicing activity of an *Anaeramoeba* singleton intein. In vitro splicing of (*A* and *B*) *A. ignava* intein *Aig* g14419 and (*C* and *D*) *Synechocystis* sp. strain PCC6803 (*Ssp*) DnaX mini-intein. Metal ion affinity-purified precursor from inclusion bodies produced by *E. coli* BL21 (DE3) transformed with plasmid GFPuv-*Aig* g14419 or GFPuv-*Ssp* DnaX in buffer B was renatured by dilution in modified buffer C as follows: (−TCEP) A small amount of purified protein (30 µL) was diluted in buffer C without tris(2-carboxyethyl)phosphine (TCEP), (+TCEP) was diluted into buffer C with 1 mM TCEP, (+TCEP, Zn^{2+}) was diluted into buffer C supplemented with 1 mM TCEP, 2 mM $ZnCl_2$ but without Ethylenediaminetetraacetic acid (EDTA), and (+TCEP, Zn^{2+}, +EDTA) was diluted in buffer C with 1 mM TECP, 2 mM $ZnCl_2$, and 4 mM EDTA. After dilution, the samples were incubated for 14 h at 20 °C and analyzed (*A* and *C*) for fluorescence upon excitation at 395 nm (*B* and *D*) for protein splicing by SDS-PAGE. Sizes of expected precursors, spliced GFPs, and excised inteins are indicated. Fluorescence intensity is in arbitrary units. The size markers are PageRuler Unstained Protein Ladder (#26614, Thermo Scientific).

While the functional distribution of nucleus-encoded intein host proteins in eukaryotes differs from that of prokaryotes (12, 13), the *Anaeramoeba* profile exhibits both eukaryotic and prokaryotic signatures (Fig. 2). For example, the functional category RNA processing and modification (A) (which includes Prp8 and Prp16) is well represented in the intein-containing host proteins of *Anaeramoeba* and other eukaryotes but not in prokaryotes. Similarly, category K (Transcription) is well represented in *Anaeramoeba* and other eukaryotes but is a minor category in prokaryotic host proteins. However, most *Anaeramoeba* inteins are found in proteins belonging to category L—Replication, recombination, and repair—which is a common feature of prokaryotic but not eukaryotic host proteins.

We also detected *Anaeramoeba* inteins in proteins belonging to functional categories that were not previously reported to contain intein-hosting proteins, namely Cytoskeleton (Z) and Chromatin structure and dynamics (B). Curiously, multiple *Anaeramoeba* proteins belonging to these categories were found to be invaded by inteins: two types of Dynein (Cytoplasmic dynein 1 heavy chain 1 and Dynein heavy chain 10, axonemal isoform) and one Kinesin for Cytoskeleton (Z); and two different Helicases in category B (Chromatin remodeling protein HARP/SMARCAL1 and Chromatin remodeling complex SWI/SNF, component SWI2) (Fig. 2). Furthermore, some of these proteins host several inteins (e.g., three in a dynein homolog of *A. ignava*). We note that these inteins are not randomly distributed within the newly identified host proteins but are instead localized to more highly conserved regions. Collectively, these results indicate that cytoskeletal and chromatin structuring proteins can be "suitable" intein hosts.

***Anaeramoeba* Inteins Are Evolutionarily Diverse.** To infer the evolutionary origins of the inteins in *Anaeramoeba*, we used a sequence similarity network-based approach. This is because inteins are often so divergent in sequence that phylogenetic methods fail to yield meaningful results (9). This is particularly problematic when comparing nonallelic inteins, i.e., those inserted in different proteins or different regions of the same protein. Consistent with the diversity of host proteins in *Anaeramoeba* and the low sequence identity shared among the inteins themselves, *Anaeramoeba* inteins have diverse origins (Fig. 3). Eighteen distinct clusters linking *Anaeramoeba* inteins to those in prokaryotes, eukaryotes or viruses were inferred. This pattern is suggestive of many independent acquisitions from diverse sources, although it is impossible to discern exactly how many due to the complex nature of the network.

Nine *Anaeramoeba* inteins are singletons—they show no meaningful similarity to any known inteins. Additionally, two inteins in BMAN, 16 in SCHOONER1, and 16 in BUSSELTON2 form clusters comprised only of inteins inserted at paralogous sites in recent species-specific duplicates. These are designated pseudo-singletons (Fig. 3). These 43 singleton/pseudo-singleton inteins reside in 13 different KOGs. Surprisingly, only one of these KOGs (AAA+-type ATPase containing a bromodomain) is not known to contain an intein in at least one other organism, while two additional

Downloaded from https://www.pnas.org by UPPSALA UNIVERSITETSBIBLIOTEK BMC-BIBLIOTEKET on January 2, 2024 from IP address 130.238.37.18.

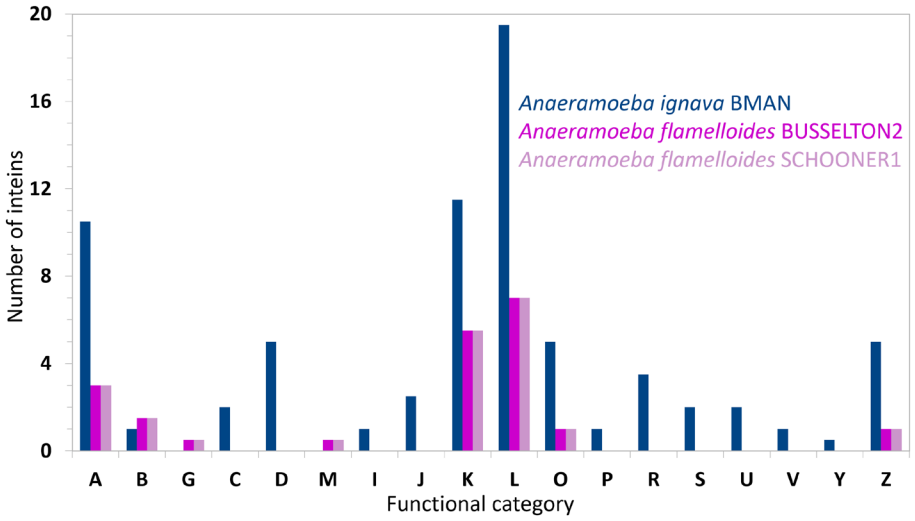

Fig. 2. Host proteins for *Anaeramoeba* inteins are functionally diverse. KOG-based functional categories of intein-containing proteins in three strains of *Anaeramoeba*. INFORMATION STORAGE AND PROCESSING: [J] Translation, ribosomal structure and biogenesis; [A] RNA processing and modification; [K] Transcription; [L] Replication, recombination and repair; [B] Chromatin structure and dynamics. CELLULAR PROCESSES AND SIGNALING: [D] Cell cycle control, cell division, chromosome partitioning; [Y] Nuclear structure; [V] Defense mechanisms; [T] Signal transduction mechanisms; [M] Cell wall/membrane/envelope biogenesis; [N] Cell motility; [Z] Cytoskeleton; [W] Extracellular structures; [U] Intracellular trafficking, secretion, and vesicular transport; [O] Posttranslational modification, protein turnover, chaperones. METABOLISM: [C] Energy production and conversion; [G] Carbohydrate transport and metabolism; [E] Amino acid transport and metabolism; [F] Nucleotide transport and metabolism; [H] Coenzyme transport and metabolism; [I] Lipid transport and metabolism; [P] Inorganic ion transport and metabolism; [Q] Secondary metabolites biosynthesis, transport and catabolism. POORLY CHARACTERIZED: [R] General function prediction only; [S] Function unknown.

KOGs have inteins but not at orthologous sites elsewhere in the tree of life [Activating Signal Cointegrator 1 Complex subunit 3 (ASC1S3) and DNApol theta/eta (POLQ)]. The 10 remaining KOGs have inteins in *Anaeramoeba* that are inserted at orthologous sites in other species. The fact that these inteins are nevertheless singletons/pseudo-singletons can be explained two ways: either they correspond to independent intein invasions of the same site by unrelated inteins, or they are allelic inteins that have diverged to the point that homology is no longer detectable (or below the threshold used for building the network). The latter scenario is the more likely explanation in at least some cases. For example, the singleton inteins hosted by the BMAN MCM8, PcrA-like helicase, CDC48 ATPase, mismatch DNA repair Msh2, RPB1, and SWI-SNF helicase proteins, as well as those in the BUSSELTON2 and SCHOONER1 MCM3 and MCM5 proteins, all show weak but consistent similarity to inteins inserted at orthologous sites in either bacteria or archaea or to other *Anaeramoeba* inteins. Also consistent with an "ancient" acquisition of the singleton/pseudo-singleton inteins is the fact that most of them lack obvious HEs, unlike most other *Anaeramoeba* inteins in the network (Fig. 3).

Twenty-nine *Anaeramoeba* intein-bearing proteins have intein-containing homologs in prokaryotes (in most cases, their insertion sites are homologous). Many of these inteins present in cross-kingdom homologs cluster together in our analysis (Fig. 3). For example, cluster 18 contains *A. flamelloides* and prokaryotic inteins inserted at homologous sites in Ribonuclease HII. Cluster 13 contains a BMAN MCM8 intein inserted at "site e" of the archaeal MCM homolog, and cluster 11 includes the BMAN MCM6 intein (insertion site c) and inteins in the homologous position of archaeal MCM proteins. Interestingly, the *Anaeramoeba* cluster 13 and 11 inteins are short—160 and 185 amino acids, respectively—consistent with ancient insertion and subsequent HE domain decay (6). Where these inteins arose and how they came to reside in the same site of homologous proteins in organisms from different domains of life is unclear. While *Anaeramoeba* have bacterial epibionts (23), we found no evidence for symbiont-to-host intein transfer as the bacterial proteomes are intein-free (23) (secondary intein loss in the epibionts is a formal possibility). Finally, it is worth noting that *Anaeramoeba* inteins clustering with prokaryotic inteins do not always reside in homologous proteins. For example, the *Anaeramoeba* RtcB intein does not cluster with prokaryotic RtcB inteins but rather with inteins in prokaryotic phosphoenolpyruvate synthase (PEPS). Interestingly, the intein insertion sites of BMAN RtcB and prokaryotic PEPS are vaguely similar (RtcB: FGST/CHGE; PEPS: GG[TRM]T/CHAA), raising the possibility of microhomology-mediated nonallelic homing (see below).

Seven network clusters link *Anaeramoeba* inteins to those of other eukaryotes. Several of these associations obey the so-called allelic rule, i.e., inteins hosted by orthologous proteins at orthologous sites cluster together. For example, the BMAN Prp8 intein in "site d" clusters with the Prp8 site d intein in the green alga *Bathycoccus* (Fig. 3, cluster 4). Other clusters are mixtures of allelic and nonallelic intein associations and, in some cases, both eukaryotic and prokaryotic host proteins (clusters 2, 17, 7, 15, 6, and 16). Interestingly, all but one of these clusters also contain viral inteins.

Viruses Mediate Intein Spread. Eight clusters (2, 5, 6, 7, 9, 15, 16, and 17) contain *Anaeramoeba* inteins and inteins from the *Nucleocytoviricota*, a phylum of large dsDNA viruses that infect eukaryotes. Five of these eight also contain inteins from at least one other eukaryote (Fig. 3). Together with the fact that several intein-bearing host proteins are common to *Anaeramoeba* and viruses (Table 1), this suggests that viruses have played a role in intein acquisition in *Anaeramoeba* and other eukaryotes (12, 27), in the same way that phages have spread inteins in prokaryotes (28, 29). Some viral intein-containing clusters include nonallelic *Anaeramoeba* inteins. The most striking example is cluster 15, which contains inteins in 24 different BMAN proteins and 28 sites (four sites in Prp8, two in VMA) (Dataset S1), as well as in nine different viral proteins. Such intein promiscuity could be the result of viral HEs with very relaxed DNA recognition specificities, which could conceivably facilitate nonallelic intein spread (30) (*SI Appendix*, section 8 and Fig. S14). In the case of Anaeramoebae at least,

Downloaded from https://www.pnas.org by UPPSALA UNIVERSITETSBIBLIOTEK BMC-BIBLIOTEKET on January 2, 2024 from IP address 130.238.37.18.

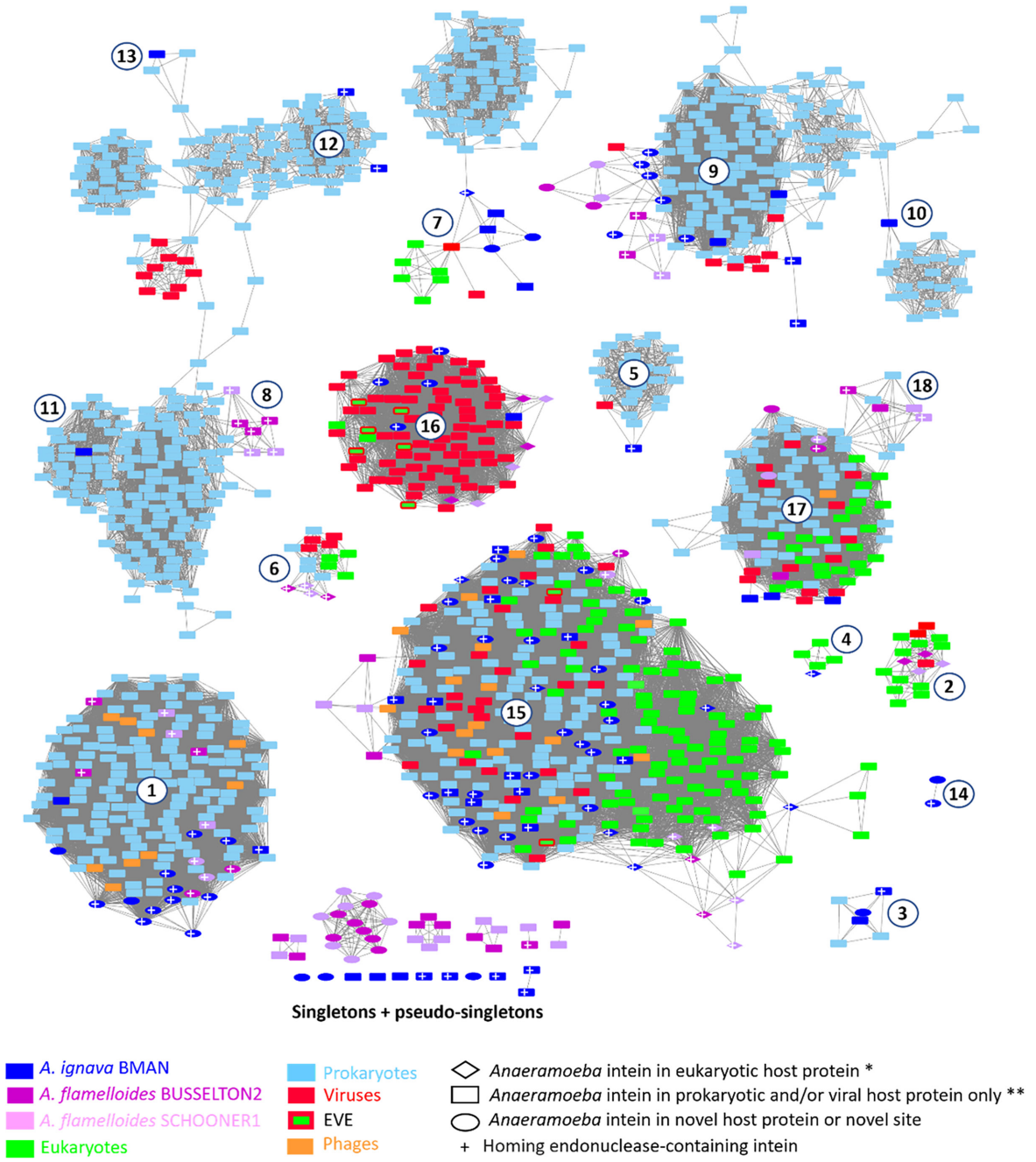

Fig. 3. *Anaeramoeba* inteins are evolutionarily diverse. Sequence Similarity Network (SSN) of inteins in *A. ignava* strain BMAN, *A. flamelloides* strains BUSSELTON2 and SCHOONER1, and related inteins found in public databases. The edges between the nodes are a measure of the similarity between nodes (score threshold corresponding to 30% identity on average). Anaeramoebae inteins belong to 18 different clusters corresponding to at least as many acquisition events. Some inteins do not show obvious sequence similarity with any other known inteins and are singleton or "pseudo-singleton" clusters (see main text). Inteins in viral-like sequences in eukaryotic genomes are designated EVEs (Endogenous Viral Elements). Note: "+": HE domain detected in *Anaeramoeba* intein using one or both methods (HEs in the inteins of other organisms are not highlighted); *In orthologous site; **In homologous site.

this could be exacerbated by the DNA repair machinery using nonhomologous templates for recombination, a possibility that deserves further investigation [cluster 15 proteins include many of those in the DNA replication and repair pathways and are likely repaired by nonhomologous end joining (NHEJ); see below].

Anaeramoeba inteins belonging to clusters 1 and 9 have atypical N- and C-terminal residues, respectively. Crucially, both features are also found in viruses (in phages for cluster 1 and *Nucleocytoviricota* for cluster 9), (*SI Appendix*, sections 4 and 5), once again suggesting that viruses contribute to the diversification of intein alleles. Another remarkable feature of *Anaeramoeba* cluster 1 inteins is that they appear to be "Class 3" inteins, which have thus far only been described in bacteria and phages (29, 31, 32). These inteins are also found in several unrelated protein hosts, which is consistent with nonallelic spread (*SI Appendix*, Fig. S2).

Virus-mediated lateral transfer of intein-containing host proteins might also explain the origins and spread of inteins in eukaryotes. We detected previously undescribed inteins in *Klebsormidium nitens* (Viridiplantae), *Tetrabaena socialis* (Viridiplantae), a multicellular alga (33), and *Honda fermentalgiana* (Labyrinthulomycetes), some of which appeared to reside in DNApol proteins; this has not been reported previously in eukaryotes. Detailed investigation revealed that the intein host proteins in *K. nitens*, *H. fermentalgiana,* and *T. socialis* are in fact viral-like DNA polymerases whose genes reside

Downloaded from https://www.pnas.org by UPPSALA UNIVERSITETSBIBLIOTEK BMC-BIBLIOTEKET on January 2, 2024 from IP address 130.238.37.18.

on bona fide nuclear scaffolds (*SI Appendix*, Fig. S3). The presence of viral genes in eukaryotic nuclear chromosomes has been documented previously (34–36), including in *K. nitens* (37) and *T. socialis* (38), where they have acquired spliceosomal introns (37, 38). While LGT has played a role in the evolution of *Anaeramoeba* genomes (23), none of the intein-hosting proteins in the three strains analyzed herein are of obvious viral ancestry, even though some of the inteins themselves appear to be. Regardless, LGT of intein-bearing host proteins from viruses is a potential avenue for allelic and nonallelic homing into eukaryotic genes.

Intragenomic Mobility of *Anaeramoeba* Inteins. The BUSSELTON2 and SCHOONER1 strains of *A. flamelloides* share recent common ancestry (23); it is thus not surprising that their inteins share high percentages of amino acid identity, reside in the same sites in orthologous proteins, and cluster together in our network (Fig. 3). By contrast, the intein host protein and insertion site profiles of *A. flamelloides* and *A. ignava* are very different (Fig. 2 and Dataset S1). Only eight orthologous insertion sites are shared between the two species [one in MCM8, one in DNA-directed RNA polymerase 1 subunit 1 (RPA1), one in DNA-directed RNA polymerase 3 subunit 2 (RPC2), two in Prp8, one in a Kinesin, one in a chromo domain-containing protein (SNF2), and one in the chromatin-remodeling protein HARP/SMARCAL1], and only five of 18 network clusters contain inteins from both species. This shows dramatic intein evolution since the two species diverged; it is unclear whether this is due primarily to differential intein acquisition or loss.

We analyzed the *A. ignava* and *A. flamelloides* inteins belonging to the same network clusters (clusters 1, 9, 15, 16, and 17) to determine whether they were acquired independently in the two species or inherited vertically from a common ancestor. Phylogenetic analysis shows that in the case of RPC2 (cluster 17), HE-lacking inteins found at orthologous sites in *A. ignava* and both strains of *A. flamelloides* branch together, consistent with vertical inheritance (Fig. 4*A*). The evolutionary histories of most other inteins shared between *A. ignava* and *A. flamelloides* are more complicated. Remarkably, this includes instances of intein spread to different locations within the same protein, spread among paralogous genes/proteins, and even between nonhomologous proteins. Such instances of intragenomic mobility are also apparent for inteins exclusive to each species.

Phylogenetic analysis of cluster 15 inteins revealed a potential intragenic transfer. Inteins present at sites b and d in the Prp8 of *A. flamelloides* strains branch together (Fig. 4*B*), which is surprising given that their amino acid motifs are different (site b = VMDL/CQVF; site d = EAFR/CWKA) and that sites b and d are both known to harbor inteins in other organisms. Intein homing between ancient paralogs in *Anaeramoeba* was also observed (here ancient refers to: i) paralogs that duplicated prior to the divergence of *A. ignava* and *A. flamelloides* but after *Anaeramoeba* diverged from other eukaryotes, ii) paralogs shared with a subset of other eukaryotes, as well as iii) truly ancient paralogs common to all/most extant eukaryotes). For example, multiple MCM paralogs in *A. ignava* and *A. flamelloides* have homologous inteins at paralogous sites, although they are different paralogs in each species (Fig. 5 *C* and *D*). A similar situation exists in DNApol (Fig. 5*B*) and examples of differential intein insertion were also found for the first and second subunits of RNApols I, II, and III (*SI Appendix*, Fig. S6). Several *A. ignava* Msh paralogs also have homologous inteins (Fig. 5*C* and *SI Appendix*, Fig. S7). We analyzed the intein similarity network (Fig. 3) and built phylogenies to determine whether these inteins were acquired independently from exogenous sources and/or if interparalog transfers occurred. Evidence for both scenarios was found (*SI Appendix*, section 6).

The fact that some ancient paralogs in *Anaeramoeba* have inteins while others do not led us to examine the DNA corresponding to

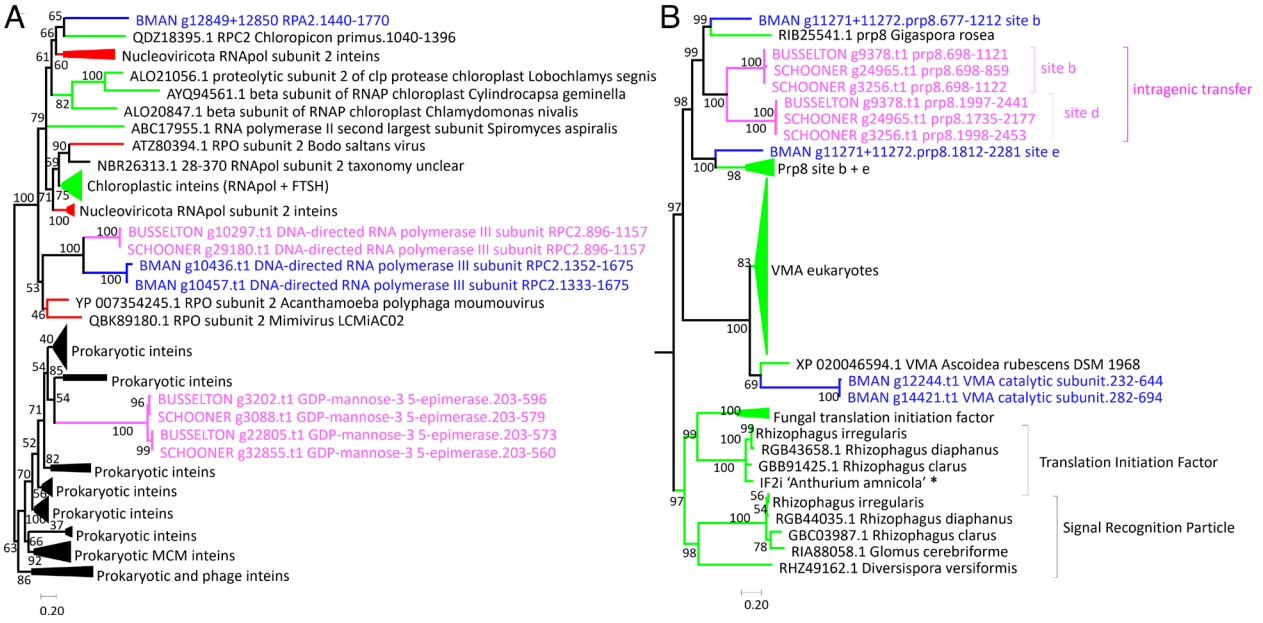

Fig. 4. Modalities of intein spread in *Anaeramoeba* are diverse. (*A*) Phylogenetic tree of cluster 17 inteins inferred under the C20+G4 mixture model and 1,000 ultrafast bootstraps from 191 sites. *A. ignava* BMAN inteins are in blue, *A. flamelloides* inteins are in pink, viral inteins are in red, and eukaryotic inteins are in green. Viral inteins are interspersed among inteins from diverse eukaryotes including *Anaeramoeba,* suggesting that viruses are involved in the spread of inteins between eukaryotes. The expanded tree (*SI Appendix*, Fig. S4) shows that viruses could also be involved in the transfer of inteins between nuclear and chloroplast homologs (*SI Appendix*, section 9). (*B*) Phylogenetic tree of cluster 15 inteins inferred under the C20+G4 mixture model and 1,000 ultrafast bootstraps. The sequences were aligned with mafft-linsi and 300 sites with <20% gaps were selected. An expanded tree is presented in *SI Appendix*, Fig. S5. Color-coding is the same as in part *A*. A potential intragenic intein transfer between site b and site d in *A. flamelloides* Prp8 is marked in pink. An intragenomic intein transfer appears to have taken place between the Signal Recognition Particle Receptor (SRP) and Translation initiation factor IF-2P (TIF) in a *Rhizophagus* ancestor. *The source of the intein labeled "*Anthurium amnicola*" is unclear.

Downloaded from https://www.pnas.org by UPPSALA UNIVERSITETSBIBLIOTEK BMC-BIBLIOTEKET on January 2, 2024 from IP address 130.238.37.18.

the intein insertion sites. We found that in several instances, when inteins are absent, spliceosomal introns are found in the DNA near or at the intein insertion sites in another paralog (Fig. 5 and *SI Appendix*, Figs. S6, S8, and S9). Intron acquisition might thus have served to prevent intein homing by disrupting the HE recognition sites. This has been proposed in the case of Prp8 inteins in *Eupenicillium* species (39), and intron-intein mutual exclusivity was described in Prp8 in strains of *Bathycoccus* (27). The case of DNApol paralogs in *Anaeramoeba* is particularly intriguing by its symmetry. Whereas *A. ignava* DNApol alpha has an intein in the same place as *A. flamelloides* DNApol delta does, the *A. ignava* DNApol delta has an intron precisely where the *A. flamelloides* DNApol alpha has one, three nucleotides after the intein target site (Fig. 5*B*) (The surprising absence of a DNApol zeta paralog in *Anaeramoeba* has important functional implications and is discussed in *SI Appendix*, section 7). This pattern could be due to parallel intron gains, illegitimate recombination events, or intein "decay," as proposed in *Bathycoccus* (27). Intron dynamics aside, whether or not these examples of interparalog intein spread can be said to respect the "allelic rule" (6) is debatable. In any case, interparalog intein transfer has to our knowledge not been formally described before. Even more striking is our identification of intein homing to the phosphate-binding P-loops (Walker A motifs) of nonhomologous proteins (Fig. 6*B*) and that of homologous inteins in the common BRR2 helicase domains of two otherwise unrelated proteins (Fig. 6*A* and *SI Appendix*, section 8). In *Anaeramoeba*, inteins appear to be capable of spreading among proteins that share very little amino acid sequence identity.

Discussion

The spread of inteins between and within genomes is a long-standing puzzle in evolutionary biology. Often considered parasitic genetic elements (6), inteins are patchily distributed across the tree of life; when present, they reside in a restricted set of host proteins and only one or a few inteins are typically present in the genome of a given organism [note that some prokaryotic genomes have more than a dozen (13)]. We have shown that members of the newly described protist phylum Anaeramoebae harbor an exceptional abundance of intein-encoding regions in their nuclear genomes (103 in *A. ignava*), many of which appear to bend or break the "rules" of intein biology. Numerous novel intein host proteins were identified, including those in functional categories not known to harbor inteins in other organisms (e.g., cytoskeletal proteins and those involved in chromatin structure/dynamics). Taken as a whole, the intein content profiles of Anaeramoebae share similarities with both prokaryotes and eukaryotes, suggesting a role for LGT in their evolution and blurring the distinction between allelic and nonallelic homing.

As applied to inteins, the classical "homing cycle" of a parasitic element (41) begins with the acquisition of a HE-containing intein that spreads through a population by homing until no intein-lacking alleles remain. The HE then becomes inactive and disappears due to reduced selective pressure. The HE-free intein is lost and the empty site becomes a target for reinvasion. While appealing in its simplicity, this model ignores the obvious difficulty of intein loss, which must be very precise if the host protein is to retain functionality (3, 6), and it does not specify how HE-bearing inteins are acquired in the first place (14, 18, 42). Our results support and greatly extend those of previous studies (12, 43, 44) in showing that LGT—within and between the three domains of life—can be an important source of inteins in eukaryotic nuclear genomes. In the case of *Anaeramoeba*, inteins appear to have come directly or indirectly from diverse prokaryotic sources. These anaerobic protists live in close association with diverse prokaryotes and a parallel study of the *Anaeramoeba* nuclear genome provides robust evidence for prokaryote-to-eukaryote LGT (23). Viruses are also clearly part of the *Anaeramoeba* intein picture. Some *Nucleocytoviricota* are thought to have a nuclear stage and thus could conceivably deliver intein donor alleles straight to the nucleus. The HE itself could also be encapsidated in a viral

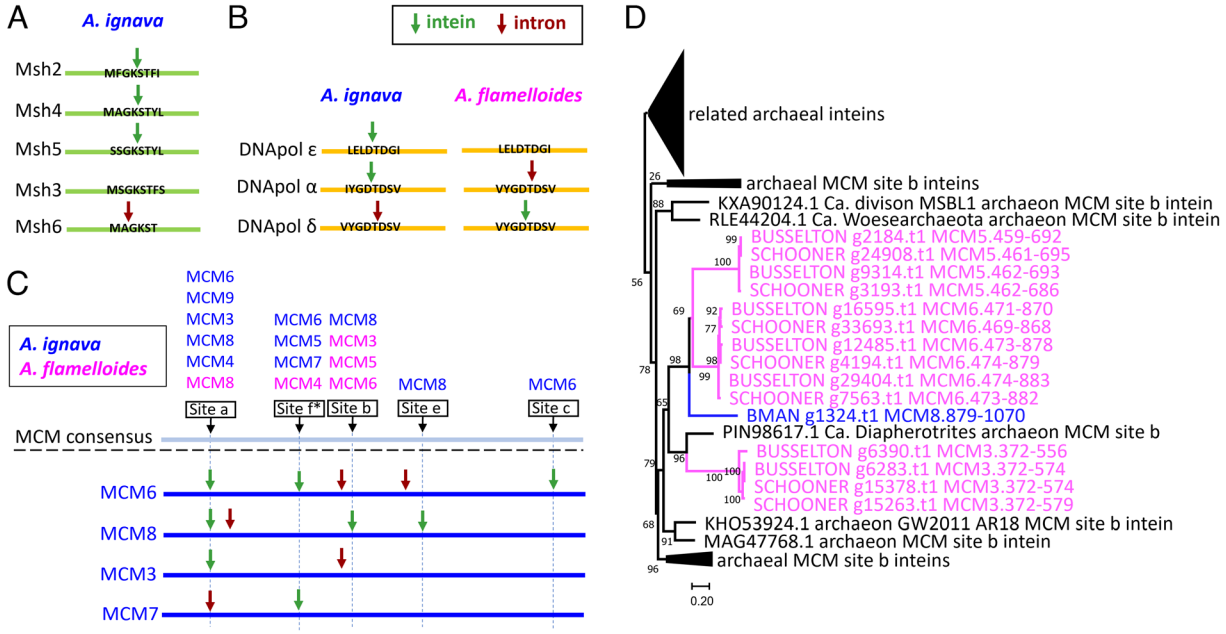

Fig. 5. Location of inteins and introns in paralogous proteins in *Anaeramoeba*. (*A*) Msh family proteins (Msh2, 3, 4, 5, and 6). (*B*) Family B-DNA polymerases (DNApol epsilon, alpha, and delta). (*C*) MCM family proteins (MCM3, 4, 5, 6, 7, 8, and 9; note that in the lower half of the panel, only MCMs with introns near intein insertion sites are shown; all such introns are in *A. ignava*). *Novel site described in this study. (*D*) Phylogenetic reconstruction of MCM site b inteins inferred under the C20+G4 mixture model and 1,000 ultrafast bootstraps. Sequences were aligned with mafft-linsi, and 145 sites with <10% gaps were selected. *A. ignava* BMAN inteins are in blue, *A. flamelloides* inteins for both strains are in pink.

Downloaded from https://www.pnas.org by UPPSALA UNIVERSITETSBIBLIOTEK BMC-BIBLIOTEKET on January 2, 2024 from IP address 130.238.37.18.

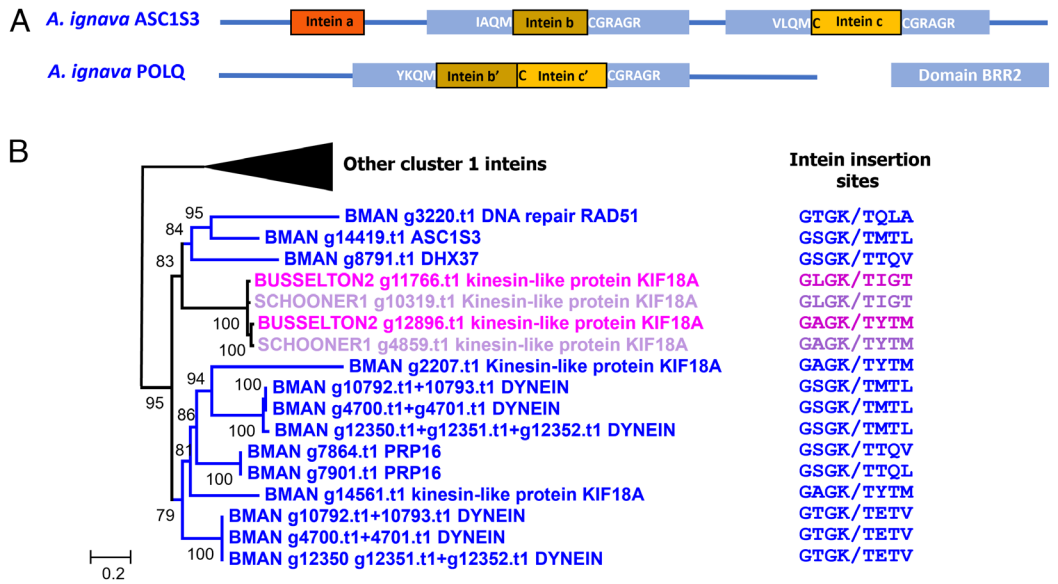

Fig. 6. Intragenomic intein transfers have occurred in *Anaeramoeba*. (*A*) Inteins c and c' form their own network cluster (cluster 14) and are inserted at the same site in the BRB2 helicase domain of activating signal cointegrator 1 complex subunit 3 (ASC1S3). Inteins b and b' are singletons but are best matches to each other, suggesting that they have a common origin. They are also inserted at the same site in the BRB2 domain. DNApol theta/eta (POLQ) inteins are next to each other in the BRB2 domain, while the ASC1S3 inteins are each present in one of the two BRB2 domains encoded by ASC1S3. (*B*) All *Anaeramoeba* cluster 1 inteins are inserted in P-loop motifs (shown at right) and branch together in the phylogenetic tree. The ML phylogeny was inferred under the C20+G mixture model with 1,000 ultrafast bootstrap replicates. The sequences were aligned with MUSCLE (40) (default parameters) and 115 amino acid sites manually selected. These data suggest intein transfers between P-loops in unrelated proteins.

particle and delivered at the same time as intein donor DNA. The observation that several viral proteins with intein domains were detected in the CroV virion proteome of the protist *Cafeteria roenbergensis* (45) is in line with this possibility.

Most unexpected is the level of intein dynamism seen at the species level within *Anaeramoeba*. The intein distribution profiles of the three strains analyzed herein are exceedingly complex. Whereas some inteins are best explained as having been present in the common ancestor of *A. ignava* and *A. flamelloides*, many others are likely the product of more recent LGT. Having landed in the *Anaeramoeba* nuclear genome, many of these inteins appear to have spread to different locations within the same protein and to different proteins—both homologous and non-homologous—with recent intron gain acting as a barrier to both allelic and nonallelic homing. At present we can only speculate as to how and why this has occurred. But it is noteworthy that unlike other Metamonada (46), *Anaeramoeba* spp. appear capable of NHEJ and have an unusual suite of DNA repair enzymes, many of which contain inteins (*SI Appendix*, section 10 and Fig. S15). It may be that at least some of the HEs in Anaeramoebae are highly promiscuous, and/or that the protein machineries underlying homology-directed gene conversion (4, 5) and/or NHEJ are unusually conducive to illegitimate intein homing. Regardless of the underlying reason(s), it remains to be seen whether Anaeramoebae are genuine outliers or represent the tip of an intein diversity iceberg. In any case, the promiscuous inteins described herein are of considerable practical use to protein biochemists working in the genetic engineering and biotechnology sectors (e.g., refs. 47–49).

Materials and Methods

Intein Searches. Two methods were used to detect inteins in the three *Anaeramoeba* predicted proteomes. First, BLAST (50) v2.9.0 was used (e-value <= 1e−5) with the inteins published in InBase and those described by Green et al. (12) as queries. Second, searches with Hidden Markov Models (HMMs) were performed using Hmmer (51) v. 3.3 using HMM profiles built using the N-terminal and C-terminal splicing domains of reference inteins (smart 00305, smart00306, PS50817, PS50818, PF14623, PF13403, and cd00081). Inteins detected during this first round of searching were used as queries for additional BLASTp searches. tBLASTn was used to detect potential pseudogenized inteins using the same queries as previously as well as the newly extracted complete set of *Anaeramoeba* full-length inteins.

Intein Splicing. The splicing of inteins in *Anaeramoeba* was investigated experimentally using a combination of shotgun proteomics (*A. flamelloides* BUSSELTON2) and GFP-based in vitro expression analyses in *E. coli* (*A. ignava* BMAN). Detailed methodologies can be found in *SI Appendix*, section 2. Proteomic data can be found in the PRoteomics IDEntifications (PRIDE) database (awaiting final access links).

Phylogenetic Analysis. Unless specified otherwise, intein sequences were aligned with mafft (52) v7.407 using default parameters, sites were selected with bmge (53) v1.12 with the -m BLOSUM30 parameter, and IQ-TREE (54) v1.6.3 was used to reconstruct phylogenies with 1,000 ultrafast bootstrap replicates (55) under the C20+G4 mixture model.

Sequence Similarity Network. *Anaeramoeba* inteins were used as queries to retrieve homologous inteins from the *nr* database (2020). For each intein, the 25 best matching inteins were selected. The EFI EST software (56) was used to construct the network from the splicing domains of retrieved inteins using default parameters. For filtration into clusters, a threshold score corresponding to an average of 30% identity was used, which is an approximate threshold by which allelic and nonallelic inteins can be discriminated (9). The resulting network was visualized and edited with Cytoscape (57).

HE Domain Detection. Two methods were used to detect HE domains, Interproscan v5.32-71.0 and HMM searches. In the latter case, HMM profiles constructed from alignments of HEs in the PFAM (PF00961, PF03161, PF05204, PF09062, PF10298, PF14527, PF14528, PF01223, PF01844, PF02945, PF03165, PF04231, PF05315, PF05551, PF05766, PF06147, PF07510, PF09230, PF12639, PF13391, PF13392, PF13395, PF13930, PF14279, PF14410, PF14411, PF14412, PF14414, PF16784, PF16786, and PF18275) and Prosite (PS50164) databases were used as queries.

Downloaded from https://www.pnas.org by UPPSALA UNIVERSITETSBIBLIOTEK BMC-BIBLIOTEKET on January 2, 2024 from IP address 130.238.37.18.

DNA Replication and Repair Systems. One hundred thirty-one human and yeast proteins from the replisome and NHEJ, HR, BER, NER, and MMR DNA repair pathways were used to identify their best PANTHER v15 family HMM model using PANTHER Score tool (58). These HMM models were then used to identify homologs in the predicted proteomes of the three *Anaeramoeba* genomes as well as those of *T. vaginalis*, *M. exilis*, *C. frisia*, *C. membranifera*, *K. bialata*, *G. intestinalis* WB, *S. salmonicida*, *N. gruberi*, and *D. discoideum*. Query proteins and retrieved homologs were then aligned with a random taxonomic sample from the same PANTHER family of interest for ortholog validation. Alignments were trimmed and subjected to phylogenetic reconstructions similarly to those described in ref. 46.

Data, Materials, and Software Availability. Protein sequence alignments of intein sequences have been deposited in DRYAD (DOI: 10.5061/dryad.q2bvq83rg). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD045607. Previously published data were used for this work (23). All other data are included in the manuscript and/or supporting information.

ACKNOWLEDGMENTS. We thank B. Curtis for bioinformatic assistance throughout the project and P.X.-Q. Liu and two reviewers for comments on an earlier version of the manuscript. We also thank the Proteomics Core Facility, Sahlgrenska Academy, Gothenburg University, with financial support from SciLifeLab and BioMS. This work was supported by Natural Sciences and Engineering Research Council of Canada grants RGPIN-2019-05058 awarded to J.M.A. and RGPIN-2022-05430 awarded to A.J.R., as well as a Czech Academy of Sciences grant 21-30563S awarded to I.Č. C.W.S. is supported by the Swedish Research Council (Vetenskapsrådet starting grant 2020-05071).

Author affiliations: [a]Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia B3H 4R2, Canada; [b]Institute for Comparative Genomics, Dalhousie University, Halifax, Nova Scotia B3H 4R2, Canada; [c]Microbiology and Immunology, Department of Cell and Molecular Biology, Biomedical Centre, Uppsala University, Uppsala 751 24, Sweden; [d]Microbiology Group, Department of Biology, Lund University, Lund 223 62, Sweden; and [e]Department of Zoology, Charles University, Prague 128 00, Czech Republic

1. C. W. Lennon, M. Belfort, Inteins. *Curr. Biol.* **27**, R204–R206 (2017).
2. F. B. Perler *et al.*, Protein splicing elements: Inteins and exteins–a definition of terms and recommended nomenclature. *Nucleic Acids Res.* **22**, 1125–1127 (1994).
3. X. Q. Liu, Protein-splicing intein: Genetic mobility, origin, and evolution. *Annu. Rev. Genet.* **34**, 61–76 (2000).
4. M. Belfort, M. E. Reaban, T. Coetzee, J. Z. Dalgaard, Prokaryotic introns and inteins: A panoply of form and function. *J. Bacteriol.* **177**, 3897–3903 (1995).
5. F. S. Gimble, J. Thorner, Homing of a DNA endonuclease gene by meiotic gene conversion in *Saccharomyces cerevisiae*. *Nature* **357**, 301–306 (1992).
6. J. P. Gogarten, E. Hilario, Inteins, introns, and homing endonucleases: Recent revelations about the life cycle of parasitic genetic elements. *BMC Evol. Biol.* **6**, 94 (2006).
7. S. M. Soucy, M. S. Fullmer, R. T. Papke, J. P. Gogarten, Inteins as indicators of gene flow in the halobacteria. *Front. Microbiol.* **5**, 299 (2014).
8. K. S. Swithers, A. G. Senejani, G. P. Fournier, J. P. Gogarten, Conservation of intron and intein insertion sites: Implications for life histories of parasitic genetic elements. *BMC Evol. Biol.* **9**, 303 (2009).
9. F. B. Perler, G. J. Olsen, E. Adam, Compilation and analysis of intein sequences. *Nucleic Acids Res.* **25**, 1087–1093 (1997).
10. R. Hirata *et al.*, Molecular structure of a gene, VMA1, encoding the catalytic subunit of H(+)-translocating adenosine triphosphatase from vacuolar membranes of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **265**, 6726–6733 (1990).
11. P. M. Kane *et al.*, Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase. *Science* **250**, 651–657 (1990).
12. C. M. Green, O. Novikova, M. Belfort, The dynamic intein landscape of eukaryotes. *Mob. DNA* **9**, 4 (2018).
13. O. Novikova *et al.*, Intein clustering suggests functional importance in different domains of life. *Mol. Biol. Evol.* **33**, 783–799 (2016).
14. O. Novikova, N. Topilina, M. Belfort, Enigmatic distribution, evolution, and function of inteins. *J. Biol. Chem.* **289**, 14490–14497 (2014).
15. C. Bernstein, Deoxyribonucleic acid repair in bacteriophage. *Microbiol. Rev.* **45**, 72–98 (1981).
16. A. Lopes, J. Amarir-Bouhram, G. Faure, M.-A. Petit, R. Guerois, Detection of novel recombinases in bacteriophage genomes unveils Rad52, Rad51 and Gp2.5 remote homologs. *Nucleic Acids Res.* **38**, 3952–3962 (2010).
17. C. Weigel, H. Seitz, Bacteriophage replication modules. *FEMS Microbiol. Rev.* **30**, 321–381 (2006).
18. J. P. Gogarten, A. G. Senejani, O. Zhaxybayeva, L. Olendzenski, E. Hilario, Inteins: Structure, function, and evolution. *Annu. Rev. Microbiol.* **56**, 263–287 (2002).
19. T. L. Pavankumar, Inteins: Localized distribution, gene regulation, and protein engineering for biological applications. *Microorganisms* **6**, 19 (2018).
20. C. W. Lennon, M. Stanger, N. K. Banavali, M. Belfort, Conditional protein splicing switch in hyperthermophiles through an intein-extein partnership. *mBio* **9**, e02304–17 (2018).
21. C. W. Stairs *et al.*, Anaeramoebae are a divergent lineage of eukaryotes that shed light on the transition from anaerobic mitochondria to hydrogenosomes. *Curr. Biol.* **31**, 5605–5612.e5 (2021).
22. P. Táborský, T. Pánek, I. Čepička, Anaeramoebidae fam. nov., a novel lineage of anaerobic amoebae and amoeboflagellates of uncertain phylogenetic position. *Protist* **168**, 495–526 (2017).
23. J. Jerlström-Hultqvist *et al.*, A novel symbiosome in an anaerobic single-celled eukaryote. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.03.03.530753 (Accessed 5 March 2023).
24. F. B. Perler, InBase: The intein database. *Nucleic Acids Res.* **30**, 383–384 (2002).
25. C. M. Green *et al.*, Spliceosomal Prp8 intein at the crossroads of protein and RNA splicing. *PLoS Biol.* **17**, e3000104 (2019).
26. X.-Q. Liu, J. Yang, Q. Meng, Four inteins and three group II introns encoded in a bacterial ribonucleotide reductase gene. *J. Biol. Chem.* **278**, 46826–46831 (2003).
27. A. Monier, S. Sudek, N. M. Fast, A. Z. Worden, Gene invasion in distant eukaryotic lineages: Discovery of mutually exclusive genetic elements reveals marine biodiversity. *ISME J.* **7**, 1764–1774 (2013).
28. D. S. Kelley, C. W. Lennon, M. Sea-Phages, O. Novikova. Belfort, Mycobacteriophages as incubators for intein dissemination and evolution. *mBio* **7**, e01537–16 (2016).
29. K. Tori, F. B. Perler, Expanding the definition of class 3 inteins and their proposed phage origin▽. *J. Bacteriol.* **193**, 2035–2041 (2011).
30. L. Gallot-Lavallee, G. Blanc, J.-M. Claverie, Comparative genomics of *Chrysochromulina Ericina* Virus (CeV) and other microalgae-infecting large DNA viruses highlight their intricate evolutionary relationship with the established Mimiviridae family. *J. Virol.* **91**, e00230–17 (2017), 10.1128/JVI.00230-17.
31. M. W. Southworth, J. Benner, F. B. Perler, An alternative protein splicing mechanism for inteins lacking an N-terminal nucleophile. *EMBO J.* **19**, 5019–5026 (2000).
32. K. Tori *et al.*, Splicing of the mycobacteriophage Bethlehem DnaB intein: Identification of a new mechanistic class of inteins that contain an obligate block F nucleophile. *J. Biol. Chem.* **285**, 2515–2526 (2010).
33. J. Featherston *et al.*, The 4-Celled *Tetrabaena socialis* nuclear genome reveals the essential components for genetic control of cell number at the origin of multicellularity in the volvocine lineage. *Mol. Biol. Evol.* **35**, 855–870 (2018).
34. J. Filée, Multiple occurrences of giant virus core genes acquired by eukaryotic genomes: The visible part of the iceberg? *Virology* **466–467**, 53–59 (2014).
35. F. Maumus, G. Blanc, Study of gene trafficking between acanthamoeba and giant viruses suggests an undiscovered family of amoeba-infecting viruses. *Genome Biol. Evol.* **8**, 3351–3363 (2016).
36. F. Maumus, A. Epert, F. Nogué, G. Blanc, Plant genomes enclose footprints of past infections by giant virus relatives. *Nat. Commun.* **5**, 4268 (2014).
37. L. Gallot-Lavallée, G. Blanc, A glimpse of nucleo-cytoplasmic large DNA virus biodiversity through the eukaryotic genomics window. *Viruses* **9**, 17 (2017).
38. M. Moniruzzaman, A. R. Weinheimer, C. A. Martinez-Gutierrez, F. O. Aylward, Widespread endogenization of giant viruses shapes genomes of green algae. *Nature* **588**, 141–145 (2020).
39. S. Elleuche, C. Pelikan, N. Nolting, S. Pöggeler, Inteins and introns within the prp8 -gene of four Eupenicillium species. *J. Basic Microbiol.* **49**, 52–57 (2009).
40. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
41. M. R. Goddard, A. Burt, Recurrent invasion and extinction of a selfish gene. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 13880–13885 (1999).
42. M. Belfort, R. J. Roberts, Homing endonucleases: Keeping the house in order. *Nucleic Acids Res.* **25**, 3379–3388 (1997).
43. X. Q. Liu, Z. Hu, A DnaB intein in *Rhodothermus marinus*: Indication of recent intein homing across remotely related organisms. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 7851–7856 (1997).
44. K. S. Swithers, S. M. Soucy, E. Lasek-Nesselquist, P. Lapierre, J. P. Gogarten, Distribution and evolution of the mobile vma-1b intein. *Mol. Biol. Evol.* **30**, 2676–2687 (2013).
45. M. G. Fischer, I. Kelly, L. J. Foster, C. A. Suttle, The virion of Cafeteria roenbergensis virus (CroV) contains a complex suite of proteins for transcription and DNA repair. *Virology* **466–467**, 82–94 (2014).
46. D. E. Salas-Leiva *et al.*, Genomic analysis finds no evidence of canonical eukaryotic DNA processing complexes in a free-living protist. *Nat. Commun.* **12**, 6003 (2021).
47. P. Tornabene *et al.*, Intein-mediated protein trans-splicing expands adeno-associated virus transfer capacity in the retina. *Sci. Transl. Med.* **11**, eaav4523 (2019).
48. R. López-Igual, J. Bernal-Bayard, A. Rodríguez-Patón, J.-M. Ghigo, D. Mazel, Engineered toxin–intein antimicrobials can selectively target and kill antibiotic-resistant bacteria in mixed populations. *Nat. Biotechnol.* **37**, 755–760 (2019).
49. C. Sarmiento, J. A. Camarero, Biotechnological applications of protein splicing. *Curr. Protein Pept. Sci.* **20**, 408–424 (2019).
50. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
51. S. R. Eddy, Multiple alignment using hidden Markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3**, 114–120 (1995).
52. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
53. A. Criscuolo, S. Gribaldo, BMGE (Block Mapping and Gathering with Entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
54. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
55. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFboot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
56. R. Zallot, N. Oberg, J. A. Gerlt, The EFI web resource for genomic enzymology tools: Leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. *Biochemistry* **58**, 4169–4182 (2019).
57. P. Shannon *et al.*, Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
58. H. Mi *et al.*, Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat. Protoc.* **14**, 703–721 (2019).

Downloaded from https://www.pnas.org by UPPSALA UNIVERSITETSBIBLIOTEK BMC-BIBLIOTEKET on January 2, 2024 from IP address 130.238.37.18.