Title: Actinobacteriophage Inteins: Host Diversity, Local Dissemination, and Non-Canonical Architecture

Authors: Sophia P. Gosselin[1], Danielle Arsenault[1], Johann Peter Gogarten[1,2*]

1: Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06268-3125, USA

2: Institute for Systems Genomics, University of Connecticut, Storrs, CT 06268-3125, USA

*: Corresponding author: Johann Peter Gogarten, Department of Molecular and Cell Biology, University of Connecticut, 91 North Eagleville Rd, Storrs, CT 06268-3125, USA; Phone 1 860 465 6267;
E-mail: gogarten@uconn.edu.

Classification: Biological, Health, and Medical Sciences/Microbiology
Biological, Health, and Medical Sciences/Evolution

Keywords: Intein, homing endonuclease, gene drive, Actinobacteriophages, divergent homologs

Abstract

Intein presence within Actinobacteriophages (within PhagesDB) was last surveyed in 2016, and despite a 5-fold increase in the size of the database, has not been updated since. To address this, we present a modern survey of the current iteration of the PhagesDB database. We developed a new algorithm — Iterative Cluster Expansion BLAST (ICE-BLAST) — to expand our search to more divergent sequences. Nearly 800 inteins were retrieved through this process; the majority of which were previously unreported. We describe the nature of these inteins, their classes, integration target sites, distribution within phage clusters, and explore the geographical location of nearly identical intein sequences found in divergent exteins. Our findings suggest that these inteins recently invaded local phage populations. We also find two instances of a Cas4 exonuclease intein evolving from a terminase intein, and propose a model by which one of these inteins was able to utilize sequence similarity conferred by a shared nucleotide binding site to jump between genes. Additionally, we find inteins with never-before reported homing endonucleases, and inteins with homing endonucleases encoded in a reading frame separate from that which encodes the extein and the intein's self-splicing domains. We provide predicted structures for these elements and hypothesize on their evolution and relation to free-standing homing endonucleases within phage genomes. Finally, we provide evidence that these "non-canonical" inteins are still transferring between host genomes, in a fashion similar to other inteins with canonical homing endonucleases within the dataset.

Introduction

Inteins are self-splicing and site-specific mobile genetic elements, which invade protein encoding genes (1–3). The first intein was sequenced in 1987; however, due to sequence availability in databases at the time, the intein was not recognized as an insertion sequence, and the host protein was not identified as a vacuolar ATPase subunit (4). In 1990 this intein was identified as an insertion sequence which persisted at the mRNA level, and was only spliced out following translation (5, 6). In the following three and a half decades, significant insight has been gained into the biochemistry of these elements, their life cycle, and their evolution. Inteins are present throughout all three domains of life, and are found more often within bacterial, archaeal, and viral genomes relative to those of most eukaryotes (1, 7). Within eukaryotes, inteins are found more frequently in fungi (8), and amoebae (9).

An intein begins its journey as an intervening DNA sequence within the coding sequence of a host gene — the functional host gene product is referred to as the extein. The intein's coding sequence, being in-frame with the host gene, is transcribed and translated together with this host gene. Once translated, the intein will begin the process of removing itself from the host protein via the use of its two splicing domains (1, 2). One of these splicing domains sits at the N terminus of the intein sequence and the other at the C terminus. These two splicing domains come together at the protein level and auto-catalytically remove the intein from the host protein while ligating the host protein back together (10, 11). The specific mechanism for this splicing has been divided into one of three classes depending on the initial steps of splicing, the residues involved, and differences in reaction chemistry (1). Class 1 inteins are the most common among the three classes, but all three end with the same final cyclization and hydrolysis steps. Most of the time, these two splicing domains are divided by a third domain the homing endonuclease (HEN) domain.

The HEN domain allows the intein to transfer between host genomes, and self-propagate. If the intein has a functional HEN domain, it can initiate propagation of the intein allele into an un-invaded host allele (if one exists) by homing to a target site, making a double stranded cut, and relying on host DNA repair machinery to repair the now broken un-invaded allele by using the intein containing allele as a template (2, 12). Intein alleles can thus propagate via any method of horizontal gene transfer (or zygote formation (13)), so long as a host with an un-invaded allele uptakes DNA with the intein coding sequence. Not all inteins contain a HEN domain, and the splicing domains will function without one — inteins lacking said HEN domain are referred to as mini-inteins (14). Hypothesized to be the product of purifying selection that led to the decay of the HEN domain, these mini-inteins, if left to continue decaying without recombining with an intein-less allele, can lead to the creation of split inteins. Split inteins are inteins which acquire a stop codon within their HEN domain, leading to the host gene being split into two separate open reading frames (ORFs). Crucially each ORF contains one half of the host protein attached to a functioning splicing domain. Since the splicing domains still function, the ORF protein products can be made into a functional protein via the trans splicing and ligating activity of the two intein splicing domains attached to the partial extein proteins (15–17).

Several different HEN domains have been reported in inteins. Most are "canonical" LAGLIDADG homing endonuclease, which are contiguous with the rest of the intein and present across all domains of life (18, 19). Others include: inteins with an HNH endonuclease (20), split-inteins within phages

possessing VSR-like endonucleases (a variant of the PD-(D/E)xK endonuclease) (15), and two LAGLIDADG HENs which were reported to be encoded on the anti-sense strand instead (2, 19). Interestingly, most of these “non-canonical” inteins are found within phage genomes, with mycobacteriophages serving as a particular hot-spot in research.

Mycobacteriophages, and more generally, actinobacteriophages have become an exciting topic of research within the medical field as potential therapeutics to tackle multi-drug resistant tuberculosis, and other *Mycobacteria* based diseases (21–23). In 2003, there were only 14 mycobacteriophage genomes available for study, greatly limiting the potential avenues of research (24). This number rapidly expanded over the following decades as several teaching programs (SEA-PHAGES and PHIRE) established citizen-science based approaches to isolate and sequence phages (25). These actinobacteriophage genomes are now stored in PhagesDB (25). This database has grown tremendously in unison with the growth of the SEA-PHAGES program; exploding to, at the time of writing, 5075 fully sequenced phage genomes available for download (based on their “Live Phage Stats” sidebar).

It has been known for over a decade that inteins are present within actinobacteriophages (26–28). However, since the last survey of these inteins (26) the PhagesDB database has expanded from 841 to 5075 genomes; providing a unique resource to study distribution and evolution of inteins. Furthermore, the previous work’s queries were limited to intein sequences known at the time, and we have since discovered additional phage inteins not seen in the previous study. Hence, we have conducted a new survey of PhagesDB as it was in the Fall of 2023. To capture more divergent inteins within the database, we developed an algorithm, ICE-BLAST, which uses a combination of PSI-BLAST and USEARCH clustering to iteratively search an amino acid database using the centroid sequences of previous searches. Herein we discuss the development and testing of this algorithm, and the inteins we discovered through it. We investigate their distribution within phage clusters, variation in terms of associated exteins, insertion sites within host proteins, and geographic dissemination. Through the latter we found evidence supporting events of rapid, and geographically localized propagation of intein alleles into new populations. We discovered two cases of class 3 inteins within terminase exteins which have a common ancestor with inteins that have invaded a Cas4 exonuclease. Additionally we report several cases of inteins with HEN domains or architectures that have rarely been reported on, or were previously unseen. These include HENs encoded within a separate ORF on the same strand as the intein, and HENs that are on a separate ORF on the anti-sense strand.

Methods

Code Availability

All code used in this project is available at https://github.com/sophiagosselin/intein_distribution along with usage and more specific information regarding each script. The two exceptions to this are ICE-BLAST and Invader-Sim which are detailed below, and kept in their own separate repositories (see the relevant sections).

Iterative Cluster Expansion BLAST: ICE-BLAST

ICE-BLAST (Iterative Cluster Expansion BLAST) is a Perl (v 5.36.0) based program that captures highly divergent homologous protein sequence matches to a sequence of interest. The program can be found at https://github.com/sophiagosselin/ICE_BLAST along with detailed usage, dependencies, and other information. The Algorithm is detailed below, and parameters used in this study are described in the Dataset Curation section.

ICE-BLAST loops through the following steps (described visually in Figure 1 - A). First, the input sequence is used as the query in a PSI-BLAST (v 2.12.0+) (29) search versus a training database known to contain sequences related to the query (but it is not the database of interest, e.g. Uniref50 (30)). The resulting position specific scoring matrix (PSSM) is then used to run a second PSI-BLAST search against the protein database of interest (out-database in Figure 1 - A). The resulting matches from this second search are filtered based on e-value, percent identity, and query coverage, after which the filtered matches are saved into memory. These filtered matches are clustered using USEARCH (v11.0.667) (31) at the percent sequence identity specified by the user (default 70%) using cluster_fast mode. Finally, the centroids of these sequence clusters are used as the query sequences for the next iteration of the algorithm. This loop continues until (a) no new matches are found in the second PSI-BLAST search, (b) the clustering step produces no new centroid sequences, or (c) the maximum number of iterations (specified by the user) is reached.

ICE-BLAST also has a domain specific mode which can be enabled by the user; this mode treats the query sequence as a sub-sequence of any subject proteins. The algorithm proceeds through the same loop as before, but instead of using the entire protein match from the second PSI-BLAST search, only the portion matching the query is used for clustering. Importantly, this match is confined to an upper and lower bound set by the user when invoking domain specific mode. They are equivalent to the average length of the query sequence(s) plus or minus the user specified percentage of the average length (i.e. an average sequence length of 100aa and a domain specific mode cutoff of 20% would lead to an upper bound of 120aa and a lower bound of 80aa).

Sequence Simulations – Invader Sim

ICE-BLAST was tested using simulated sequence data. Simulation of intein invasion into extein sequences was completed using the Perl program invader_sim.pl (available at https://github.com/sophiagosselin/invader_sim), which uses PAML-EVOLVER (v4.9j) (32) and a Markov chain to simulate the invasion process (described visually in Figure 1 – B).

The program simulates phylogenies of inteins and exteins using PAML-EVOLVER. Invader Sim then pairs these phylogenies 1-1 at random such that each intein tree is paired with one extein tree. Intein invasion is simulated by randomly selecting one tip from the intein phylogeny to "invade" a randomly selected tip from the extein phylogeny. The intein tip with the closest patristic distance in the intein phylogeny to the previous intein tip is chosen as the next invading sequence. The target of this invasion is chosen by a Markov chain using an exponential distribution and two uniform random variables for selection and rejection. This process continues until all inteins have “invaded” an extein sequence and have been paired. Note that this need not mean all extein sequences have been invaded. For testing purposes some extein sequences were always left unpaired (see “Parameters Used In Testing” section below for more detail).

Once the simulated invasion process is complete, amino acid sequences are simulated for the intein and extein phylogenies using PAML-EVOLVER. Intein sequences are inserted into their paired extein counterpart at an insertion site randomly selected within the middle two quartiles of the extein's amino acid sequence (this position is used for all invading inteins within the current extein-intein phylogeny pair). This process is repeated for every sample the user requests.

A modification was made for a second test such that only a subsample of the intein tips simulated are used to invade the extein sequences (to provide a more realistic database with only partial sampling of the true intein population). This subsample of intein sequences (specified by the user) is chosen at random from the entire intein phylogeny, and only those randomly selected sequences will be used in the subsequent steps of the algorithm. All other tips in the intein phylogeny are ignored in regards to selecting the closest tip during the invasion process, and are not used to simulate sequences.

Parameters Used In Testing

Invader Sim simulations in this manuscript used the following parameters.

The extein phylogenies were simulated with a birth rate of 1, a death rate of 1, 1000 tips, a sampling fraction of 0.1, and a mutation rate of 0.1. The intein phylogenies were simulated with a birth rate of 2, a death rate of 1, a sample fraction of 0.5 and a mutation rate of 1. The number of tips for the intein phylogenies differed between the sub-sample and original methodology for dataset construction. In the latter we simulated a phylogeny with 100 intein tips, whereas the former simulated a phylogeny with 1000 intein tips but randomly selected 100 of said tips for invasion.

Intein invasion into exteins was simulated with a window size of 100, and an exponential probability distribution with a lambda of 0.5.

Extein amino acid sequences were simulated to be 500 amino acids in length, using an alpha parameter of 0.5, 4 rate categories, and a proportional model. Intein sequences were simulated with a length of 300 amino acids, an alpha parameter of 1, 4 rate categories, and a proportional model.

Dataset Curation

PhagesDB was used to construct our BLAST databases for this study (25). Two local databases were constructed, one from amino acid sequence files, and one from whole genomes, with associated metadata collected from the PhagesDB API. The genomic database was downloaded and formatted on

September 28 2023, and the amino acid database was downloaded and formatted on November 2 2023. Given the nature of PhagesDB (updates causing reverse compatibility issues), these databases were not updated over the course of this work.

The initial ICE-BLAST query for this study consisted of 78 different intein sequences. These inteins were drawn from three sources: InBase (as of its final update in 2010) (33), known Actinobacteriophage inteins from the Kelley et al. 2016 manuscript (26), and a sample of published and unpublished inteins found by the Gogarten Lab in the PhagesDB database between 2020 and 2024 (34). For additional information and accessions of these query inteins see Supplemental Table 1.

The initial PSSMs for ICE-BLAST were trained on a local copy of Uniref50 (30) database downloaded on November 7th of 2021.

Intein Curation and Classification

An ICE-BLAST search was conducted on the local PhagesDB database using the Uniref50 database to train the initial PSSM. A percent identity of 60%, an e-value of 1e-20 and a domain specific cutoff bound threshold of 25% were used as inputs for filtering.

The initial ICE-BLAST results of putative intein sequences were clustered using USEARCH (v11.0.667) (31) at 70% identity using the cluster_fast algorithm. The associated complete exteins with intein sequences were then extracted from the BLAST database using blastdbcmd. Putative intein sequences were aligned to the entire sequence using MAFFT (v7.490) (35) and visually inspected. For any sequences whose nature was unclear, a combination of BLAST against NCBI's non-redundant database (nrdb) (36), and HHPred (using the following databases: PDB, Pfam, and NCBI CDD) (37), was used to determine whether the identified candidate sequence was actually an intein, and where its boundaries laid. False positives were removed from the dataset at this point. Extein sequences were re-annotated using BLAST searches against the NCBI nrdb, and HHPred (using the same databases as before). Inteins were classified as class I, II, or III, based on comparisons to intein splicing domain consensus sequences (see Tori et. Al 2010) (28). Finally, in-house Perl script were used to extract associated coding sequences from the local PhagesDB genome database. Amino acid and nucleotide fasta files were created through this process for: inteins only, exteins only, and exteins with inteins.

Intein HENs were classified using the method above, with two exceptions. First, we used the PHROG (38) database within HHPred (especially for more divergent sequences) to point us in the direction of potentially related HEN domains as specified in the PFAM section of its entries. Second, we used InterPro's InterProScan feature (as of August 2024) (39) when HHPred or BLAST alone did not provide meaningful matches.

Geographic Maps

Phage isolation locations were extracted from PhagesDB's API as mentioned in the Database Curation section. Since our principal interest was identifying inteins which recently entered into new phages and environments, we processed our metadata as follows. First intein nucleotide sequences were clustered using USEARCH as described prior, but with a cutoff of 97.5% sequence identity (resulting in 91 clusters). We then excluded intein clusters with less than four members as a meaningful phylogenetic

tree cannot be generated with less than four sequences (this removed 61 clusters). For each cluster that passed this filter step we then compared the exteins which inteins had invaded. Clusters were filtered out if exteins within a cluster were more self-similar than their intein sequences were (this removed another 13 clusters). This final set of 13 intein sequence clusters were then used to build nucleotide phylogenetic trees, and their extein sequences (along with intein-less homologs from PhagesDB) were used to construct a second phylogenetic tree. These phylogenies were compared, and only clusters whose intein and extein trees were in-congruent were selected for further study.

Resulting extein phylogenies were marked for intein presence or absence, and the inteins present were then mapped onto their geographic isolation location. Maps of intein isolation locations were created using R (v4.4.1) and the following libraries: sf (v1.0-16) (40), ggplot2 (v3.5.1) (41), maps (v3.4.2) (42), mapdata (v2.3.1) (43), mapproj (v1.2.11) (44), rnaturalearthdata (1.0.0) (45), and rnaturalearthhires (v1.0.0.9000) (46).

Phylogenetic Trees

All phylogenies in this study were built from MAFFT sequence alignments. Unless specified, MAFFT was allowed to select the best alignment algorithm for a given set of sequences (strategy: 1. --auto).

The phylogenies in this study were constructed using IQTREE (v2.0.7) using ModelFinder (-m MFP) and 1000 non-parametric bootstraps (-bb 1000) (47). Models selected for specific trees are included in the respective figure descriptions. Trees were visualized in Dendroscope (v3.8.10) (48) and Figtree (v1.4.4) (49), and their readability improved using Inkscape (v1.3.2) (50). Clans with extremely similar or identical sequences were collapsed in some phylogenies for the sake of readability. Un-collapsed versions are included in the supplemental material (see Figure 2, and Supplemental Figure 1 plus Supplemental Figure 2, for examples of collapsed and un-collapsed respectively).

Predicted Protein Structures

Protein structures were predicted for non-canonical inteins and their associated secondary ORFs using the AlphaFold (51) feature built into UCSF ChimeraX (52, 53). The structures were colored according to the predicted local distance difference test (pLDDT) value per-site. Each predicted structure was used as the query in a DALI search of the PDB database (54).

AlphaFold (51) was also used to generate all predicted structures in the Cas4/terminase intein evolution study. UCSF ChimeraX (52, 53) was used to compare the predicted structure of the LittleE terminase large subunit to the solved structure of a phage large terminase subunit bound to ATP (55).

Intein insertion site consensus sequences

WebLogo (56) was used for any consensus sequence plots generated in this study.

AU Tests

AU tests (57) were performed in study via the method built into IQTree2 (47).

Results

ICE-BLAST Algorithm

In this study, we developed an algorithm to assist in intein discovery. This algorithm, called “Iterative Cluster Expansion BLAST”, or ICE-BLAST for short, uses PSI-BLAST (29) as the principal sequence search tool combined with USEARCH clustering (31) to identify new queries from previous search iteration. ICE-BLAST following the flow chart depicted in Figure 1 – A. Detailed information on the algorithm can be found in the methods section. In short, ICE-BLAST is an iterative algorithm which searches amino acid databases with PSI-BLAST, clusters filtered outputs using USEARCH and then uses the centroids of those sequence clusters as the next iteration of query sequences.

Simulations of Intein Evolution

Since ICE-BLAST involves multiple individual searches, one must take proper precautions against the inclusion of false-positives. They are undesirable within the final output, and can corrupt a search — introducing errors that will propagate through all downstream iterations. Hence, it was important to test the software against known data to check for overall sensitivity and specificity. We developed a sequence simulation algorithm, Invader-Sim, for this purpose. The algorithm is depicted in Figure 1 – B, with additional information available in the methods section. The program uses a combination of PAML-Evolver and a Markov chain to simulate datasets of intein containing extein sequences. Our results showed that ICE-BLAST performed as well as PSI-BLAST on the simulated datasets. Specific results and additional discussion can be found in supplementary materials (see “Results from ICE-BLAST tests using simulated data”).

Intein Distribution in Mycobacteriophages

We conducted our ICE-BLAST search on the PhagesDB database as of September 28th 2023. Inteins from several sources were used as query sequences in this study (Supplemental Table 1). Initial results from ICE-BLAST contained 1994 sequences which were then clustered by USEARCH at 70% identity. Each cluster was visually validated to confirm that it was (1.) a genuine intein, and (2.) had correct borders for the N and C termini. This initial dataset was pruned down to 784 intein containing sequences (Supplemental Table 2). The large reduction in putative matches was attributed entirely to a false positive match to a DNA Primase sequence; causing over 1000 matches to DNA primase extein domains which were falsely incorporated into the dataset. This was likely the result of a partial intein from InBase included within our query set, or a match to portions of the N-extein leading to the propagation of false positive hits.

The remaining putative intein matches were re-annotated using BLAST and HHPred. We found that the 784 putative inteins invaded 18 distinct proteins. Unsurprisingly, the majority of our putative intein sequences were found in extein sequences which have previously been reported as containing inteins. These include: Cas4 exonuclease, DNA helicase, DNA methylase, DNA primase, nucelotidyltransferase, phosphodiesterase, portal protein, terminase large subunit, and thymidylate synthase. The remaining proteins are new to this study within actinobacteriophages, and include: DNA

polymerase, minor capsid protein, ribonucleotide reductase subunits alpha and beta, RtcB-like RNA ligase, tape measure protein, and tRNA splicing RNA ligase. Finally we recovered several hypothetical proteins in this search; most were able to be classified into one of the protein categories above. However, two types of intein containing exteins could not be definitively annotated. Based on results from HHPred and the NCBI conserved domain database, we assigned putative annotations of: a baseplate wedge domain containing protein, and an integrin protein. The former, based on the annotations of other closely related intein containing exteins, could be a portal protein of some sort, but does not share considerable similarity to known portal proteins in any NCBI database. The intein present in the putative integrin was previously reported on in Kelley et al. 2016, with fewer intein containing members (26).

The putative inteins were clustered at 70% sequence identity. For each intein cluster, we determined the intein splicing class (1, 2, or 3), and the PhagesDB phage clusters that they have been found in (Table 1). The majority of putative inteins described in this study were class 1 (497) while the remainder (287) were class 3. The class 1 inteins displayed a much wider range of host exteins (16 different proteins) versus the class 3 inteins which were only found within 4 different extein hosts. We found that more than half of the inteins within our dataset (431) invaded just two proteins, with 293 inteins within terminase sequences, and 138 within Cas4 exonucleases (Figure 3). The inteins in this study span a wide variety of PhagesDB phage clusters, with the vast majority within clusters C1 (415), A1 (92), and E (68). It is interesting to note that all but one of the 68 inteins found in E cluster phages were found in terminase genes, with one in a DNA methylase. All other inteins were distributed across 31 phage clusters and several unique singletons. None of those 31 phage clusters contained more than 35 intein sequences and most contained less than 10.

Intein Homing Endonucleases

Most of the inteins in our database possessed classic intein architecture: an N-terminal splicing domain, a C-terminal splicing domain, and a “canonical” LAGLIDADG HEN domain in-between. There were some mini-inteins (at least 9), and several decaying or highly divergent LAGLIDADG containing inteins. We also found 4 inteins with rare or previously unseen architectures. The distributions of these “non-canonical” inteins are described in Figure 3. We found 13 inteins which contained an HNH HEN instead of LAGLIDADG, which has been observed in phages (20, 58). The 13 HNH HEN containing inteins were the least bizarre of the 4 “non-canonical” inteins, as the other 3 cases possessed HEN’s in secondary open reading frames.

First, we found 7 inteins which contained a LAGLIDADG HEN on the anti-sense strand (Figure 4). This anti-sense LAGLIDADG containing intein has only been reported on twice prior, in passing, in papers from 1998 and 2002 (2, 19). Second, we found 16 inteins with a VSR-like HEN on the anti-sense strand. This anti-sense encoded HEN lies within the region of the intein that should contain an in-frame HEN. The VSR-like HEN was reported in the context of split inteins with no other known intein-associated reports (15). Finally, we found inteins which contained an endonuclease VII-like HEN in a separate +3 ORF within the region of the intein that should contain a HEN. These inteins are the only ones in our dataset which exhibit secondary ORF’s of any meaningful length — all others exhibited many stop codons across all 5 alternative reading frames.

The specific protein structure of these non-canonical inteins is unknown. AlphaFold predicted structures suggest that the portion of the intein which should be an in-frame HEN does not make any recognizable structures, and cannot be matched to any known structure using the DALI server PDB search tool (Supplemental Figure 3). However, predicted structures of the secondary ORF for the anti-sense LAGLIDADG HEN, VSR-like HEN, and endonuclease VII-like HEN, all matched to their associated predicted domains (Supplemental Figure 4). It is important to note that these matches were only to the core catalytic domain of these HENs; significant portions of these structures could not be matched to known proteins. For the anti-sense LAGLIDADG HEN, the LAGLIDADG core sits between residues 40 and 185, matching to PDB structure 2o7m (CreI — a known LAGLIDADG HEN) with a Z-score of 14.8. The VSR-like HEN predicted structure matched to PDB structure 1cw0 (VSR) with a Z-score of 10.6 across the residues 144 to 267. Finally, the endonuclease VII-like HEN matches to PDB structure 1e7d (a recombination endonuclease VII) with a Z-score of 6.5 across residue 1 to residue 67.

Based on the predicted structures for the endonuclease VII-like and anti-sense LAGLIDADG HENs, coupled with the fact that these HENs can be dimeric, the unknown terminal domains of these structures may be linker domains or secondary DNA binding domains (59, 60). The VSR-like HEN appears to contain a long stretch of linked turns and alpha helices not present in normal VSR endonuclease structures (61). The predicted local distance difference test (pLDDT) for this stretch is relatively high, especially within helices, but the region lacked meaningful matches against PDB structures. However, an HHPred search found a significant match (e-value of 0.000046) against protein structure 3R3P from Bacillus phage 0305phi8-36, which indicates the VSR-like HEN may be an EDxHD HEN (62).

Actinobacteriophage Intein Phylogenetics

To understand more about actinobacteriophage intein evolution, we constructed phylogenies from the CDSs of all inteins in our database separated by HEN type with the canonical LAGLIDADG HEN inteins further separated by splicing class.

The class 1 inteins with a canonical LAGLIDADG HEN produced a phylogeny which provided evidence for several independent events of inteins host diversification (Figure 2 – A). For example, the terminase inteins form a polyphyletic clan with at least four separate groups. Each group is separated divided from the others by multiple highly supported splits along the inner nodes. Similarly, portal protein inteins and DNA methylase inteins form polyphyletic clans with five and three separate groups respectively.

The class 3 inteins with canonical LAGLIDADG HENs are almost entirely found within Cas4 exonuclease and terminase exteins (Figure 2 – B). Within this phylogeny appears to be two cases of Cas4 associated inteins diverging from terminase inteins, which are discussed further below

For the VSR-like HEN intein phylogeny, many of the splits are poorly supported, especially those associated with phages: Sejanus, REQ3 and Malisha. However, there are two nodes with 100% bootstrap support that separate the inteins of DE1 cluster phages from one another (Supplemental

Figure 5). While they are all found in the same protein and the same phage cluster, they are split into 2 well supported clans.

While light inferences could be drawn from the phylogenies of the endonuclease-VII-like, HNH, and anti-sense LAGLIDADG inteins, current sampling is ultimately too sparse to make any definite claims from the intein phylogenies alone (Supplemental Figure 6, Supplemental Figure 7, Supplemental Figure 8).

Extein Switching

The class 3 intein phylogeny (Figure 2 B) depicts two cases of a clade of Cas 4 exonuclease inteins emerging from a clade of terminase large subunit (abbreviated to terminase following) inteins. A subtree was generated to better depict these two cases Figure 5). Case 1 involves phage Frankenweenie (FW), while Case 2 involves phage ScoobyDoobyDoo (SDD).

To assess whether this topology was a phylogenetic artifact, an approximately unbiased (AU) tree topology test (57) was used to gauge the strain induced when one attempts to force the Case 1 and Case 2 inteins to group by extein type. The same determined by IQ-TREE's ModelFinder for the original class 3 intein phylogeny (WAG+F+R4) was used for the AU test. The test produced little support for the grouping of these inteins by extein type, (bootstrap proportions of 0.998 and 0.002 for tree constrained by Case and the tree constrained by extein type respectively (Supplemental Table 3)). In other words, the Case 1 Cas4 inteins are more similar to the Case1 terminase inteins than they are to the Case 2 Cas4 inteins, and vice versa.

Case 1

Comparing the intein insertion site nucleotide sequence across the Case 1 Cas4 and Case 1 terminase genes did not reveal striking similarity beyond AAG on the 5' side of the insertion site and xCxA on the 3' side. Phages selected to represent the Case 1 Cas4 inteins (FW and Dexes) produced near identical predicted intein structures. All Case 1 terminase inteins exhibit sequence degradation in the HEN region, indicating that they may be becoming mini inteins. This is reflected in the predicted intein structures for the phages representing this group (FW and Kela) Figure 5). In spite of the comparison difficulties induced by decay of the terminase intein HENs, the splicing domains of these inteins align well across the Cas4 and terminases. To assess whether the Case 1 terminase inteins are still being horizontally transferred, two phylogenies were generated using the Case 1 terminase exteins and intein-free terminases with a similar nucleotide binding site retrieved from BLAST searches against NCBI's nrdb. The first tree was built from the amino acid sequence alignment of the described terminases (Supplemental Figure 9 - A). Domain shuffling in terminases is common (63) making sequence alignment difficult, and potentially complicating the phylogeny. Since there is a contiguous region of similarity across all the terminases used for this analysis at the nucleotide binding site, which is also the intein insertion site for the Case 1 terminases, a second phylogeny was generated from an the +/-75 nucleotides surrounding the intein insertion site (Supplemental Figure 9 - B).The Case 1 terminases grouped together in both with strong bootstrap support (100% and 96% respectively). While increased sampling would improve the resolution of this story, phylogenetic analysis of the currently available

sequences provides support for the Case 1 terminase inteins, which exhibit potentially decaying HENs, being primarily vertically inherited.

Case 2

Comparing the +/-30 nucleotides at the intein insertion site across the LittleE/Omega terminase, SDD Cas4, and non-SDD Case 2 Cas4 sequences reveals regions of striking similarity (Supplemental Figure 15 - A). Both insertion sites reside in important nucleotide binding sites for their respective proteins (letting | denote the insertion site: consensus HEFK|TA for terminases, VEFK|TQ for SDD Cas 4, and consensus VELK|TQ for non-SDD Cas4). Comparison to a solved phage terminase structure bound to ATP (64) hows the intein insertion is on a helix within the Rossmann fold constituting the nucleotide binding site. EFKT is the RecB-like motif-III in the Cas4 exonuclease, and is important for nucleotide binding (65). Alignment of the predicted intein structures generated for representatives of Case 2 and closely related group of terminase inteins (Figure 5) reveals them all to be near identical, barring APunk which exhibits a slight structural difference in its HEN. This difference likely reflects how the APunk terminase intein recognizes a different insertion site than the Case 2 terminase inteins. The high similarity between all Case 2 predicted intein structures even in their HEN further supports the notion that these inteins can recognize and cut an extremely similar intein insertion site. To explain the lack of intein in SDD's terminase, we later propose and discuss a model in which the SDD terminase was not a viable host for the LittleE/Omega terminase intein, and the SDD Cas4 provided a viable alternate landing site.

Intein Dissemination within Local Areas

We previously found inteins which appeared to be proliferating within local areas into divergent phage extein sequences (66). We chose to investigate our dataset for additional cases of local dissemination and proliferation of inteins into new phage populations. We filtered our data in two stages: first to find clusters of closely related inteins (97.5% similarity), then to discover clusters within the filtered set whose inter-extein sequence similarity was less than that of the inteins. These clusters were used to construct phylogenetic trees, and their geographic positions were plotted (using metadata from PhagesDB). Of the 13 clusters of intein sequences that passed this criterion, we focused on three cases which displayed high levels of in-congruence between the extein and intein phylogenies during a visual inspection.

First, we investigated a set of inteins which were found in nucleotidyltransferases (Figure 6). These inteins were split into two groups based on their similarity: one of which passed the second filtering step and one that didn't. Their presence was plotted onto an extein phylogeny, and isolation locations for phages containing inteins from the former group were plotted Figure 6 - F). This group of inteins had a global distribution, and of particular interest to us, two groups of co-localized inteins. One group of 5 inteins isolated from California, and a second group found in Durban South Africa. In both of these locations, inteins exhibited nearly identical nucleotide sequences, but were found in comparatively divergent exteins. Furthermore, these extein sequences were separated in the extein phylogeny by non-intein containing sequences.

We repeated this analysis with a second group of inteins from Ohio and Pennsylvania. These 7 inteins were found within minor capsid proteins, and turned out to be the group of inteins which contain the anti-sense LAGLIDADG HEN (Figure 7). These inteins, despite being identical to each-other (for 5 of the seven), are once again found in divergent extein sequences which were separated from each-other on the phylogenetic tree with non-intein containing taxa, and taxa with completely different inteins.

Finally, we investigated the inteins found within the putative integrins exteins. These inteins fell into three different groups (by similarity). Many were found in divergent exteins within phages isolated primarily within the Maryland and DC area (Figure 8). Once again, these inteins were found in divergent exteins separated on the phylogenetic tree by non-intein containing taxa. Many of these inteins were found specifically within the grounds of Howard University in Washington DC; however, despite the small geographic distance between the phages found there, the exteins showed significant divergence. Furthermore, inteins from each of the three described clans were found on this campus. We found that the vast majority of the inteins from the bottom clan of the intein phylogeny (colored in rose-gold) were present predominantly on the eastern shoreline of the USA. The clan at the top (colored in blue) had a much wider distribution — mostly confined to the Mississippi River Valley, and the Midwest. The last of the three clans (colored in red) was found across the USA, but only within the confines of urban areas and suburbs.

For each of these cases, we investigated the geographical distribution of all the extein sequences within each of the extein phylogenies (Supplemental Figure 10, Supplemental Figure 11, Supplemental Figure 12). We found repeatedly that the geographic distribution of intein containing phages were not a 1 to 1 replica of the phages' overall geographic distribution. While the presence of intein containing phages was higher in locations where more phages were isolated, we suspect this to be a product of sampling bias. SEA-PHAGES operates a large number of class in universities and colleges on the Eastern Shoreline of the USA, and, more generally, students tend to isolate their phages within close proximity to university grounds; hence any distribution pattern will be limited based on the sampling locations.

Discussion

Overall Trends

As in previous studies the inteins we discovered within actinobacteriophages are found predominantly within DNA-binding extein sequences (26). While the majority are within terminases, the inteins we discovered show a remarkable range of host proteins. In addition to what are considered essential phage genes, we find inteins within phage genes with more limited distributions such as the thymidylate synthase, or ribonucleotide reductase. Even genes that lack nucleotide associated enzymatic activity, the minor capsid proteins, were found within the set of invaded viral host proteins. Those minor capsid protein invading inteins were also all exclusively non-canonical inteins — a fact that continues to puzzle.

Most inteins in our dataset had disjunct distributions in extein phylogenies, and phylogenies of the intein sequences frequently conflicted with the phylogeny of their extein. Ergo, we believe that many of these inteins are actively moving between different hosts. However, exceptions do exist in cases of mini inteins (an example of which can be found in Supplemental Figure 13). In addition inteins that persist in clans of extein sequences, at first sight might be considered to have decaying HENs, and like mini-inteins be vertically inherited (see the purple intein group in Figure 7). However, while the exteins show some sequence divergence, the inteins are identical likely illustrates a recent invasion event (see below for further discussion).

The many intein-free homologs accompanying intein-containing genes in our dataset indicate that intein homing into new empty target site must be balanced by the loss of inteins in other homologs. Given the seemingly rapid dissemination of inteins into local phage populations, intein loss might also be assumed to occur frequently. This could occur via gradual decay and recombination of the intein, or be the consequence of a fitness cost imparted by the intein on the phage. Inteins are known to cause a drop of fitness; in *Haloferax volcanii*, a DNA polymerase intein caused a 7% drop in fitness for the host archaeon (67). Thus, the alternative loss scenario would act at the population level — the intein imparts enough of a fitness cost such that the intein-containing phage is out-competed by intein free phages, and the intein containing phages would be slowly selected out of the population. This process would be enhanced if the non-invaded phages evolve at least a partial resistance to intein propagation (perhaps via mutations to the HEN target site). This potential impact on host fitness is a mechanism worth exploring in future work.

The Nature of Non-Canonical Inteins

In this study we described several inteins with previously unseen, or extremely rare architectures in relation to their HEN domains. To our knowledge, only the anti-sense LAGLIDADG HEN has been reported on in prior literature within the context of a fully functional intein (2, 19). These anti-sense LAGLIDADG HEN containing inteins were discovered in non-phage genomes within the *dnaX* gene of a *Synechocystis* species, and the SpoVR like protein from *Chloroflexus aurantiacus* J-10-fl (now stored on InBase (33)). Hence, while not entirely novel, we show that this intein architecture is present within

at least one other extein sequence, and importantly, that the intein and anti-sense HEN appear to work together to facilitate ongoing intein invasion events.

While the VSR-like HEN containing inteins have been reported on previously, they were only found within split-inteins (15). Our findings show that anti-sense encoded VSR-like HEN containing inteins also exist and are not merely a point of speculation. If the theory of split-intein evolution outlined by Dassa et al. 2009 is true, then perhaps these newly discovered inteins represent one of the potential off-tracks within their model. It should also be noted here, that Dassa et al. 2009 report that the VSR-like HENs they discovered were related to freestanding HENs within their database (a fact whose importance will be discussed below). Finally, the endonuclease VII-like HEN containing intein appears to be completely novel, with no know reports. Like the VSR-like HENs, this endonuclease has been reported prior in a free-standing context (68).

Given: the ubiquity of freestanding HEN genes within actinobacteriophages, the frequency with which they seem to proliferate (68–72), and the presence of related free-standing HENs to those found within our intein sequences, we believe that the three intein architectures described above originate via invasion of an intein sequence with a free-standing HEN. Given the seeming rarity of these elements, we further propose (in line with the model laid out by Dassa et al. 2009 (15)), that successful integration of these HEN into an intein endonuclease domain is extremely rare. The breaking of, and subsequent integration into the intein CDS by a free-standing HEN in this fashion may be relatively common, but most of the time this integration is likely to result in split-inteins, or introduce nonsense mutations which render post-translational rescue of function via intein splicing an impossibility.

Assuming the free-standing HEN's integration into the intein CDS is successful, the path forward is unclear. It is possible the HEN becomes a pseudo-gene, but there have been prior examples of phage genes with nested ORF's (a HEN to be specific) that still translate both proteins (73, 74). In our cases the secondary ORF's seem to be lie mostly outside the bounds of the intein splicing domains, and given the splicing domains' level of conservation, it is probable that the spicing domains still function as intended. Additionally, many of these non-canonical inteins were found in the crucial terminase gene, and it is therefore unlikely that the intein is not splicing itself out — these phages would otherwise be unable to properly package their DNA into their capsid (a certain end for that phage lineage). Our phylogenetic reconstructions of the anti-sense LAGLIDADG containing inteins provided evidence for continued invasion. These inteins are present in multiple locations within divergent host exteins, and importantly are separated with meaningful support by other evolutionary close relatives that lack the intein. A pattern like that is not reasonably explained via vertical inheritance and loss, and it is therefore almost a certainty that the splicing domains and anti-sense HEN domain of that particular intein still function as a single unit. This may also be true of two groups of inteins with VSR-like HENs (cluster IDs 60 and 91) which likewise show a pattern of inheritance more in-line with horizontal gene transfer. However, the other cases of VSR-like HEN containing inteins exhibit patterns suggestive of vertical inheritance (or were singletons and thus no evolutionary inference could be drawn). The endonuclease VII-like HEN containing inteins presented a more ambiguous distribution. While most are clustered together, there are a number of closely related extein alleles without an intein. This, plus the sequence similarity of these inteins (for example, the inteins of Cluster 47 — which includes phages Angelicase, Affeca, Keitabear, and Sedona — are all identical with the exception of Angelicase), suggests that these

inteins may still be transferring between phages horizontally (Supplemental Figure 14, Supplemental Figure 6).

If these inteins can still splice, but can no longer successfully jump, then selection will likely lead to these secondary ORF's very quickly accumulating stop codons and ceasing to function. This would be especially likely in the case of anti-sense ORF's since they would not benefit from the purifying selection acting upon the host extein to remove nonsense mutations. Mini-inteins and split-inteins would be the most likely observable outcome of this process. While we do find nine inteins which appear to have decaying HEN domains, and a further nine mini-inteins, a survey strategy fine-tuned towards finding split and mini inteins would be needed to look for these potential relics of decay. Potentially, the secondary ORF's encoded on the same strand could slide into the extein's reading frame via a single base pair indel, resulting in a fully contiguous intein with a novel HEN.

Inteins created by the mechanism described above thereby provide meaningful support for a "protein splicing first" theory of intein evolution (2, 8, 75). The number of intein associated HEN's (now including HNH, VSR-like, GIY-YIG, LAGLIDADG, etc. (15, 18–20)) suggests that there have been multiple separate invasions of intein CDSs with free-standing HENs (75). Perhaps most, if not all, of these HENs came into intein sequences via a free-standing HEN, proliferated through the intein's self-splicing machinery, and were shifted by an indel or an inversion into a contiguous orientation.

Geographic Spread of Inteins

There is a wide geographic distribution of inteins within actinobacteriophages which follows the global distribution of their phage hosts. Our data are primarily rooted in the United States of America, reflecting SEA-PHAGE's sampling bias (global participation is less dense than that of USA-based) and is present in unrelated earlier studies of the dataset (76). Even within the context of this bias, there are multiple instances of highly similar intein sequences existing within close geographic proximity whose host extein sequences exhibit a higher level of divergence then their associated inteins. We argue here, as we have in prior research (66), that these observations reflect recent invasion events of these inteins into new phage populations.

Since these intein sequences are extremely similar on the nucleotide level (and in the cases described, often identical), these inteins must have diverged from one another relatively recently. This presents two option: rapid host proliferation with direct vertical inheritance, or rapid intein proliferation via horizontal gene transfer. For some of the inteins in our dataset, the former may be considered a possibility (see the clan of intein containing exteins colored purple in Figure 7); however, where the intein containing exteins form a well-supported clan with limited geographic distribution, the observation that the inteins are identical, whereas the exteins are divergent, suggests that the inteins have recently invaded the divergent exteins instead of being vertically inherited. These findings suggest that intein invasion represents an example of a highly effective gene drive acting on the local level following the introduction of a new founder intein allele. The forces that limit this drive (after all, not all homologous exteins within a given location contain the intein) need further study. Phages, because of their rapid propagation and large population sizes provide an excellent opportunity to study intein population dynamics in more detail.

Frequent Horizontal Transfer of Inteins

In cases where we find inteins within extein sequences that lack a direct common ancestor and are separated on the phylogenetic tree by many nodes leading to other non-intein contain extein sequences, a vertical inheritance pattern would require a significant number of loss events to be an accurate model given these phylogenies. That is without accounting for the differing evolutionary histories supported by phylogenetic reconstructions of the intein and extein sequences, which often provide further evidence of a non-vertical inheritance pattern.

Since the inteins discussed above seem to possess both a wide (and oftentimes global) distribution, the next question inevitably becomes how did they get there? While humans are almost certainly acting as a secondary or tertiary vector for the intein, we propose that these inteins are likely dispersing in part via the phages we find them in and in part, via the phage's host bacteria. Previous literature, and our own anecdotal observations, find that many phage associated inteins are often found inside of the bacterial genomes, most likely within prophages (25). This provides the intein an effective secondary host and confers the intein a high degree of genetic mobility by being able to take advantage of both the viral and bacterial vector. In a bacterial cell, the intein could propagate not only when the bacterial cell replicates, but also through the numerous avenues of horizontal gene transfer that bacterial cells engage in (77). For example, intein alleles could be transferred via exogenous DNA uptake into an already existing non-invaded prophage gene, or the intein could transfer into a new phage genome when a novel phage co-infects the bacterial cell. This is in addition to the already high degree of genetic mobility granted to an intein by being inside a phage.

Nucleotide binding sites as an avenue for extein-switching

The secondary case of Cas4 inteins emerging from terminase inteins on the class 3 intein phylogeny provides a compelling argument for the use of highly similar nucleotide binding sites as a means of intein movement between otherwise dissimilar genes. The stretch of high nucleotide sequence similarity at the intein insertion site across the Case 2 Cas4 and the terminase sequences (10 bases on the 5' side, 2 bases on the 3' side) in conjunction with the other shared positions within the +/-30bp surrounding the site (Supplemental Figure 15) may be sufficient for recognition and invasion by the same intein. The proposed range for HEN recognition sites is ~12-40bp, with smaller genomes such as phage genomes being on the lower end (78). This in conjunction with the many types of phage recombinases that can tolerate some dissimilarity between regions used for homologous recombination (79) suggests this intein could possibly invade either site if given the chance. To further complete the picture and investigate why the SDD terminase does not contain the intein, we analyzed the two sites in the SDD terminase gene with highest probability for recognition and invasion by the LittleE/Omega terminase intein (Supplemental Figure 15 - A) and find SDD's Cas4 gene to be a more hospitable landing pad for the LittleE/Omega terminase intein by metrics of insertion site recognition and residues at the insertion site required for intein self-splicing (aspartate and proline at tested sites would inhibit intein splicing (Supplemental Figure 15 - B) (80, 81). Additionally, LittleE and Omega are temperate phages of the same strain of bacteria that SDD is a lytic phage of (*Mycobacterium smegmatis* mc2155) (25). The presented evidence taken together suggest a model in which the recipient genome was exposed to the donor's terminase intein during co-infection or infection of a host with the donor's

prophage. Then, either homing of the donor terminase intein into the recipient's terminase gene was thwarted by the lack of similarity between these terminases, or homing was successful but nonviable due to lack of a splicing-compliant residue on the C-terminal side of the insertion site. Luckily, a region of similarity to the donor terminase present in the recipient Cas4 conferred by a shared nucleotide binding motif provided a viable alternate insertion site for the intein (Figure 9). Being in a brand-new context also lets this copy of the intein begin a fresh evolutionary trajectory, as the majority of homologous genes in the population will initially lack this intein. We observe this as the high degree invasion of SDD's Cas4 intein into other phages Figure 5). As PhagesDB and other sources of phage genome sequences continue to grow, increased sampling will tell a clearer story of the events that unfolded in the evolution of these inteins.

Conclusions

The work outlined above shines a light on the previously unknown breadth of inteins within Actinobacteriophage genes. Throughout this research we repeatedly find disjunct distributions of intein containing exteins in extein phylogenies, and propose that the conflict between extein and intein phylogenies suggests that inteins within Actinobacteriophages frequently home into target sites within intein-free genes. We find evidence for multiple cases of rapid geographically local episodes of intein invasion that support the prior assertion. In addition, we describe several previously unseen or rare intein architectures, and provide a hypothesis to explain their potential occurrence and apparent continued dissemination. We also describe a model and potential cases for how inteins may home into new target sites, within previously un-invaded gene families, based on the intein phylogenies and the similarities and dissimilarities between target sites.

Finally, we wish to stress that these non-canonical inteins may be significantly more common than even what we have outlined here. Given their overlapping ORFs, these inteins are likely missed in automated annotation pipelines, especially within phages — current SEA-PHAGES training argues for "...typically only one frame in one strand is used for a protein-coding gene…" (82). Even to the trained eye, these inteins may currently be glossed over as 'decaying' canonical inteins or misclassified as mini-inteins. Significant work will be needed to confirm that inteins listed as decaying are correctly classified, and to discover the true extent of these inteins within phages and their bacterial hosts.

Acknowledgments and Funding Sources

We would like to thank the students and faculty of the SEA-PHAGES program and community for their work isolating, sequencing, and annotating actinobacteriophages, for maintaining PhagesDB, and for helpful discussions. We would like to thank Dr. Jonathan Klassen, Dr. Daniel Gage, and Dr. Carolyn Teschke at the University of Connecticut for helpful comments, and insights into the non-canonical inteins. Additional thanks to the Computational Biology Core, Institute for Systems Genomics, University of Connecticut, for the use of computational resources.

This work was supported through grants from the National Science Foundation within the BSF-NSF joint research program, NSF/MCB 1716046, and UConn's OVPR Research Excellence Program (REP-0000000833). The funders had no role in study design, data collection and analysis, decision to publish, or the writing of this manuscript.

Citations

1. Nanda A, Nasker SS, Mehra A, Panda S, Nayak S. 2020. Inteins in Science: Evolution to Application. Microorganisms 8:2004.

2. Gogarten JP, Senejani AG, Zhaxybayeva O, Olendzenski L, Hilario E. 2002. Inteins: structure, function, and evolution. Annu Rev Microbiol 56:263–287.

3. Lennon CW, Belfort M. 2017. Inteins. Curr Biol 27:R204–R206.

4. Shih C-K, Wagner R, Feinstein S, Kanik-Ennulat C, Neff N. 1988. A Dominant Trifluoperazine Resistance Gene from Saccharomyces cerevisiae Has Homology with F0F1 ATP Synthase and Confers Calcium-Sensitive Growth. Mol Cell Biol 8:3094–3103.

5. Hirata R, Ohsumk Y, Nakano A, Kawasaki H, Suzuki K, Anraku Y. 1990. 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.

6. Kane PM, Yamashiro CT, Wolczyk DF, Neff N, Goebl M, Stevens TH. 1990. Protein Splicing Converts the Yeast TFP1 Gene Product to the 69-kdDSubunit of the Vacuolar H+-Adenosine Triphosphatase. Science 250:651–657.

7. Novikova O, Jayachandran P, Kelley DS, Morton Z, Merwin S, Topilina NI, Belfort M. 2016. Intein Clustering Suggests Functional Importance in Different Domains of Life. Mol Biol Evol 33:783–799.

8. Green CM, Novikova O, Belfort M. 2018. The dynamic intein landscape of eukaryotes. Mob DNA 9:4.

9. Gallot-Lavallée L, Jerlström-Hultqvist J, Zegarra-Vidarte P, Salas-Leiva DE, Stairs CW, Čepička I, Roger AJ, Archibald JM. 2023. Massive intein content in Anaeramoeba reveals aspects of intein mobility in eukaryotes. Proc Natl Acad Sci 120:e2306381120.

10. Paulus H. 2001. Inteins as Enzyme. Bioorganic Chem 29:119–129.

11. Mills KV, Perler FB. 2005. The Mechanism of Intein-Mediated Protein Splicing: Variations on a Theme. Protein Pept Lett 12:751–755.

12. Belfort M, Roberts RJ. 1997. Homing endonucleases: keeping the house in order. Nucleic Acids Res 25:3379–3388.

13. Gimble FS, Thorner J. 1992. Homing of a DNA endonuclease gene by meiotic gene conversion in Saccharomyces cerevisiae. Nature 357:301–306.

14. Iwaï H, Mikula KM, Oeemig JS, Zhou D, Li M, Wlodawer A. 2017. Structural Basis for the Persistence of Homing Endonucleases in Transcription Factor IIB Inteins. J Mol Biol 429:3942–3956.

15. Dassa B, London N, Stoddard BL, Schueler-Furman O, Pietrokovski S. 2009. Fractured genes: a novel genomic arrangement involving new split inteins and a new homing endonuclease family. Nucleic Acids Res 37:2560–2573.

16. Tavassoli A, Benkovic SJ. 2007. Split-intein mediated circular ligation used in the synthesis of cyclic peptide libraries in E. coli. Nat Protoc 2:1126–1133.

17. Hoffmann S, Terhorst TME, Singh RK, Kümmel D, Pietrokovski S, Mootz HD. 2021. Biochemical and Structural Characterization of an Unusual and Naturally Split Class 3 Intein. Chembiochem Eur J Chem Biol 22:364–373.

18. Gimble FS. 2000. Invasion of a multitude of genetic niches by mobile endonuclease genes. FEMS Microbiol Lett 185:99–107.

19. Gorbalenya AE. 1998. Non-canonical inteins. Nucleic Acids Res 26:1741–1748.

20. Dalgaard JZ, Klar AJ, Moser MJ, Holley WR, Chatterjee A, Saira Mian I. 1997. Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and

identification of an intein that encodes a site-specific endonuclease of the HNH family. Nucleic Acids Res 25:4626–4638.

21. Yang F, Labani-Motlagh A, Bohorquez JA, Moreira JD, Ansari D, Patel S, Spagnolo F, Florence J, Vankayalapati A, Sakai T, Sato O, Ikebe M, Vankayalapati R, Dennehy JJ, Samten B, Yi G. 2024. Bacteriophage therapy for the treatment of Mycobacterium tuberculosis infections in humanized mice. Commun Biol 7:1–13.

22. Allué-Guardia A, Saranathan R, Chan J, Torrelles JB. 2021. Mycobacteriophages as Potential Therapeutic Agents against Drug-Resistant Tuberculosis. Int J Mol Sci 22:735.

23. Zeynali kelishomi F, Khanjani S, Fardsanei F, Saghi Sarabi H, Nikkhahi F, Dehghani B. 2022. Bacteriophages of Mycobacterium tuberculosis, their diversity, and potential therapeutic uses: a review. BMC Infect Dis 22:957.

24. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, Jacobs-Sera D, Falbo J, Gross J, Pannunzio NR, Brucker W, Kumar V, Kandasamy J, Keenan L, Bardarov S, Kriakov J, Lawrence JG, Jacobs WR, Hendrix RW, Hatfull GF. 2003. Origins of Highly Mosaic Mycobacteriophage Genomes. Cell 113:171–182.

25. Russell DA, Hatfull GF. 2017. PhagesDB: the actinobacteriophage database. Bioinformatics 33:784–786.

26. Kelley DS, Lennon CW, Belfort M, Novikova O. 2016. Mycobacteriophages as Incubators for Intein Dissemination and Evolution. mBio 7:e01537-16.

27. Tori K, Perler FB. 2011. Expanding the Definition of Class 3 Inteins and Their Proposed Phage Origin▿. J Bacteriol 193:2035–2041.

28. Tori K, Dassa B, Johnson MA, Southworth MW, Brace LE, Ishino Y, Pietrokovski S, Perler FB. 2010. Splicing of the Mycobacteriophage Bethlehem DnaB Intein. J Biol Chem 285:2515–2526.

29. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

30. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, the UniProt Consortium. 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31:926–932.

31. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.

32. Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24:1586–1591.

33. Perler FB. 2002. InBase: the Intein Database. Nucleic Acids Res 30:383–384.

34. Gosselin SP, Arsenault DR, Jennings CA, Gogarten JP. 2023. The Evolutionary History of a DNA Methylase Reveals Frequent Horizontal Transfer and Within-Gene Recombination. 2. Genes 14:288.

35. Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30:772–780.

36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

37. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. 2018. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol 430:2237–2243.

38. Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio RE, Mom R, Toussaint A, Petit M-A, Enault F. 2021. PHROG: families of prokaryotic virus proteins clustered using remote homology.

NAR Genomics Bioinforma 3:lqab067.

39. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240.

40. Pebesma E. 2018. Simple Features for R: Standardized Support for Spatial Vector Data. R J 10:439.

41. Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org.

42. Deckmyn OS code by RAB and ARWR version by RBE by TPM and A. 2023. maps: Draw Geographical Maps (3.4.2).

43. Brownrigg OS code by RAB and ARWR version by R. 2022. mapdata: Extra Map Databases (2.3.1).

44. Minka DMP for R by RB and TP, Bivand transition to P 9 codebase by R. 2023. mapproj: Map Projections (1.2.11).

45. South A, Michael S, Massicotte P. 2024. rnaturalearthdata: World Vector Map Data from Natural Earth Used in “rnaturalearth” (1.0.0).

46. South A, Michael S, Massicotte P. 2024. rnaturalearthhires: High Resolution World Vector Map Data from Natural Earth used in rnaturalearth. https://docs.ropensci.org/rnaturalearthhires.

47. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol 37:1530–1534.

48. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8:460.

49. Rambaut A. 2012. FigTree v1. 4.

50. Developers IW. Inkscape - Draw Freely. | Inkscape. https://inkscape.org/. Retrieved 2 October 2024.

51. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. 7873. Nature 596:583–589.

52. Meng EC, Goddard TD, Pettersen EF, Couch GS, Pearson ZJ, Morris JH, Ferrin TE. 2023. UCSF ChimeraX: Tools for structure building and analysis. Protein Sci 32:e4792.

53. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE. 2021. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci Publ Protein Soc 30:70–82.

54. Holm L, Laiho A, Törönen P, Salgado M. 2023. DALI shines a light on remote homologs: One hundred discoveries. 1. Protein Sci 32:e4519.

55. Zhao H, Christensen TE, Kamau YN, Tang L. 2013. Structures of the phage Sf6 large terminase provide new insights into DNA translocation and cleavage. Proc Natl Acad Sci 110:8075–8080.

56. Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Res 14:1188–1190.

57. Shimodaira H. 2002. An Approximately Unbiased Test of Phylogenetic Tree Selection. Syst Biol 51:492–508.

58. DALGAARD JZ, MOSER MJ, HUGHEY R, MIAN IS. 1997. Statistical Modeling, Phylogenetic Analysis and Structure Prediction of a Protein Splicing Domain Common to Inteins and Hedgehog Proteins. J Comput Biol 4:193–214.

59. Raaijmakers H, Vix O, Törõ I, Golz S, Kemper B, Suck D. 1999. X-ray structure of T4 endonuclease VII: a DNA junction resolvase with a novel fold and unusual domain-swapped dimer architecture. EMBO J 18:1447–1458.

60. McMurrough TA, Brown CM, Zhang K, Hausner G, Junop MS, Gloor GB, Edgell DR. 2018. Active site residue identity regulates cleavage preference of LAGLIDADG homing endonucleases. Nucleic Acids Res 46:11990–12007.

61. Hennecke F, Kolmar H, Bründl K, Fritz H-J. 1991. The vsr gene product of E. coli K-12 is a strand- and sequence-specific DNA mismatch endonuclease. Nature 353:776–778.

62. Taylor GK, Heiter DF, Pietrokovski S, Stoddard BL. 2011. Activity, specificity and structure of I-Bth0305I: a representative of a new homing endonuclease family. Nucleic Acids Res 39:9705–9719.

63. Leavitt JC, Gilcrease EB, Wilson K, Casjens SR. 2013. Function and horizontal transfer of the small terminase subunit of the tailed bacteriophage Sf6 DNA packaging nanomotor. Virology 440:117–133.

64. Zhao H, Christensen TE, Kamau YN, Tang L. 2013. Structures of the phage Sf6 large terminase provide new insights into DNA translocation and cleavage. Proc Natl Acad Sci 110:8075–8080.

65. Shiimori M, Garrett SC, Graveley BR, Terns MP. 2018. Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. Mol Cell 70:814-824.e6.

66. Gosselin SP, Arsenault DR, Jennings CA, Gogarten JP. 2023. The Evolutionary History of a DNA Methylase Reveals Frequent Horizontal Transfer and Within-Gene Recombination. 2. Genes 14:288.

67. Naor A, Altman-Price N, Soucy SM, Green AG, Mitiagin Y, Turgeman-Grott I, Davidovich N, Gogarten JP, Gophna U. 2016. Impact of a homing intein on recombination frequency and organismal fitness. 32. Proc Natl Acad Sci 113:E4654–E4661.

68. Pope WH, Jacobs-Sera D, Best AA, Broussard GW, Connerly PL, Dedrick RM, Kremer TA, Offner S, Ogiefo AH, Pizzorno MC, Rockenbach K, Russell DA, Stowe EL, Stukey J, Thibault SA, Conway JF, Hendrix RW, Hatfull GF. 2013. Cluster J Mycobacteriophages: Intron Splicing in Capsid and Tail Genes. PLOS ONE 8:e69273.

69. Edgell DR, Gibb EA, Belfort M. 2010. Mobile DNA elements in T4 and related phages. Virol J 7:290.

70. Barth ZK, Dunham DT, Seed KD. 2023. Nuclease genes occupy boundaries of genetic exchange between bacteriophages. 3. NAR Genomics Bioinforma 5:lqad076.

71. Zhang L, Xu D, Huang Y, Zhu X, Rui M, Wan T, Zheng X, Shen Y, Chen X, Ma K, Gong Y. 2017. Structural and functional characterization of deep-sea thermophilic bacteriophage GVE2 HNH endonuclease. Sci Rep 7:42542.

72. Hatfull GF. 2020. Actinobacteriophages: Genomics, Dynamics, and Applications. 1. Annu Rev Virol 7:37–61.

73. Wright BW, Molloy MP, Jaschke PR. 2021. Overlapping genes in natural and engineered genomes. Nat Rev Genet 23:154.

74. Gibb EA, Edgell DR. 2007. Multiple Controls Regulate the Expression of mobE, an HNH Homing Endonuclease Gene Embedded within a Ribonucleotide Reductase Gene of Phage Aeh1. J Bacteriol 189:4648–4661.

75. Derbyshire V, Belfort M. 1998. Lightning strikes twice: Intron–intein coincidence. Proc Natl Acad Sci 95:1356–1357.

76. Jacobs-Sera D, Abad LA, Alvey RM, Anders KR, Aull HG, Bhalla SS, Blumer LS, Bollivar DW, Bonilla JA, Butela KA, Coomans RJ, Cresawn SG, D'Elia T, Diaz A, Divens AM, Edgington NP, Frederick GD, Gainey MD, Garlena RA, Grant KW, Gurney SMR, Hendrickson HL, Hughes LE, Kenna MA, Klyczek KK, Kotturi H, Mavrich TN, McKinney AL, Merkhofer EC, Parker JM, Molloy SD, Monti DL, Pape-Zambito DA, Pollenz RS, Pope WH, Reyna NS, Rinehart CA, Russell DA, Shaffer CD, Sivanathan V, Stoner TH, Stukey J, Sunnen CN, Tolsma SS, Tsourkas PK, Wallen JR, Ware VC, Warner MH, Washington JM, Westover KM, Whitefleet-Smith JL, Wiersma-Koch HI, Williams DC, Zack KM, Hatfull GF. 2020. Genomic diversity of bacteriophages infecting Microbacterium spp. PLOS ONE 15:e0234636.

77. Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. 8. Nat Rev Genet 16:472–482.

78. Jurica MS, Stoddard BL. 1999. Homing endonucleases: structure, function and evolution. Cell Mol Life Sci CMLS 55:1304–1326.

79. Lopes A, Amarir-Bouhram J, Faure G, Petit M-A, Guerois R. 2010. Detection of novel recombinases in bacteriophage genomes unveils Rad52, Rad51 and Gp2.5 remote homologs. Nucleic Acids Res 38:3952–3962.

80. Mills KV, Johnson MA, Perler FB. 2014. Protein splicing: how inteins escape from precursor proteins. J Biol Chem 289:14498–14505.

81. Turgeman-Grott I, Arsenault D, Yahav D, Feng Y, Miezner G, Naki D, Peri O, Papke RT, Gogarten JP, Gophna U. 2023. Neighboring inteins interfere with one another's homing capacity. PNAS Nexus 2:pgad354.

82. Guiding Principles of Annotation. https://seaphagesbioinformatics.helpdocsonline.com/guiding-principles. Retrieved 24 October 2024.

Figures and Tables

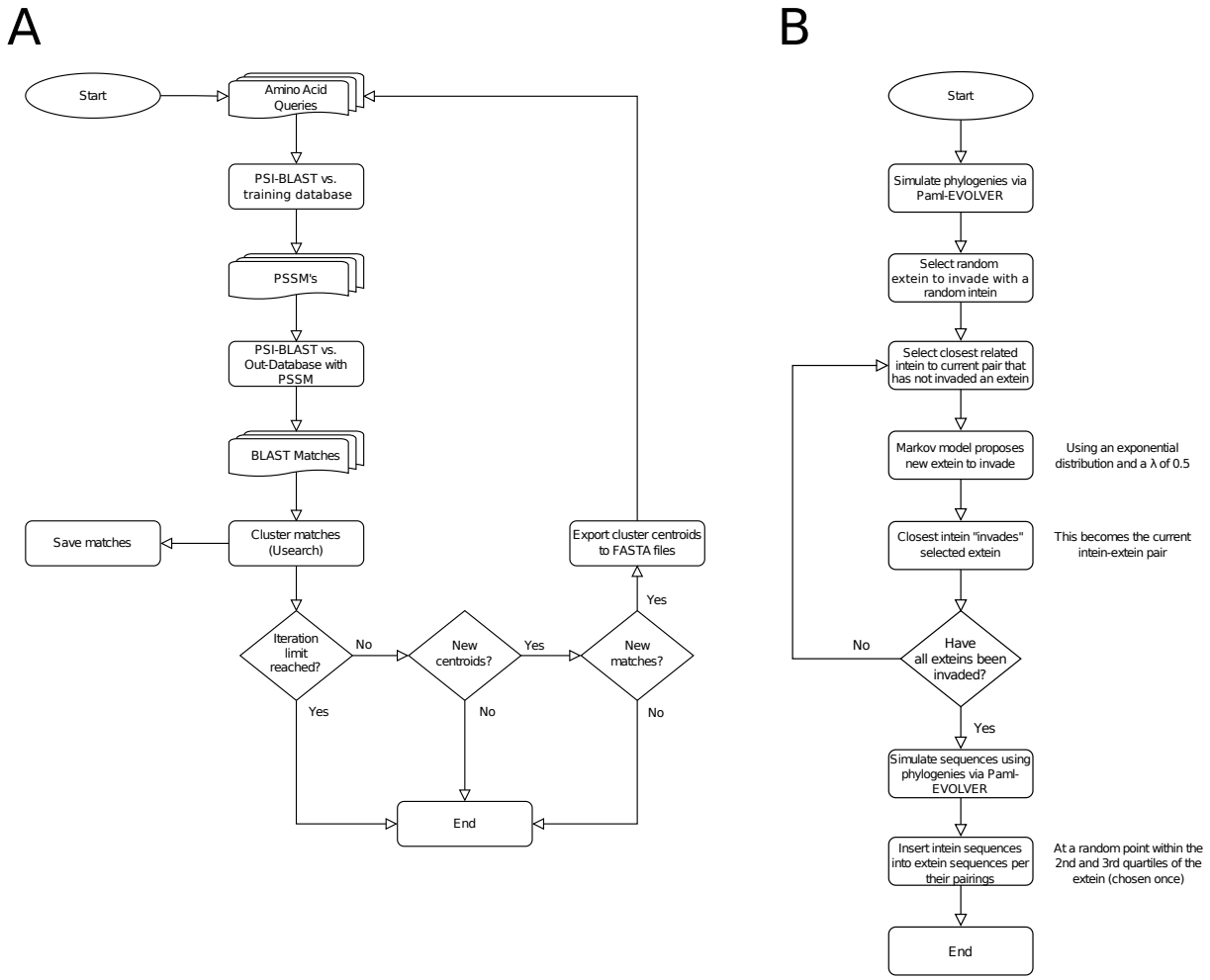

Figure 1: Flow charts describing the algorithms of (A) ICE-BLAST, and (B) Invader-Sim.

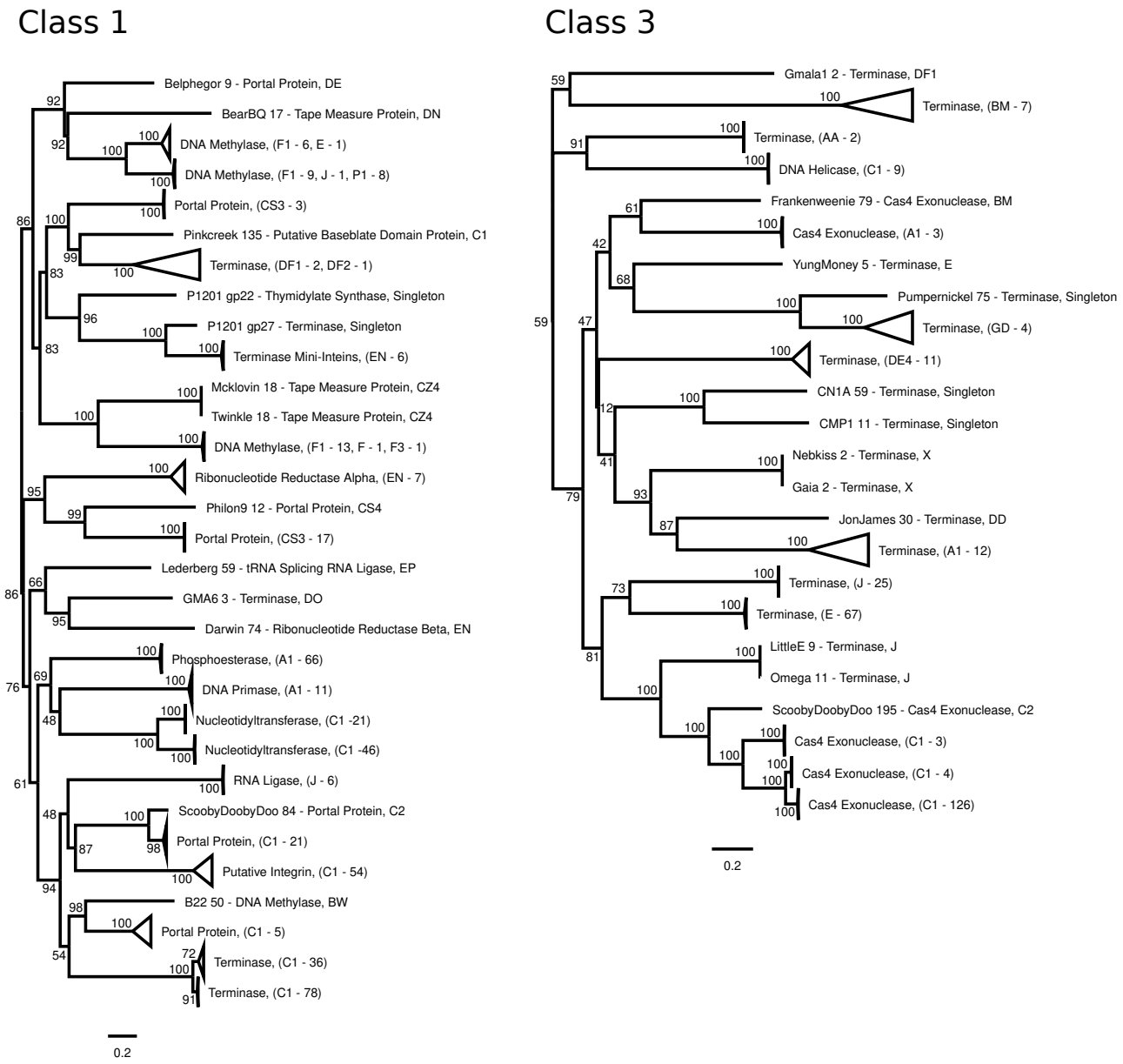

Figure 2: Maximum likelihood phylogenies of the intein encoding DNA sequences for the class 1 (A) and class 3 (B) inteins with canonical LAGLIDADG HEN's. Clans of inteins within the same extein are collapsed down for the sake of readability; the number of inteins, and which phage clusters they are found in, are noted in the names for said collapsed nodes. Substitution models used were (A) TIM2+F+R3 and (B) TPM2+F+R4. Non-parametric bootstrap support is listed on internal nodes. Scale bar measures substitutions per site.

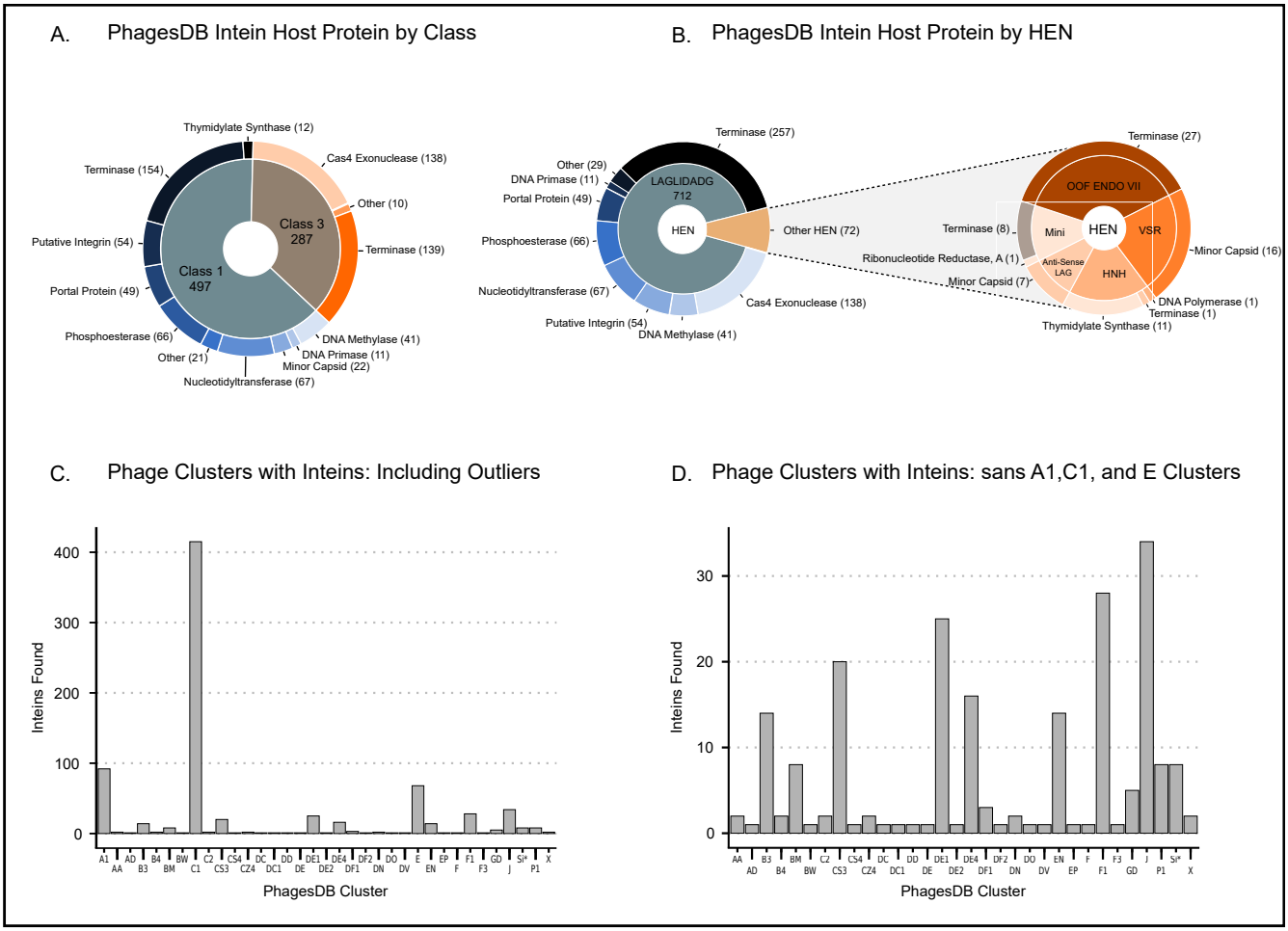

Figure 3: (A) Donut chart describing the splicing class (inner circle), and associated extein sequence (outer circle) for the inteins in our dataset. (B) The donut chart on the left describes the class of HEN (inner circle), and associated extein (outer circle) for the inteins in our dataset. For ease of reading, the non-LAGLIDADG HEN's are shown in the chart on the right without the canonical LAGLIDADG HEN inteins. (C) Bar chart of the number of inteins found within each phage cluster. (D) The same chart as C but with clusters A1, C1, and E removed.

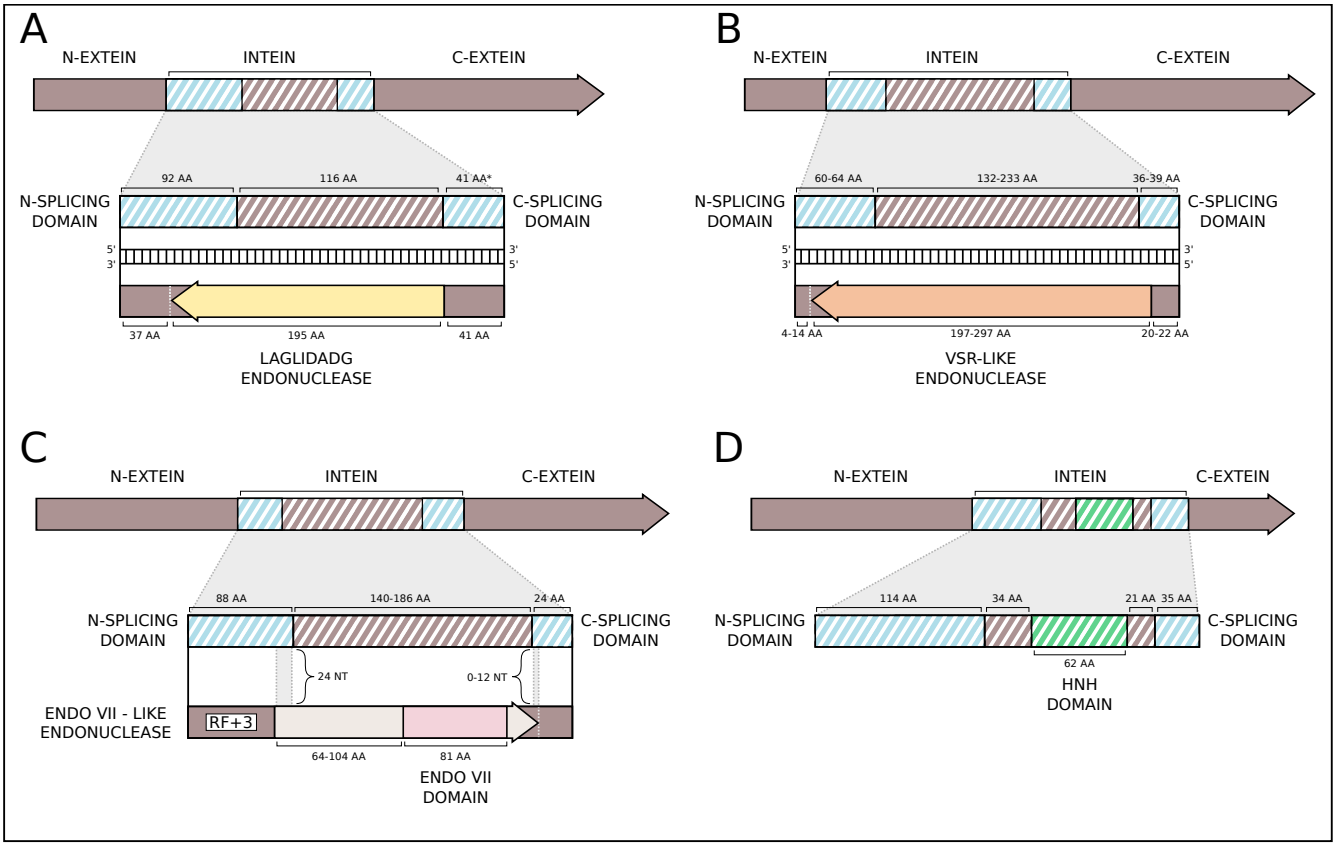

Figure 4: Abstract representations of intein architectures which contain non-canonical HEN's, either at the amino acid level or nucleotide level. Any ranges shown consist of the maximum and minimum lengths, seen in different taxa, for the domains of the in-frame intein, the HEN, or the non-coding sequences surrounding the HEN CDS. (A) Intein with an anti-sense LAGLIDADG HEN shown with its position relative to the intein CDS. (B) Intein with an anti-sense VSR-like HEN shown in the same fashion as A. (C) Intein with an endonuclease-VII-like HEN in the third reading frame. (D) Intein with an in-frame HNH HEN.

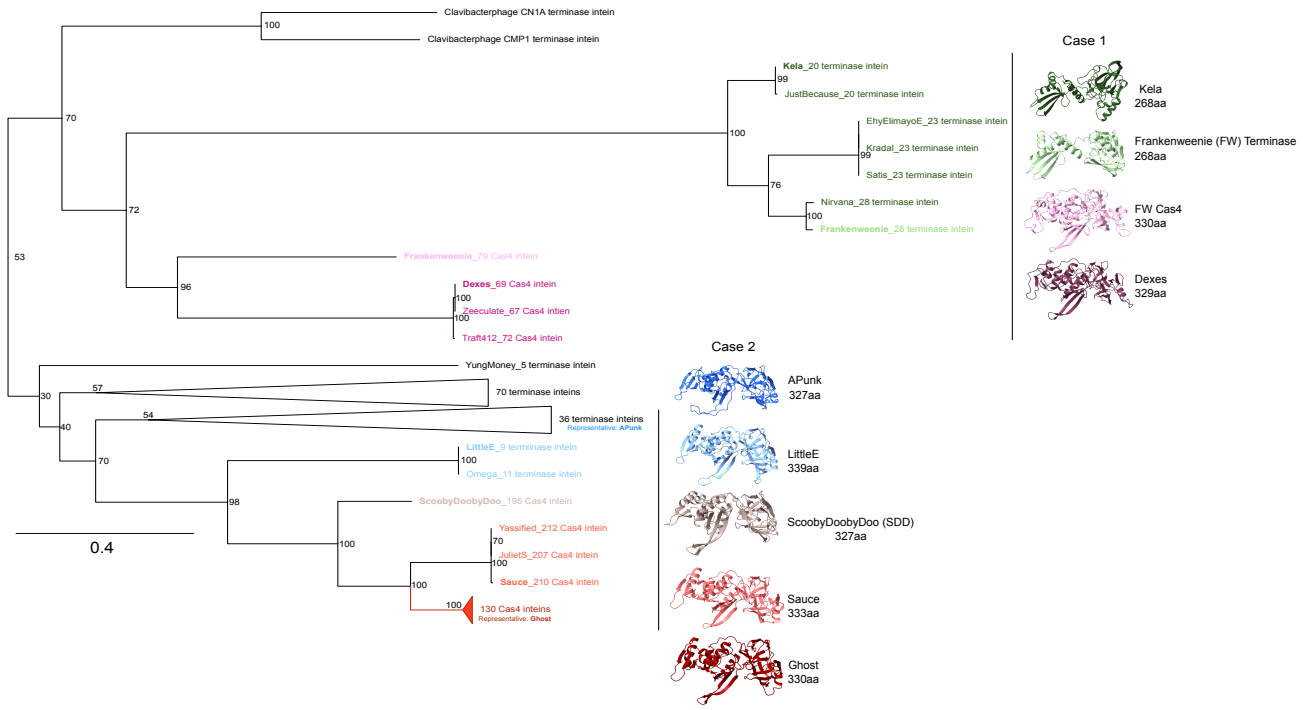

Figure 5: Unrooted maximum-likelihood phylogeny (using WAG+F+R4 as the substitution model) of class 3 inteins (Figure 2 – B) depicting two cases in which a Cas4 intein emerges from a group of terminase inteins. Structures of representatives from each group of inteins were predicted with AlphaFold2. Intein length is listed next to each structure. Non-parametric bootstrap support listed on each node. Scale bar indicates amino acid substitutions per site.

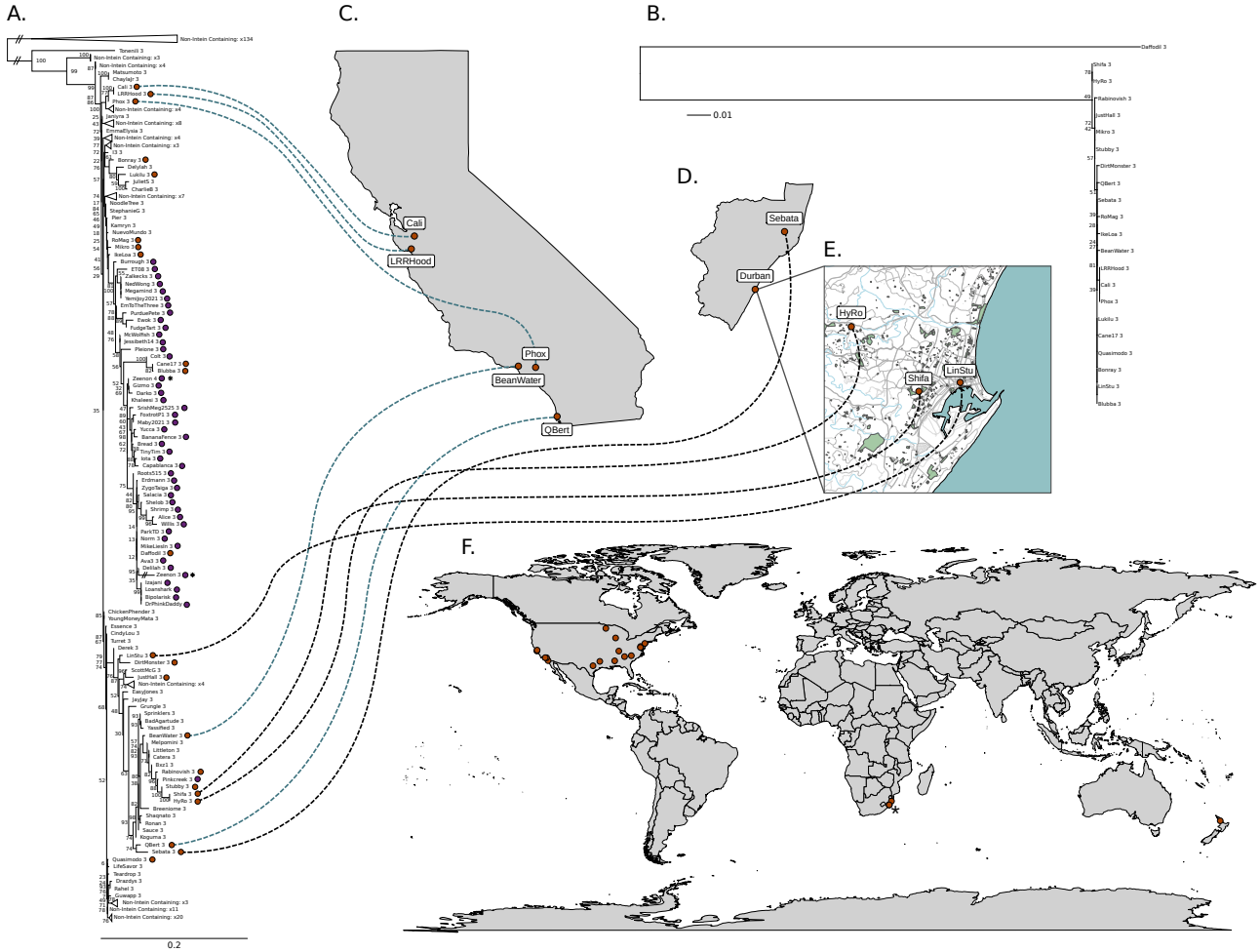

Figure 6: (A) Maximum likelihood phylogeny of in intein free and intein containing nucleotidyltransferase extein encoding nucleotide sequences, using TIM+F+R5 as the substitution model. Double dashes indicate branches shortened for the sake of readability. Colored circles to the right of taxa names indicate intein presence, and the color indicate which of the two clusters the inteins within these exteins belong to. Non-parametric bootstrap support is listed on internal nodes. There is a split intein in this dataset – indicated by a + sign. Scale bar measures substitutions per site. (B) Maximum likelihood phylogeny of the intein encoding sequences from the orange cluster, using HKY+F+I as the substitution model. Bootstrap support is shown as it was in A. Geographic maps show the distribution of inteins from the orange cluster within: California (C), Natal South Africa (D), the city of Durban (E), the whole world (F).

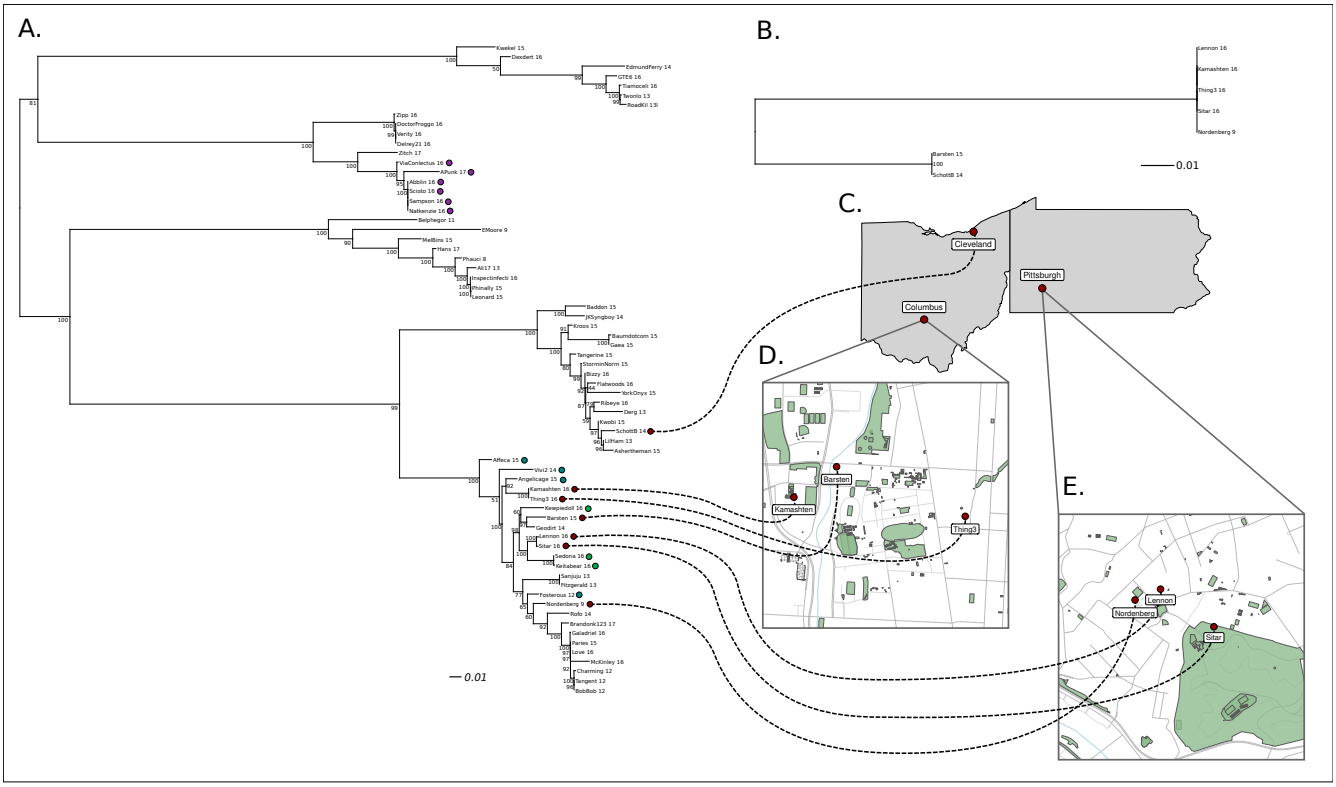

Figure 7: (A) Maximum likelihood phylogeny of the nucleotide sequences of intein containing and intein free minor capsid proteins using TN+F+R3 as the substitution model. Colored circles to the right of taxa names indicate intein presence, and each color corresponds to a different intein cluster. Non-parametric bootstrap support is listed on internal nodes. Scale bar measures substitutions per site. (B) Maximum likelihood phylogeny of the intein nucleotide sequences from the red cluster only using F81+F as the substitution model. Bootstrap support is shown as it was previously. The geographic maps indicate the distribution of inteins from the red cluster in the Ohio Pennsylvania area (C), Ohio State University and surrounding area (D), and the University of Pittsburgh and its surrounding area (E).

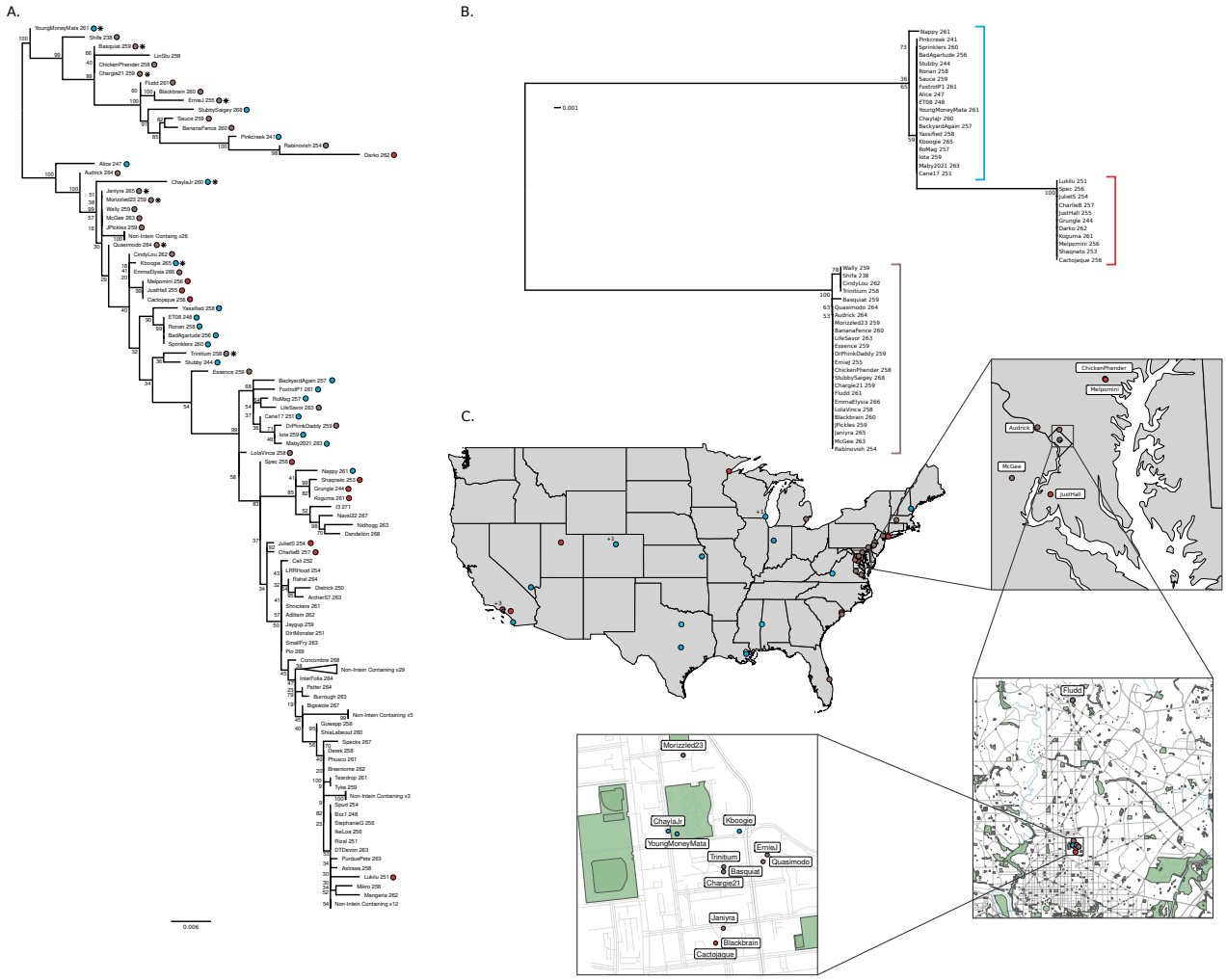

Figure 8: (A) Maximum likelihood phylogeny based on the extein encoding nucleotide sequences of intein free and intein containing putative integrins using TIM2+F+I+G4 as the substitution model. Colored circles to the right of taxa names indicate intein presence, and the color indicates which clan said intein belongs to (see B). Non-parametric bootstrap support is listed on internal nodes. The * indicates that these taxa were isolated within Washington DC. Scale bar measures substitutions per site. (B) Maximum likelihood phylogeny of all intein encoding nucleotide sequences from A, using TN+F as the substitution model. Colored bars on the three clans correspond to the intein colors in A. Bootstrap support is shown as it was previously. (C) The geographic maps indicate the distribution of these inteins. First within the entire United States of America, and then gradually zooming in on Washington DC, and eventually Howard University. Isolation locations where more than one intein containing phage were collected are indicated with a plus sign and the corresponding number of additional inteins. The color of the circles corresponds to the clans in B.

Figure 9: Three different scenarios for attempted intein invasion (right) by the LittleE/Omega (LO) terminase intein (left). Top depicts Cas4/RecB-like gene found in LO lacking the same nucleotide-binding motif that ScoobyDoobyDoo (SDD)'s Cas4 gene has, thus homing of LO terminase intein into this gene is inhibited by lack of similarity. Middle depicts no extremely similar nucleotide binding motif shared between LO terminase and SDD terminase, thus invasion fails. Third depicts LO terminase intein using the nucleotide binding site shared between LO terminase and SDD Cas4 which is a viable intein insertion site to invade SDD's Cas4. Structure of LittleE/Omega's terminase intein predicted using AlphaFold2.

Cluster	Centroid Phage	CDS ID	Members	PhagesDB Clusters	Reviewed Extein AA Annotation	Splice Class	HEN Type
0	P1201	gp22	1	Singleton	thymidylate synthase	1	LAGLIDADG
1	LRRHood	3	22	C1	nucleotidyltransferase	1	LAGLIDADG
2	Lahirium	8	3	CS3	portal protein	1	LAGLIDADG
3	Pinkcreek	135	1	C1	baseplate wedge domain protein	1	LAGLIDADG
4	Job42	69	15	F, F1, F3	DNA methylase	1	LAGLIDADG
5	Belphegor	9	1	DE	portal protein	1	LAGLIDADG
6	Darwin	74	1	EN	ribonucleotide reductase, beta subunit	1	LAGLIDADG
7	Magnito	50	66	A1	phosphoesterase	1	LAGLIDADG
8	Settecandela	2	2	AA	terminase, large subunit	3	LAGLIDADG
9	YungMoney	5	1	DC1	terminase, large subunit	3	LAGLIDADG
10	Kamryn	213	133	C1	cas4 exonuclease	3	LAGLIDADG
11	CN1A	59	1	Singleton	terminase, large subunit	3	LAGLIDADG
12	ZygoTaiga	92	11	C1	thymidylate synthase	1	HNH
13	C3PO	75	7	EN	ribonucleotide reductase, alpha subunit	1	LAGLIDADG
14	GordDuk1	2	1	DF1	terminase, large subunit	1	LAGLIDADG *
18	BobBob	2	7	DE1	terminase, large subunit	1	Endonuclease-VII-like

19	ChaylaJr	260	54	C1	integrin, putative	1	LAGLIDADG
20	JayJay	92	22	C1, C2	portal protein	1	LAGLIDADG
35	ShiLan	65	7	F1, E	DNA methylase	1	LAGLIDADG
36	BlueCrab	67	18	P1, F1, J	DNA methylase	1	LAGLIDADG
37	Petp2012	56	11	A1	DNA primase	1	LAGLIDADG *
38	Kalah2	10	25	J	terminase, large subunit	3	LAGLIDADG
39	Cantare	3	1	Singleton	terminase, large subunit	1	N/A – Mini-Intein
40	JonJames	30	1	DD	terminase, large subunit	3	LAGLIDADG
41	Command613	10	67	E	terminase, large subunit	3	LAGLIDADG
42	ViaConlectus	2	11	DE2, DE4	terminase, large subunit	3	LAGLIDADG
43	Malisha	5	1	DN	minor capsid, MuF-like	1	VSR-like
44	Sejanus	8	1	AD	minor capsid, putative	1	VSR-like
46	Cece	59	1	GD	terminase, large subunit	3	N/A – Mini-Intein
47	Sedona	2	4	DE1	terminase, large subunit	1	Endonuclease-VII-like
48	Astraea	249	114	C1	terminase	1	LAGLIDADG
53	BearBQ	17	1	DN	tape measure protein	1	LAGLIDADG

54	Gizmo	94	5	C1	portal protein	1	LAGLIDADG
55	GordTnk2	2	1	DF1	terminase, large subunit	1	LAGLIDADG *
56	Gmala1	2	1	DF1	terminase, large subunit	3	LAGLIDADG
57	ScoobyDooby Doo	195	1	C2	cas4 exonuclease	3	LAGLIDADG
58	Anglerfish	12	12	A1	terminase, large subunit	3	LAGLIDADG
59	CMP1	11	1	Singleton	terminase, large subunit	3	LAGLIDADG
60	Vivi2	14	4	DE1	minor capsid, MuF-like	1	VSR-like
62	PauloDiaboli	72	4	GD	terminase, large subunit	3	LAGLIDADG
63	Phlyer	6	14	B3	terminase, large subunit	1	Endonuclease-VII-like
69	Colt	3	45	C1	nucleotidyltransferase	1	LAGLIDADG
70	B22	50	1	BW	DNA methylase	1	LAGLIDADG
71	GMA3	2	1	DF2	terminase, large subunit	1	LAGLIDADG *
72	Scioto	16	6	DE4	minor capsid, MuF-like	1	VSR-like
73	Phoebus	158	6	J	Rtcb-like RNA ligase	1	LAGLIDADG
74	Kela	20	7	BM	terminase, large subunit	3	LAGLIDADG +
79	Lederberg	59	1	EP	tRNA splicing RNA ligase	1	LAGLIDADG +
80	Nebkiss	2	2	X	terminase, large	3	LAGLIDADG

					subunit		
81	Traft412	72	3	A1	cas4 exonuclease	3	LAGLIDADG
82	REQ3	37	1	Singleton	minor capsid, MuF-like	3	VSR-like
83	Sticker17	8	17	CS3	portal protein	1	LAGLIDADG
84	Philon9	12	1	CS4	portal protein	1	LAGLIDADG
87	GMA6	3	1	DO	terminase, large subunit	1	LAGLIDADG
88	Omega	11	2	J	terminase, large subunit	3	LAGLIDADG
89	Shaqnato	180	9	C1	DNA helicase	3	LAGLIDADG
90	Frankenweenie	79	1	BM	cas4 exonuclease	3	LAGLIDADG
91	Kewpiedoll	16	3	DE1	minor capsid, MuF-like	1	VSR-like
92	Pumpernickel	75	1	Singleton	terminase, large subunit	3	LAGLIDADG +
102	P1201	27	7	EN, Singleton	terminase, large subunit	1	LAGLIDADG – Mini and Full
103	P1201	78	1	Singleton	ribonucleotide reductase, alpha subunit	1	N/A – Mini-Intein
104	Mcklovin	18	2	CZ4	tape measure protein	1	LAGLIDADG
105	Vincenzo	5	2	B4	terminase, large subunit	1	Endonuclease-VII-like
106	Wojtek	72	1	DV	DNA polymerase, E-like	1	HNH

107	Nordenberg	9	7	DE1	minor capsid, MuF-like	1	LAGLIDADG Anti-sense
108	Clown	5	1	DC	terminase, large subunit	1	HNH

Table 1: Descriptive table summarizing information regarding the centroid sequences for the intein clusters generated during this study. * indicates inteins where the HEN might be a LAGLIDADG HEN, but evidence was inconclusive. + indicates inteins where the HEN might be decaying, but evidence was inconclusive.