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

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 — Email: gogarten@uconn.edu.

Author Contributions: Conceptualization: S.P.G., D.A., J.P.G.; Methodology: S.P.G., D.A., J.P.G.; Software: S.P.G.; Validation: D.A.; Formal Analysis: S.P.G., D.A.; Investigation: S.P.G, D.A.; Data Curation: S.P.G., D.A.; Writing: S.P.G., D.A., J.P.G.; Editing & Review: S.P.G., D.A., J.P.G.; Visualization: S.P.G., D.A.; Supervision: J.P.G.

Competing Interest Statement: The authors claim no competing interests.

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 large subunit 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 domain. 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.

33 ## Introduction

34 Inteins are self-splicing, site-specific, mobile genetic elements that invade protein-encoding genes
35 across all domains of life and viruses, including bacterial viruses (phages) (Gogarten et al. 2002; Lennon
36 and Belfort 2017; Nanda et al. 2020). An intein begins its journey as an intervening DNA sequence at a
37 specific insertion site within the coding sequence of a host gene. Unlike the intervening DNA sequences
38 known as introns which are post-transcriptionally removed at the RNA level, an intein is transcribed and
39 translated together with its host gene. The intein then seamlessly excises itself post-translationally from
40 the host protein, referred to as the extein, through the unique process of protein splicing (Shih et al. 1988;
41 Hirata et al. 1990; Kane et al. 1990). Protein splicing, performed by the self-splicing domain of the intein,
42 yields a functional extein and the free intein. Inteins are divided into three classes depending on their
43 initial steps of splicing, the residues involved, and differences in reaction chemistry (Tori et al. 2010;
44 Nanda et al. 2020).

45 Once excised from the extein, the intein uses its homing endonuclease (HEN) domain to engage
46 in the self-propagation method known as homing (Goddard and Burt 1999). When in the presence of a
47 copy of its host gene lacking the intein's DNA sequence at its particular insertion site, the HEN makes a
48 double-strand break at the intein insertion site. Through homologous recombination-based repair using
49 the intein-containing copy of the gene as the template, the intein sequence is copied into the previously
50 vacant insertion site (Belfort and Roberts 1997). Homing allows inteins to rapidly rise to saturation in
51 populations, largely through horizontal transfer-based homing, at which point there is no longer selective
52 pressure to maintain the HEN (Goddard and Burt 1999). The HEN degrades, turning the full intein into a
53 mini intein which contains only the self-splicing domain (Iwaï et al. 2017). Sometimes, an intein may
54 acquire an internal stop codon in the HEN as it degrades, generating a split intein (Tavassoli and
55 Benkovic 2007; Dassa et al. 2009; Hoffmann et al. 2021). Eventually, the mini and split inteins are lost
56 and the intein-free allele once again dominates. Newer models propose the co-existence of full, mini,
57 split, and empty alleles as opposed to synchronized progression through the phases (Gogarten and
58 Hilario 2006; Yahara et al. 2009; Barzel et al. 2011). As an alternate evolutionary endpoint to degradation
59 and loss, some inteins have also been inferred to have acquired novel functions, acting as regulatory

60 elements under stressful conditions (Novikova et al. 2016; Lennon et al. 2018) or evolving into mating-
61 type switching endonucleases such as HO of *Saccharomyces cerevisiae* (Coughlan et al. 2020).

62 Several different HEN domain types have been reported in inteins. Most inteins possess
63 canonical LAGLIDADG homing endonucleases, which are in-frame with the rest of the intein (Gorbalenya
64 1998; Gimble 2000). Others include: inteins with an in-frame HNH endonuclease (Dalgaard, Klar, et al.
65 1997), split-inteins possessing VSR-like endonucleases (Dassa et al. 2009), and LAGLIDADG HENs
66 encoded on the anti-sense strand (Gorbalenya 1998; Gogarten et al. 2002). Interestingly, most of these
67 non-canonical inteins are found within phage genomes, with actinobacteriophages serving as a particular
68 hot-spot in research.

69 Actinobacteriophages are promising therapeutic alternatives to antibiotics in the treatment of
70 multi-drug resistant actinobacterial infections (Allué-Guardia et al. 2021; Zeynali kelishomi et al. 2022;
71 Dedrick et al. 2023; Yang et al. 2024). However, much is still unknown about actinobacteriophage
72 structure, function, and genomic evolution, necessitating increased studies in these areas. In recent
73 years, teaching programs (SEA-PHAGES and PHIRE) employing citizen-science based approaches to
74 isolate and sequence actinobacteriophages have added thousands of genomes to databases such as
75 PhagesDB (Russell and Hatfull 2017). PhagesDB thus serves as an ideal candidate for large-scale
76 computational analyses aiming to dissect phage genomic evolutionary dynamics, with inteins serving as
77 particularly powerful tools for tracing genetic exchange between phage populations (Soucy et al. 2014).

78 Inteins are present within actinobacteriophages (Tori et al. 2010; Tori and Perler 2011; Kelley et
79 al. 2016). However, since the last survey of these inteins (Kelley et al. 2016) the PhagesDB database has
80 expanded from 841 to 5075 genomes; providing a unique resource to study the distribution and evolution
81 of phage inteins. Hence, we conducted a new survey of PhagesDB, accessed in September of 2023. To
82 capture more divergent inteins within the database, we developed an algorithm, ICE-BLAST, which uses
83 a combination of PSI-BLAST and USEARCH clustering to iteratively search an amino acid database
84 using the representative sequences of clusters (centroids) of matches from previous searches. Herein we
85 discuss the development and testing of this algorithm, and the inteins discovered through its application
86 to PhagesDB. We investigate intein distribution within phage clusters, variation in host genes, and

87 geographic dissemination. Through the latter, we found evidence supporting events of rapid
88 geographically localized propagation of intein alleles into new populations. Our phylogenetic analyses
89 also uncovered two occurrences of class 3 inteins of terminase large subunits sharing a common
90 ancestor with Cas 4 exonuclease inteins. Additionally, we report several cases of inteins with HEN
91 domains and architectures that have rarely been reported on, or were previously unseen. These include
92 HENs encoded within a separate ORF on the same strand as the intein, and HENs in a separate ORF on
93 the anti-sense strand.
94

Results

Intein Distribution across Actinobacteriophage Clusters

Using our program ICE-BLAST, tested using simulated data from our program Invader-Sim, we identified 784 putative intein sequences in PhagesDB (details on the intein retrieval and validation process are provided under Materials and Methods). A table containing information on all 784 inteins including their host phage, host protein (extein), splicing class, and HEN type is available (Supplemental Table 1). These 784 inteins span a wide variety of PhagesDB phage clusters (Figure 1), with the majority being in clusters C1 (415), A1 (92), and E (68), all of which are listed as mycobacteriophage clusters. The other 209 inteins were distributed across 31 phage clusters and several singleton phages (phages not assigned to a cluster). Of the 31 phage clusters, the maximum number of inteins observed in one cluster was 35, and the majority of clusters contained less than 10 inteins. The most intein-rich genomes were all from cluster C1 phages, with phage Cane17 having seven inteins, followed by phages Blubba and Colt with six inteins each.

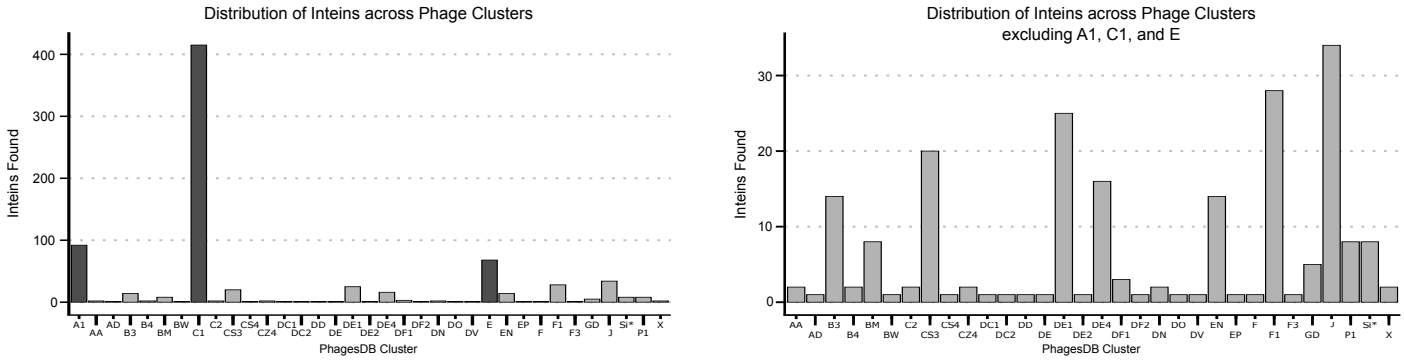


Figure 1. Total number of inteins found in each PhagesDB phage cluster, depicted with (left) and without (right) intein-rich clusters A1, C1, and E. A1, C1, and E are depicted with dark gray bars and contained 92, 415, and 68 inteins respectively. Singletons are denoted as Si*.

Wide Range of Actinobacteriophage Proteins Invaded by Inteins

By re-annotating the host protein sequences of the 784 putative inteins, we found that they invade 19 distinct proteins (Figure 2). Unsurprisingly, the majority were found in proteins previously reported as containing inteins. These include: Cas4 exonuclease (InterPro IPR013343), DNA helicase

116 (InterPro IPR006500), DNA methylase N-6/N-4 (Pfam PF01555) and C-5 (InterPro IPR001525), DNA
117 primase (InterPro IPR006171), nucleotidyltransferase (InterPro IPR018775), phosphoesterase (InterPro
118 IPR029052), portal protein (InterPro IPR009279), terminase large subunit (Pfam PF04466), and
119 thymidylate synthase (InterPro IPR003669). The remaining host proteins are new to this study within
120 actinobacteriophages, and include: DNA polymerase III alpha subunit (InterPro IPR011708), minor capsid
121 protein (InterPro IPR006528), ribonucleotide reductase subunits alpha (InterPro IPR026459) and beta
122 (InterPro IPR026494), RtcB-like RNA ligase (InterPro IPR021122), tape measure protein (InterPro
123 IPR013491), and tRNA splicing RNA ligase (InterPro IPR036025). Two types of intein-containing exteins
124 identified could not be definitively annotated, but have best matches to baseplate wedge domain
125 containing proteins (UniProt M9MUF4) and integrin-like proteins respectively. Based on the annotations
126 of other closely related intein-containing exteins, the former could be a portal protein of some sort, but
127 does not share considerable similarity to known portal proteins in any NCBI database. The inteins present
128 within the putative integrin-like exteins were previously reported on in Kelley et al. 2016, but with fewer
129 intein-containing exteins (Kelley et al. 2016).

130 **Variation in Intein Splicing Domains**

131 The putative inteins were clustered at 70% sequence identity. For each intein cluster, we
132 determined the intein splicing class (1, 2, or 3). Most of the inteins were class 1 (497) while the remainder
133 (287) were class 3. The class 1 inteins displayed a much wider range of host proteins (16 different
134 proteins) than the class 3 inteins which were only found within 4 different extein hosts. More than half of
135 the inteins within our dataset (431) invaded just two proteins, with 293 inteins within terminase large
136 subunit sequences, and 138 within Cas4 exonucleases (Figure 2). The size of class 1 inteins ranged from
137 385 amino acids (aa) (terminase large subunit of cluster DC2 phage Clown) to 134 aa (terminase large
138 subunit of cluster EN phage C3PO). The size of class 3 inteins ranged from 359 aa (DNA helicase of
139 cluster C1 phage Alice) to 225 aa (minor capsid of singleton phage REQ3). In addition to classifying the
140 splicing domain of each intein, we classified their HEN domains, which revealed further diversity and
141 previously undocumented intein architectures.

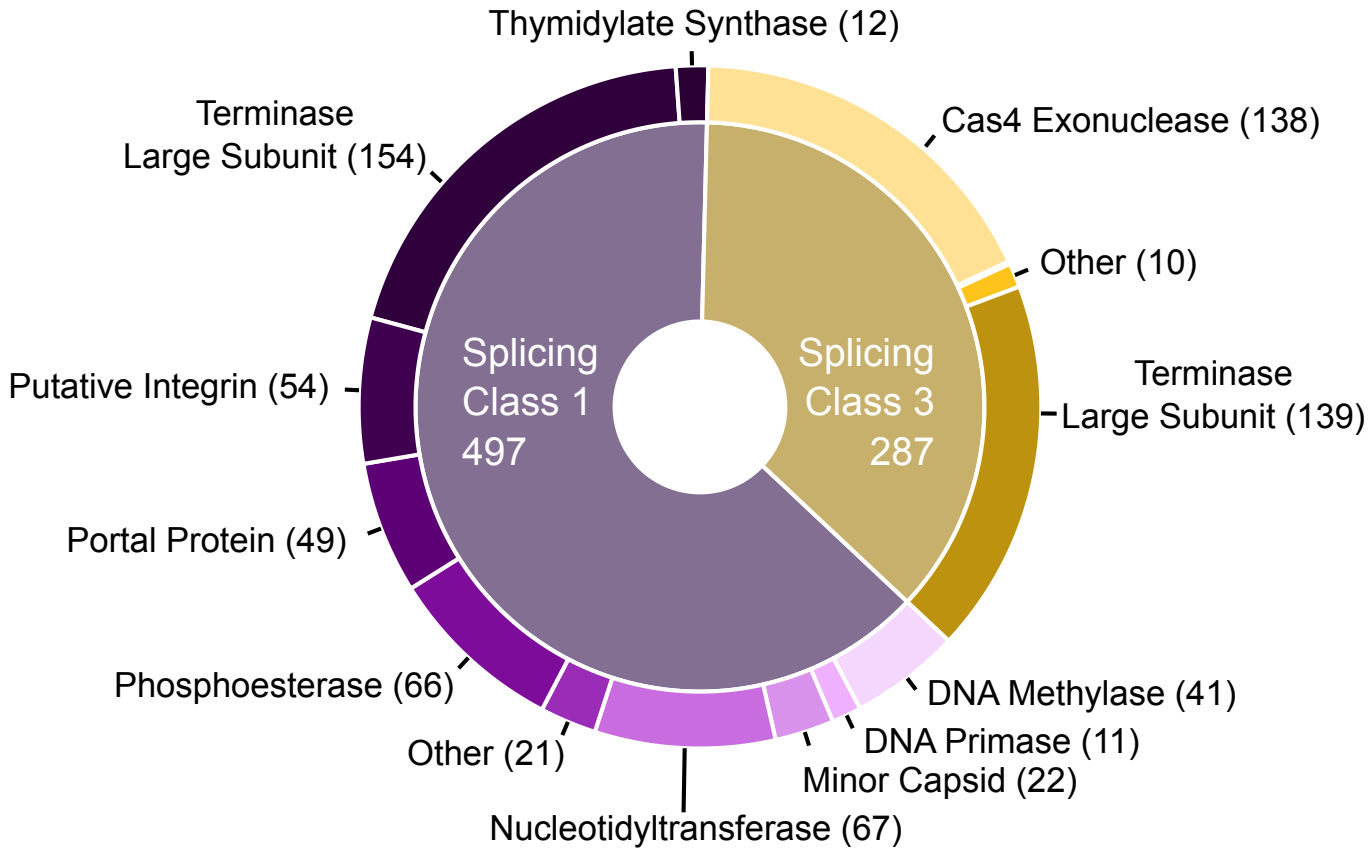


Figure 2. Identified inteins, which fall into class 1 and class 3 based on their splicing domains, are distributed across 19 different host genes: baseplate wedge domain protein (InterPro M9MUF4), Cas4 exonuclease (InterPro IPR013343), DNA helicase (InterPro IPR006500), DNA methylase N-6/N-4 (Pfam PF01555) and C-5 (InterPro IPR001525), DNA polymerase III alpha subunit (InterPro IPR011708), DNA primase DnaG-like (InterPro IPR006171), putative integrin-like protein, minor capsid protein (InterPro IPR006528), nucleotidyltransferase DNA polymerase beta superfamily (InterPro IPR018775), phosphoesterase (InterPro IPR029052), portal protein (InterPro IPR009279) ribonucleotide reductase alpha subunit (InterPro PR026459) and beta subunit (InterPro IPR026494), RtcB-like RNA ligase (InterPro IPR021122), tape measure protein (InterPro IPR013491), terminase large subunit (Pfam PF04466), thymidylate synthase (InterPro IPR003669), and tRNA-splicing ligase (InterPro IPR036025).

Variation in Intein Homing Endonuclease (HEN) Domains

Most of the inteins in our database (712) possess classic intein architecture: a self-splicing domain and a canonical LAGLIDADG (InterPro IPR004860) HEN domain. Of the remaining 72 inteins, nine have decaying or completely degraded HENs (mini inteins), and 63 possess unique non-canonical HENs. These non-canonical HENs include in-frame HNH (InterPro IPR002711) HENs (Figure 3A),

158 LAGLIDADG or VSR-like (InterPro IPR004603) HENs encoded on the anti-sense strand (Figure 3B-C),
159 and out-of-frame endonuclease VII-like (InterPro IPR004211) HENs (Figure 3D). The inteins possessing
160 these non-canonical HENs are distributed across four host protein types: DNA polymerase, minor capsid,
161 terminase large subunit, and thymidylate synthase (Figure 3E). Based on current knowledge, these four
162 non-canonical HEN architectures are rare, with some being reported on for the first time with this work.
163 The non-canonical HENs which have already been reported on include the in-frame HNH HEN which is
164 observed in phages (Dalgaard, Klar, et al. 1997; Dalgaard, Moser, et al. 1997), the anti-sense
165 LAGLIDADG HENs (Gorbalenya 1998; Gogarten et al. 2002), and the VSR-like HEN which was reported
166 in the context of split inteins with no other known intein-associated reports (Dassa et al. 2009). The fourth
167 type of non-canonical HEN we found, an endonuclease VII-like HEN encoded within the HEN region but
168 in frame 3 relative to the rest of the intein, has not yet been reported to our knowledge.

169 To further assess the identities of the out-of-frame and anti-sense non-canonical HENs,
170 AlphaFold-predicted structures of the inteins (Supplemental Figure 1) and the HENs in their respective
171 frames (Supplemental Figure 2) were compared to solved structures. Based on the predicted structures
172 for the endonuclease VII-like and anti-sense LAGLIDADG HENs, coupled with the fact that these HENs
173 can be dimeric, the terminal domains of these structures may be linker domains or secondary DNA
174 binding domains (Raaijmakers et al. 1999; McMurrough et al. 2018). The VSR-like HEN appears to
175 contain a long stretch of linked turns and alpha helices not present in normal VSR endonuclease
176 structures (Hennecke et al. 1991). The predicted local distance difference test (pLDDT) — a per-residue
177 measure of local prediction confidence (Jumper et al. 2021) — for this stretch is relatively high, especially
178 within helices, but the region lacked meaningful matches against PDB structures. However, an HHPred
179 search found a significant match (e-value of 0.000046) against protein structure 3R3P from Bacillus
180 phage 0305phi8-36, which indicates the VSR-like HEN may be an EDxHD HEN (Taylor et al. 2011).

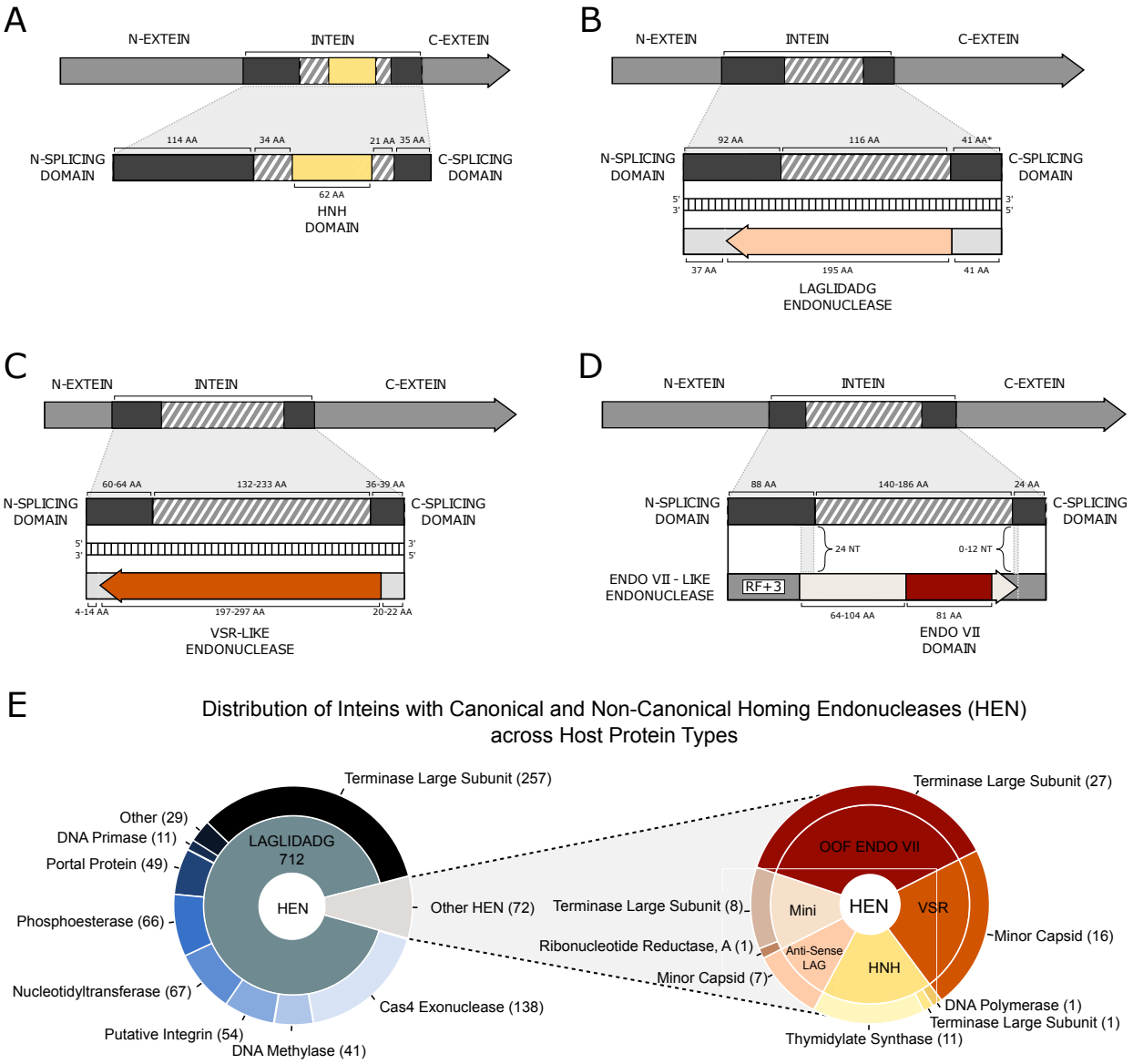


Figure 3. A-D. Gene-map depictions of the four observed non-canonical intein HEN organizations. Regions shaded with diagonal lines indicate HEN-encoding region. **A.** in-frame HNH HEN. **B.** anti-sense LAGLIDADG HEN. **C.** anti-sense VSR-like HEN. **D.** endonuclease VII-like HEN in reading frame 3 relative to intein. **E.** Distribution of inteins with canonical and non-canonical HEN architectures across host protein types.

Actinobacteriophage Intein Phylogenetics

Having characterized the structural architectures of each intein, we sought to investigate how these inteins have evolved with respect to their host proteins and host phages. Performing such evolutionary characterizations of inteins can often unveil pathways for genetic exchange within

191 populations that may have gone unnoticed when focusing on genetic material with primarily orthodox
192 vertical inheritance (Soucy et al. 2014; Turgeman-Grott et al. 2023). Thus, we constructed phylogenies
193 from the DNA sequences of all inteins in our database. The extracted subsets of intein sequences used to
194 generate these phylogenies were separated by HEN type, with the largest set (canonical HENs) split into
195 two sets based on splicing class (class 1 or class 3). By analyzing the grouping patterns of inteins within
196 the phylogenies compared to the types of host proteins they invade, potential instances of intein host
197 protein switching (e.g. a terminase large subunit-targeting intein evolving into a Cas4 exonuclease-
198 targeting intein) can be identified.

199 While some inferences could be drawn from the phylogenies of the inteins with non-canonical
200 HEN domains, current sampling is ultimately too sparse to make any definite claims from the intein
201 phylogenies alone (Supplemental Figures 3-6). In contrast, the inteins with canonical HENs produced
202 more informative phylogenies.

203 The class 1 inteins with a canonical LAGLIDADG HEN produced a phylogeny which provided
204 evidence for several independent events of intein integration site diversification events (Figure 4). For
205 example, the terminase large subunit inteins form a polyphyletic clan with at least four separate groups.
206 Each group is separated, divided from the other terminase large subunit groups by multiple highly
207 supported splits along the inner nodes. Similarly, portal protein inteins and DNA methylase inteins form
208 polyphyletic clans with five and three separate groups respectively.

209 The class 3 inteins with canonical LAGLIDADG HENs are almost entirely found within Cas4
210 exonuclease and terminase large subunit exteins (Figure 4). Within this phylogeny are two instances of
211 Cas4 associated inteins diverging from terminase large subunit inteins. One of these groupings had
212 particularly high bootstrap support, and was thus further investigated as a potential case of extein
213 switching (an intein evolving to target an insertion site inside of a new type of host protein).

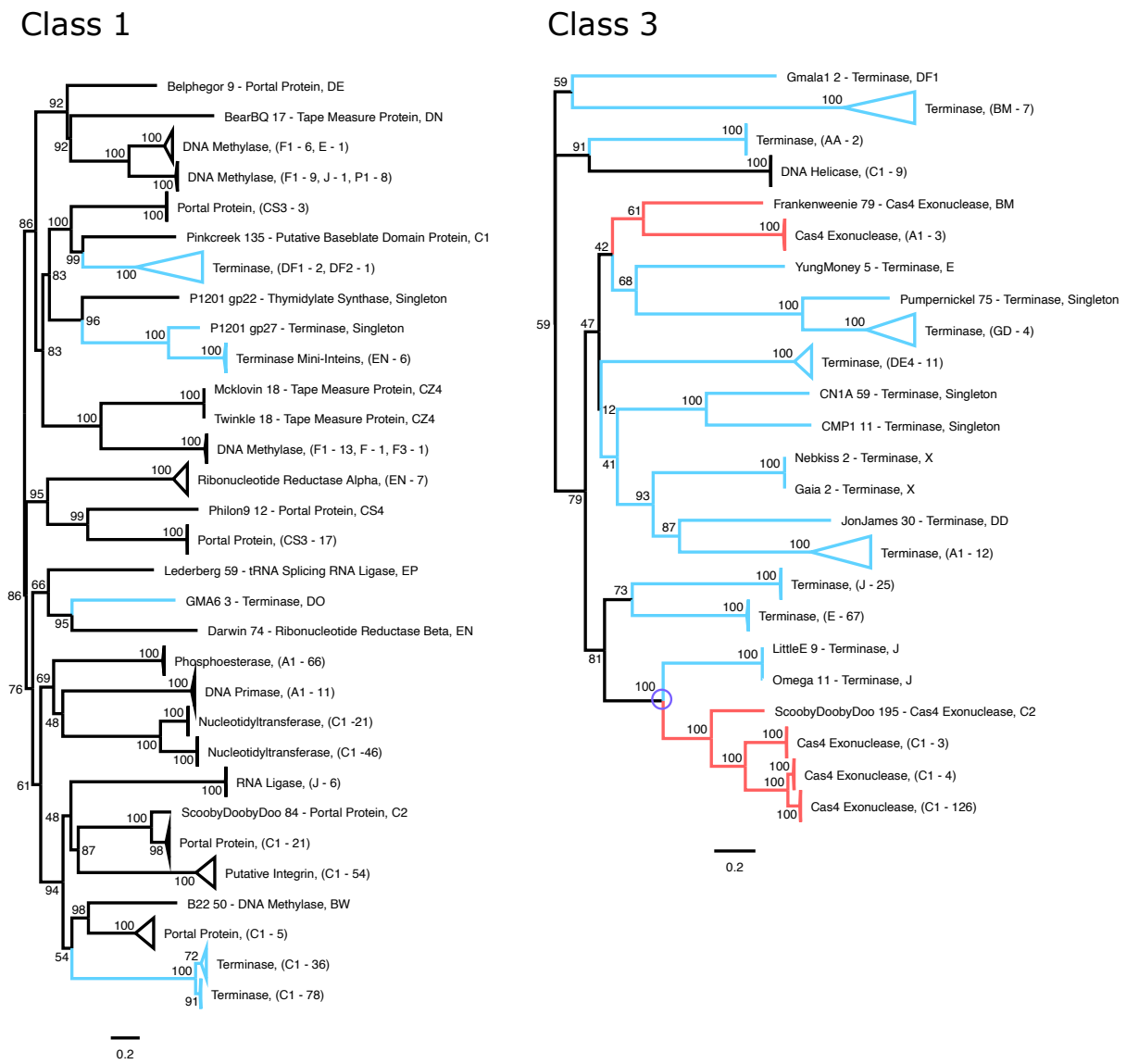


Figure 4. The DNA sequences of the inteins with canonical LAGLIDADG HEN domains were split into two subsets by splicing class (class 1 and class 3), with each subset then used for maximum likelihood phylogenetic reconstruction. Substitution models used were TIM2+F+R3 (class 1) and TPM2+F+R4 (class 3). Ultrafast bootstrap support is listed on internal nodes. Scale bar measures substitutions per site. Clans of inteins within the same extein are collapsed; clade labels indicate the number of inteins in the collapsed group, and which phage clusters they are found in. Branches leading to terminase large subunit inteins are colored in blue, and branches leading to Cas4 exonuclease inteins are colored in red. On the class 3 phylogeny, the potential extein switching event between terminase large subunit and Cas4 exonuclease inteins (involving phages LittleE and ScoobyDoobyDoo) further investigated is indicated with a purple circle.

225 **Extein Switching**

226 The class 3 intein phylogeny (Figure 4) depicts two instances of a clade of Cas 4 exonuclease
227 inteins grouping inside of a clade of terminase large subunit inteins. This grouping suggests that an intein
228 which now targets an insertion site in Cas4 exonucleases evolved from an intein which targets an
229 insertion site in terminase large subunits. We will refer to this type of evolutionary event as extein
230 switching. To assess whether this extein switching topology was a phylogenetic artifact, an
231 approximately unbiased (AU) tree topology test (Shimodaira 2002) was used to gauge the strain induced
232 when one attempts to force the terminase large subunit and Cas4 exonuclease inteins to group by extein
233 type, as opposed to the observed topology. The same model determined by IQ-TREE's ModelFinder for
234 the original class 3 intein phylogeny (WAG+F+R4) was used for the AU test. The test produced little
235 support for the grouping of these inteins by extein type, (bootstrap proportions of 0.998 and 0.002 for tree
236 constrained by extein switching event and the tree constrained by extein type respectively (Supplemental
237 Table 2)). After investigating both putative extein switching events, the event involving phages LittleE and
238 ScoobyDoobyDoo provided a more complete picture of how the extein switching may have occurred. The
239 terminase large subunit inteins of the other extein switching event (involving phage Frankenweenie
240 (Figure 4)) all exhibited degradation in their HEN domains, making comparison of predicted intein
241 structures for this group less informative. Thus, the event involving phages LittleE and ScoobyDoobyDoo
242 will be the focus of the presented investigation that follows.

243 **Extein Switching from Terminase Large Subunit to Cas4 Exonuclease**

244 Comparing the +/-30 nucleotides at the intein insertion site across the LittleE terminase large
245 subunit and ScoobyDoobyDoo (SDD) Cas4 exonuclease reveals regions of striking similarity
246 (Supplemental Figure 7A). Both insertion sites reside in stretches of DNA sequence which encode
247 important nucleotide binding sites for their respective proteins (Supplemental Figure 7B). Comparison to
248 the solved structure of a phage terminase large subunit bound to ATP (Zhao et al. 2013) shows the intein
249 insertion site encodes a region on an alpha-helix within the Rossmann fold constituting the nucleotide
250 binding site within the ATPase domain of the terminase large subunit. The SDD Cas4 exonuclease intein
251 insertion site encodes the motif EFKT, which is the RecB-like motif-III in the Cas4 exonuclease, and is

252 important for nucleotide binding (Shiimori et al. 2018). In addition to assessing insertion site sequence
253 similarity, we compared the predicted protein structures for the inteins. Alignment of the predicted
254 structures generated for the terminase large subunit intein of LittleE, the Cas4 exonuclease intein of SDD,
255 and the terminase large subunit intein of cluster DE4 phage APunk (representing a clade of terminase
256 large subunit inteins which do not group with LittleE's terminase large subunit intein) reveals them all to
257 be near identical, barring APunk (Figure 5). These results align with the expected results based on the
258 phylogeny. The APunk intein does not align as strongly to the LittleE and SDD inteins as they do to each
259 other, specifically exhibiting a slight structural difference in its HEN (Figure 5). This difference in HEN
260 structure between the APunk and LittleE terminase large subunit inteins in particular reflects how they
261 target two separate regions of the terminase large subunit gene (both intein insertion site regions encode
262 alpha-helices in the ATPase domain, but two distinctly different helices). In contrast, the high similarity
263 between the HEN domains of the SDD and LittleE predicted intein structures further supports the notion
264 that these inteins can recognize and cut an extremely similar intein insertion site. To explain the lack of
265 intein in SDD's terminase large subunit, we later propose a model in which the SDD terminase large
266 subunit was not a viable host for the LittleE terminase large subunit intein, and the SDD Cas4
267 exonuclease provided a viable alternate insertion site.

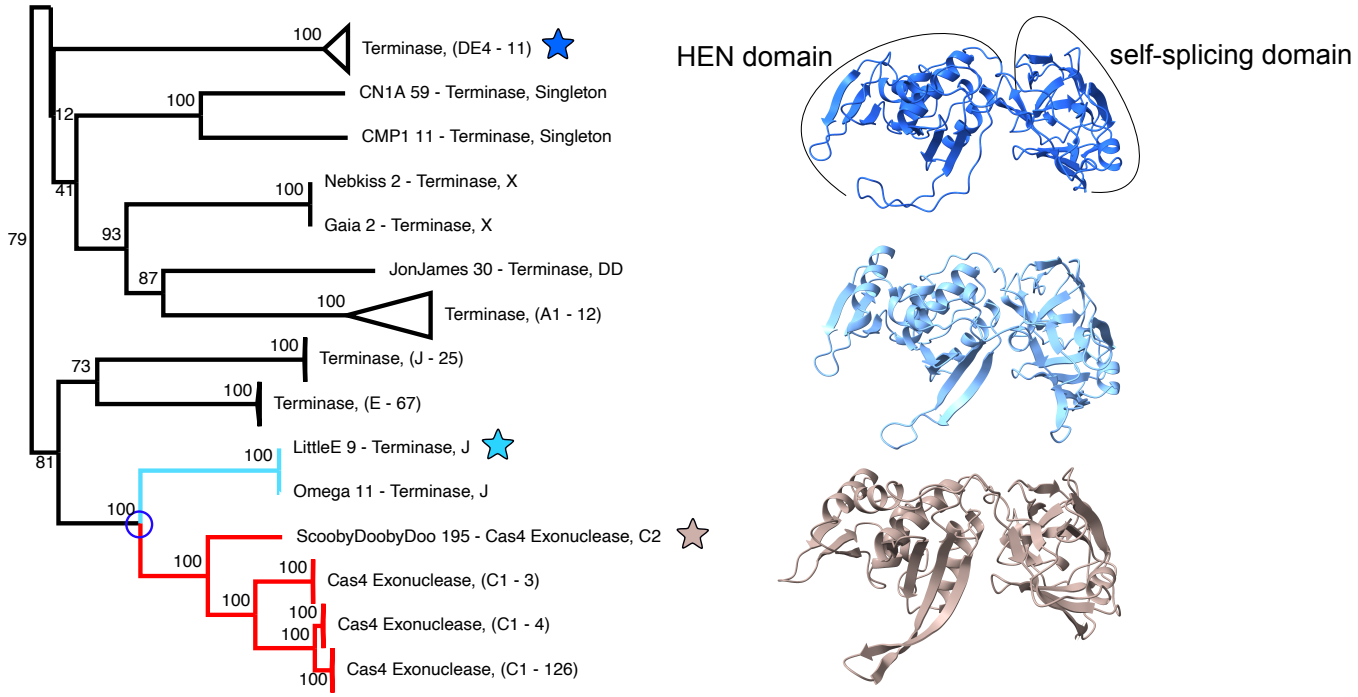


Figure 5. Subtree from phylogeny of class 3 inteins (Figure 4) depicting one instance in which a Cas4 exonuclease intein emerges from a group of terminase large subunit inteins. For each intein indicated with a star (cluster DE4 representative phage APunk terminase large subunit intein, LittleE terminase large subunit intein, and ScobyDoobyDoo (SDD) Cas4 exonuclease intein) a predicted protein structure was generated with AlphaFold. The predicted intein structures (right) are APunk (top) LittleE (middle) and SDD (bottom). Alignments of the predicted intein structures using the Matchmaker tool in ChimeraX produce the following results: LittleE aligned to SDD yields a sequence alignment score of 995.6, and RMSD of 2.716; APunk aligned to LittleE yields a sequence alignment score of 644.9, and RMSD of 3.097; APunk aligned to SDD yields a sequence alignment score of 643.1, and RMSD of 2.994.

Intein Dissemination within Local Areas

Typically, copies of a particular intein-invaded extein in different individuals will have higher sequence similarity than the inteins that invade them. This is due to the types of genes invaded by inteins often being of vital function to their hosts and thus highly conserved, as well as the increased evolutionary rates experienced by homing elements such as inteins (Goddard and Burt 1999). We further discuss the relative sequence conservation and selection pressures on intein and extein sequences in Supplemental

285 Materials under the section Supplemental Results and Discussion: Sequence Conservation in Inteins and
286 Exteins. Thus, two individuals having less similar exteins but nearly identical inteins is cause for further
287 investigation, as this can indicate a recent transfer event since the intein has not yet had time to diverge
288 from its origin copy despite the higher evolutionary rates of inteins. Geographic data regarding host phage
289 isolation paired with these events can be further investigated to assess the locality of such recent
290 invasion events.

291 Following this line of analysis, we previously found inteins which appeared to be proliferating into
292 divergent phage extein sequences within local geographic areas (Gosselin et al. 2023). We chose to
293 investigate our dataset for additional cases of local dissemination and proliferation of inteins into new
294 phage populations. We filtered our data in two stages: first to find clusters of closely related inteins
295 (97.5% similarity), then to discover clusters within the filtered set whose extein sequences were more
296 divergent from each other than their respective intein sequences. These clusters were used to construct
297 phylogenetic trees, and their geographic positions were plotted using metadata from PhagesDB. Of the
298 resulting 13 clusters of intein sequences, we focused on three cases which displayed high levels of
299 incongruence between the extein and intein phylogenies during a visual inspection, suggesting recent
300 intein transfer between divergent hosts.

301 First, we investigated a set of inteins which were found in nucleotidyltransferases (Figure 6). This
302 group of inteins had a global distribution, and of particular interest to us, two groups of co-localized
303 inteins. One group of 5 inteins isolated from California, and a second group found in Durban, South
304 Africa. In both of these locations, inteins exhibited nearly identical nucleotide sequences, but were found
305 in comparatively divergent exteins. Furthermore, these extein sequences were separated in the extein
306 phylogeny by non-intein containing sequences.

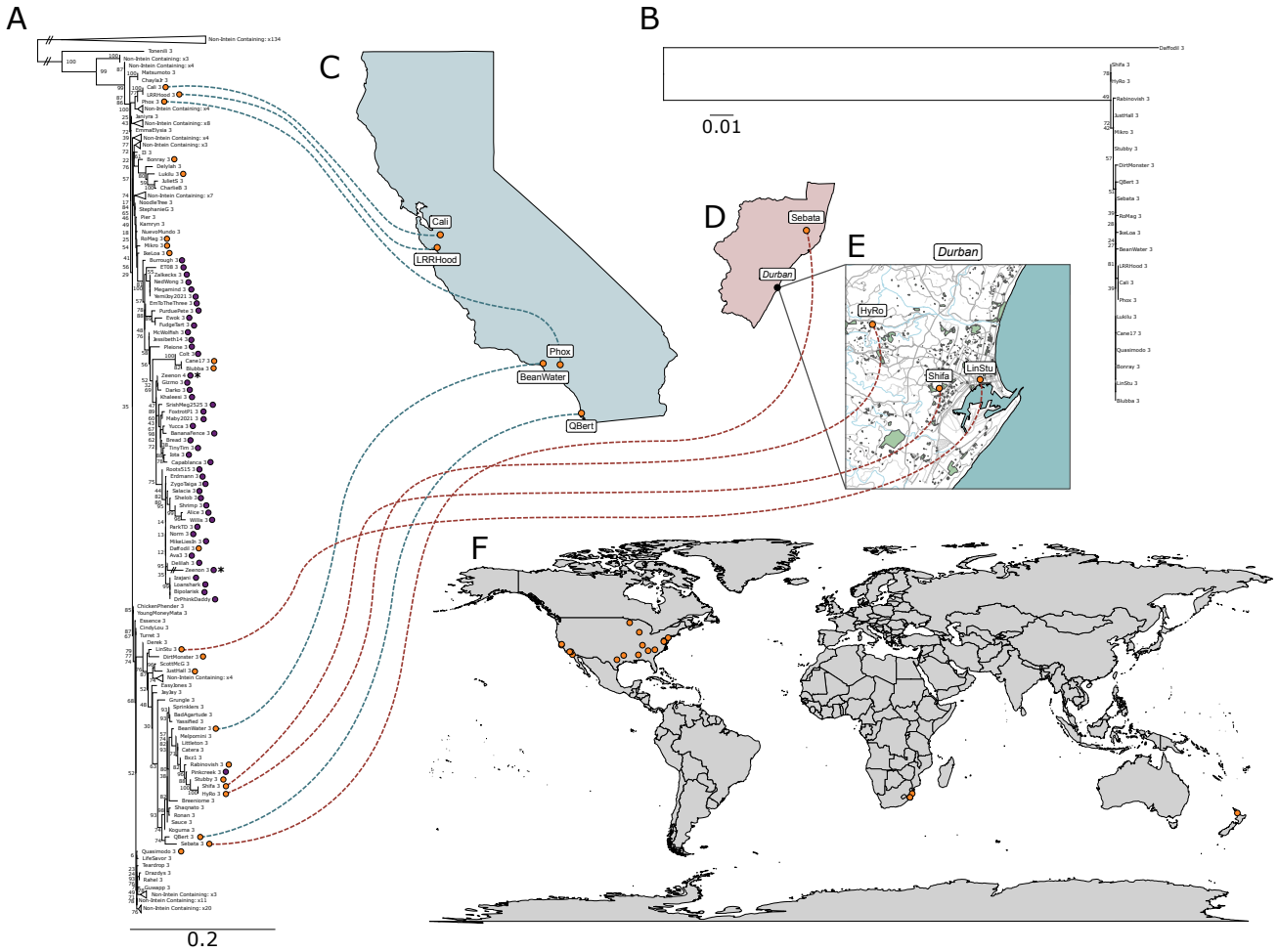


Figure 6. **A.** Maximum likelihood phylogeny of intein-free and intein-containing nucleotidyltransferase extein-encoding nucleotide sequences, using TIM+F+R5 as the substitution model. Double slashes indicate branches shortened for readability. Circles to the right of taxa names indicate intein presence, with orange and purple denoting two different intein clusters. Ultrafast bootstrap support is listed on internal nodes. There is a split intein in this dataset, from phage Zeenon, indicated by asterisks on the extein phylogeny since its extein is split into two separate open reading frames by the split intein. 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**), and the whole world (**F**). Note that the phages from South Africa (**E**) have identical or nearly identical intein sequences (**B**) but that their exteins do not group together and are separated by exteins harboring no intein, or inteins from a different cluster (**A**). Similarly, the phages isolated in California have nearly identical inteins (**B**), but their exteins are divergent and do not group together (**A**).

We repeated this analysis with a second group of inteins from Ohio and Pennsylvania, found in minor capsid proteins (Figure 7, red cluster). These 7 inteins possess an anti-sense LAGLIDADG HEN. 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.

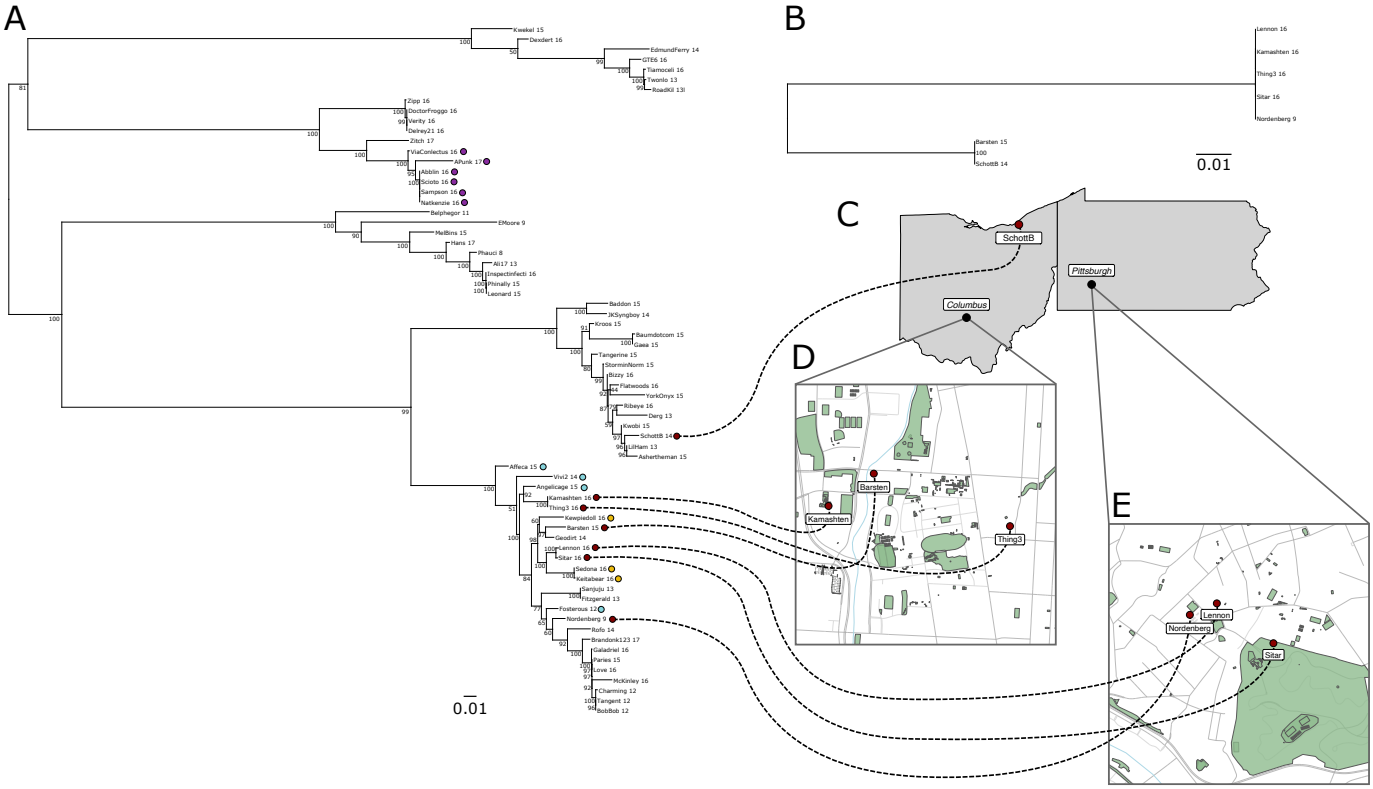


Figure 7. **A.** Maximum likelihood phylogeny of the nucleotide sequences of intein-containing and intein-free minor capsid proteins using TN+F+R3 as the substitution model. Colored circles to the right of taxa names indicate intein presence, and each color corresponds to a different intein cluster. Ultrafast 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**). Note that the three phages depicted in panel **E** have identical intein sequences, but that the extein from phage Nordenberg diverges from the exteins in Lennon and Sitar and that the three exteins do not form a clan.

339 Finally, we investigated the inteins found within exteins annotated as putative integrin-like
340 proteins (Figure 8). Once again, these inteins were found in divergent exteins separated on the
341 phylogenetic tree by non-intein containing taxa. Many of these inteins were found specifically within the
342 grounds of Howard University in Washington DC; however, despite the small geographic distance
343 between the phages found there, the exteins showed significant divergence. Furthermore, inteins from
344 each of the three clusters were found on this campus. The vast majority of the inteins from the bottom
345 clan of the intein phylogeny (Figure 8B, gold) were present predominantly on the eastern shoreline of the
346 USA. The clan at the top (Figure 8B, blue) had a much wider distribution — mostly confined to the
347 Mississippi River Valley, and the Midwest. The last of the three clans (Figure 8B, red) was found across
348 the USA.

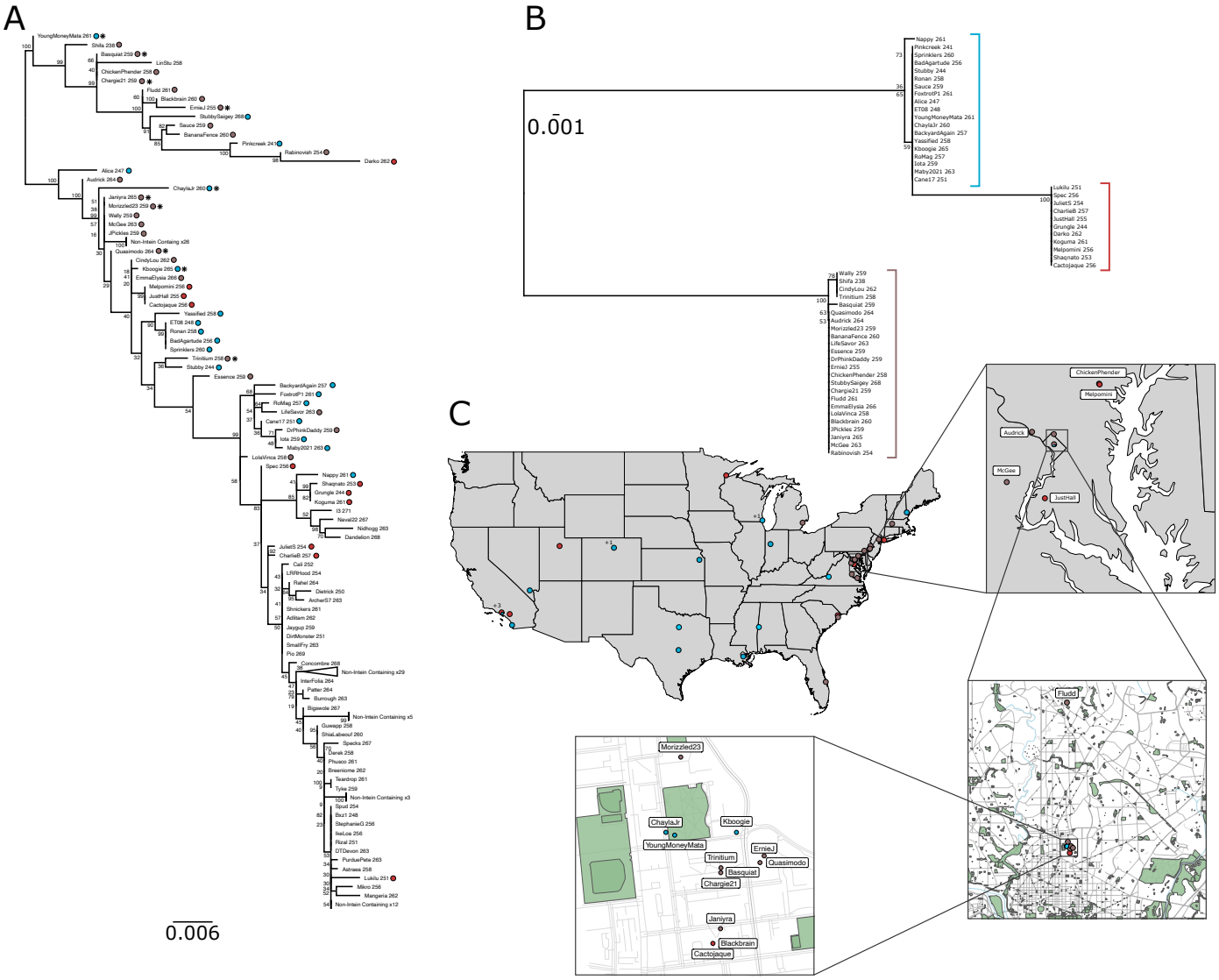


Figure 8. **A.** Maximum likelihood phylogeny based on the extein-encoding nucleotide sequences of intein-free and intein-containing putative integrin-like proteins using TIM2+F+I+G4 as the substitution model. Colored circles to the right of taxa names indicate intein presence, and the color indicates which cluster (**B**) said intein belongs to. Ultrafast bootstrap support is listed on internal nodes. An asterisk indicates that the phage was 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.** Geographic maps indicating 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

360 indicated with a plus sign and the corresponding number of additional inteins. The color of the circles
361 corresponds to the clans in **B**.

362

363 ## Discussion

364 **Overarching Trends**

365 As in previous studies, the inteins we discovered within actinobacteriophages are found
366 predominantly within DNA-binding extein sequences (Kelley et al. 2016). While the majority are within
367 terminase large subunits, the inteins we discovered show a remarkable range of host proteins. In addition
368 to ubiquitous phage genes (such as the terminase large subunit), we find inteins within phage genes with
369 more limited distributions such as thymidylate synthase, or ribonucleotide reductase subunits. Even
370 genes that lack nucleotide-associated enzymatic activity, the minor capsid proteins, were found within the
371 set of invaded phage proteins. Interestingly, the inteins invading minor capsid proteins were exclusively
372 inteins with non-canonical HENs.

373 Many inteins in our dataset had sporadic distributions across their extein phylogenies, and
374 phylogenies of the intein sequences frequently conflicted with the phylogeny of their extein. We believe
375 that many of these inteins are actively moving between different hosts. However, exceptions do exist in
376 cases of mini inteins (an example of which can be found in Supplemental Figure 8). In addition, inteins
377 that persist in clans of extein sequences, at first sight might be considered to have decaying HENs, and
378 like mini-inteins be vertically inherited (see the purple intein group in Figure 7). However, while the exteins
379 show some sequence divergence, the fact that the inteins are identical likely illustrates a recent invasion
380 event (see below for further discussion).

381 The many intein-free homologs accompanying intein-containing genes in our dataset indicate that
382 intein homing into new empty target site must be balanced by the loss of inteins in other homologs. Given
383 the seemingly rapid dissemination of inteins into local phage populations, intein loss might also be
384 assumed to occur frequently. This could occur via gradual decay and recombination of the intein, or be
385 the consequence of a fitness cost imparted by the intein on the phage. Inteins are known to cause a drop

386 in fitness; in the archaeon *Haloferax volcanii,* a DNA polymerase intein caused a 7% drop in fitness for
387 the host (Naor et al. 2016). Thus, the alternative loss scenario would act at the population level — the
388 intein imparts enough of a fitness cost such that the intein-containing phage is out-competed by intein-
389 free phages, and the intein containing phages would be slowly selected out of the population. This
390 process would be enhanced if the non-invaded phages evolve at least a partial resistance to intein
391 propagation (perhaps via mutations to the HEN target site). This potential impact on host fitness is a
392 mechanism worth exploring in future work.

393 **The Nature of Non-Canonical Inteins**

394 In this study we described several inteins with previously unseen, or extremely rare architectures
395 in relation to their HEN domains. To our knowledge, only the anti-sense LAGLIDADG HEN has been
396 reported on in prior literature within the context of a fully functional intein (Gorbalenya 1998; Gogarten et
397 al. 2002). These anti-sense LAGLIDADG HEN containing inteins were discovered in non-phage genomes
398 within the *dnaX* gene of a *Synechocystis* species, and the SpoVR-like protein from *Chloroflexus*
399 *aurantiacus* J-10-fl (sequences deposited in the InBase database (Perler 2002)). Hence, while not entirely
400 novel, we show that this intein architecture is present within at least one other extein sequence, and
401 importantly, that the intein and anti-sense HEN appear to work together to facilitate ongoing intein
402 invasion events.

403 VSR-like HEN-containing inteins have been reported on previously, within split-inteins (Dassa et
404 al. 2009). Our findings show that anti-sense VSR-like HEN-containing full inteins also exist and are not
405 merely a point of speculation. If the theory of split-intein evolution outlined by Dassa et al. 2009 is true,
406 then perhaps these newly discovered inteins represent one of the potential off-tracks within their model. It
407 should also be noted here, that Dassa et al. 2009 report that the VSR-like HENs they discovered were
408 related to freestanding HENs within their database (a fact whose importance will be discussed below).

409 Finally, the endonuclease VII-like HEN containing intein appears to be completely novel, with no
410 known reports. Like the VSR-like HENs, this endonuclease has been reported prior in a free-standing
411 context (Pope et al. 2013).

412 Given the ubiquity of freestanding HEN genes within actinobacteriophages, the frequency with
413 which these genes seem to proliferate (Edgell et al. 2010; Pope et al. 2013; Zhang et al. 2017; Hatfull
414 2020; Barth et al. 2023), and the presence of free-standing HENs related to those found within our intein
415 sequences, we believe that the three intein architectures described above originate via invasion of an
416 intein sequence by a free-standing HEN. Given the seeming rarity of these elements, we further propose
417 (in line with the model laid out by Dassa et al. 2009 (Dassa et al. 2009)), that successful integration of
418 these HENs into the region of an intein between splicing associated blocks is extremely rare. The
419 breaking of, and subsequent integration into the intein CDS by a free-standing HEN in this fashion may
420 be relatively common, but most of the time this integration is likely to result in split-inteins, or introduce
421 nonsense mutations which prevent post-translational rescue of extein function via intein splicing.

422 Assuming the free-standing HEN's integration into the intein CDS is successful, the path forward
423 is unclear. It is possible the HEN becomes a pseudo-gene, but there have been prior examples of phage
424 genes with nested HEN open reading frames (ORFs) that still translate both proteins (Gibb and Edgell
425 2007; Wright et al. 2021). In our cases, the secondary ORFs typically lie outside the bounds of the intein
426 splicing domain, and given the splicing domain's high conservation, it is probable that the spicing domain
427 still functions properly. Additionally, many of these non-canonical inteins were found in the crucial
428 terminase large subunit gene, and it is therefore unlikely that the intein is not splicing itself out — these
429 phages would otherwise be unable to properly package their DNA into their capsid (a certain end for that
430 phage lineage). Our phylogenetic reconstructions of the anti-sense LAGLIDADG-containing inteins
431 provided evidence for continued invasion. These inteins are present in multiple locations within divergent
432 host exteins, and importantly are separated with meaningful support by relatives that lack the intein. Such
433 a pattern is not reasonably explained via vertical inheritance and loss, and it is therefore likely that the
434 splicing domain and anti-sense HEN domain of that particular intein still function as a single unit. This
435 may also be true of two groups of inteins with VSR-like HENs (Supplemental Table 1, clusters 60 and 91)
436 which likewise show a pattern of inheritance more in-line with horizontal gene transfer. However, the
437 other cases of VSR-like HEN-containing inteins exhibit patterns suggestive of vertical inheritance (or were
438 singletons and thus no evolutionary inference could be drawn). The endonuclease VII-like HEN-
439 containing inteins presented a more ambiguous distribution. While most are clustered together, there are

440 a number of closely related extein alleles without an intein. This, plus the sequence similarity of these
441 inteins (for example, the inteins of cluster 47 — which includes phages Angelicase, Affeca, Keitabear,
442 and Sedona — are all identical with the exception of Angelicase), suggests that these inteins may still be
443 transferring between phages horizontally (Supplemental Figure 9, Supplemental Figure 4).

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

453 Inteins created by the mechanism described above thereby provide meaningful support for a
454 "protein splicing first" theory of intein evolution (Derbyshire and Belfort 1998; Gogarten et al. 2002; Dassa
455 et al. 2009; Green et al. 2018). The number of intein-associated HEN's (including HNH, VSR-like, GIY-
456 YIG, LAGLIDADG, etc. (Dalgaard, Klar, et al. 1997; Gorbalenya 1998; Gimble 2000; Dassa et al. 2009))
457 suggests that there have been multiple separate invasions of intein coding sequences by free-standing
458 HENs (Derbyshire and Belfort 1998). Perhaps most, if not all, of these HENs came into intein sequences
459 via a free-standing HEN, proliferated through the intein's self-splicing machinery, and were shifted by an
460 indel or an inversion into a contiguous orientation.

461 **Frequent Horizontal Transfer of Inteins**

462 Extein phylogenies with intein presence mapped onto them, such as in Figures 6A, 7A, and 8A,
463 can reveal likely cases of vertical inheritance, recent dissemination, and horizontal transfer of inteins.
464 Using an extein phylogeny with intein presence mapped will differentiate between horizontal transfer and
465 vertical inheritance or recent dissemination; however, intein phylogenies and assessment of intein and

466 extein conservation are required to differentiate between vertical inheritance with frequent intein loss
467 events and recent dissemination.

468 A set of intein-containing exteins which all group together initially indicates vertical inheritance of
469 that intein, such as the group in Figure 7A containing phage ViaConlectus. While the phylogeny is
470 unrooted, if one follows the presented arrangement one would infer that ViaConlectus was invaded by the
471 purple cluster intein, and this intein was then vertically inherited by phages APunk, Abblin, Scioto,
472 Sampson, and Natkenzie. To assess whether the event is a case of vertical inheritance or recent rapid
473 dissemination, the degree of similarity between inteins in these cases must be further analyzed, as well
474 as geographic proximity between hosts when such information is available. In such cases, extremely
475 similar inteins in geographically-close but less similar exteins may indicate a recent intein dissemination
476 event.

477 In contrast to the phylogenetic patterns suggesting vertical inheritance or recent local
478 dissemination are cases where we find inteins within extein sequences that lack a direct common
479 ancestor and are separated on the extein phylogeny by many nodes leading to other non intein-
480 containing extein sequences. In such cases, a vertical inheritance pattern would require many loss events
481 to be an accurate model given these phylogenies. To supplement findings from extein phylogenies with
482 intein presence mapping, the phylogeny of the intein of interest can be reconstructed and its topology
483 compared to that of the extein phylogeny to identify incongruencies, which often provide further evidence
484 of a non-vertical inheritance pattern. For example, in the extein phylogeny with intein presence mapping
485 in Figure 7A, we see a well-supported (99% ultrafast bootstrap support) split between a group of exteins
486 from phages including phage SchottB, and a group of exteins from phages including phages Kamashten,
487 Thing3, Barsten, Lennon, Sitar, and Nordenberg. Despite the evolutionary distance between exteins,
488 these phages all contain near identical inteins at the same position in their extein (inteins from the red
489 cluster in Figure 7A). To supplement this observation, Figure 7B depicts the intein phylogeny for this red
490 cluster of inteins which shows the SchottB intein grouping closely with the Barsten intein, despite the
491 large well-supported evolutionary distance between their exteins. The most parsimonious model
492 explaining this observation is a horizontal transfer event between SchottB and a member of the
493 Kamashten group, as opposed to vertical inheritance with many independent loss events. Applying this

494 phylogenetic analysis to each of our extein-intein phylogeny pairings thus reveals a multitude of cases for
495 the horizontal transfer of inteins between phages.

496 For highly similar inteins which possess a wide geographic distribution, the next question
497 inevitably becomes how did they get there? While humans are almost certainly acting as a secondary or
498 tertiary vector for the intein, we propose that these inteins are likely dispersing in part via the phages we
499 find them in and in part, via the phage's host bacteria. Previous literature, and our own anecdotal
500 observations, find that many phage-associated inteins are found inside of bacterial genomes, most likely
501 within prophages (25). Thus, inteins experience increased genetic mobility, taking advantage of both the
502 viral and bacterial vectors. As a prophage, the intein can propagate not only when the bacterial cell
503 replicates, but also through the numerous avenues of horizontal gene transfer that bacterial cells engage
504 in (Soucy et al. 2015), and by transfer into closely related phages upon infection.

505 **Geographic Spread of Inteins**

506 Our dataset contains several groups of inteins which are extremely similar on the nucleotide level
507 (and in the cases described, often identical), these inteins must have diverged from one another relatively
508 recently. This presents two options: rapid host proliferation with direct vertical inheritance, or rapid intein
509 proliferation via horizontal gene transfer. For some of the inteins in our dataset, the former may be
510 considered a possibility (see the clan of intein containing exteins colored purple in Figure 7); however,
511 where the intein containing exteins form a well-supported clan with limited geographic distribution, the
512 observation that the inteins are identical, whereas the exteins are divergent, suggests that the inteins
513 have recently invaded the divergent exteins instead of being vertically inherited. This argument is
514 strengthened by the observation that inteins found in divergent exteins experience less selection for
515 sequence conservation than the exteins that harbor them (see Supplemental Materials: Supplemental
516 Results and Discussion – Sequence Conservation in Inteins and Exteins). The finding of nearly identical
517 inteins found in divergent exteins in close geographical proximity suggests that intein invasion represents
518 an example of a highly effective gene drive acting on the local level following the introduction of a new
519 founder intein allele. The forces that limit this drive (after all, not all homologous exteins within a given
520 location contain the intein) need further study (see the above discussion of fitness cost to the host that

521 can be associated with inteins). Phages, because of their rapid propagation and large population sizes
522 provide an excellent opportunity to study intein population dynamics in more detail.

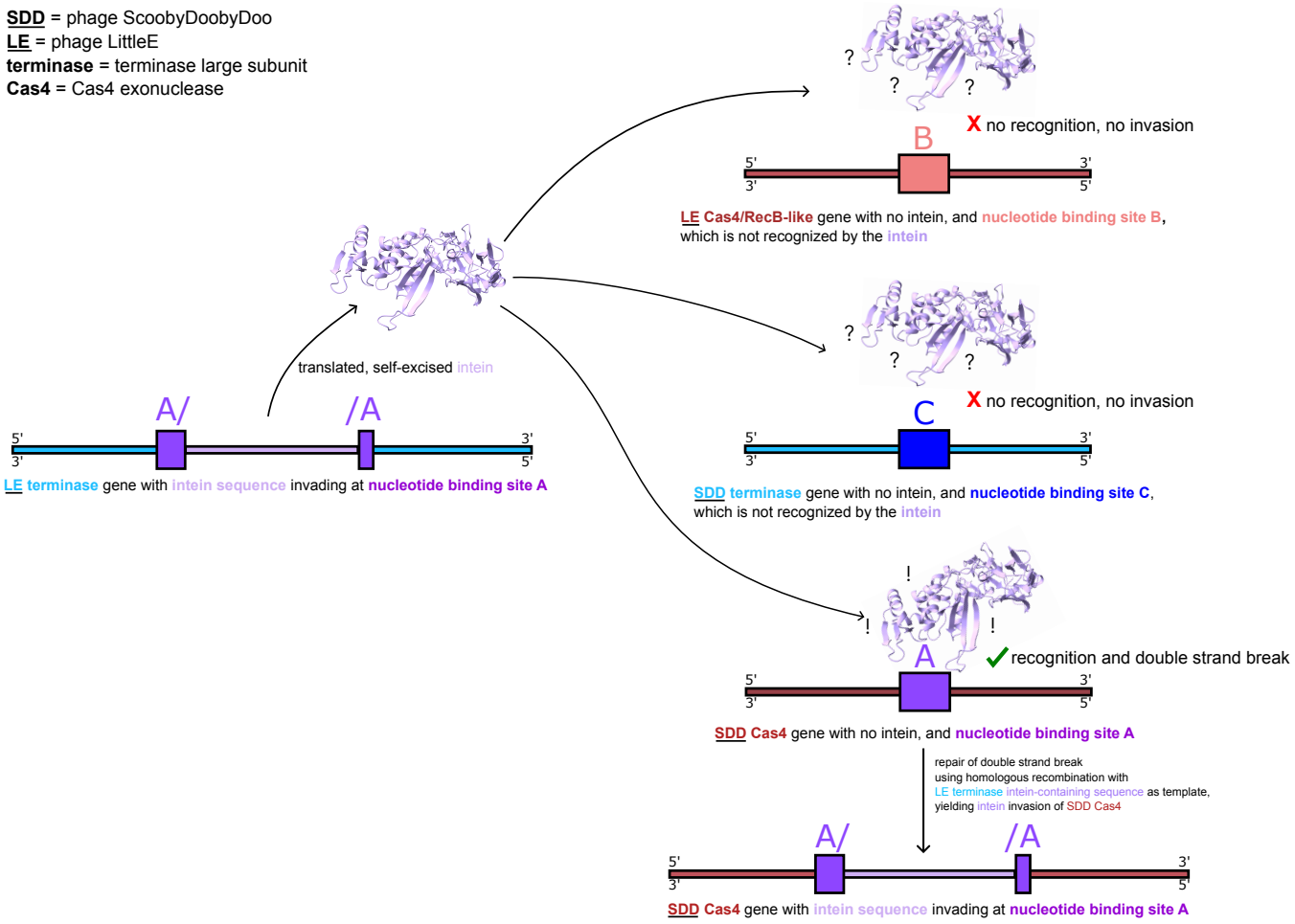
523 **Nucleotide binding sites as an avenue for extein-switching**

524 The instance of Cas4 exonuclease inteins evolving from terminase large subunit inteins on the
525 class 3 intein phylogeny involving phages LittleE and ScoobyDoobyDoo (SDD) (Figure 4) provides a
526 compelling argument for the use of highly similar nucleotide binding sites as a means of intein movement
527 between otherwise dissimilar genes. The stretch of identical nucleotides between the two intein insertion
528 sites (10 bases on the 5' side, 2 bases on the 3' side) in conjunction with the other shared positions within
529 the +/-30bp surrounding the site (Supplemental Figure 7A) may be sufficient for recognition and invasion
530 by the same intein. The proposed range for HEN recognition sites is ~12-40bp, with recognition sites in
531 smaller genomes such as phage genomes being on the lower end of the range (Jurica and Stoddard
532 1999). Additionally, many types of phage recombinases can tolerate some dissimilarity between regions
533 used for homologous recombination (Lopes et al. 2010) suggesting this intein could possibly invade either
534 site if given the chance.

535 To further complete the picture and investigate why the SDD terminase large subunit and LittleE
536 Cas4/RecB-like exonuclease are intein-free, we analyzed the regions corresponding to the intein-insertion
537 sites on the intein-free proteins (Supplemental Figure 7B) and find SDD's Cas4 exonuclease gene to be a
538 more hospitable landing pad for the LittleE terminase large subunit intein by metrics of insertion site
539 recognition and residues at the insertion site required for intein self-splicing (the SDD terminase large
540 subunit encodes an alanine at the position which would need to be a serine, threonine, or cysteine for
541 successful intein splicing to occur) (Supplemental Figure 7B) (Mills et al. 2014; Turgeman-Grott et al.
542 2023). Additionally, LittleE is a temperate phage of the same strain of bacteria that SDD is a lytic phage of
543 (*Mycobacterium smegmatis* mc2155) (Russell and Hatfull 2017).

544 The presented evidence taken together suggest a model in which the recipient genome was
545 exposed to the donor's terminase large subunit intein during co-infection or infection of a host containing
546 the donor's prophage. Then, either homing of the donor terminase large subunit intein into the recipient's
547 terminase large subunit gene was thwarted by the lack of similarity between these terminase large

548 subunits, or homing was successful but nonviable due to lack of a splicing-compliant residue on the C-
549 terminal side of the insertion site. By happenstance, a region of similarity to the donor terminase large
550 subunit present in the recipient Cas4 exonuclease gene conferred by a shared nucleotide binding motif
551 provided a viable alternate insertion site for the intein (Figure 9). Being in a brand-new context also lets
552 this copy of the intein begin a fresh evolutionary trajectory, as the majority of homologous genes in the
553 population will initially lack this intein. This may explain the presence of SDD's Cas4 exonuclease intein in
554 over 130 other phage Cas4 exonucleases (Figure 4). As PhagesDB and other phage genome sequence
555 resources continue to grow, increased sampling will tell a clearer story of the events that unfolded in the
556 evolution of these inteins.



557

558 **Figure 9**: Three different scenarios for attempted intein invasion by the LittleE terminase large subunit
559 (TLS) intein, explaining simultaneous presence of intein in LittleE TLS and ScoobyDoobyDoo (SDD) Cas4

560 exonuclease genes, but lack of intein in LittleE Cas4/RecB-like endonuclease and SDD TLS genes. TLS
561 exteins are depicted as thin blue boxes, Cas4 exonuclease exteins are depicted as thin red boxes, and
562 the intein sequence within invaded genes is depicted as a thin purple box. Nucleotide binding site A
563 (depicted as a thick purple box) is the insertion site recognized by the LittleE TLS intein, and is present in
564 both the LittleE TLS gene and the ScoobyDoobyDoo (SDD) Cas4 exonuclease gene. Nucleotide binding
565 site B (depicted as a thick light red box) is not recognized by the LittleE TLS intein, and is present in the
566 LittleE Cas4/RecB-like exonuclease gene in place of nucleotide binding site A. Thus, the LittleE TLS
567 intein cannot engage in homing with the LittleE Cas4/RecB-like exonuclease gene and invasion cannot
568 occur. Nucleotide binding site C (depicted as a thick dark blue box) is not recognized by the LittleE TLS
569 intein, and is present in the SDD TLS gene in place of nucleotide binding site A. Thus, the LittleE TLS
570 intein cannot engage in homing with the SDD's TLS gene and invasion cannot occur. Since the SDD
571 Cas4 exonuclease gene contains nucleotide binding site A, the LittleE TLS intein can engage in homing,
572 allowing the invasion of the LittleE TLS intein into the SDD Cas4 exonuclease gene.

573

574 ## Conclusions

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

585 Finally, we wish to stress that these non-canonical inteins may be significantly more common
586 than even what we have outlined here. Given their overlapping ORFs, these inteins are likely missed in
587 automated annotation pipelines, especially within phages — current SEA-PHAGES training argues for
588 "...typically only one frame in one strand is used for a protein-coding gene…" (SEA PHAGES). Even to
589 the trained eye, these inteins may currently be glossed over as 'decaying' canonical inteins or

590 misclassified as mini-inteins. Significant work will be needed to confirm that inteins listed as decaying are

591 correctly classified, and to discover the true extent of these inteins within phages and their bacterial hosts.

592 ## Materials and Methods

593 **Code Availability**

594 All code used in this project is available at https://github.com/sophiagosselin/intein_distribution
595 along with usage information. The two exceptions to this are ICE-BLAST and Invader-Sim which are kept
596 in separate repositories detailed below.

597 **Iterative Cluster Expansion BLAST: ICE-BLAST**

598 To increase efficiency in mass intein sequence retrieval, we developed "Iterative Cluster Expansion
599 BLAST" (ICE-BLAST). ICE-BLAST is a Perl (v 5.36.0) based program using PSI-BLAST (Altschul et al.
600 1997) as the principal sequence search tool and USEARCH clustering (Edgar 2010) to identify new queries
601 from previous iterations. Through this process, ICE-BLAST retrieves highly divergent homologous protein
602 sequence matches. In short, ICE-BLAST iteratively searches amino acid databases with PSI-BLAST,
603 clusters filtered outputs using USEARCH, then uses the representative sequence of each match cluster
604 (centroids) as the next query sequences (Figure 10A).

605 ICE-BLAST is available at https://github.com/sophiagosselin/ICE_BLAST along with detailed
606 usage, dependencies, and other information. Algorithm details and parameters used for this study are
607 available in Supplemental Materials.

608 **Simulations of Intein Evolution: Invader-Sim**

609 Since ICE-BLAST involves multiple individual searches, one must take proper precautions against
610 false-positives. False-positives can introduce errors that will propagate through all downstream iterations.
611 Hence, it was important to test ICE-BLAST against known data to assess overall sensitivity and specificity.
612 We developed a Perl (v 5.36.0) based sequence simulation program, Invader-Sim, to generate sequences
613 simulating intein invasion into extein sequences. The program uses a combination of PAML-EVOLVER
614 (v4.9j) (Yang 2007) and a Markov chain to simulate datasets of intein-containing extein sequences (Figure
615 10B). ICE-BLAST performed as well as PSI-BLAST on the simulated datasets. Results and additional
616 discussion can be found in Supplementary Materials.

Invader-Sim is available at https://github.com/sophiagosselin/invader_sim. Information on the method and parameters used for this study are available in the Supplemental Materials.

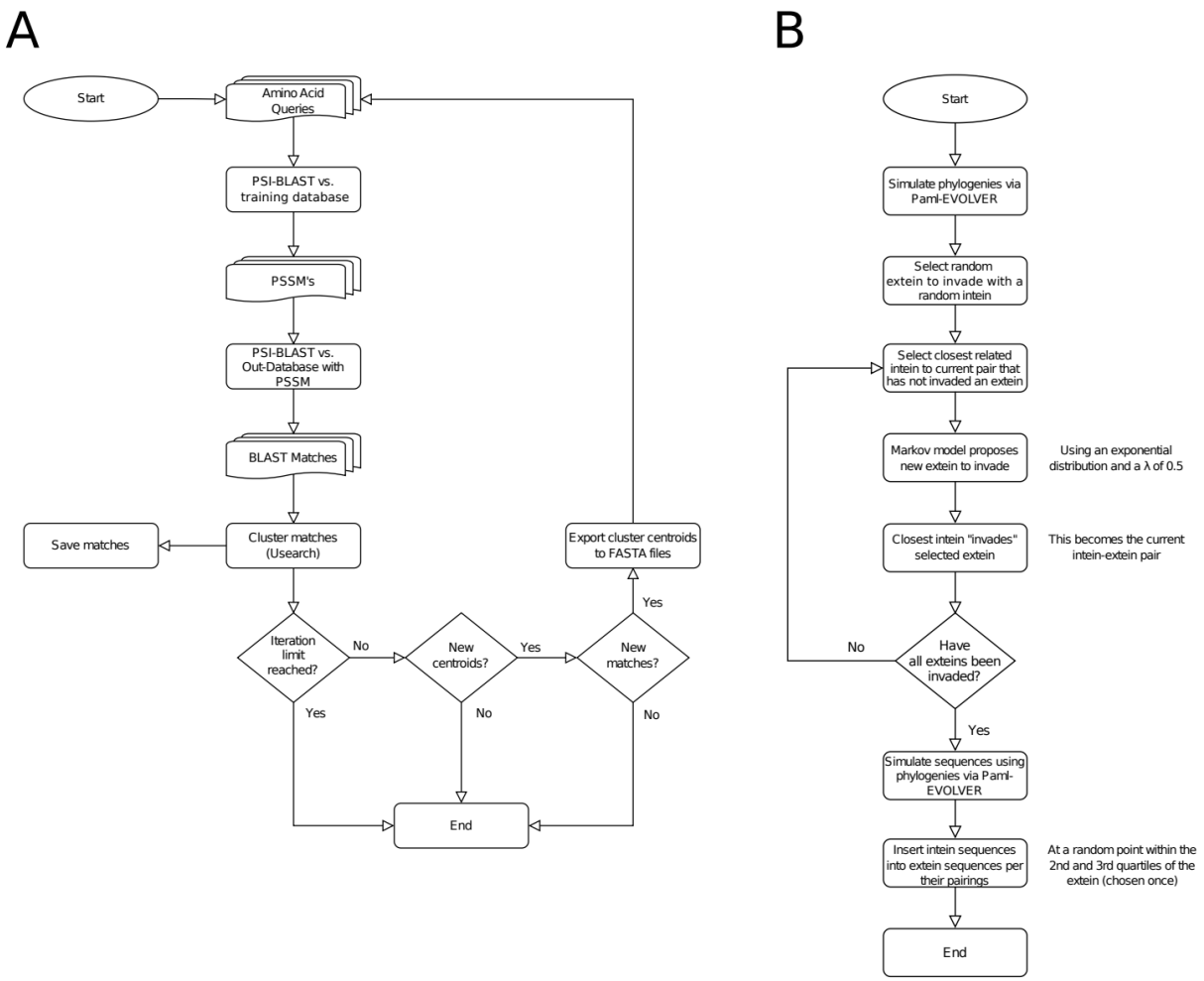


Figure 10. Flow charts describing the algorithms of ICE-BLAST and Invader-Sim. **A.** ICE-BLAST begins by taking amino acid query sequences and iteratively searching against a training amino acid sequence databank using PSI-BLAST. The matches resulting from this iterative search are clustered by sequence similarity, where a representative amino acid sequence is selected from each cluster (referred to as the centroid sequence for that cluster). The centroids of each cluster are then used as input for a new round of PSI-BLAST searches, with the process repeated until either the iteration limit is reached, there are no new centroids, or there are new centroids but no new matches over all. **B.** Invader-Sim begins by using PAML-EVOLVER to simulate an extein phylogeny and an intein phylogeny. From the phylogeny of potential exteins, a random individual is chosen to be invaded by an intein from the intein phylogeny. The closest relative of the randomly selected intein is given a new extein to invade, with the extein selection

630 proposed by a Markov model. The process is repeated until all exteins are invaded by an intein.
631 Simulated sequences are then generated using PAML-EVOLVER.

632 **Dataset Curation: Applying ICE-BLAST to PhagesDB**

633 Using PhagesDB data (Russell and Hatfull 2017), two databases were constructed: one from
634 amino acid sequence files, and one from whole genomes, with associated metadata from the PhagesDB
635 API. The genomic and amino acid databases were downloaded and formatted on September 28 2023
636 and November 2 2023 respectively. The databases were not updated over the course of this work, due to
637 reverse incompatibility of PhagesDB updates.

638 The initial ICE-BLAST query set for this study consisted of 78 intein sequences drawn from three
639 sources: InBase (final update 2010) (Perler 2002), known Actinobacteriophage inteins (Kelley et al.
640 2016), and inteins found by the Gogarten Lab in the PhagesDB database between 2020 and 2024
641 (Gosselin et al. 2023) (Supplemental Table 3). Initial PSSMs were trained on Uniref50 (Suzek et al. 2015)
642 downloaded November 7 2021. For the ICE-BLAST search, a percent identity of 60%, an e-value of 1e-
643 20, and a domain specific cutoff bound threshold of 25% were used as inputs for filtering.

644 Initial results from ICE-BLAST contained 1994 sequences which were then clustered by
645 USEARCH (v11.0.667) (Edgar 2010) at 70% identity using the cluster_fast algorithm. Associated intein-
646 invaded extein sequences were then extracted from the BLAST database using blastdbcmd. Putative
647 intein sequences were aligned to the invaded extein sequences using MAFFT (v7.490) (Katoh and
648 Standley 2013). When intein identity was uncertain, BLAST against NCBI's non-redundant database
649 (nrdb) (Altschul et al. 1990), and HHPred against PDB, Pfam, and NCBI CDD (Zimmermann et al. 2018),
650 were used to determine whether the sequence was actually an intein, and where its boundaries laid.
651 False positives were removed from the dataset at this point. This screening reduced the dataset to 784
652 confirmed intein-containing sequences (Supplemental Table 1). The large was attributed entirely to a
653 false positive match to a DNA Primase sequence; causing over 1000 matches to DNA primase extein
654 domains which were falsely incorporated. Extein sequences were re-annotated using BLAST and HHPred
655 as described above. Inteins were classified as class 1, 2, or 3, by comparing to self-splicing domain
656 consensus sequences (Tori et al. 2010). In-house Perl scripts were used to extract associated coding

657 sequences from the mentioned PhagesDB genome database. Intein and extein sequence files are
658 available in Supplemental Data.

659 HENs were classified using the method above, with two exceptions. First, we used the PHROG
660 (Terzian et al. 2021) database within HHPred (Zimmermann et al. 2018) (especially for more divergent
661 sequences) to identify potentially related HENs as specified in the PFAM section of its entries. Second,
662 we used InterProScan (as of August 2024) (Jones et al. 2014) when BLAST and HHPred were
663 inconclusive.

664 **Geographic Maps**

665 First, intein nucleotide sequences were clustered using USEARCH (97.5% sequence identity)
666 resulting in 91 clusters. Clusters with less than four members were excluded, as a meaningful
667 phylogenetic tree cannot be generated. If exteins within a cluster were more self-similar than their intein
668 sequences were, the cluster was removed. Remaining clusters were then used to build intein nucleotide
669 phylogenies, and their extein sequences (with intein-less homologs from PhagesDB) were used to
670 construct an extein phylogeny. Only clusters whose intein and extein phylogenies were in-congruent were
671 selected for further study.

672 Extein phylogenies were marked for intein presence, and the inteins present were then mapped
673 onto their geographic isolation location. Phage isolation locations were extracted from PhagesDB's API.
674 Maps of intein isolation locations were created using R (v4.4.1) and the following libraries: sf (v1.0-16)
675 (Pebesma 2018), ggplot2 (v3.5.1) (Wickham 2016), maps (v3.4.2) (Deckmyn 2023), mapdata (v2.3.1)
676 (Brownrigg 2022), mapproj (v1.2.11) (Minka and Bivand 2023), rnaturalearthdata (1.0.0) (South et al.
677 2024a), and rnaturalearthhires (v1.0.0.9000) (South et al. 2024b).

678 **Phylogenetic Trees and AU Tests**

679 Phylogenies in this study were built from MAFFT sequence alignments. Unless specified, MAFFT
680 was allowed to select the best alignment algorithm for a given set of sequences (strategy: 1. --auto). IQ-
681 TREE (v2.0.7) using ModelFinder (-m MFP) and 1000 ultrafast bootstraps (-bb 1000) (Minh et al. 2020)
682 was used for phylogenetic reconstruction. Models selected are included in phylogeny figure legends.

683 Trees were visualized using Dendroscope (v3.8.10) (Huson et al. 2007), Figtree (v1.4.4) (Rambaut 2012),
684 and Inkscape (v1.3.2) (Developers). Clans with extremely similar or identical sequences were collapsed in
685 some phylogenies for readability. Un-collapsed versions are included in the supplemental material (see
686 Figure 4 and Supplemental Figure 10 for examples of collapsed and un-collapsed respectively).

687 AU tests (Shimodaira 2002) were performed via the method built into IQ-TREE2 (Minh et al.
688 2020).

689 **Predicted Protein Structures**

690 Protein structures were predicted for non-canonical inteins and their associated secondary ORFs
691 using AlphaFold (Jumper et al. 2021) through UCSF ChimeraX (Pettersen et al. 2021; Meng et al. 2023).
692 Structures were colored according to the predicted local distance difference test (pLDDT) value per-
693 residue. Each predicted structure was used as query in a DALI search of the PDB database (Holm et al.
694 2023).

695 AlphaFold (Jumper et al. 2021) was also used to generate predicted structures in the extein
696 switching study. UCSF ChimeraX (Pettersen et al. 2021; Meng et al. 2023) was used to compare the
697 predicted structure of the LittleE terminase large subunit to the solved structure of a phage terminase
698 large subunit bound to ATP (Zhao et al. 2013).

699 **Using Predicted Structures to Investigate and Classify Non-Canonical HENs**

700 For the inteins with non-canonical HENs, predicted structures suggest that the portion of the
701 intein which should be an in-frame HEN does not make any recognizable structures (Supplemental Figure
702 1), and cannot be matched to any known structure using the DALI server PDB search tool. However,
703 predicted structures of the anti-sense LAGLIDADG HEN, anti-sense VSR-like HEN, and out-of-frame
704 endonuclease VII-like HEN, all matched to their associated predicted domains (Supplemental Figure 2). It
705 is important to note that these matches were only to the core catalytic domain of these HENs; significant
706 portions of these structures could not be matched to known proteins. For the anti-sense LAGLIDADG
707 HEN, the LAGLIDADG core sits between residues 40 and 185, matching to PDB structure 2O7M (Prieto
708 et al. 2007) (CreI — a known LAGLIDADG HEN) with a Z-score of 14.8. The VSR-like HEN predicted
709 structure matched to PDB structure 1CW0 (Tsutakawa et al. 1999) (VSR) with a Z-score of 10.6 across

710 the residues 144 to 267. Finally, the endonuclease VII-like HEN matches to PDB structure 1E7D
711 (Raaijmakers et al. 2001) (a recombination endonuclease VII) with a Z-score of 6.5 across residue 1 to
712 residue 67.

713 **Intein Insertion Site Consensus Sequence Plots**

714 WebLogo (Crooks et al. 2004) was used to generate consensus sequence plots in this study.

Acknowledgments

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.

729 **References**

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

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

734 Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and
735 PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*
736 25:3389–3402.

737 Barth ZK, Dunham DT, Seed KD. 2023. Nuclease genes occupy boundaries of genetic exchange
738 between bacteriophages. *NAR Genomics and Bioinformatics* 5:lqad076.

739 Barzel A, Obolski U, Gogarten J, Kupiec M, Hadany L. 2011. Home and away- the evolutionary dynamics
740 of homing endonucleases. *BMC Evol Biol* 11:324.

741 Belfort M, Roberts RJ. 1997. Homing endonucleases: keeping the house in order. *Nucleic Acids*
742 *Research* 25:3379–3388.

743 Brownrigg OS code by RAB and ARWR version by R. 2022. mapdata: Extra Map Databases. Available
744 from: https://cran.r-project.org/web/packages/mapdata/index.html

745 Coughlan AY, Lombardi L, Braun-Galleani S, Martos AA, Galeote V, Bigey F, Dequin S, Byrne KP, Wolfe
746 KH. 2020. The yeast mating-type switching endonuclease HO is a domesticated member of an
747 unorthodox homing genetic element family.Moses AM, Wittkopp PJ, Rusche L, editors. *eLife*
748 9:e55336.

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

751 Dalgaard JZ, Klar AJ, Moser MJ, Holley WR, Chatterjee A, Saira Mian I. 1997. Statistical modeling and
752 analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein
753 that encodes a site-specific endonuclease of the HNH family. *Nucleic Acids Research* 25:4626–
754 4638.

755 Dalgaard JZ, Moser MJ, Hughey R, Mian SI. 1997. Statistical Modeling, Phylogenetic Analysis and
756 Structure Prediction of a Protein Splicing Domain Common to Inteins and Hedgehog Proteins.
757 *Journal of Computational Biology* 4:193–214.

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

761 Deckmyn OS code by RAB and ARWR version by RBE by TPM and A. 2023. maps: Draw Geographical
762 Maps. Available from: https://cran.r-project.org/web/packages/maps/index.html

763 Dedrick RM, Smith BE, Cristinziano M, Freeman KG, Jacobs-Sera D, Belessis Y, Whitney Brown A,
764 Cohen KA, Davidson RM, Van Duin D, et al. 2023. Phage Therapy of *Mycobacterium* Infections:
765 Compassionate Use of Phages in 20 Patients With Drug-Resistant Mycobacterial Disease.
766 *Clinical Infectious Diseases* 76:103–112.

767 Derbyshire V, Belfort M. 1998. Lightning strikes twice: Intron–intein coincidence. *Proceedings of the*
768 *National Academy of Sciences* 95:1356–1357.

769 Developers IW. Inkscape - Draw Freely. | Inkscape. Available from: https://inkscape.org/

770 Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460–
771 2461.

772 Edgell DR, Gibb EA, Belfort M. 2010. Mobile DNA elements in T4 and related phages. *Virology Journal*
773 7:290.

774 Gibb EA, Edgell DR. 2007. Multiple Controls Regulate the Expression of mobE, an HNH Homing
775 Endonuclease Gene Embedded within a Ribonucleotide Reductase Gene of Phage Aeh1.
776 *Journal of Bacteriology* 189:4648–4661.

777 Gimble FS. 2000. Invasion of a multitude of genetic niches by mobile endonuclease genes. *FEMS*
778 *Microbiology Letters* 185:99–107.

779 Goddard MR, Burt A. 1999. Recurrent invasion and extinction of a selfish gene. *Proc. Natl. Acad. Sci.*
780 *U.S.A.* 96:13880–13885.

781 Gogarten JP, Hilario E. 2006. Inteins, introns, and homing endonucleases: recent revelations about the
782 life cycle of parasitic genetic elements. *BMC Evol Biol* 6:94.

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

785 Gorbalenya AE. 1998. Non-canonical inteins. *Nucleic Acids Research* 26:1741–1748.

786 Gosselin SP, Arsenault DR, Jennings CA, Gogarten JP. 2023. The Evolutionary History of a DNA
787 Methylase Reveals Frequent Horizontal Transfer and Within-Gene Recombination. *Genes*
788 14:288.

789 Green CM, Novikova O, Belfort M. 2018. The dynamic intein landscape of eukaryotes. *Mobile DNA* 9:4.

790 Hatfull GF. 2020. Actinobacteriophages: Genomics, Dynamics, and Applications. *Annu. Rev. Virol.* 7:37–
791 61.

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

794 Hirata R, Ohsumk Y, Nakano A, Kawasaki H, Suzuki K, Anraku Y. 1990. Molecular structure of a gene,
795 VMA1, encoding the catalytic subunit of H(+)-translocating adenosine triphosphatase from
796 vacuolar membranes of Saccharomyces cerevisiae. *Journal of Biological Chemistry* 265:6726–
797 6733.

798 Hoffmann S, Terhorst TME, Singh RK, Kümmel D, Pietrokovski S, Mootz HD. 2021. Biochemical and
799 Structural Characterization of an Unusual and Naturally Split Class 3 Intein. *Chembiochem: A*
800 *European Journal of Chemical Biology* 22:364–373.

801 Holm L, Laiho A, Törönen P, Salgado M. 2023. DALI shines a light on remote homologs: One hundred
802 discoveries. *Protein Science* 32:e4519.

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

805 Iwaï H, Mikula KM, Oeemig JS, Zhou D, Li M, Wlodawer A. 2017. Structural Basis for the Persistence of
806 Homing Endonucleases in Transcription Factor IIB Inteins. *Journal of Molecular Biology*
807 429:3942–3956.

808 Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G,
809 et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30:1236–
810 1240.

811 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek
812 A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*
813 596:583–589.

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

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

819 Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in
820 Performance and Usability. *Molecular Biology and Evolution* 30:772–780.

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

823 Lennon CW, Belfort M. 2017. Inteins. *Current Biology* 27:R204–R206.

824 Lennon CW, Stanger M, Banavali NK, Belfort M. 2018. Conditional Protein Splicing Switch in
825 Hyperthermophiles through an Intein-Extein Partnership. *mBio* 9:e02304-17.

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

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

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

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

836 Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-
837 TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.
838 *Molecular Biology and Evolution* 37:1530–1534.

839 Minka DMP for R by RB and TP, Bivand transition to P 9 codebase by R. 2023. mapproj: Map
840 Projections. Available from: https://cran.r-project.org/web/packages/mapproj/index.html

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

843 Naor A, Altman-Price N, Soucy SM, Green AG, Mitiagin Y, Turgeman-Grott I, Davidovich N, Gogarten JP,
844 Gophna U. 2016. Impact of a homing intein on recombination frequency and organismal fitness.
845 *Proceedings of the National Academy of Sciences* 113:E4654–E4661.

846 Novikova O, Jayachandran P, Kelley DS, Morton Z, Merwin S, Topilina NI, Belfort M. 2016. Intein
847 Clustering Suggests Functional Importance in Different Domains of Life. *Molecular Biology and*
848 *Evolution* 33:783–799.

849 Pebesma E. 2018. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*
850 10:439.

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

852 Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE. 2021. UCSF
853 ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* 30:70–
854 82.

855 Pope WH, Jacobs-Sera D, Best AA, Broussard GW, Connerly PL, Dedrick RM, Kremer TA, Offner S,
856 Ogiefo AH, Pizzorno MC, et al. 2013. Cluster J Mycobacteriophages: Intron Splicing in Capsid
857 and Tail Genes. *PLOS ONE* 8:e69273.

858 Prieto J, Redondo P, Padró D, Arnould S, Epinat J-C, Pâques F, Blanco FJ, Montoya G. 2007. The C-
859 terminal loop of the homing endonuclease I-CreI is essential for site recognition, DNA binding and
860 cleavage. *Nucleic Acids Research* 35:3262–3271.

861 Raaijmakers H, Törö I, Birkenbihl R, Kemper B, Suck D. 2001. Conformational flexibility in T4
862 endonuclease VII revealed by crystallography: implications for substrate binding and cleavage1
863 1Edited by K. Morikawa. *Journal of Molecular Biology* 308:311–323.

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

867 Rambaut A. 2012. FigTree v1. 4.

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

869 SEA PHAGES. Guiding Principles of Annotation. Available from:
870 https://seaphagesbioinformatics.helpdocsonline.com/guiding-principles

871 Shih C-K, Wagner R, Feinstein S, Kanik-Ennulat C, Neff N. 1988. A Dominant Trifluoperazine Resistance
872 Gene from Saccharomyces cerevisiae Has Homology with F0F1 ATP Synthase and Confers
873 Calcium-Sensitive Growth. *Molecular and Cellular Biology* 8:3094–3103.

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

876 Shimodaira H. 2002. An Approximately Unbiased Test of Phylogenetic Tree Selection. *Systematic*
877 *Biology* 51:492–508.

878 Soucy SM, Fullmer MS, Papke RT, Gogarten JP. 2014. Inteins as indicators of gene flow in the
879 halobacteria. *Front. Microbiol.* [Internet] 5. Available from:
880 http://journal.frontiersin.org/article/10.3389/fmicb.2014.00299/abstract

881 Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. *Nat Rev Genet*
882 16:472–482.

883 South A, Michael S, Massicotte P. 2024a. rnaturalearthdata: World Vector Map Data from Natural Earth
884 Used in "rnaturalearth." Available from: https://cran.r-
885 project.org/web/packages/rnaturalearthdata/index.html

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

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

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

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

896 Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio RE, Mom R, Toussaint A, Petit M-A, Enault
897 F. 2021. PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR*
898 *Genomics and Bioinformatics* 3:lqab067.

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

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

903 Tsutakawa SE, Jingami H, Morikawa K. 1999. Recognition of a TG Mismatch: The Crystal Structure of
904 Very Short Patch Repair Endonuclease in Complex with a DNA Duplex. *Cell* 99:615–623.

905 Turgeman-Grott I, Arsenault D, Yahav D, Feng Y, Miezner G, Naki D, Peri O, Papke RT, Gogarten JP,
906 Gophna U. 2023. Neighboring inteins interfere with one another's homing capacity. *PNAS Nexus*
907 2:pgad354.

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

910 Wright BW, Molloy MP, Jaschke PR. 2021. Overlapping genes in natural and engineered genomes.
911 *Nature Reviews. Genetics* 23:154.

912 Yahara K, Fukuyo M, Sasaki A, Kobayashi I. 2009. Evolutionary maintenance of selfish homing
913 endonuclease genes in the absence of horizontal transfer. *Proc. Natl. Acad. Sci. U.S.A.*
914 106:18861–18866.

915 Yang F, Labani-Motlagh A, Bohorquez JA, Moreira JD, Ansari D, Patel S, Spagnolo F, Florence J,
916 Vankayalapati A, Sakai T, et al. 2024. Bacteriophage therapy for the treatment of Mycobacterium
917 tuberculosis infections in humanized mice. *Commun Biol* 7:1–13.

918 Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*
919 24:1586–1591.

920 Zeynali kelishomi F, Khanjani S, Fardsanei F, Saghi Sarabi H, Nikkhahi F, Dehghani B. 2022.
921 Bacteriophages of Mycobacterium tuberculosis, their diversity, and potential therapeutic uses: a
922 review. *BMC Infectious Diseases* 22:957.

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

926 Zhao H, Christensen TE, Kamau YN, Tang L. 2013. Structures of the phage Sf6 large terminase provide
927 new insights into DNA translocation and cleavage. *Proceedings of the National Academy of*
928 *Sciences* 110:8075–8080.

929 Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva
930 V. 2018. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at
931 its Core. *Journal of Molecular Biology* 430:2237–2243.

932