A Selection for Intein Splicing in

Phage Assisted Continuous Evolution

By

Andrew Hennes

B.S. Chemistry, Computer Science and Molecular Biology
Massachusetts Institute of Technology, 2023

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF ENGINEERING IN COMPUTER SCIENCE AND MOLECULAR BIOLOGY

at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Authored by: Andrew Hennes
Department of Electrical Engineering
and Computer Science, Chemistry.
May 12, 2023

Certified by: Xiao Wang
Assistant Professor of Chemistry, Thesis Supervisor

Certified by: David Liu
Professor of Natural Science, Thesis Supervisor

Accepted by: Katrina LaCurts
Chair, Master of Engineering Thesis Committee

A Selection for Intein Splicing in Phage Assisted Continuous Evolution

by

Andrew Hennes

Submitted to the Department of Electrical Engineering and Computer Science
on May 12, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Computer Science and Molecular Biology

Abstract

Developing tools for target manipulation of protein chemistry remains a longstanding goal in chemical biology. Trans-splicing inteins fulfill a crucial niche in this area by enabling post translational splicing of separate polypeptides, introduction of post translational modifications, and other chemistry on the amide backbone. This has motivated engineering and evolution of bespoke intein properties, including extein constraints, split site, kinetics, and orthogonality. Rational approaches are highly biased and preexisting directed evolution methods are laborious and often struggle with guaranteeing splicing dependance in a selection. To accelerate intein evolution we introduce a phage assisted continuous evolution (PACE) for intein properties. We show this selection is strictly splicing dependant, discriminates between intein splicing rates from one minute to several hours, discriminates between preferred over unpreferred extein contexts, supports propagation of phage in a multi-passage PANCE format, and circumvents recombination-driven phage cheating through a recombination-resistant helper strain. We anticipate this selection will enable facile, rapid evolution of inteins with bespoke properties.

Thesis Supervisor: Xiao Wang
Title: Assistant Professor of Chemistry. Core Member of the Broad Institute.

Thesis Supervisor: David Liu
Title: Professor of the Natural Sciences. Core Member of the Broad Institute.

Contents

List of Figures

Chapter 1

Introduction

Inteins are single-turnover enzymes which catalyze the splicing together of separate polypeptides through a native trans-amide linkage[1]. Since their initial discovery as genetic parasites of protein coding sequences[2], inteins have been the focus of intense biotechnological interest for their applications in canonical and noncanonical protein synthesis[1,3], understanding of mammalian biology[4,5] and synthetic control of protein function[6,7]. Most inteins are natively contiguous peptide sequences which splice in-cis with their host coding sequence to excise themselves[1]. Manipulation of this conserved intein splicing mechanism has enabled trapping and control of progression through the sequential splicing intermediates for useful synthetic[3], chromatographic[1], and chemical biology[6] applications. However, the observation that a subset of inteins are natively or can be artificially split has been most important in recent applications of inteins in chemical and synthetic biology[3,4,5,7,8]. This is due to protein trans-splicing enabling orthogonal and efficient protein reconstitution[7], more robust splicing-dependent control of protein function[6], and therapeutic reconstitution of proteins from separate coding sequences in vivo[8].

However applications of trans-splicing inteins have been limited by the role of extein sequences in the catalytic mechanism[9], irreversibly-formed off-pathway covalent products[10], incorrect folding pathways[11], the insolubility of individually expressed inteins[12], the synthesizability of intein fragments[13], and generally slow on-pathway splicing kinetics[7]. The diverse distribution of inteins in all three domains of life[1]

has motivated biotechnology-driven efforts which have unearthed inteins with broadened extein promiscuity[14] and accelerated splicing kinetics and efficiency[15] in vitro and in vivo. However, despite such successes many applications such as native protein reconstitution[3] require highly bespoke extein requirements or splicing properties absent from native inteins. To address these challenges both kinetics[14] and extein promiscuity[16] have been rationally improved. However, such rational engineering is often prohibitively laborious and lacks the unbiasedness and throughput of properly designed directed evolution selections[17].

Directed evolution of intein splicing has previously been shown to be capable of tailoring certain bespoke properties of intein splicing. In response to the challenge of rational protein engineering[18], direct coupling of intein splicing to bacterial growth was achieved by identification of splicing-dependent split sites in kanamycin aminoglycosidase (KanR) [19]. While this[19,20] and similar[20] approaches have yielded improvements in promiscuity of select inteins[19] and small molecule control of splicing[20], intein directed evolution has shown limited success compared to bioinformatic mining and discovery in identifying intein with desired properties. In particular, alternating sequential mutagenesis by PCR and selection based on resistance or cell sorting selection is laborious, limited in library size and mutation rate, and limited in applications which likely require extensive campaigns, such as recoding an intein fragment.

Phage assisted continuous evolution (PACE) addresses the discontinuity by directly coupling replication of the filamentous M13 bacteriophage to the activity of a phage encoded gene[21]. This is implemented through a host bacterial circuit which couples a desired biochemical activity to production of propagation-competent M13 pIII, which is strictly required for phage infection[22]. PACE has been used to evolve modified T7 RNA polymerase promoter recognition[23], amino-acyl tRNA transferase activity[24], recoded protease substrate recognition[25], DNA deaminase activity[26], simultaneous protein affinity and solubility[27], CRISPR-Cas endonuclease activity promiscuity[28], and de-novo deaminase activity[29]. PACE has shown broad biotechnological utility as a rapid and generalizable system for protein evolution. Here, we propose, develop, and validate a PACE circuit architecture for intein evolution

that is strictly splicing dependant, discriminates between intein splicing rates from one minute to several hours, discriminates between preferred over unpreferred extein contexts, supports propagation of phage in a multi-passage PANCE format, and circumvents recombination-driven phage cheating through a recombination-resistant helper strain.

Chapter 2

Results

2.1 Using a pIII And-Logic Gate to Engineer a PACE Circuit for Splicing Selection

2.1.1 Repurposing And-Logic Gate Split pIII for Intein Evolution

Previously a pIII allele split using the trans-splicing NpuDnaE intein was reported which requires expresses of both N- and C-terminal pIII fragment to support phage propagation. We hypothesized that expression of one of the pIII fragments from the M13 bacteriophage would place selective pressure for mutations that enhance intein splicing when the other intein–pIII fragment were expressed in the bacterial host (Fig. 2-1).

2.1.2 The Signal Peptide Split Site is Tolerant of Diverse Hexapeptides

Inteins are endogenously inserted into diverse sequence contexts in native host proteins. Significant divergence from the native chemical extein environment often interferes with or completely inhibits intein splicing activity[9]. Biotechnologically useful inteins are frequently identified from diverse intein families beyond the DnaE family[15].

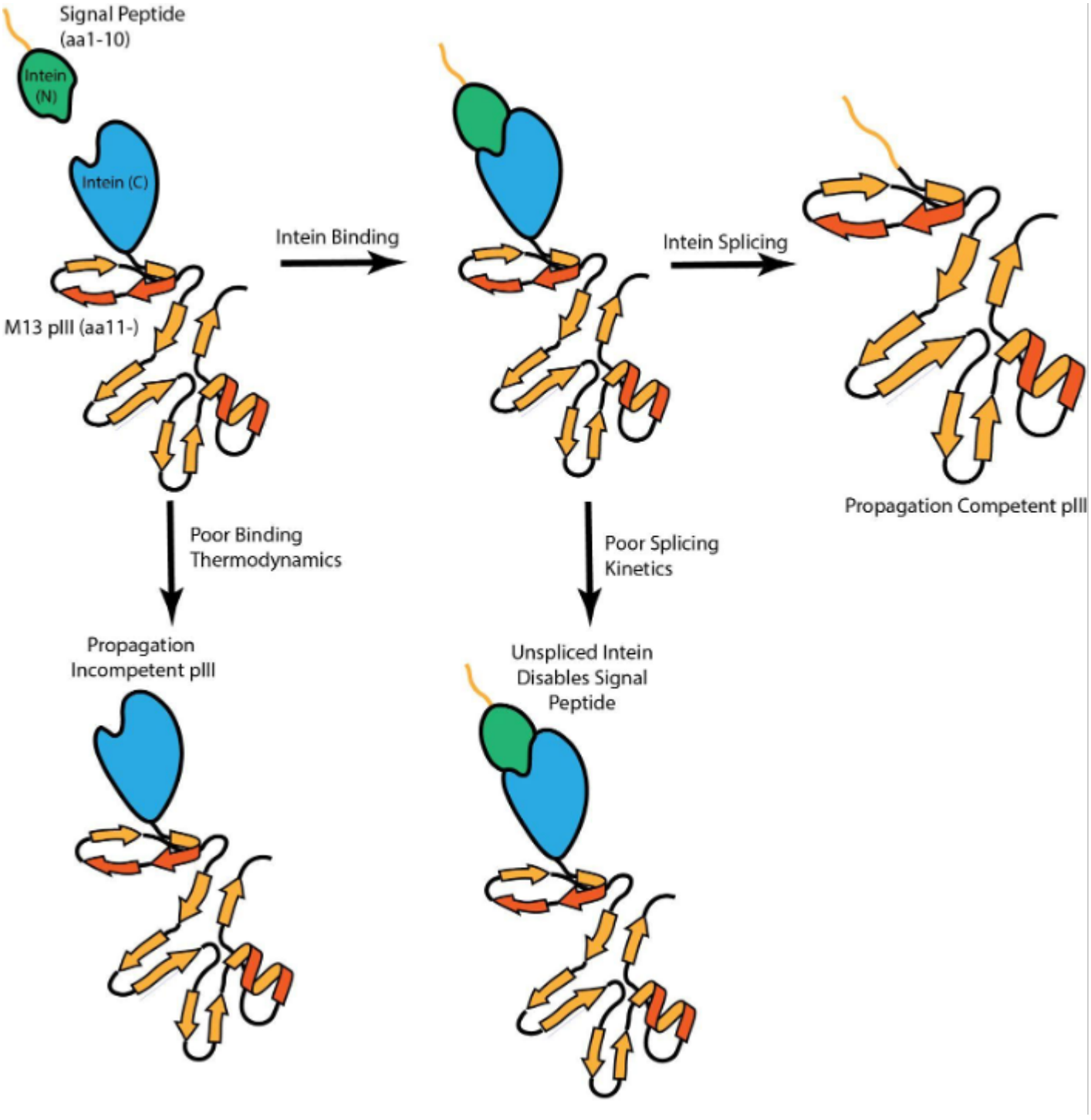

Figure 2-1: Production of propagation competent pIII should be dependant on efficient association and splicing of pIII split at the signal peptide sequence. This is associated with M13 fitness through the requirement of pIII incorporation into M13 bacteriophage particles to support phage infectivity[21]

Therefore an ideal splicing selection circuit would tolerate arbitrary extein sequences or a subset of extein which capture most common a wide chemical range of local extein environments. As pdrevious studies have reported limited mutational tolerance at residues adjacent to protein split sites[20], we sought to test the effect of extein scar sequence on phage propagation. We expressed pIII with three amino acidsfrom the N-terminal extein and three from the C-terminal extein, which have been shown to be most important in influencing intein splicing kinetics [9]. We chose extein contexts from phylogenetically diverse inteins and tested their effect on phage propagation by incorporating them into the S2208[21] helper strain and testing the modified strains ability to support M13 phage propagation expressing maltose binding protein (MBP) (Fig. 2-2). The effect of various extein scars on M13 phage propagation was no more than 2 to 3-fold which is insignificant in the context of PA(N)CE. This shows that the signal peptide split sites is relatively tolerant of diverse extein scars which could be inserted into pIII at the signal peptide.

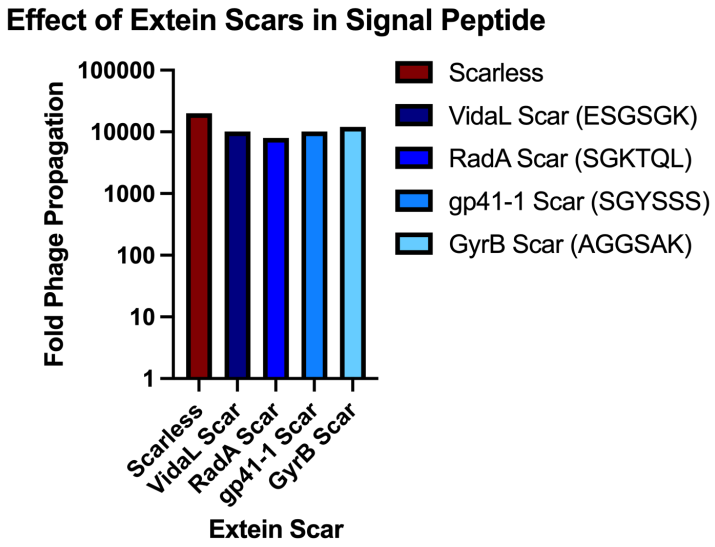

Figure 2-2: Propagation of M13 bacteriophage expressing MBP on the S2208 helper strain with different extein scars inserted into the signal peptide.

2.1.3 Splicing Dependence of pIII Split at the Signal Peptide

Protein split sites often show incomplete or completely absent dependance on splicing of correctly bound split inteins[20]. However, proper intein association and folding is necessary but not sufficient for intein splicing, as mutations can preserve the former while inactivating the latter. To measure the dependence of phage propagation on splicing progression through the conserved mechanism of intein splicing[1], phage expressing N terminal NpuDnaE were used to infect host strain expressing C terminal NpuDnaE intein. The third (N136A), second (C+1A), or first (C1A) step in sequential intein splicing was inhibited by mutation to alanine either on the phage or host strain as appropriate. Another intein, gp41, which is known to be completely orthogonal to NpuDnaE intein was used as a non-binding intein negative control. While the most active splicing inactive variant supported positive phage propagation, the level reported was still 100 fold below phage able to complete intein splicing. This suggested that the circuit likely selected for intein splicing over non splicing phage.

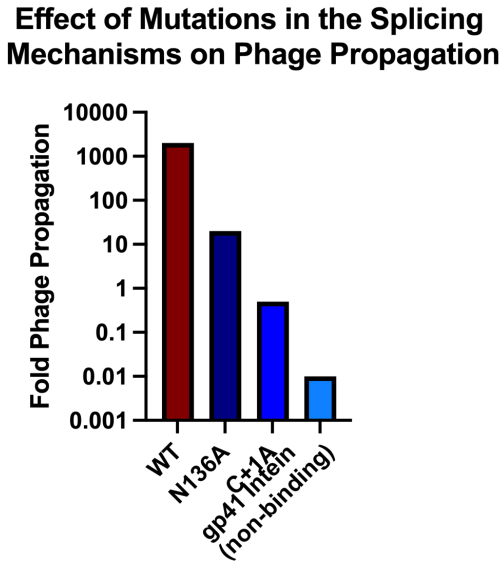

Figure 2-3: Propagation of M13 bacteriophage expressing N-terminal NpuDnaE or N-terminal gp41 on host strain expressing C-terminal NpuDnaE. Mutations are in the splicing mechanism of NpuDnaE using standard intein and extein residue counting notation[1].

2.1.4 Selection for Splicing over Non Splicing Intein Variants

To explicitly show splicing intein variants are selected over non-splicing variants, a multipassage PANCE-like experiment at 1e8 pfu/mL with NpuDnaE(C1A) in 1000 fold excess to wildtype NpuDnaE expressing phage. Phage were sequentially passaged at 1:100 over 6 passages, each passage was plaqued, and four such plaques were sequenced to identify whether they were splicing inactive (C1A) or splicing active (WT). After a single round of selection 0 our of 4 plaques when sequenced showed the presence of the catalytically inactive NpuDnaE(C1A). This same result was observed for four additional passages. This suggests selection through reconstitution of the signal peptide can select for and reliably maintain splicing active phage over splicing inactive phage variants. This experiment was run with six replicates in separate lagoons which all showed the same aforementioned result.

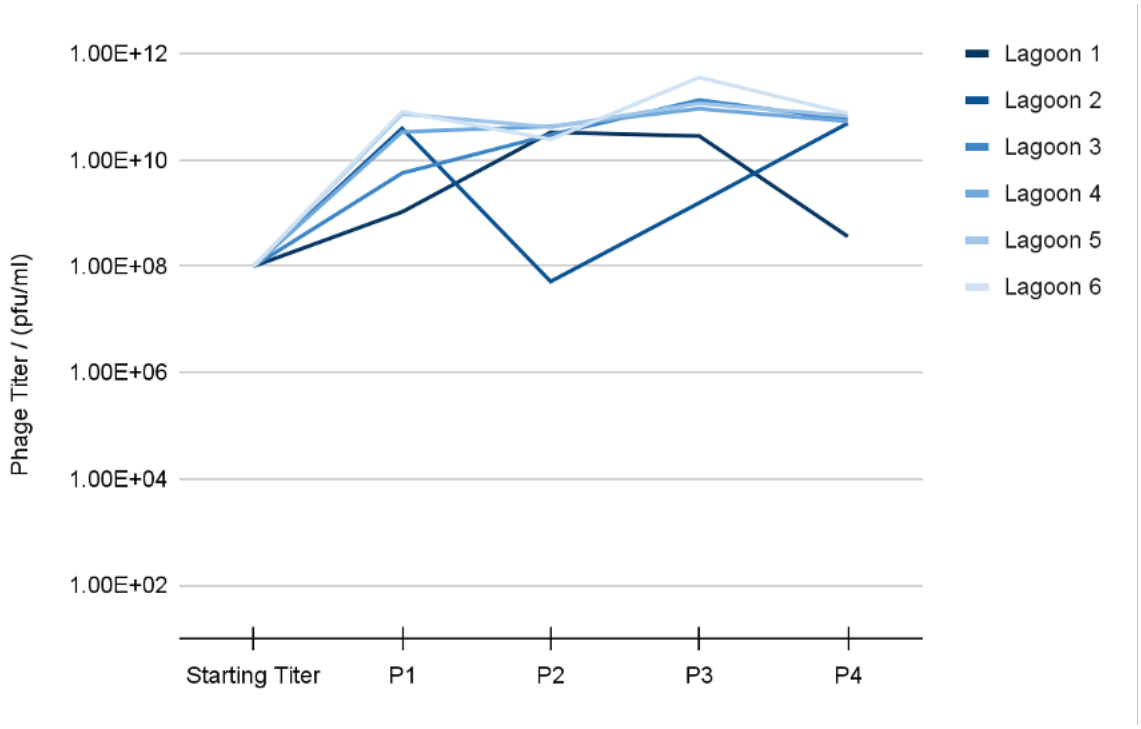

Figure 2-4: Selection of splicing active and splicing inactive phage over five passages.

2.1.5 Evolution of Kinetically-Impaired Engineered Intein Variants

Previous efforts to engineer NpuDnaE intein, included engineered a "charge swapped" (CS) NpuDnaE variant that is orthogonal to its wild type sequence [11]. However, rational engineering of intein properties can inadvertently harm other intein properties, such as splicing kinetics [1, 11]. In particular, CS NpuDnaE exhibits worsened splicing kinetics with its other engineered half.

To improve splicing activity of engineered orthogonal "charge swapped" CD NpuDnaE variants, we evolved the N-terminal NpuCS to improve splicing with its cognate, host-expressed C-terminal CS NpuDnaE fragment (Fig. 5). The increase in phage titer at passage three reflected mutagenic drift in the 2208 strain. Other changes in phage propagation reflect the effect of improved phage fitness and user controlled changes in dilution rate from passage to passage. The selection was run in triplicate with three separately passaged PANCE lagoons.

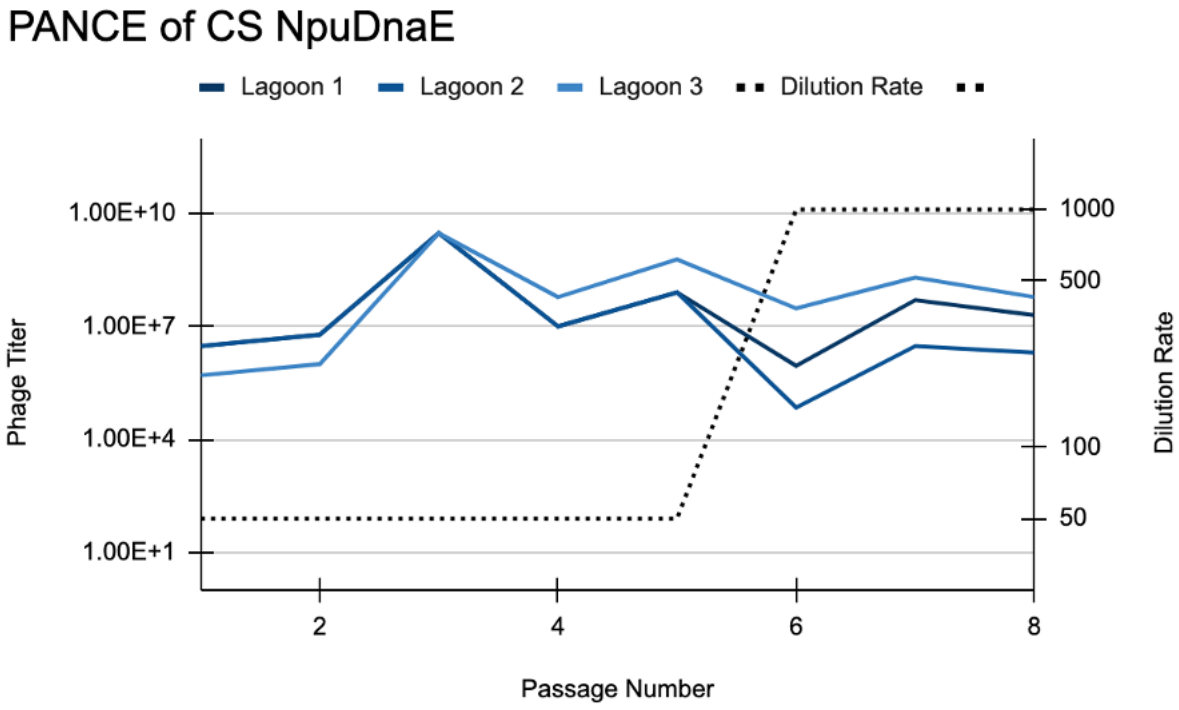

Figure 2-5: Selection of NpuDnaE CS Phage

2.1.6 Evolution of CS NpuDnaE Intein Identifies Structurally Interpretable Evolutionary Solutions

The three lagoons identified separate sets of evolutionary solutions. Each lagoon contained at least three amino acids in each of the lagoons as determined by at least 3/4 of sequenced phage containing that mutation. The observation of multiple non-identical solutions between lagoons does not necessarily suggest differences in selection conditions but rather can be explained by the stochasticity of mutational events and the effect of clonal interference[17]. Of the six mutations which converged in a lagoon, four were present adjacent to a "charge swapped residue" (E5K, I8V, L15P, and D93A/N) and two swapped locations of a putative salt bridge (K24E and E27K). This suggests that most observed conserved mutations are accommodating rationally engineered charge swapped mutations.

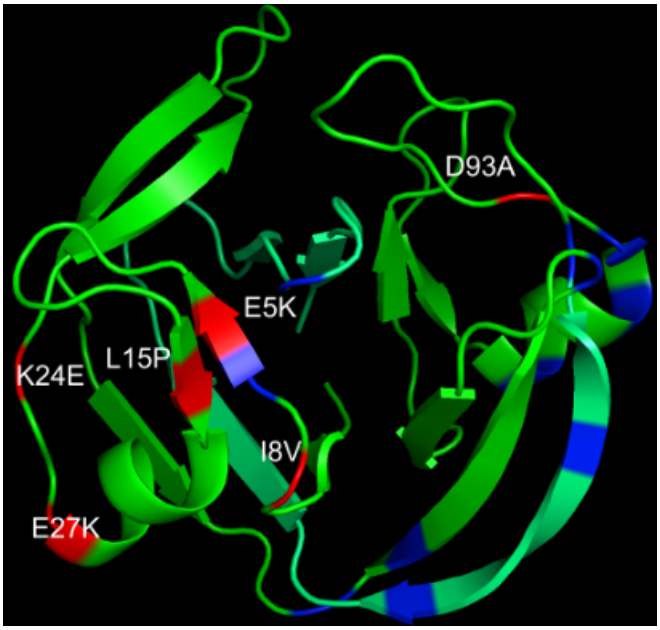

Figure 2-6: Location of CS mutations and PANCE evolved mutations mapped onto a crystal structure of NpuDnaE intein.

2.1.7 Evolved Phage Show Improved Propagation and Fitness

Fold phage propagation is expected to improve as higher fitness phage are selected in PANCE. As M13 phage which complete their life cycle more quickly and with more resulting infection-competent M13 phage particles are more fit, measurements of improved phage propagation should reflect increases in fitness. To measure improvement in phage propagation of evolved phage populations, propagation of evolved and starting genotype CS NpuDnaE were measured. Phage propagation was observed to improve at least 100 fold from the starting phage genotype, which suggests a large improvement of phage fitness over <1 week of directed evolution.

	E5	I8	E12	L15	K24	E27	V44	D62	D78	K89	D93
Lagoon 1: Plaque 1					E	K					A
Lagoon 1: Plaque 2					E	K					A
Lagoon 1: Plaque 3					E					E	A
Lagoon 1: Plaque 4					E	K					A
Lagoon 2: Plaque 1					E						A
Lagoon 2: Plaque 2	K			P	E				N		N
Lagoon 2: Plaque 3	K		K	P			G	N			
Lagoon 2: Plaque 4	K			P	E						N
Lagoon 3: Plaque 1		V				K					A
Lagoon 3: Plaque 2		V				K					A
Lagoon 3: Plaque 3		V			E	K					N

Figure 2-7: Mutational adaptions after six passages in PANCE.

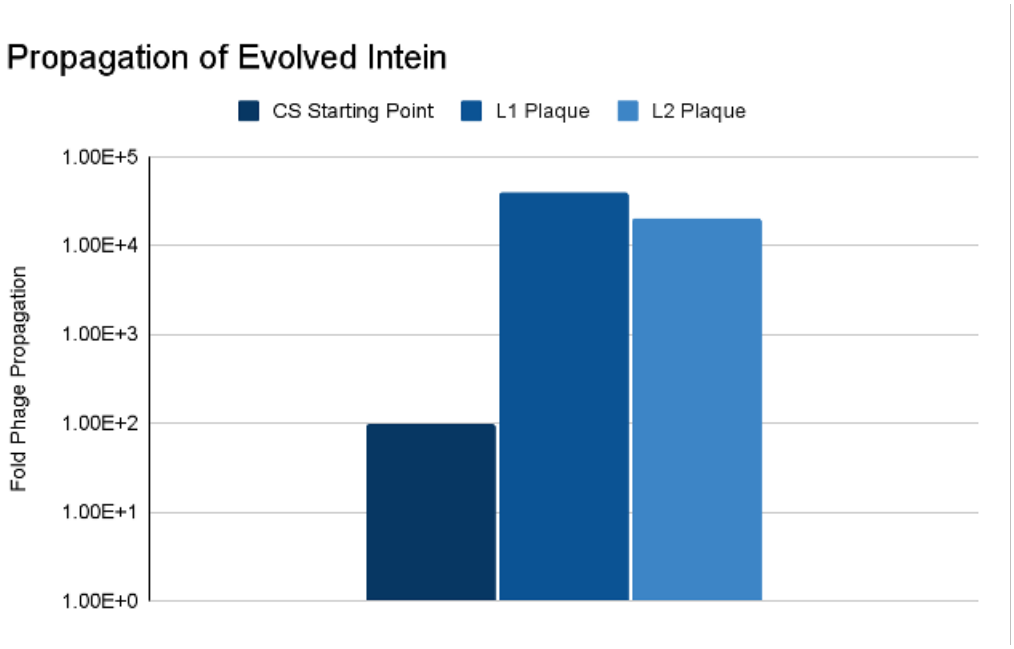

Figure 2-8: Location of CS mutations and PANCE evolved mutations mapped onto a crystal structure of NpuDnaE intein.

2.1.8 Propagation of Diverse Intein-Expressing Phage with Controlled Stringency

Control of phage propagation is commonly implemented by tuning the strength of the phage shock promoter or ribosomal binding site (RBS) for the pIII coding sequence in the PA(N)CE host strain. Under the assumption of pseudo first order intein kinetics, which describes intein splicing[1], the rate of concentration of spliced pIII is proportional to the concentration of correctly assembled split intein–pIII fusion and the rate of intein splicing [1]. When host expressed intein is decreased (assuming this is less then phage-expressed intein), the maximal concentration of the correctly assembled split intein will be proportionately decreased. If this results in a decrease in the rate of pIII production or pseudo steady state concentration of pIII below the level required for maximally rapid phage propagation, it would be expected decreasing the

[1] $[pIII] = \frac{k_{p3}}{t_{\frac{1}{2}}} * \frac{k_{splice}}{k_{splice}+\alpha}$ up to a constant under the assumption of pseudo steady-state, where k_{p3} is the rate of pIII production of the phage- or plasmid-encoded pIII-fragment (whichever is lower), $t_{1/2}$ is the rate bacterial division, and α is function of $t_{1/2}$. Note that in the limit of inefficient splicing kinetics this is linear in both the amount of pIII produced and the rate splicing.

expression of host encoded pIII could be a mechanism for controlling stringency of the intein splicing selection.

However, it is unclear whether the relationship between intein kinetics and phage fitness is permissive to propagation of slow splicing inteins. The C terminal domain of pIII is known to be necessary and sufficient for assembly of pIII into intact phage particles[30]. Thus equal incorporation of unspliced, correctly-assembled split pIII with correctly spliced pIII into a phage particle might result in uniformly infection-incompetent phage [2]. If the results were a greater than linear dependence on the fraction of intein spliced, this would select especially strongly against slowly splicing intein variants.

2.1.9 Initial Fold-Propagation of Phylogentically and Kinetically Diverse Inteins

To test these ideas in inteins of diverse phylogenies and known splicing half lives, we expressed gp41, RadA intein, and GyrB intein from phage with the cognate fragment-pIII fusion expressed in the 2060 host strain. Fast splicing gp41 showed similar propagation to the similarly fast splicing NpuDnaE intein. The level of phage propagation could be tuned by decreasing the level of expression of the C terminal intein fragment. This suggests controlling the level of host expressed intein can challenge even rapidly splicing intein for improved fitness. RadA and GyrB, by contrast, have much slower splicing kinetics on the order of an hour or longer. Of three RadA split sites tested and one GyrB split tested, only the atypically split C terminal RadA split showed positive in the lowest stringency.

2.1.10 Evolution of gp41 Split Intein

To determine whether fast splicing gp41 intein could propagate under conditions not permissive to the wild type sequence, we passage M13 phage expressed gp41 under

[2]the true distribution is likely better approximated by a multinomial distribution, however the simplistic example above exhibits the same behavior.

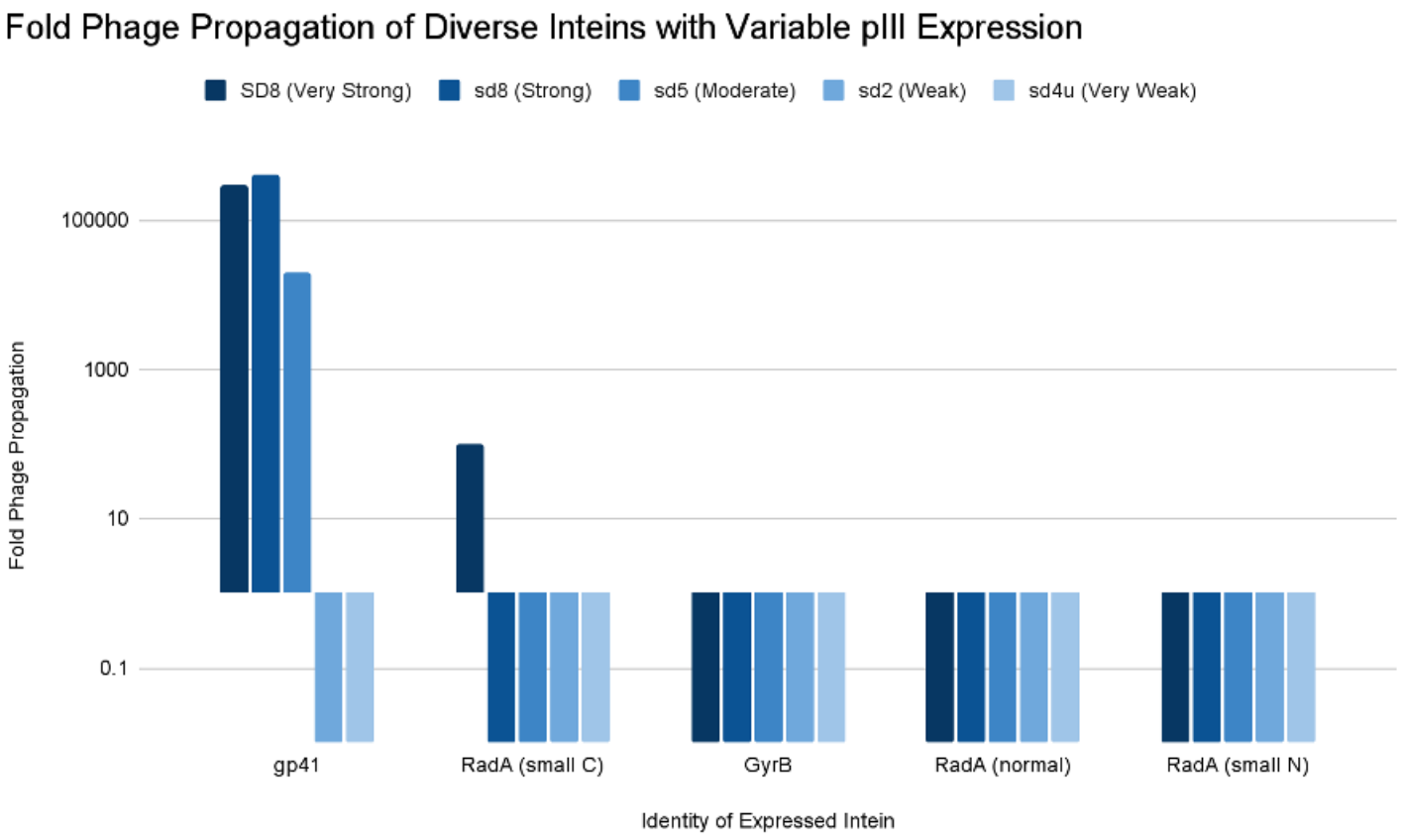

Figure 2-9: PANCE campaign on gp41-1 intein.

17 rounds of PANCE. The campaign was initially began with the permissive sd5 RBS strength however was then later switched to alternating passages of sd5 and sd2 expression level. All passages were done at a challenging 1:1000 fold dilution.

Propagation of gp41 significantly improved following mutation of S-3M at passage 3. While by the end of the campaign phage began tolerating 1:10 dilution in sd2 expressing host strain with an associated E65D/K mutation in two of the six lagoons, the ultrafast splicing activity of gp41 and the paucity of further fitness gains suggest sd2 posed fundamental expression level limitations for further improvements in phage propagation. The observation of significant improvements in splicing activity by evolution of extein suggests suggests either that the signal-peptide split site is highly sensitive to improvements in in-vivo splicing even with a few second splicing half life, or that subtle improvements in the extein sequence could still improve phage fitness.

2.1.11 Evolution of RadA C Terminally Split Intein

To determine whether we could identify mutations in PhoRadA that would improve its splicing rate, we began a PANCE campaign of C-terminally split PhoRadA. Evolution of PhoRadA over ten passages led to marginal improvements in propagation with no associated changes in PhoRadA sequence. To increase stringency passages were alternated between SD8 and sd8 RBS strength at 1:10 dilution. At the end of the evolution campaign some lagoons began to independently acquire frameshift mutations which eliminated the majority of the PhoRadA sequence. Furthermore some inteins were observed to obtain mutations in key catalytic residues. This suggested that under high selection stringency a splicing-independant cheating mechanism had developed to maximize the fitness of the unspliced complex of N and C terminally associated intein. Furthermore, propagation of GyrB and the other two splits of RadA by alternating mutagenic drift and selection failed to identify improved splicing inteins.

	S-3	E65
Lagoon 1	M	
Lagoon 2	M	
Lagoon 3	M	
Lagoon 4: plaque 1	M	D
Lagoon 4: plaque 2	M	
Lagoon 4: plaque 3	M	
Lagoon 4: plaque 4	M	
Lagoon 5	M	
Lagoon 6: Plaque 1	M	
Lagoon 6: Plaque 2	M	K
Lagoon 6: Plaque 3	M	
Lagoon 6: Plaque 4	M	K

Figure 2-10: Mutational outcomes from a PANCE campaign on gp41-1 intein.

2.1.12 Challenges Cloning PACE Strains and Exploring Intein Sequence Space by Mutagenic Drift

An intrinsic limitation of the application of canonical PACE helper strain (s2208) in intein evolution is its high rate of recombination with phage. This is due to the presence of 33 (N-terminal pIII on phage) to 1245 (C-terminal pIII on phage) nucleotide of homology between the N- or C- terminal half of pIII in phage and cooresponding segment of pIII in s2208. Cloning and propagation of intein-encoding phage in S2208s led to levels of recombinant phage of <1e2 pfu/mL as measured by plaqueing on 2060s. Coorespondingly, infection with 1e5 pfu/mL of N-terminal intein containing phage was sufficient to avoid inclusion of cheater phage. However, drifting to 1e8 pfu/mL led to cheater appearance in PANCE campaigns two passages after mutagenic drifting, which required restrating from prior passages in the PANCE campaigns.

2.2 Development of a More Optimal Circuit Architecture for Selection of Intein Splicing Activity

Advantages and limitations of the split pIII signal peptide architecture for evolution of intein splicing suggest a few properties that are likely necessary for a more optimal intein splicing PACE circuit.

First, the circuit should be strictly splicing dependent with negligible fitness for non-splicing intein variants. This suggests the mechanism for coupling intein splicing to protein activity should not rely on subtle differences in local conformational dynamics associated with splicing. Rather, the mechanism should directly select for the presence or absence of the intein domain itself which would prevent propagation of partially spliced intein variants. Great well put

Second, the circuit should be able to support intein propagation in diverse extein contexts. Previous studies have shown that the -1 and +2 extein residues immediately adjacent to catalytic splicing residues have the greatest influence on splicing half lives.

Thus, any intein selection circuit capable of propagating inteins with diverse extein preferences should be able accept a wide range of chemically distinct extein residues at these positions.

Third, the circuit should select for improved kinetics over a range of splicing half lives ranging from a few minutes to multi-hour splicing half lives. While evolution of engineered charge-swapped NpuDnaE intein variants in the signal-peptide split pIII circuit improved splicing kinetics in the minute time-scale, many biotechnologically useful inteins have undesirable splicing half lives ranging from 40 minutes to 4 hours. In the context of a PACE circuit based on pIII reconstitution, it is likely that unspliced intein must not only avoid forming propagation-competent phage particles but also not compete with unspliced signal-peptide split pIII to inteins much slower than the phage life cycle of approximately 30 minutes.

Fourth, constrained exploration of intein and extein sequence space by fixing residue identity can identify inteins with bespoke or promiscuous extein preferences, or reprogrammed intein target identity. However, exploration of the sequence space of both intein fragments might often be desirable in unbiased selections for improved splicing kinetics or orthogonality. To achieve a high level of plasticity regarding intein evolution architecture choice, an intein evolution circuit should enable evolution of N and C-terminally split inteins with selective choice of which fragments and exteins can be evolved.

Finally, an optimal intein selection architecture should support unbiased exploration of sequence space independent of selection for desired splicing activity. The reported split-pIII architecture is incompatible with mutagenic drift in the standard PACE 2208 drift strain due to the presence of unavoidable, high homology of phage and plasmid encoded pIII. The resulting high rate of plasmid to phage transfer of genetic information leads to a rate of cheater phage formation $>1/100$ pfu/mL.

2.2.1 A T7–Lysozyme Fusion-Based Alternative Architectures for Evolution of Intein Splicing

We considered two potential circuits that would satisfy the above five listed criteria for a more optimal splicing circuit. After extensive numerical simulation based optimization of the first circuit which used splicing of the T7–lysozyme linker to activate T7 transcription (Fig. 13), four of the five criteria above were satisfied. However, phage propagation is highly dependent on ultrafast, sub-minute splicing half-life of gp41 due to the uncompetitive mechanism of T7 inhibition by lysozyme (Fig.15), which failed to be activated even by the slower splicing NpuDnaE intein (Fig. 14).

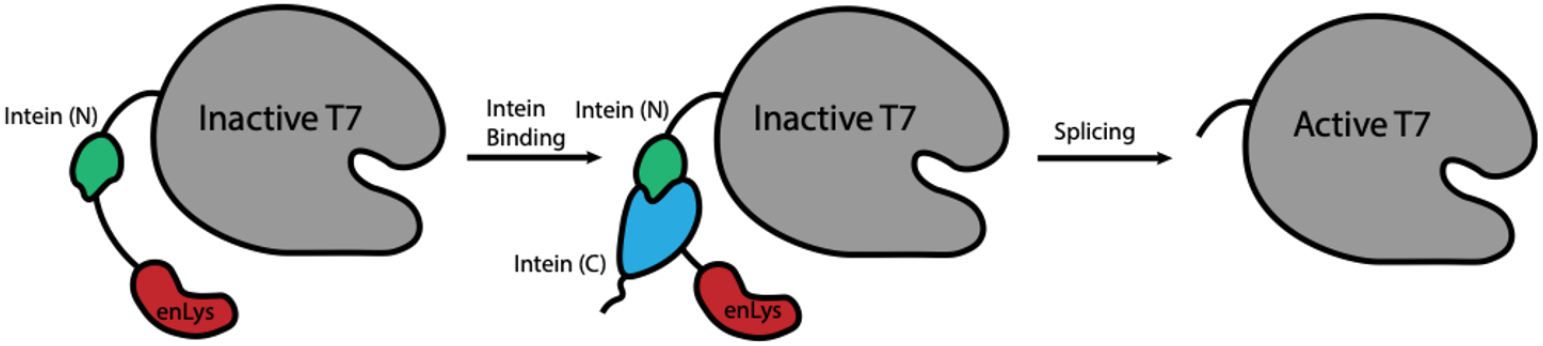

Figure 2-11: Mechanism for intein activation of T7-lysozyme circuit.

2.2.2 A Split pIII C-Terminus Based Architecture for Evolution of Intein Splicing

To satisfy the five aforementioned aspects of a more optimal intein splicing selection architecture, we chose to identify split sites in the most C-terminal domain of pIII. M13 pIII plays multiple roles in phage infection and replication, including recognition of the primary F-receptor, recognition of a necessary coreceptor, and assembly into the phage particle[30]. It has previously been shown that genetic deletion of the C1 domain, which is required for infection, can implement a dominant negative pIII allele. While the structural and functional details of all but the two most-terminal domains are

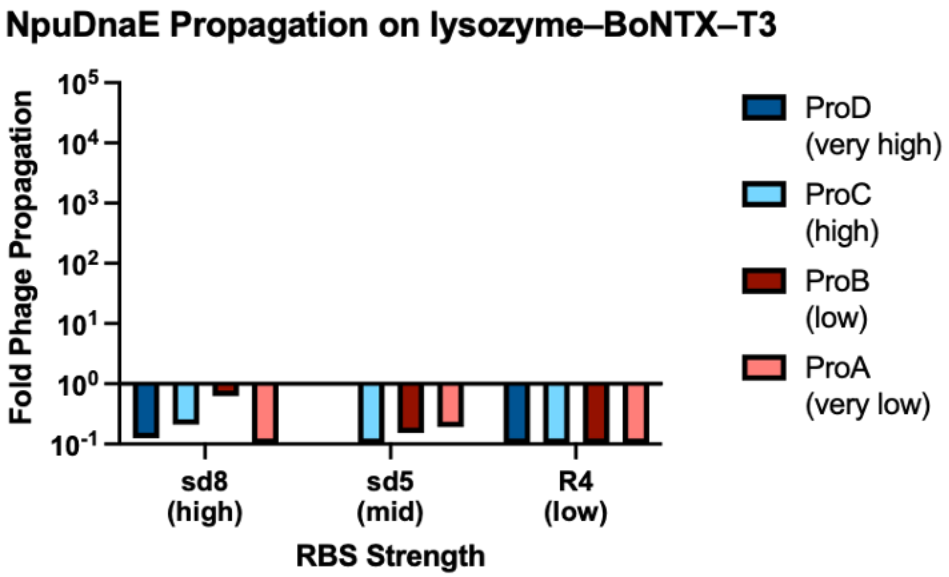

Figure 2-12: Propagation of NpuDnaE expressing phage on T7-lysozyme circuit.

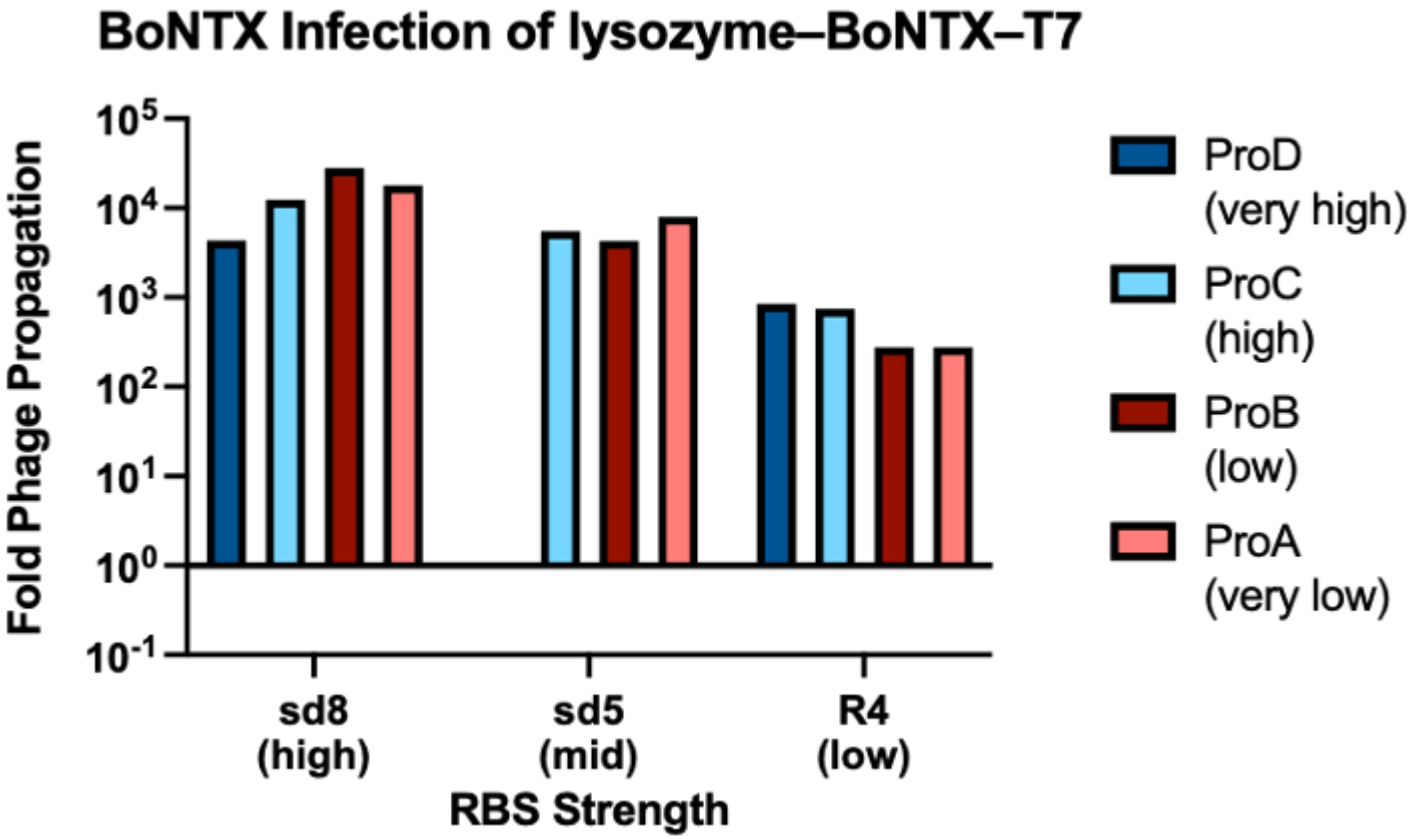

Figure 2-13: Propagation of gp41-1 expressing phage on T7-lysozyme circuit.

known, we hypothesized that the hydrophobicity and known role in phage assembly of C2 suggested it was likely buried in a hydrophobic core of the phage particle.

We hypothesized that splicing within the second C-terminal domain of pIII would select for splicing by sterically preventing assembly of unspliced pIII allele into phage particles due to the presence of the entire intein HINT[1] domain in the C2 structure. In principle, this would also prevent unspliced pIII allele from interfering with the fitness of correctly spliced allele, enabling positive propagation with lower initial splicing kinetics. The C2 domain naturally contains most residues and diverse 2-mers, which suggests desired extein contexts can be obtained through properly chosen split sites. Furthermore, the ability to express the N-terminal, C-terminal, or both intein fragments in evolving phage controls evolution of intein components. Any non-phage expressed intein fragments would be expressed in host plasmid and enable tuning of maximum spliced pIII, thus implementing control of selection stringency.

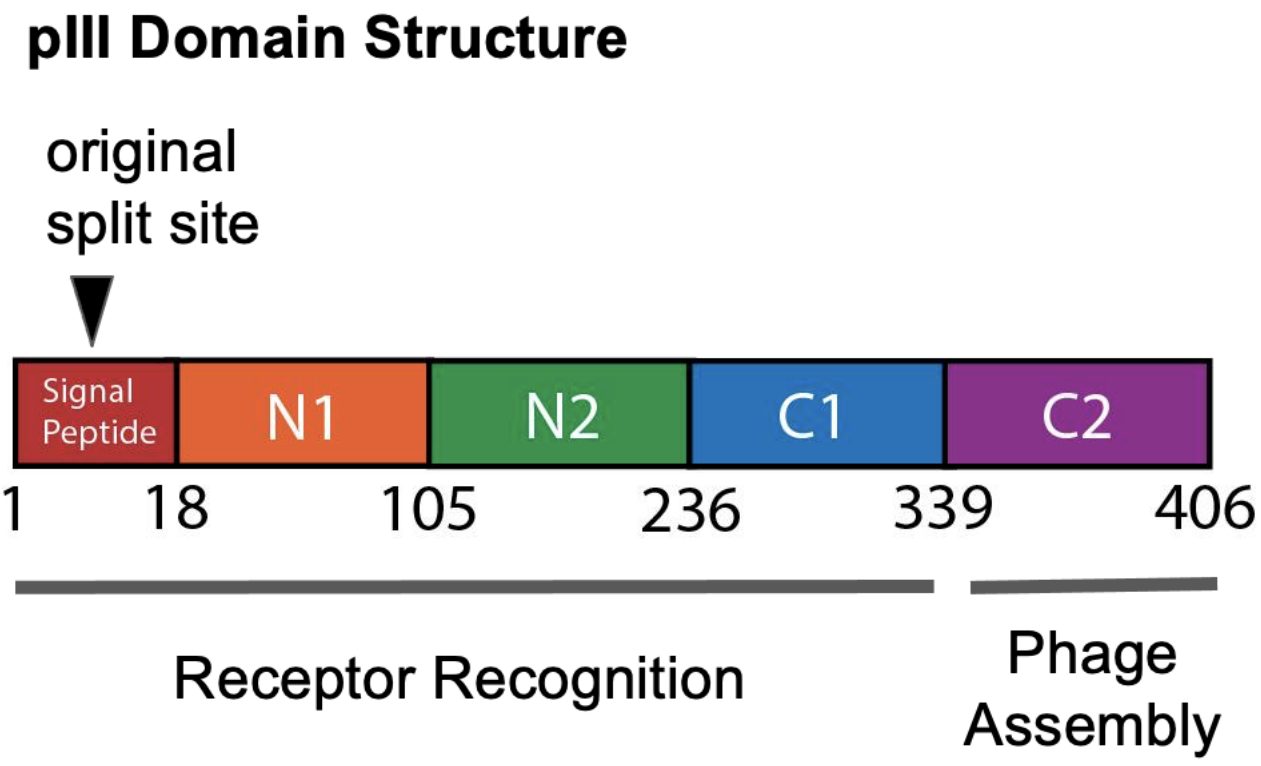

Figure 2-14: Domain organization of pIII.

2.2.3 Unavoidable Intein Catalytic Scars and an Alanine SAR Method for Identifying Site-Specific Steric-Sensitivity

An intrinsic challenge for identifying intein split sites in proteins, such as pIII, is identifying sites that tolerate the +1 catalytic residue present in all inteins. For DnaE inteins, this residue is natively a cysteine. This poses contradicting optimization challenges for any putative C2 pIII split site. First, any split site must be tolerant enough of change in sequence identity to accept a cysteine substitution that its incorporation into correctly-spliced pIII would not interfere with phage propagation. Otherwise, selection for phage propagation might be dominated by reprogramming extein contexts instead of optimizing intein properties. However, sequences which are most accepting of such mutations tend to be surface exposed residues, where unspliced intein would be most likely to not interfere with phage propagation. In previous efforts to identify split sites in KanR this has resulted in the vast majority of tolerated split sites being completely or largely independent of intein splicing. Thus, a more rational method for identifying a set of splicing dependent split sites with diverse -1/+2 extein contexts in pIII C2 is desired.

Alanine scanning is commonly used to coarsely map out structure activity relationships between side chain substitution and protein activity. Previously the relationship between the pIII C1 and C2 sequence and phage presentation has been mapped for phage display applications using alanine scanning[30]. We hypothesized that as alanine and cysteine have a similar hydrophobicity and show strong substitution tolerance in the BLOSSUM62 matrix, the tolerance of an alanine substitution in pIII C2 would likely correlate strongly with the acceptance of a cysteine substitution in the same sequence position. Furthermore, if the pIII C2 domain showed regions of strong non-tolerance of alanine substitution adjacent to individual sites accepting alanine substitution, it is likely that these sites would represent structurally conserved regions with serendipitous sites tolerant to side chain substitutions.

2.2.4 Identifying Accepted pIII C2 Split Sites

Fold drop in phage display after four rounds of selection [30] were analyzed and used as a proxy for how strongly an alanine mutation impacted the assembly of correctly assembled pIII. Values greater than one indicated de-enrichment and values less than one over-enrichment. Ten residue positions were identified that resulted in relatively low changes in phage propagation, were adjacent to sites which were intolerant to alanine substitutions, and showed at least moderate tolerance for VidaL extein residues or otherwise had a native +1 cysteine. Furthermore these sites were chosen to explore a large diversity of -1 and +2 extein sequence identities to enable selection on diverse intein starting points.

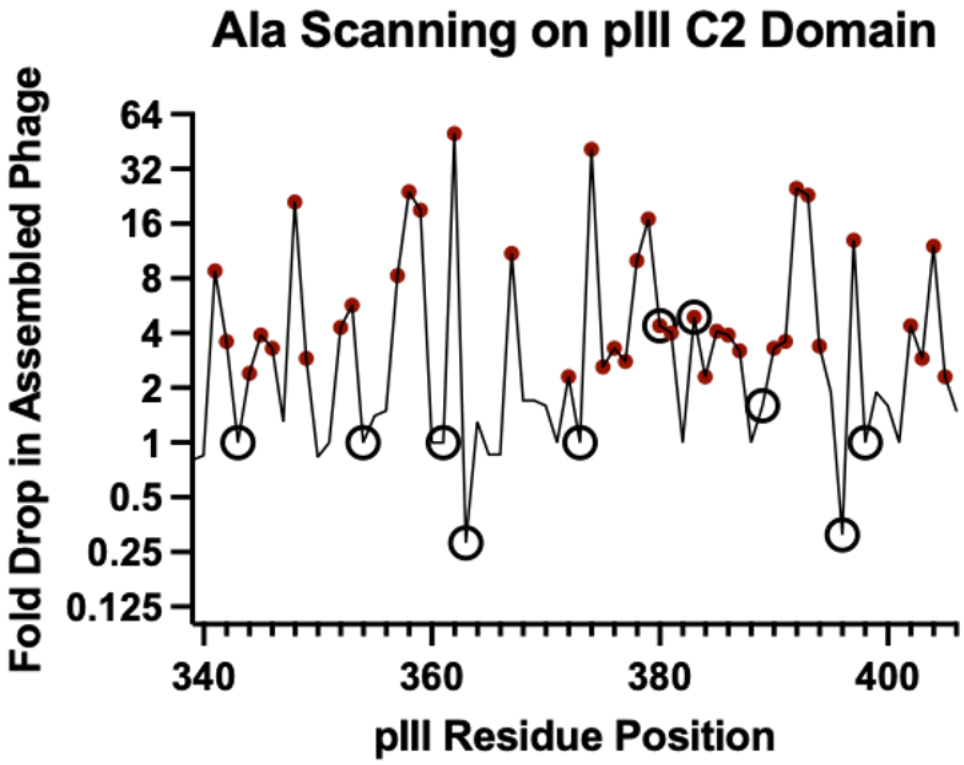

Figure 2-15: Selection of pIII Split Sites (labeled by there letter and circled in the diagram above).

2.2.5 1. Validation of Splicing Dependance of Selected pIII Split Sites

We validated that the identified split sites in the pIII C2 domain tolerate both a +1 cysteine scar and are dependent on splicing progression through the entire splicing mechanism. Intein splicing requires two separate and sequential events to ultimately splice together separate polypeptide chains. First two inteins must bind each other, usually rapidly, and then splice together with a characteristic half life. To specifically measure the ability of the correctly assembled intein HINT domain to splice and reconstitute propagation-competent pIII, Cfa was inserted in frame into the ten aforementioned split sites as the catalytically active wild type sequence or with each step in the splicing mechanism separately knocked out. M13 phage lacking pIII were then propagated on these strains and phage propagation was measured. Of these only one split site, split site H, showed phage propagation in the absence of splicing activity.

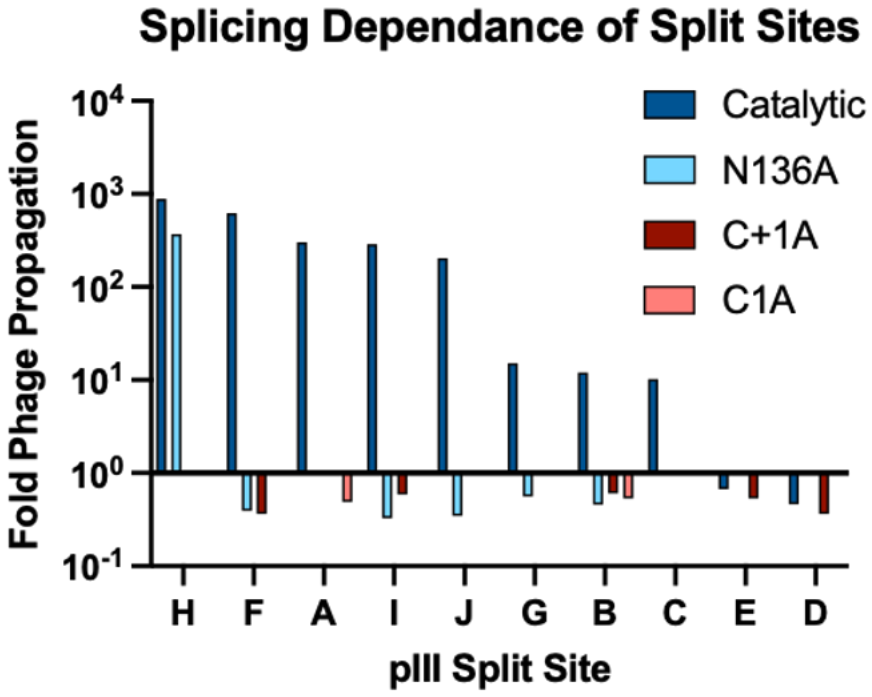

Figure 2-16: Effect of splicing inactivation on phage .

2.2.6 2. Validation of Discrimination of Splicing Kinetics Over Minute to Hour-Long Half Lives

:

We show that phage propagation is directly and strongly dependent on the splicing half life of Cfa in its split site extein context. The splicing of all inteins is highly dependent upon the local extein sequence. For Cfa intein this dependance lies completely on the +2 extein residue[31]. As split sites were partly chosen for diverse extein contexts, this enables directly testing the effect of manipulating solely the intein splicing rate of the correctly folded Cfa HINT domain on phage propagation. This approach to controls for conflating variables such as intein sequence, solubility, and thermodynamic stability which would be challenging if a diverse intein kinetics were tested by using inteins with naturally diverse splicing kinetics. Furthermore, a strong relationship between splicing kinetics and fold phage propagation would suggest that local sequence context outside of the extein has little influence on influencing intein splicing efficiency, which is a desirable characteristic.

To determine if the in vitro and in vivo splicing activity of Cfa intein explain the observation variation in phage propagation, we chose the 8 split sites which were splicing dependent and tolerated the Cfa +1 cysteine scar. We compared the in vitro splicing half life of these inteins with their measured ability to support phage propagation. A standard curve which was fit to the log-log plot of in vitro splicing kinetics and fold phage propagation strongly (R2= 0.92). Similarly, the previously well validated in-vivo KanR reconstitution assays are strongly correlated (R2= 0.96) with phage propagation. As the splicing rate of Cfa varies from one minute to three hours, these together suggest the circuit is dependant on the splicing rate of Cfa, even across split sites, and can discriminate splicing half lives from 1 minute to well over an hour by coupling these differences to changes in phage propagation.

Split	C-Extein
F	CF
I	CF
A	CY
J	CN
G	CL
B	CR
C	CG
D	CP

Figure 2-17: Extein contexts of splicing-dependant pIII split sites.

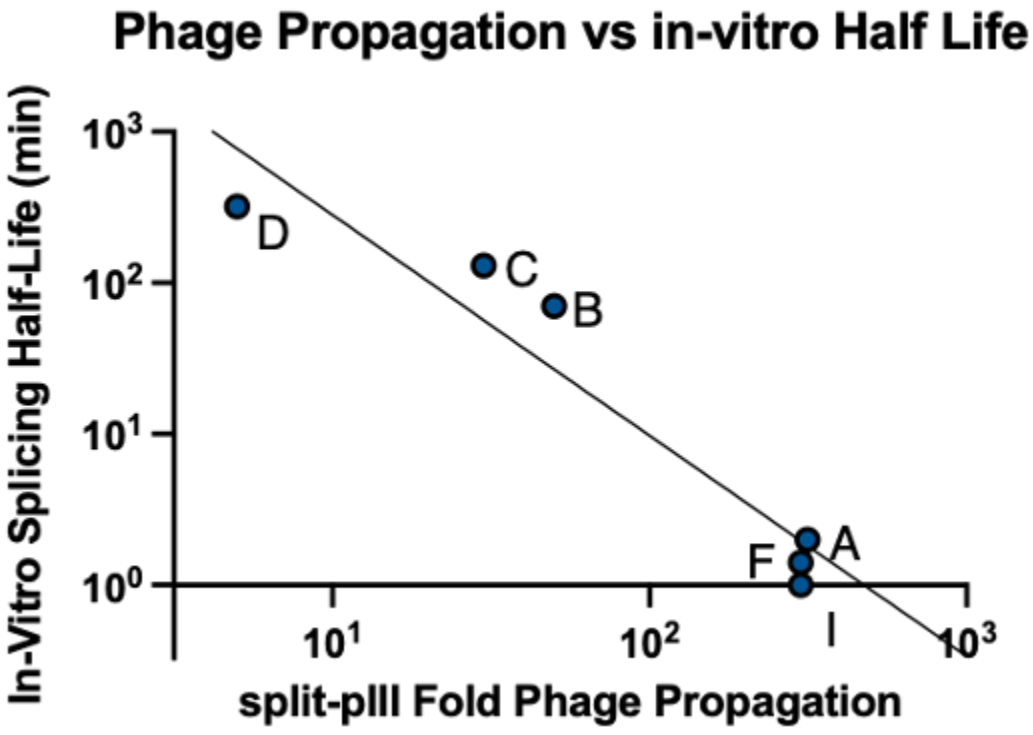

Figure 2-18: Relation between phage propagation and in vitro splicing half-life.

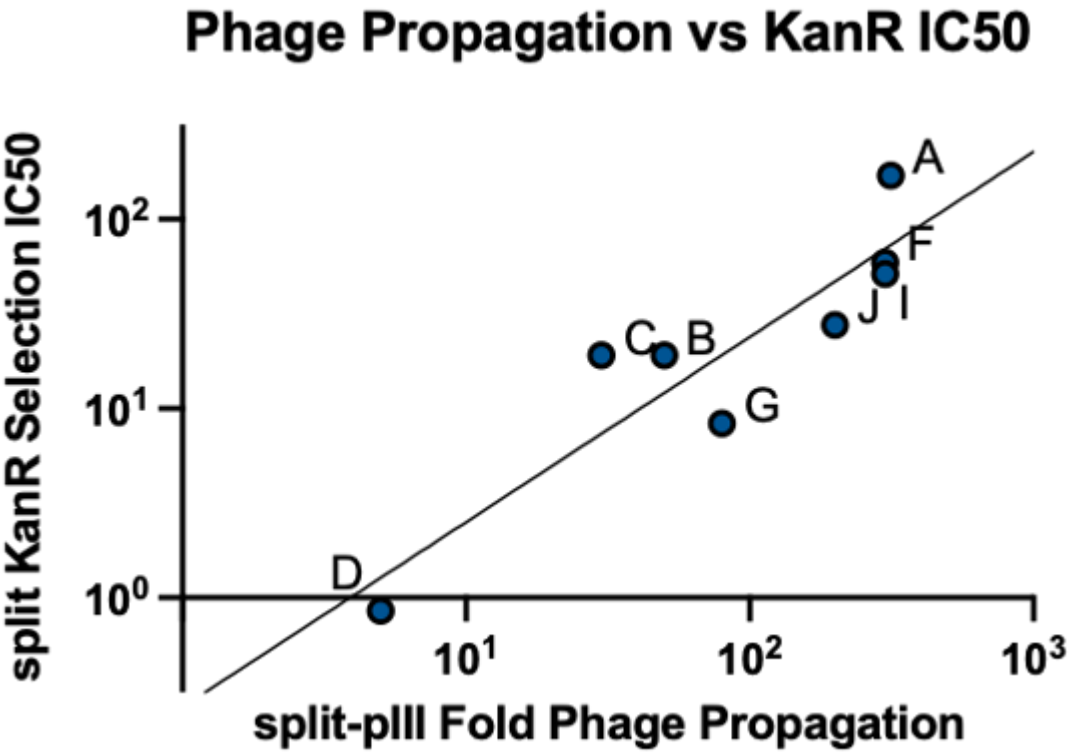

Figure 2-19: Relation between phage propagation and in vitro splicing half-life.

2.2.7 Plasmid Expressed Intein-Split pIII C2 Do Not Support Phage Propagation

To measure the effect of requiring intein association and splicing on phage propagation, N- and C- terminal split pIII were expressed from separate plasmids and spliced in-trans and assayed for their ability to support phage propagation. The DnaE intein family contains several biotechnological useful members and has been broadly characterized in the past. It is also known that most members have well-defined extein requirements similar to Cfa which should control for loss of splicing due to different extein requirements of non DnaE family inteins. Thus, the DnaE family of inteins including Npu, Cra, and Cwa were chosen and tested within a subset of extein contexts which had been previously characterized and shown to be active.

Furthermore, it has been observed several times independently that inteins are often highly insoluble when fused to split proteins. To avoid this two solubility factors,

SUMO and MBP, were employed based on their previously shown ability to solubilize inteins when fused to insoluble protein-intein fusions[32]. To avoid any incorporation of SUMO or MBP into the final pIII product spliced product, these solubility factors were fused immediately adjacent to the intein split site to avoid incorporation into the final splicing product.

To measure ability of pIII split at different splits sites and with various DnaE inteins to support propagation of phage, phage expressing the orthogonal gp41 intein were infected into corresponding strains (blue bars, below). No strain supported phage propagation above 5 fold.

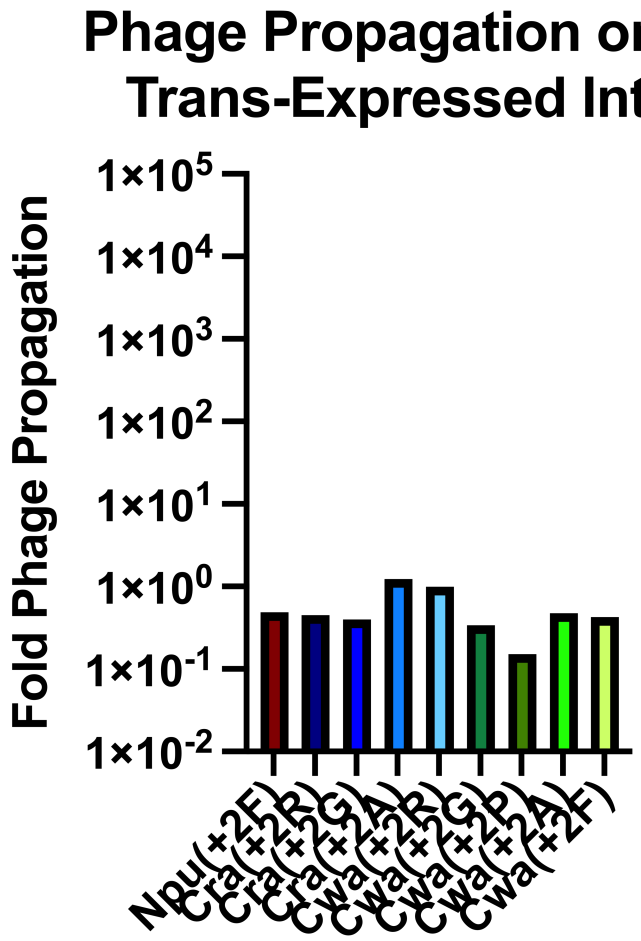

Figure 2-20: Propagation of phage lacking pIII on host strain expressing each half of pIII fused to intein.

To measure if an abscence of phage propagation was due to too little of N- or C-

terminal DnaE intein, phage expressing either N- or C-terminal Cwa intein fused to the pIII split site with a +2R extein were infected in the same strains. C-terminal intien-pIII fusion expressing phage did not lead to significant positive propagation in any strains measured. By contrast, N-terminal Cwa +2R extein expressing phage were observed to greatly improve propagation in the cooresponding split site for both Cwa and Cra intein, which is consistent with some degree of cross splicing between different members of the DnaE family. These results suggest that the N-terminal intein fargment is limiting in the intermolecular splicing reaction.

2.2.8 3. Validation of Expected Extein-Dependant Phage Propagation Behavior in-Trans

To test if Cra and Cwa intein recapitulated their known extein requirements in phage propagation, N-terminal Cra and Cwa expressing phage strains were used to infect strains expressing only the C-terminal cognate intein and the resulting fold-phage propagation was measured. Cra showed robust propagation on its wild type extein but strongly decreased propagation on the non-wild type +2R and +2G exteins. Similarly the previously well known broad extein dependence of Cwa intein was reflected in propagation relatively unaffected by potentially challenging exteins, including +2 proline which had not been measured previously. Together these show that splicing-dependent reconstitution of pIII in the C2 domain can discriminate based on preferred extein context of inteins.

2.2.9 4. Validation of DnaE Inteins in PANCE

:

To confirm that the newly designed circuit was capable of maintaining phage titer over multiple passages in PANCE, Npu, Cra, and Cwa expressing phage were propagated on strains encoding their cognate intein half over six propagations. Over this time titer was observed to stay stable about 1e9 pfu/mL for most lagoons. This suggests the developed PANCE circuit is able to function correctly in maintaining

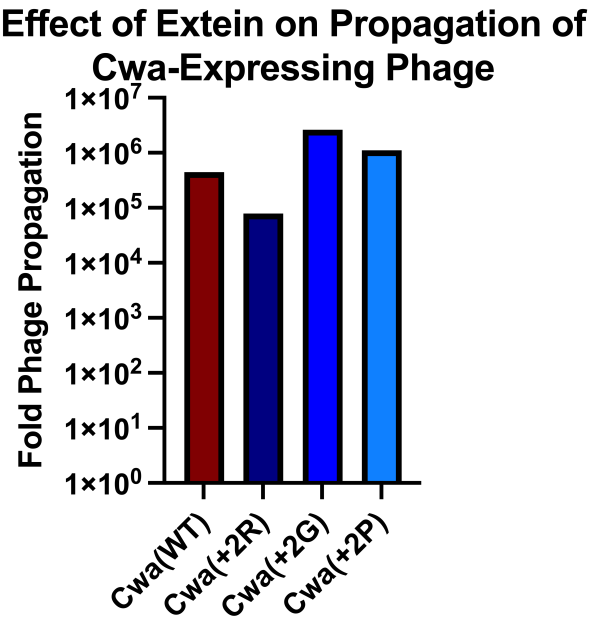

Figure 2-21: Propagation of Cwa expressing phage in diverse extein contexts.

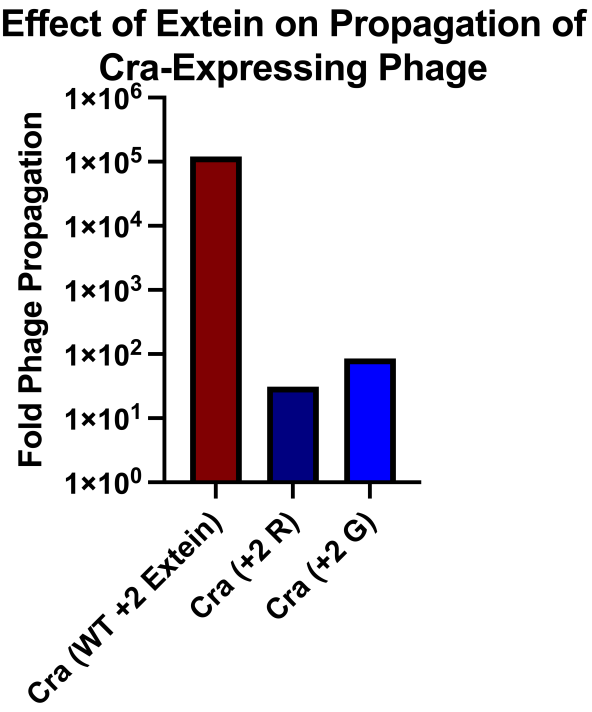

Figure 2-22: Propagation of Cra expressing phage in diverse extein contexts.

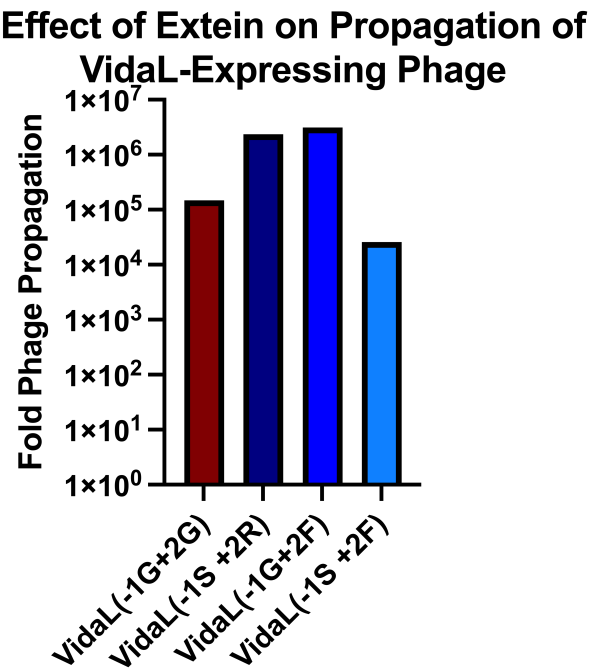

Figure 2-23: Propagation of VidaL Expressing Phage in PA(N)CE Circuit.

phage titers in PANCE.

2.2.10 Design of a Recombination Resistant Cloning Strain.

First, we sought to quantify the rate of recombination resulting from cloning and propagation S2208. The rate of recombination-driven incorporation of pIII in phage expressing C-terminal intein fragments, such as VidaL intein, was between 1 to 10 cheaters per 100 pfu/mL. This was determined by serial dilution of phage stocks of interest followed by serial propagation on 2060s[21] to look for presence of cheating phage under conditinos that are otherwise non-permissive to phage propagation. Results were similar for phage expressing N-terminal intein fragments except that the recombination rate was around 1 to 10 cheating phage per 1e8 pfu/mL.

These results in combination with observed recombination-driven generation of cheating phage in PANCE campaigns precluded manipulation of C-terminal intein-expressing phage in any propagation or evolution experiments. This suggested that a complete intein PACE architecture would require a new helper strain which was robust

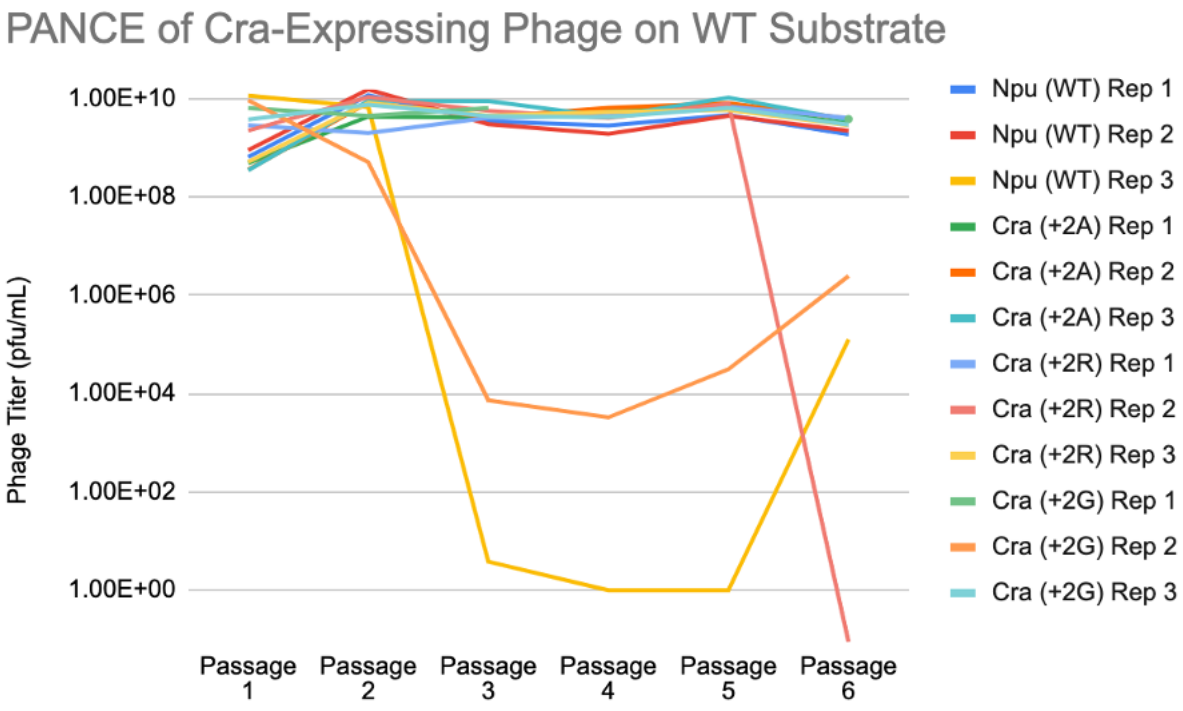

Figure 2-24: Propagation of Cra-expressing phage over multiple passages.

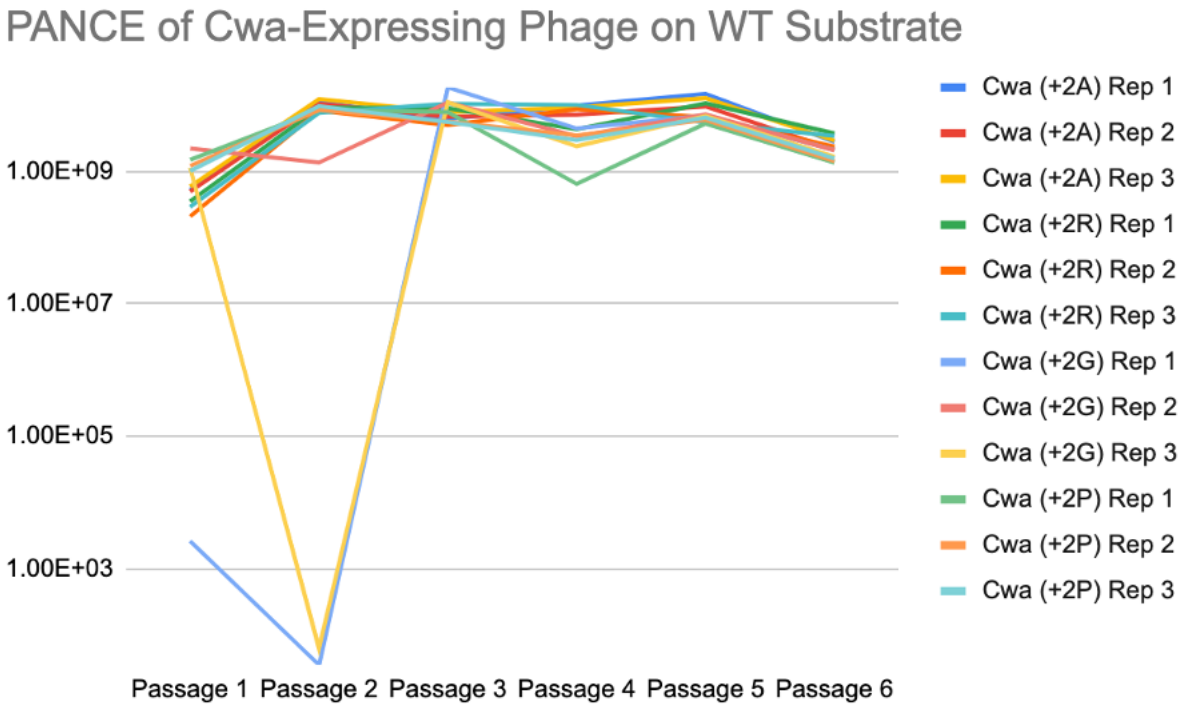

Figure 2-25: Propagation of Cwa-expressing phage over multiple passages.

to recombination with phage expressing part or all of pIII. As the observed rate of recombination was so high, this further suggested to us that modest improvements to drifting and cloning in the S2208 helper strain, such as modifying codon optimization to reduce homology or modifying propagation conditions to minimze time and number of phage genomes exposed to host strain pIII coding sequence were highly unlikely to prevent recombination.

To prevent recombination between helper strain and phage we considered that if pIII were non-contiguous and spread accross two plasmids, generation of recombinant cheating phage would require two separate low probability recombination events to occur in the same phage. To accomplish this we split pIII at its signal peptide and expressed each half of pIII fused to one half of NpuDnaE intein. This new recombination resistant strain was given the name S1040.

2.2.11 5. Validation of Recombination Resistant Cloning Strain.

The plaquing efficiency and rate of phage propagation of S2208 and S1040 were directly compared side-to-side. We considered that cloning phage would be a particularly challenging test of helper strain utility. Cloning efficacy as measured by plaques formed during cloning of three separate phage were measured. The resulting efficinecy varied from 10 to 400 plauqes formed, however the difference in plaquing efficiency between S1040 and S2208 was very minimal.

The ability to propagate phage rapidly was additionally measured. A 1e2 fold difference in phage propagation was evident at 8 hours as S1040 apparently propagated phage less quickly. This is is potentially due to the effect of intein expression on cellular growth. However, this difference largely disappeared when the allowed propagation time was increased to 16 hours, supporting this hypothesis.

S1040s were used for cloning and propagating all phage reported when testing the T7 and C2-domain split III based selection architecture. When newly cloned strains were passaged on s2060[21], no phage were observed after four passages. This continued to be the case even when starting at very high phage titers of aronud 1e10 pfu/mL. This is in contrast to the same phage cloned using S2208 which generated cheating

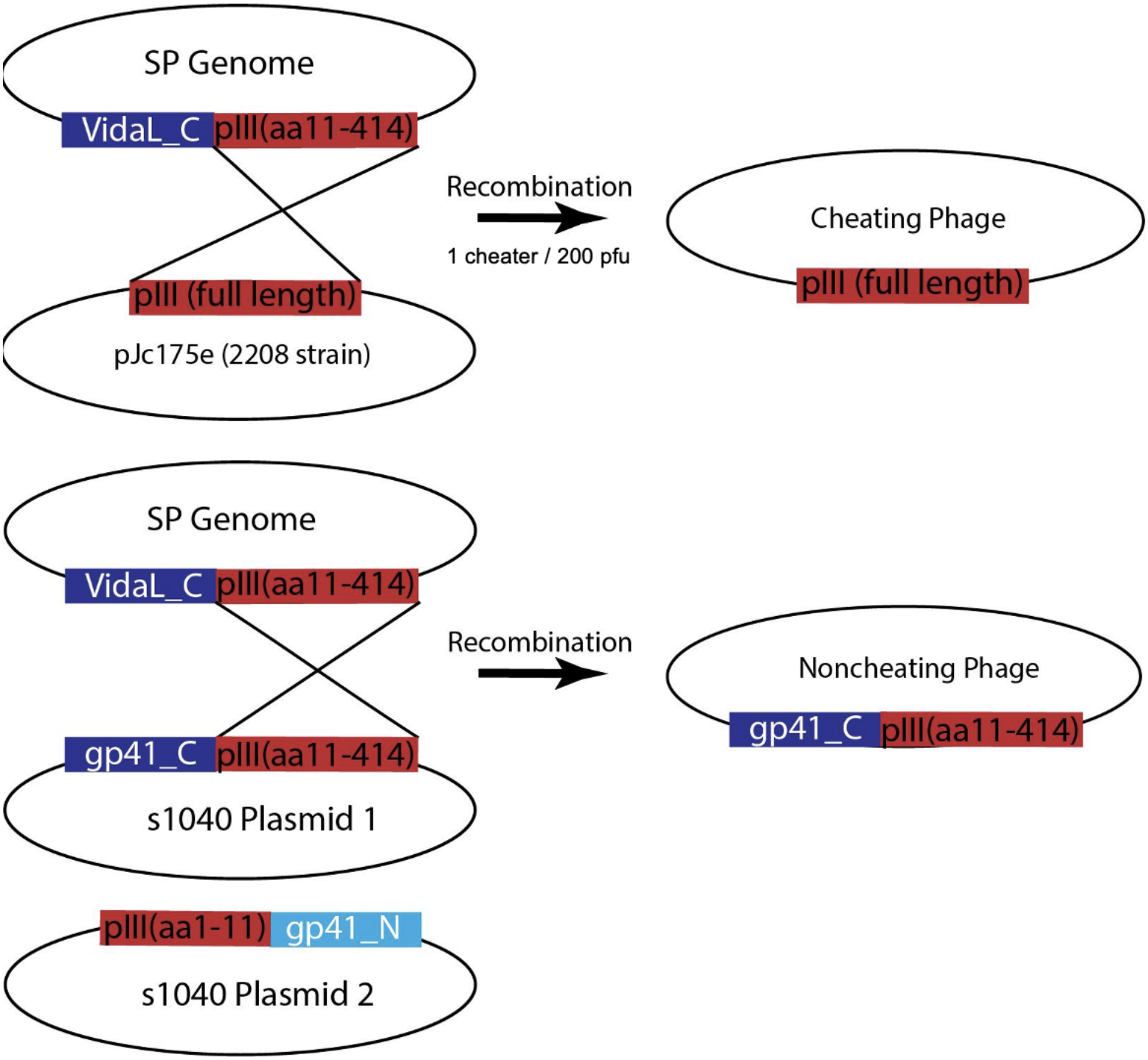

Figure 2-26: Principle behind recombination resistant helper strain.

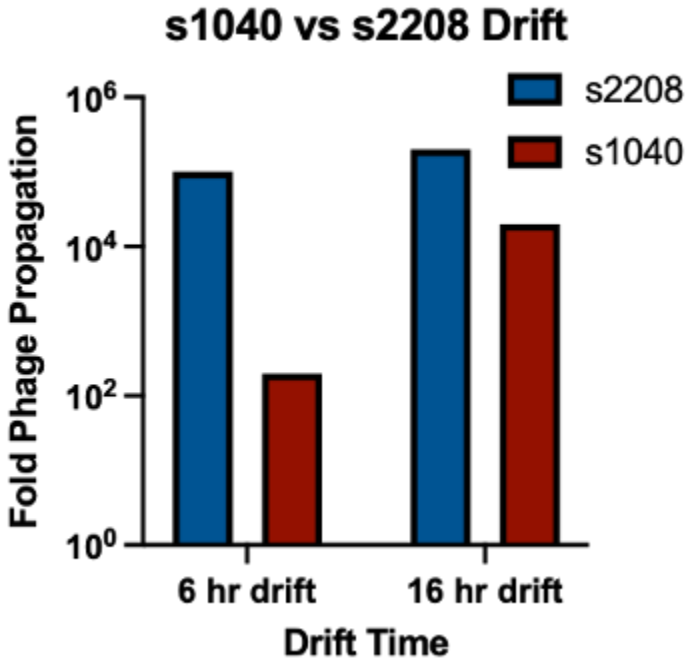

Figure 2-27: Plaquing efficiency in cloning three separate

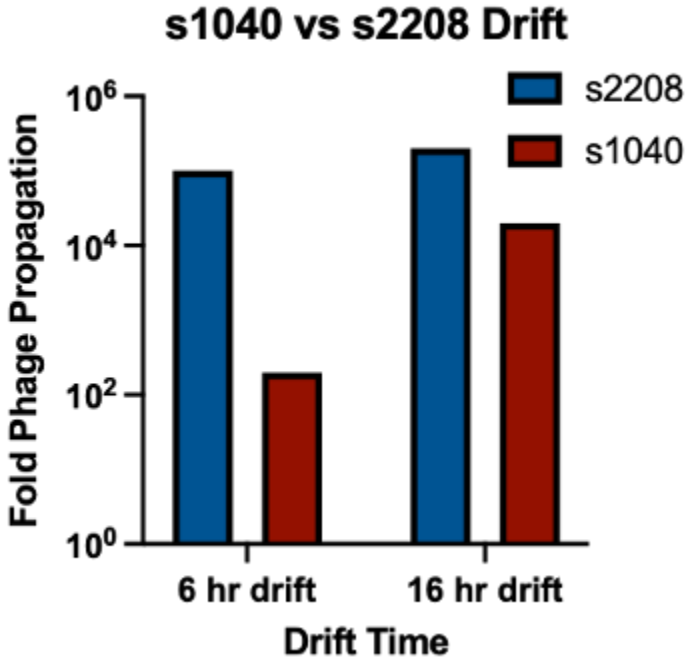

Figure 2-28: Propagation of Cwa-expressing phage over multiple passages.

phage at a level of 1e9 pfu/mL after two passages. Since adoption of S1040 as the sole helper strain no cheating phage or recombination events between phage and helper strain have been detected. This suggests that the architecture is effectively robust to recombination and can propagate phage with at least 1245 nt of homology with pIII without any generation of cheating recombinant phage. We conclude the new S1040 strain is effectively recombination resistant and a valuable addition to the PACE toolkit for propagating phage with large homology to host derived genetic material.

Chapter 3

Conclusion

3.1 Conclusion

Inteins enable targeted manipulation of protein chemistry through protein trans-splicing. Phage assisted continuous evolution (PACE) has previously shown promise for rapidly evolving bespoke protein functionality in diverse biochemical systems. We have extended PACE to select for rapid intein splicing. In particular, we have shown the intein splicing PACE circuit engineered is strictly splicing dependant, discriminates between intein splicing rates from one minute to several hours, discriminates between preferred over unpreferred extein contexts, supports propagation of phage in a multi-passage PANCE format, and circumvents recombination-driven phage cheating through a recombination-resistant helper strain. We anticipate this selection will enable facile, rapid evolution of inteins with bespoke properties.

Bibliography

1. Shah, N. H.; Muir, T. W. Inteins: Nature's Gift to Protein Chemists. Chem. Sci. 2014, 5 (2), 446–461.

2. Davis, E. C.; Jenner, P. Protein Splicing ? The Lengths Some Proteins Will Go To. 1995, 67 (2), 131–137.

3. Agouridas, V.; El Mahdi, O.; Diemer, V.; Cargoët, M.; Monbaliu, J.-C. M.; Melnyk, O. Native Chemical Ligation and Extended Methods: Mechanisms, Catalysis, Scope, and Limitations. Chemical Reviews 2019, 119 (12), 7328–7443.

4. Burton, A. J.; Haugbro, M.; Parisi, E.; Muir, T. W. Live-Cell Protein Engineering with an Ultra-Short Split Intein. Proceedings of the National Academy of Sciences 2020, 117 (22), 12041–12049.

5. Seath, C. P.; Burton, A. J.; Sun, X.; Lee, G.; Kleiner, R. E.; MacMillan, D. W. C.; Muir, T. W. Tracking Chromatin State Changes Using Nanoscale Photo-Proximity Labelling. Nature 2023, 616 (7957), 574–580.

6. Barbara Di Ventura; Mootz, H. D. Switchable Inteins for Conditional Protein Splicing. Biological Chemistry 2019, 400 (4), 467–475.

7. Pinto, F.; Thornton, E. L.; Wang, B. An Expanded Library of Orthogonal Split Inteins Enables Modular Multi-Peptide Assemblies. Nature Communications 2020, 11 (1).

8. Davis, J. R.; Banskota, S.; Levy, J. M.; Newby, G. A.; Wang, X.; Anzalone, A. V.; Nelson, A. T.; Chen, P. J.; Hennes, A. D.; An, M.; Roh, H.; Randolph, P. B.; Musunuru, K.; Liu, D. R. Efficient Prime Editing in Mouse Brain, Liver and Heart with Dual Aavs. Nature Biotechnology 2023. DOI:10.1038/s41587-023-01758-z.

9. Amitai, G.; Callahan, B. P.; Stanger, M. J.; Belfort, G.; Belfort, M. Modulation of Intein Activity by Its Neighboring Extein Substrates. Proceedings of the National Academy of Sciences 2009, 106 (27), 11005–11010.

10. Frutos, S.; Goger, M.; Giovani, B.; Cowburn, D.; Muir, T. W. Branched Intermediate Formation Stimulates Peptide Bond Cleavage in Protein Splicing. Nature Chemical Biology 2010, 6 (7), 527–533.

11. Shah, N. H.; Vila-Perelló, M.; Muir, T. W. Kinetic Control of One-Pot Trans-Splicing Reactions by Using a Wild-Type and Designed Split Intein. Angewandte Chemie International Edition 2011, 50 (29), 6511–6515.

12. Wang, Z.; Li, N.; Wang, Y.; Wu, Y.; Mu, T.; Zheng, Y.; Huang, L.; Fang, X. Ubiquitin-Intein and SUMO2-Intein Fusion Systems for Enhanced Protein Production and Purification. 2012, 82 (1), 174–178.

13. Lin, Y.; Li, M.; Song, H.; Xu, L.; Meng, Q.; Liu, X.-Q. Protein Trans-Splicing of Multiple Atypical Split Inteins Engineered from Natural Inteins. PLoS ONE 2013, 8 (4), e59516.

14. Shah, N. H.; Dann, G. P.; Vila-Perelló, M.; Liu, Z.; Muir, T. W. Ultrafast Protein Splicing Is Common among Cyanobacterial Split Inteins: Implications for Protein Engineering. Journal of the American Chemical Society 2012, 134 (28), 11338–11341.

15. Carvajal-Vallejos, P.; Pallissé, R.; Mootz, H. D.; Schmidt, S. R. Unprecedented Rates and Efficiencies Revealed for New Natural Split Inteins from Metagenomic Sources. Journal of Biological Chemistry 2012, 287 (34), 28686–28696.

16. Stevens, A. J.; Sekar, G.; Shah, N. H.; Mostafavi, A. Z.; Cowburn, D.; Muir, T. W. A Promiscuous Split Intein with Expanded Protein Engineering Applications. Proceedings of the National Academy of Sciences 2017, 114 (32), 8538–8543.

17. Packer, M. S.; Liu, D. R. Methods for the Directed Evolution of Proteins. Nature Reviews Genetics 2015, 16 (7), 379–394.

18. Lutz, S. Beyond Directed Evolution—Semi-Rational Protein Engineering and Design. Current Opinion in Biotechnology 2010, 21 (6), 734–743.

19. Lockless, S. W.; Muir, T. W. Traceless Protein Splicing Utilizing Evolved Split Inteins. Proceedings of the National Academy of Sciences 2009, 106 (27), 10999–11004.

20. Appleby-Tagoe, J. H.; Thiel, I. V.; Wang, Y.; Wang, Y.; Mootz, H. D.; Liu, X.-Q. Highly Efficient and More General Cis- and Trans-Splicing Inteins through Sequential Directed Evolution. The Journal of Biological Chemistry 2011, 286 (39), 34440–34447.

21. Miller, S. M.; Wang, T.; Liu, D. R. Phage-Assisted Continuous and Non-Continuous Evolution. Nature protocols 2020, 15 (12), 4101–4127.

22. Smeal, S. W.; Schmitt, M. A.; Pereira, R. R.; Prasad, A.; Fisk, J. D. Simulation of the M13 Life Cycle I: Assembly of a Genetically-Structured Deterministic Chemical Kinetic Simulation. Virology 2017, 500, 259–274.

23. Esvelt, K. M.; Carlson, J. C.; Liu, D. R. A System for the Continuous Directed Evolution of Biomolecules. Nature 2011, 472 (7344), 499–503.

24. Bryson, D. I.; Fan, C.; Guo, L.-T.; Miller, C.; Söll, D.; Liu, D. R. Continuous Directed Evolution of Aminoacyl-TRNA Synthetases. Nature Chemical Biology 2017, 13 (12), 1253–1260. https://doi.org/10.1038/nchembio.2474.

25. Blum, T. R.; Liu, H.; Packer, M. S.; Xiong, X.; Lee, P.-G.; Zhang, S.; Richter, M.; Minasov, G.; Satchell, K. J. F.; Dong, M.; Liu, D. R. Phage-Assisted Evolution of Botulinum Neurotoxin Proteases with Reprogrammed Specificity. Science 2021, 371 (6531), 803–810.

26. Thuronyi, B. W.; Koblan, L. W.; Levy, J. M.; Yeh, W.-H.; Zheng, C.; Newby, G. A.; Wilson, C.; Bhaumik, M.; Shubina-Oleinik, O.; Holt, J. R.; Liu, D. R. Continuous Evolution of Base Editors with Expanded Target Compatibility and Improved Activity. Nature Biotechnology 2019, 37 (9), 1070–1079.

27. Wang, T.; Badran, A. H.; Huang, T. P.; Liu, D. R. Continuous Directed Evolution of Proteins with Improved Soluble Expression. Nature Chemical Biology 2018, 14 (10), 972–980.

28. Huang, T. P.; Heins, Z. J.; Miller, S. M.; Wong, B. G.; Balivada, P. A.; Wang, T.; Khalil, A. S.; Liu, D. R. High-Throughput Continuous Evolution of Compact Cas9 Variants Targeting Single-Nucleotide-Pyrimidine PAMs. Nature Biotechnology 2022.

29. Richter, M. F.; Zhao, K. T.; Eton, E.; Lapinaite, A.; Newby, G. A.; Thuronyi, B. W.; Wilson, C.; Koblan, L. W.; Zeng, J.; Bauer, D. E.; Doudna, J. A.; Liu, D. R. Phage-Assisted Evolution of an Adenine Base Editor with Improved Cas Domain Compatibility and Activity. Nature Biotechnology 2020, 38 (7), 883–891.

30. Weiss, G. A.; Roth, T. A.; Baldi, P. F.; Sidhu, S. S. Comprehensive Mutagenesis of the C-Terminal Domain of the M13 Gene-3 Minor Coat Protein: The Requirements for Assembly into the Bacteriophage Particle. Journal of Molecular Biology 2003, 332 (4), 777–782.

31. Stevens AJ, et al. (2016) Design of a split intein with exceptional protein splicing activity. J Am Chem Soc 138:2162–2165.

32. Guerrero, F.; Ciragan, A.; Iwaï, H. Tandem SUMO Fusion Vectors for Improving Soluble Protein Expression and Purification. Protein Expression and Purification 2015, 116, 42–49.