THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 288, NO. 9, pp. 6202–6211, March 1, 2013
© 2013 by The American Society for Biochemistry and Molecular Biology, Inc. Published in the U.S.A.

Faster Protein Splicing with the *Nostoc punctiforme* DnaE Intein Using Non-native Extein Residues

Received for publication, November 2, 2012, and in revised form, January 9, 2013 Published, JBC Papers in Press, January 10, 2013, DOI 10.1074/jbc.M112.433094

Manoj Cheriyan, Chandra Sekhar Pedamallu[1], Kazuo Tori[2], and Francine Perler[3]
From the New England Biolabs, Inc., Ipswich, Massachusetts 01938

Background: Inteins are protein maturation machines that enable technologies for regulating enzymes and protein semisynthesis.
Results: Robust splicing was achieved with novel genetically selected exteins.
Conclusion: In contrast to commonly held perceptions, the natural extein was not the fastest splicing substrate.
Significance: Natural precursors balance extein- and intein-selective pressures. Specificity studies provide a predictive rubric for identifying new intein insertion sites.

Inteins are naturally occurring intervening sequences that catalyze a protein splicing reaction resulting in intein excision and concatenation of the flanking polypeptides (exteins) with a native peptide bond. Inteins display a diversity of catalytic mechanisms within a highly conserved fold that is shared with hedgehog autoprocessing proteins. The unusual chemistry of inteins has afforded powerful biotechnology tools for controlling enzyme function upon splicing and allowing peptides of different origins to be coupled in a specific, time-defined manner. The extein sequences immediately flanking the intein affect splicing and can be defined as the intein substrate. Because of the enormous potential complexity of all possible flanking sequences, studying intein substrate specificity has been difficult. Therefore, we developed a genetic selection for splicing-dependent kanamycin resistance with no significant bias when six amino acids that immediately flanked the intein insertion site were randomized. We applied this selection to examine the sequence space of residues flanking the *Nostoc punctiforme* Npu DnaE intein and found that this intein efficiently splices a much wider range of sequences than previously thought, with little N-extein specificity and only two important C-extein positions. The novel selected extein sequences were sufficient to promote splicing in three unrelated proteins, confirming the generalizable nature of the specificity data and defining new potential insertion sites for any target. Kinetic analysis showed splicing rates with the selected exteins that were as fast or faster than the native extein, refuting past assumptions that the naturally selected flanking extein sequences are optimal for splicing.

Inteins are a class of single turnover enzymes that catalyze a self-excision reaction out of a host protein-intein fusion (the precursor protein), resulting in a free intein and a mature host protein (the extein) where the splice junctions have been ligated with a standard peptide bond. Over 550 inteins have been identified in the InBase database, and they are ubiquitous in single cell organisms from all domains of life (1). Intein splicing mechanisms are diverse, with three distinct mechanisms identified to date (2). In each of these mechanisms, a series of three or four carefully coordinated displacement reactions was catalyzed in the absence of additional cofactors or energy sources by the intein plus the extein residue forming the C-terminal splice junction (2). Coordination of these multistep pathways is likely to require a precise architecture at the intein active site that can be affected by proximal extein residues.

Splicing in native host proteins is so rapid that the precursor protein is rarely, if ever, detected in nature, whereas splicing in heterologous host proteins often results in significant amounts of single splice junction cleavage byproducts or unreacted precursor. As a result, it has been proposed that the native flanking extein sequences are optimal for splicing of each intein (3, 4). However, inteins are generally located within highly conserved extein motifs and active sites of essential proteins, so presumably there is also selective pressure to maintain the residues at the intein insertion site for optimal extein function (5). Thus we hypothesize that the native precursor may not necessarily represent the most rapid splicing context for the intein but, rather, a balance of host protein sequence requirements, the specificity of the homing endonuclease (present in many inteins), and the need for a functional intein.

Because of their unique chemistry, inteins have proven useful in numerous applications such as protein tagging and purification, control of enzyme activity, assembling active proteins *in vivo*, protein semisynthesis *in vitro*, segmental isotope labeling for NMR, and many others (for review, see Refs. 6–9). The technologies enabled by inteins are growing, yet there remain some fundamental gaps in our understanding of how inteins work, especially in the specificity requirements for splicing (defined here as the flanking extein sequences that permit splicing). Previous studies characterizing the effect of varying a single amino acid flanking the N- or C-terminal splice junction demonstrated that the subset of extein residues that promote splicing or cleavage are specific to each intein (10–14). To our knowledge, no studies have examined the immense number of

⌘ *Author's Choice*—Final version full access.

[1] Present address: Dana-Farber Cancer Institute, Boston, MA, 02115 and The Broad Institute, Cambridge, MA 02142. E-mail: pcs.murali@gmail.com.

[2] Present address: Kyushu University, Fukuoka 812-8581, Japan. E-mail: hiris5211@yahoo.co.jp.

[3] To whom correspondence should be addressed: New England Biolabs, 240 County Rd, Ipswich, MA 01938-2723. Tel.: 978-380-7326; E-mail: perler@neb.com.

possible combinations when multiple extein residues are queried in a full splicing reaction. In lieu of such information, researchers design splicing experiments governed by the assumption that the natural host context will be the best and have, therefore, added three to five native flanking extein residues when placing an intein in a novel host protein. This limits the usefulness of these techniques by leaving behind a "scar" of flanking extein residues after splicing. Another approach to improving splicing in heterologous systems uses protein engineering strategies to change or relax intein specificity by intein mutation (4, 12, 15). Complementary tools that allow us to gauge the full scope of viable substrates of important inteins are highly desirable for understanding inteins as enzymes and expanding their usefulness in biotechnology by providing more options for intein insertion.

Here we present a genetic selection based on kanamycin (Kan)[4] resistance that allows for comprehensive examination of all possible combinations of three proximal N-terminal extein residues and three proximal C-terminal extein residues. Unlike previous intein kanamycin selection systems, our system was optimized for maintenance of kanamycin resistance with an extremely broad range of post-splicing six-residue insertions to minimize bias due to the reporter protein. The system was tested with the *Nostoc punctiforme* Npu DnaE, the *Mycobacterium tuberculosis* Mtu-H37Rv RecA, and the MP-Be DnaB inteins. We further applied this system to study splicing of the Npu DnaE intein in detail. This intein is naturally split and works in *trans* but can be fused to splice in *cis* as well (16). The Npu DnaE intein is one of the fastest known inteins (17–20) and is important for biotechnology. Our results show that the Npu DnaE intein is able to efficiently splice multiple sequences of far different composition than its native flanking exteins, defining new potential insertion sites preceding Cys-Trp or Cys-Met. These data expand the usefulness of this intein and provide a new framework for predictable design of protein splicing experiments.

EXPERIMENTAL PROCEDURES

All enzymes and reagents were obtained from New England Biolabs (NEB, Ipswich MA) unless otherwise noted and used as described by the manufacturer. DNA sequencing was performed either by the NEB Core Sequencing Facility or by Beckman Coulter Genomics (Danvers, MA) for single pass automated sequencing in two directions using cultures provided in 96-well microtiter dishes.

PCR and Plasmid Construction—All PCR amplification steps used Phusion DNA polymerase (NEB) according to the manufacturer's instructions. The following two thermocycling protocols were used for most PCR amplifications: Protocol A (1 cycle at 98 °C for 30 s; 25 cycles at 98 °C for 10 s, 65 °C for 20 s, 72 °C for 120 s; 1 cycle at 72 °C for 5 min); Protocol B (1 cycle of 98 °C for 30 s; 25 cycles of 98 °C for 10 s, 65 °C for 20 s, 72 °C for 25 s; 1 cycle of 72 °C for 5 min). Restriction endonucleases and T4 DNA ligase were obtained from NEB. Chemically competent *Escherichia coli* NEB 10-beta cells were used for all plasmid recovery stages.

Examination of Potential Intein Insertion Sites in Aminoglycoside Phosphotransferase (Aph) for High Throughput Assay Development—Aminoglycoside phosphotransferases confer Kan^R to various bacteria and are good targets for genetic selection. A protein blast search using the *Corynebacterium diphtheriae Aph* sequence (accession number AY061891.1) as the query was done on the NCBI server at blast.ncbi.nlm.nih.gov. The 10 most similar unrelated bacterial *Aph* sequences identified by the search were aligned, and the consensus sequence was compared with the *C. diphtheriae Aph* gene. Five locations (A: Ala-62—Asn-63; B: Phe-87—Ile-88; C: Glu-118 —Asn-119; D: Asp-144 —Asg-145; E: Glu-181—Met-182) were chosen as possible intein insertions sites based on possessing low sequence conservation and being surface-exposed (21). Primers were designed to insert six variable residues at each site or to replace the six amino acids centered around each site. Mutations were made by inverted PCR using Protocol A with pACYC177 as the template. pACYC177 encodes both an ampicillin (Amp) and a kanamycin resistance marker. After PCR, the parental DNA was digested completely with 10 units of DpnI for 3 h. The resulting amplified DNA was ligated and transformed into NEB 10-beta cells. Equal quantities of cells were spread on LB agar plates containing either 100 μg/ml Amp or 100 μg/ml Kan. The ratio of cells that survived on Kan *versus* Amp plates was calculated after overnight incubation at 37 °C. The PCR control using primers with no mutations in Kan^R approximates the efficiency of the inverted PCR reaction and subsequent plating, and so was used to normalize the data.

Construction of Kan^R Plasmids for Intein Insertion—Silent mutations were made to encode BspEI and NheI restriction sites for cloning inteins into the Kan^R site C yielding plasmid p3A. This plasmid contained an additional NheI site at a distal position, so the entire Kan resistance gene and its endogenous promoter were subcloned from p3A to pJF119 (22) using KasI and KpnI restriction sites. The resulting p3AM1 plasmid was Kan- and Amp-resistant.

To enhance Western blot analysis of splicing variants, a His$_6$ tag followed by a Gly-Gly-Ser-Gly linker was appended to the N terminus of the *Aph* gene by inverted PCR. The resulting plasmid was named pKanC^H.

Functional and inactive variants of the Npu DnaE intein *cis* construct, the Mtu RecA intein, and the MP-Be DnaB intein were cloned into site C of pKanC^H. Inactive inteins contained the following mutations at catalytic residues: Npu DnaE intein: C1A, N138D, C+1A; Mtu RecA intein: C1A, N440D, C+1A; MP-Be DnaB intein: C320S. All clones with these inactive inteins were kanamycin-sensitive.

Library Construction and Selection—Hand-mixed primers (forward, 5′-CT GAT TCC GGA GAA NNK NNK NNK TGC

[4] The abbreviations used are: Kan, kanamycin; Kan^R, kanamycin resistance protein or kanamycin resistance; Aph, aminoglycoside phosphotransferase; Amp, ampicillin; EP, a selected clone from the low stringency pilot selection experiment; I^C, the C-terminal fragment of the split Npu DnaE intein; I^C-P^{His}, I^C fused to the Δ Sal fragment of paramyosin with a C-terminal His tag; I^N, the N-terminal fragment of the split Npu DnaE intein; M or MBP, *E. coli* maltose-binding protein; M-I^N, maltose-binding protein fused to I^N; M-P^{His}, spliced product where the maltose-binding protein is fused to the ΔSal fragment of paramyosin; MIP, maltose-binding protein fused to an intein fused to the ΔSal fragment of paramyosin; MP-Be, *Mycobacteriophage Bethlehem*; Mtu, *Mycobacterium tuberculosis*; NEB, New England Biolabs; Npu, *Nostoc punctiforme*; RBS, ribosome binding site.

CTG AGC TAT GA; reverse, 5′-A CAC TGC TAG CGC ATC AAC AAT ATT MNN MNN GCA ATT AGA GGC AAT AA) were ordered from Integrated DNA Technologies (Coralville, IA) to introduce three variable codons flanking the intein N terminus and two variable codons flanking the intein C terminus after a fixed extein Cys residue (Fig. 1). In these primers N = A, T, G, or C, K = G or T, and M = C or A. Primers also included restriction enzyme sites for cloning into site C in the kanamycin resistance gene and sequences that matched the template, which was the *cis* construct of the Npu DnaE intein (16). PCR products using these primers were cloned into plasmids that contained mutations at the natural *Aph* ribosome binding site (RBS) changing GGGGTGTTATG to G**AA**GGT-GTT**C**ATG, which reduces the level of expression by ~20-fold.[5]

The resulting DNA was transformed into electrocompetent NEB 10-beta cells using a standard bacterial transformation protocol with up to 200 ng of DNA in a Gene Pulser Xcell (Bio-Rad). The cells were immediately resuspended in a total of 4 ml of LB media and allowed to recover for 1.5 h at 37 °C with shaking.

Freshly transformed cells were pelleted after recovery and resuspended in 1 ml of LB media before plating on Kan-selective plates. To estimate the size of the library, serial dilutions were plated on 100 μg/ml Amp plates, which yielded a library size of $\sim 5 \times 10^7$. The rest of the library was plated on 40 μg/ml Kan plates and incubated overnight at 37 °C. Colonies that grew on Kan-selective plates were arrayed into 96-well microtiter dishes containing LB media + 10% glycerol + 50 μg/ml Amp + 40 μg/ml Kan using the Genetix QPix II automated colony picker (Molecular Devices, Sunnyvale, CA). These microtiter plates were incubated overnight at 37 °C, a replica plate was made, and all plates were stored at −20 °C. Replica plates were sent to Beckman Coulter Genomics for automated sequencing.

Sequencing results were examined using an in-house automated program that compared the sequences of the RBS, the N- and C-terminal Kan^R fragments, and the intein. Only sequences that had unambiguous nucleotide assignments that did not contain any mutations and that maintained the Kan^R RBS knockdown sequence were included in our analysis of specificity. Only 483 of 768 clones sequenced yielded results that met these stringent criteria, mostly due to ambiguous nucleotide assignments.

Cloning into Luciferase and Western Blot Analysis of Expressed Proteins—To aid in cloning, XhoI and KpnI sites were introduced by silent mutagenesis to flank the firefly luciferase site Asn-230—Asp-234 using inverted PCR with the pEK4 (23) template.

Insert DNA containing the Npu DnaE intein and the desired extein sequences were prepared using the pKanC^H plasmid containing the Npu DnaE *cis*-intein (pKanC^H_Npu) as template. The PCR primers used were forward (5′-GCAT GCT CGA GAT CCT ATT TTT GGC AAT *XXX XXX XXX* TGC CTG AGC TAT GAA ACC GAA AT) and reverse (5′-GAA TGG TAC CAC ACT TAA AAT CGC AGT *XXX XXX* GCA ATT AGA GGC AAT AAA GCC ATT TTT, where *XXX* represents the appropriate sequence to encode for selection hits E2, E4, E6, E8, E9, E12, E15 E16, E18, and E20 (see Table 2). The resultant plasmids were sequenced to verify the presence of the firefly luciferase and the intein with the desired flanking extein sequence.

Western blots were done to assess the degree of splicing with each flanking extein sequence. Plasmids were transformed into NEB Express cells and incubated at 37 °C to an $A_{600\text{ nm}}$ of 0.6. Two ml of each culture was pelleted, resuspended in 250 μl of water, and then adjusted to an $A_{600\text{ nm}}$ of 5. Fifteen μl of these cells were mixed with 3× SDS sample buffer + DTT (NEB), boiled for 10 min, and electrophoresed on SDS-PAGE followed by transfer to nitrocellulose membranes using a Bio-Rad *Trans*-blot S.D. apparatus. After transfer to nitrocellulose, bands containing the firefly luciferase were detected with mouse IgG anti-firefly luciferase antibody (Abcam, Cambridge, MA) and IRDye 680 goat anti-mouse secondary antibody (LI-COR, Lincoln, NE).

Cloning Inteins into MBP-Intein-Paramyosin (MIP) Constructs for Trans-splicing Experiments—DNA encoding the Npu DnaE intein (*cis* construct) with three native extein residues flanking each side was amplified by PCR Protocol B using pKanC^H_Npu as the template and primers containing XhoI and SpeI sites. Plasmid pMIP contains the *E. coli* maltose-binding protein (M) as the N-extein and the paramyosin-Δ Sal fragment (P) as the C-extein (2). The PCR insert and plasmid pMIP_MP-Be_His (2) were both digested with XhoI and SpeI, gel-purified, and ligated to yield plasmid pMIP_Npu, which was sequenced to verify that the intein was fused in between MBP and paramyosin-Δ Sal.

The *cis*-splicing precursor was then converted into a *trans*-splicing precursor. Plasmids containing M-I^N and $\text{I}^C\text{-P}^{His}$ fragments were generated by inverted PCR with pMIP_Npu as the template and thermocycling Protocol A using M-I^N primers (forward (5′-TGA CTG GGA AAA CCC TGG CGT TAC) and reverse (5′-ATT CGG CAG ATT ATC AAC ACG CAT CA)) and $\text{I}^C\text{-P}^{His}$ primers (forward (5′-ATG ATC AAA ATT GCC ACC CGT AAA TAT CTG) and reverse (5′-AAT CTA TGG TCC TTG TTG GTG AAG TGC TCG T)). After DpnI digestion, ligation, and transformation, plasmids pNpuDnaE_MBP_I^N and pNpuDnaE_I^C_para were obtained.

The desired N- and C-terminal flanking extein sequences were introduced using a final round of inverted PCR with primers for M-I^N (forward (5′-GAG *XXX XXX XXX* TGC CTG AGC TAT GAA ACC GAA) and reverse (5′-GAG CGT ACC CCT TCC CTC GAT)) and for I^C-Para (forward (5′-*XXX XXX* ACT AGT GGT GGC ATG CTT ACA GG) and reverse (5′-GCA ATT AGA GGC AAT AAA GCC ATT TTT CAG)), where *XXX* represents the appropriate codon for each flanking extein sequence listed in Table 3.

Trans-splicing Kinetic Assay in MIP—Plasmids containing M-I^N and $\text{I}^C\text{-P}^{His}$ were transformed into NEB Express I^q cells and grown on LB + Amp plates. Ten-ml cultures were inoculated with single colonies and grown at 37 °C to an $A_{600\text{ nm}}$ of 0.6–0.8 at which point the temperature was reduced to 25 °C, and 0.5 mM isopropyl 1-thio-β-D-galactopyranoside was added. The cultures were then incubated for five more hours. One

[5] M. Cheriyan and F. Perler, unpublished data.

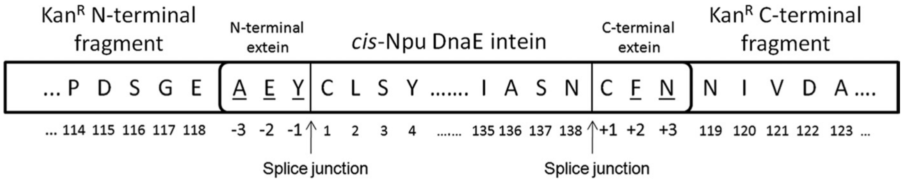

FIGURE 1. **Kan^R selection construct used to study intein specificity for flanking extein sequences.** Kan^R was used as a reporter for protein splicing with the Npu DnaE intein and three additional extein residues flanking both its N and C termini. The sequence of the native extein is shown inserted into Kan^R site C. *Underlined* amino acids were simultaneously randomized. In the numbering scheme used in this paper, residues in the Kan^R protein, the variable extein positions, and the intein were numbered independently as indicated. Positions in the Kan^R protein were numbered as in the original protein, ignoring insertions.

milliliter of each culture was harvested, and the cell pellet was stored frozen at −80 °C until kinetic assays were performed. Cell pellets were resuspended in 300 μl of reaction buffer K (100 mM Tris. pH 7.5, 5 mM DTT, 1 mM EDTA), lysed by sonication, and centrifuged for 1 min at maximum speed in a benchtop microcentrifuge. Because the M-I^N fragment expresses many fold more efficiently than the I^C-P^{His} fragment, the kinetic assays were done by mixing M-I^N and I^C-P^{His}-containing lysates in a 1:5 ratio at 30 °C over a 10-min time course. Under these conditions the amount of the M-I^N fragment observed in the Western blot remains constant, whereas the I^C-P^{His} fragment is consumed. Each 20-μl time point was stopped by the addition of 10 μl of 3 × SDS sample buffer + DTT. After boiling for 10 min, the samples were run on SDS-PAGE followed by transfer to a nitrocellulose membrane. The membrane was probed with Mouse IgG anti-His tag antibody (Novagen, EMD Chemicals, San Diego, CA) to detect I^C-P^{His} or M-P^{His} and rabbit anti-MBP antisera (NEB) to detect M-I^N or M-P^{His}. Probing with IRDye 680 goat anti-mouse and IRDye 800 Goat anti-Rabbit secondary antibodies (LI-COR) allowed for quantification of the spliced product on the LI-COR Odyssey. The percent conversion was calculated as 1–M-P^{His} formation divided by the initial substrate (I^C-P^{His}) concentration. These data were fit to a first order decay reaction using KaleidaGraph (Synergy, Reading, PA) with the equation $A = A_o e^{-kt}$, where A is the substrate concentration at t, A_o is the initial substrate concentration, k is the apparent first order rate constant, and t is time. The $t_{1/2}$ was calculated as $t_{1/2} = \ln 2/k$. At least two independent replicates of each mutant were done, and the standard deviation of the fits was calculated.

RESULTS

Developing a Selection System to Probe Extein Specificity—We set out to develop an assay system that could rapidly examine variable extein sequences of up to 10^8 members. Because of the ease of use and high stringency of antibiotic-based selections, we focused our efforts on identifying sites within *C. diphtheriae* aminoglycoside phosphotransferase (*Aph*), an enzyme responsible for kanamycin resistance (Kan^R). Based on comparison of this protein to an alignment of other *Aph* proteins, five regions were chosen as candidates to develop a splicing selection system (A: Ala-62—Asn-63; B: Phe-87—Ile-88; C: Glu-118—Asn-119; D: Asp-144—Asg-145; E: Glu-181—Met-182). These locations were surface-exposed and possess low sequence conservation, suggesting that they would be able to tolerate multiple mutations and still retain Kan^R (data not shown). A plasmid was constructed that contained β-lactamase to confer Amp resistance and the *C. diphtheriae Aph* gene under the control of its endogenous promoter. Using Phusion polymerase-based inverted PCR, we created 10 libraries where either (*a*) six codons at a chosen site were replaced with six consecutive in-frame NNK codons or (*b*) six in-frame NNK codons were inserted into the middle of the chosen sequence, where N represents any nucleotide and K represents G or C. These libraries represent the scar that would be left in the Kan^R protein after splicing when three flanking amino acids are included at each end of the intein. An ideal site would have no sequence preference and no selection bias due to the scar sequence in Kan^R. The permissibility of each site for random sequence replacement or insertion was quantified based on the percent of colonies that recovered on Kan *versus* Amp plates. Two sites were tolerant for the randomized cassettes: site C with a 6-amino acid insertion between Glu-118 and Asn-119 had 91 ± 8% permissibility, and site D with a replacement of the codons between Asn-142 and Phe-147 had 47 ± 1% permissibility. Insertion site C had been previously identified as a useful site for intein insertion (24). Other sites in the *Aph* gene have also been used for intein insertion (4, 15) but were not tested for permissibility of random scar sequences.

Inteins flanked by three native N- and C-terminal extein residues were cloned separately into site C or D and tested for the ability of splicing to restore Kan^R (Fig. 1). Catalytically inactivated inteins (see “Experimental Procedures”) were likewise inserted into the same sites to ensure that Kan^R required splicing. The inteins tested were the Mtu-H37Rv RecA intein flanked by N-terminal Lys-Asn-Lys and C-terminal Cys-Ser-Pro, the MP-Be DnaB intein flanked by N-terminal Gln-Asp-Gln and C-terminal Thr-Lys-Asn, and a *cis* construct of the naturally split Npu PCC73102 DnaE intein (16, 25) flanked by N-terminal Ala-Glu-Tyr and C-terminal Cys-Phe-Asn (Fig. 2 and data not shown). Kan^R was observed with the native inteins but not with the inactive inteins. These results validate the use of these two sites for assessing intein functionality. However, site C showed far greater tolerance for variable amino acid incorporation than site D and would thus make the best system for high throughput analysis of extein sequences.

Library Selection for Npu DnaE Intein Substrate Specificity—We constructed a library containing the intein flanked by three variable N-terminal extein residues labeled −3, −2, and −1 and two variable C-terminal extein residues labeled +2 and +3 (Fig. 1). NNK and MNN nucleotide sequences within the forward and reverse primers, respectively, were used to incorporate all possible amino acids at each variable codon. The +1 position

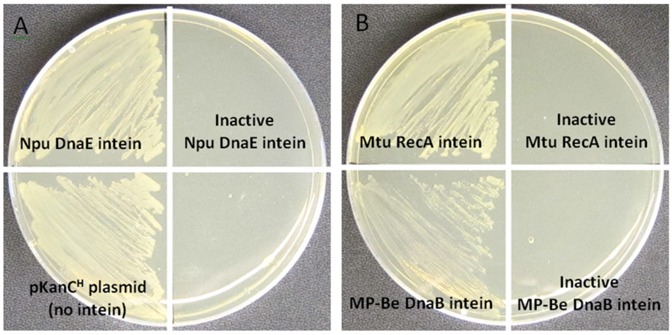

FIGURE 2. **Kanamycin resistance is dependent on intein splicing in Kan^R site C.** Three inteins were tested to ensure that resistance to kanamycin required a functional intein. *E. coli* were transformed with plasmids carrying the Kan^R gene with either an active or a catalytically inactive intein in site C. Growth after overnight incubation was then tested on plates with 40 μg/ml Kan. *A*, shown is growth at 37 °C with cells transformed with plasmids carrying the native Npu DnaE intein, an inactive Npu DnaE intein (C1A, N138D, C+1A), or a no intein control. *B*, shown is growth at 30 °C with cells transformed with plasmids carrying the native Mtu RecA intein, an inactive Mtu RecA intein (C1A, N440D, C+1A), the native MP-Be DnaB intein, or an inactive MP-Be DnaB intein (C320S).

TABLE 1

Aggregate data for the plasmid population that survived Kan^R selection

Data for 483 randomly chosen clones selected for kanamycin resistance after the Npu DnaE intein with three flanking N- and C-terminal extcin amino acids at Kan^R site C was expressed using the RBS knockdown sequence. The theoretical occurrence value represents the number of times an amino acid should be observed in an unbiased library of 483 clones based on codon redundancy using NNK primers. The native flanking extein residues for the Npu DnaE intein residue are labeled with asterisks.

		N-terminal Extein			Intein	C-terminal Extein		
Amino Acid	**Theoretical Occurrence**	**(-3)**	**(-2)**	**(-1)**		**Fixed (+1)**	**(+2)**	**(+3)**
D	15.1	21	23	0			0	19
E	15.1	19	*53	1			0	83
N	15.1	14	19	33			0	*130
Q	15.1	21	13	5			0	4
H	15.1	16	19	34			0	24
K	15.1	24	14	67			1	0
R	45.3	82	9	39			0	0
S	45.3	45	56	62			0	3
C	15.1	7	6	12		*483	1	1
T	30.1	19	27	48			0	29
P	30.1	24	46	0			0	0
G	30.1	141	81	5			0	1
A	30.1	*19	25	58			0	7
V	30.1	17	15	7			1	3
I	15.1	1	8	0			0	6
L	45.3	5	26	34			0	37
M	15.1	2	17	13			33	22
F	15.1	2	11	19			*1	34
Y	15.1	3	9	31*			2	79
W	15.1	1	6	15			444	1
STOP	15.1	0	0	0			0	0
Total	***483***	***483***	***483***	***483***		***483***	***483***	***483***

cysteine (which immediately follows the intein C terminus) is a crucial nucleophile in the splicing mechanism, so this position was fixed. The maximal complexity of a library of 5 randomized amino acid is 3.2 × 10^6. A pilot study using the endogenous Kan^R promoter and the Npu DnaE intein resulted in ~100,000 surviving colonies from a library of 1 × 10^7 clones when grown on 80 μg/ml Kan at 37 °C. Because we sought highly efficient splicing, we increased the stringency of the selection by altering the sequence of the ribosome binding site to reduce expression. By limiting production of the Kan^R precursor, we required efficient splicing to achieve the threshold needed to confer antibiotic resistance. Under these more stringent selection conditions, plating 5 × 10^7 transformants on 40 μg/ml Kan at 37 °C yielded ~5000 colonies. To assess the diversity of this library, 124 plasmids were isolated from colonies that were not subjected to Kan selection. These non-selected sequences show all possible amino acids at each of the five variable positions including the naturally occurring extein residues, which indicates that the library was sufficiently diverse (data not shown).

Because it is not practical to individually sequence all of the selected clones, the entire *Aph* and Npu DnaE intein precursor gene was sequenced from a representative, randomly chosen set of clones. Using an automated program we verified that 483 of the selected plasmids contained unambiguous sequences with no mutations in the *Aph* and the Npu DnaE intein genes. Examination of these extein sequences showed evidence of selection at some positions but not others. The largest specificity preferences were observed at the +2 and +3 positions of the C- extein (Table 1 and Fig. 3*A*). The natural splice site (Phe+2 and Asn+3) was not observed in the 483 sequenced clones. Instead, 92% of the selected sequences possessed Trp+2, and 7% contained Met+2 (Table 1 and Fig. 3). Phe, Val, Cys, and Lys were observed once, and Tyr was observed twice at the +2 position. The non-selected library confirmed that all residues were present at the +2 position in the expected quantities (data not shown), but Trp was selected over the naturally occurring Phe+2. The preference at the +3 position was for Tyr and residues similar to the natural Asn+3, with Asn+3, Asp+3, Gln+3, and Glu+3 representing 49% of the total hits (Table 1). The range of N-terminal extein sequences that allow splicing was very broad; however, it was evidently important to maintain flexibility at the N terminus because the selected hits have an abundance of Gly at the −3 and −2 positions (Table 1 and Fig. 3*D*).

Direct Observation of Splicing of the Kan^R-Npu DnaE Intein Fusion—Twenty positive hits representative of the range of sequence diversity found at the N- and C-terminal exteins were chosen for further analysis (Table 2). Cells containing these plasmids were grown in LB media plus Amp overnight, and Western blotting was used to directly confirm whether splicing occurred. Antibody probing based on the presence of an N-terminal His tag was used to detect precursor and the mature Kan^R product. All hits tested displayed robust *in vivo* splicing (~95%) with no detectable amount of cleavage products (Fig. 4 and data not shown).

Transferability of Selected Sequences to Two Unrelated Precursors—To address the question of whether the Kan^R selection results provide details about the actual specificity of the intein or reflect the unique selective pressures found within Kan^R, we cloned a subset of the selection hits into firefly luciferase. These included the Npu DnaE intein plus the three natural N- and C-terminal extein residues, an inactive intein mutant, and nine selection hits (Fig. 4*C*). Western blot analysis with anti-luciferase antibody was used to ascertain the degree of splicing when precursors were expressed in NEB Express cells at 37 °C. All Kan^R selection hits tested showed a high degree of splicing in this unrelated enzyme.

Five selection hits that represent the most commonly selected motifs in our library were cloned into a third context, a split MIP system (described below). Here again, the selected

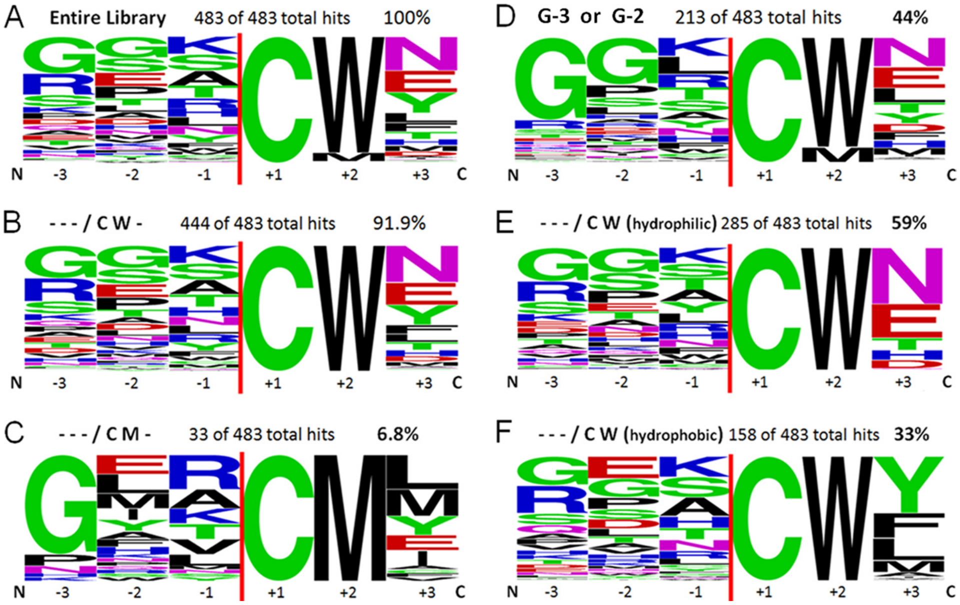

FIGURE 3. **Logo diagrams of flanking extein sequences for 483 randomly chosen positive hits selected for splicing of the Npu DnaE intein in Kan^R site C were made using the Weblogo program (28).** The *vertical red line* indicates the position of the intein. The +1 position was fixed as Cys. *A*, shown is the entire library of 483 clones. *B*, shown are all clones with Trp+2. *C*, shown are all clones with Met+2. *D*, shown are all clones with Gly at −3 or −2. *E*, shown are all clones with Trp+2 and a hydrophilic residue at +3. *F*, shown is all clones with Trp+2 and a hydrophobic residue at +3.

extein sequences promoted efficient protein splicing (Table 3), indicating that these hits represent general specificity criteria for the Npu DnaE intein rather than a solution specific to Kan^R site C.

Kinetics of Trans-splicing of the Npu DnaE Intein with Selected Exteins—To further validate the selected extein sequences, we examined the kinetics of *trans*-splicing of the naturally split Npu DnaE intein using either the native flanking extein residues or sequences that represent commonly selected motifs (Table 3). As controls, an inactive Npu DnaE intein mutant and a variant identified in the low stringency pilot selection (EP) were also tested (Fig. 5). M-I^N and I^C-P^{His} fragments (Fig. 6*A*) were expressed separately in NEB Express cells by induction with 0.5 mM isopropyl 1-thio-β-D-galactopyranoside at 25 °C for 4 h. Cells were harvested, and the kinetics of *trans*-splicing were assayed at 30 °C over a 10-min time course using the soluble fraction of the lysate. Samples taken at various time points were electrophoresed on SDS-PAGE followed by Western blot analysis to quantify spliced products (Fig. 6*B*). In the homologous Ssp DnaE intein, association of the I^N and I^C fragments was shown to approach diffusion-controlled limits (26, 27). Assuming that the Npu DnaE intein I^N and I^C fragments associate in a similarly rapid manor, the kinetics of *trans*-splicing can be fit to a first order decay reaction when the concentration of M-I^N is maintained at excess over the concentration of I^C-P^{His}.

In our system the Npu DnaE intein with native flanking extein residues has fast kinetics for *trans*-splicing, with an apparent first order rate constant of 0.63 ± 0.07 min^{-1} and a $t_{1/2}$ of 66 s at 30 °C (Table 3). This value is in good agreement with the previously published rate constant of 0.66 ± 0.12 min^{-1} and a $t_{1/2}$ of 63 s at 37 °C (18, 20). The inactive intein showed no observable M-P^{His} spliced product or disappearance of starting materials M-I^N or I^C-P^{His} over a 1-h time course. The EP sample from the low stringency pilot study yielded incomplete splicing with a k of 0.06 ± 0.004 min^{-1} and a $t_{1/2}$ of 11.5 min (Table 3). The majority of the assayed hits (E2, E4, E6, and E16) displayed comparable splicing rates to the Npu DnaE intein with its native extein sequence (Table 3). The apparent first order rate constant for E6 (extein sequence Arg-Gly-Lys/Cys-Trp-Glu) is 1.6 ± 0.2 min^{-1} with a $t_{1/2}$ of 28 s at 30 °C, which is 2.5 times faster than that of the native extein.

DISCUSSION

One of the greatest challenges in the study of inteins is separating the contribution to splicing rates of the extein *versus* the intein. To address this problem, we designed a high throughput system where a random six-amino acid substitution did not detectably inhibit host protein function, hypothesizing that this would allow us to study intein specificity requirements without significant bias from the extein. However, finding such sites is complicated, as most locations that meet these criteria fail for other reasons: (*a*) splicing does not occur, (*b*) activity of the host protein is not disrupted when the intein is present, or (*c*) splicing does not rescue host protein activity. In a study of 12 sites in firefly luciferase (9 replacement and 11 insertion libraries) none of the tested sites fully satisfied these criteria (data not shown). However one site in Kan^R met all criteria for a specificity assay, which was critical to the success of this study.

The Npu DnaE intein specificity results in Kan^R were validated by (*a*) the absence of bias as evidenced by 91% of Kan^R C site random cassettes maintaining kanamycin resistance, (*b*) transfer of selected extein sequences to two unrelated host proteins promoted similar levels of splicing, and (*c*) splicing rates equivalent to or greater than the naturally occurring exteins in *trans*-splicing experiments in a different extein protein.

Our specificity data provide a predictive model to scan available target protein sequence space to find or engineer intein

TABLE 2

Selected extein sequences

The flanking extein sequences of clones representing the sequence diversity of selection hits are listed along with the flanking extein sequences in the native extein and in an inactive intein with mutations in essential residues C1A, the intein C-terminal Asn and C+1A. Cys+1 was held constant in the selection.

Name	N-extein	C-extein
	-3 -2 -1	+1 +2 +3
Native	A E Y	C F N
E1	W T A	C C Q
E2	G W N	C M L
E3	G Y A	C M M
E4	G L T	C M Y
E5	G S K	C W D
E6	R G K	C W E
E7	S A F	C W E
E8	G V T	C W E
E9	E M N	C W N
E10	P G A	C W N
E11	G G H	C W N
E12	G A L	C W N
E13	G P L	C W H
E14	D N K	C W T
E15	R E R	C W V
E16	A E K	C W L
E17	R E K	C W M
E18	S G H	C W F
E19	H G K	C W Y
E20	G R S	C W Y
EP	G D L	C K Q
Inactive control	A E Y	A F N

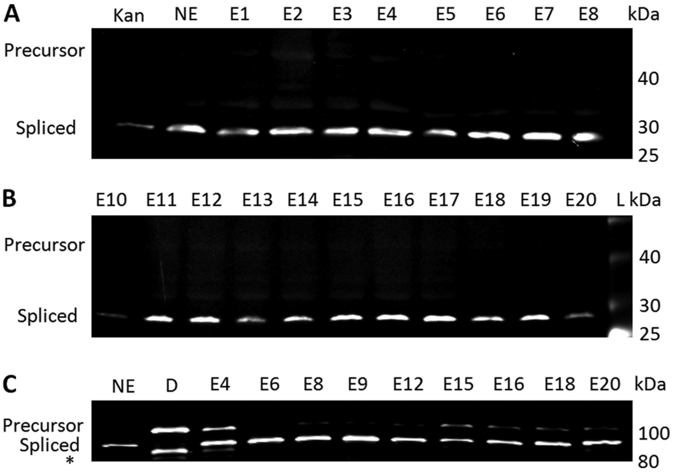

FIGURE 4. **Splicing of the Npu DnaE intein with the extein sequences listed in Table 2 fused to either Kan^R or luciferase.** *A*, shown are Kan^R-Npu DnaE intein fusions E1–E8, the native extein (*NE*), or Kan^R with an N-terminal His tag but no intein (Kan). *B*, shown are Kan^R-Npu DnaE intein fusions E10–E20. *Lane L* contains the NEB prestained broad range (10–230 kDa) Ladder. Western blots *A* and *B* were probed with Mouse IgG anti-His tag antibody followed by detection with LI-COR IRDye 680 goat anti-mouse secondary antibody. *C*, shown are luciferase-Npu DnaE intein fusions with selected or native (*NE*) flanking extein sequences. *Lane D*, a mutated, catalytically inactive intein fusion is shown. The *asterisk* band is the result of proteolysis of the precursor in the inactive intein sample. Western blot C was probed with Mouse IgG anti-firefly luciferase antibody followed by detection with LI-COR IRDye 680 goat anti-mouse secondary antibody.

TABLE 3

Kinetics of *trans*-splicing by the Npu DnaE intein with the listed flanking extein sequences in M-I^N and I^C-P^{His}

Kinetic parameters were measured as described under "Experimental Procedures" for selection hits (E2, E4, E6, E12, and E16), the native extein, the EP clone selected from a low stringency pilot experiment with the native RBS, and an inactive intein with mutations in essential residues C1A in M-I^N and the intein C-terminal Asn plus C+1A in I^C-P^{His}.

Name	N-extein	C-extein	*K*	$t_{1/2}$
	-3 -2 -1	+1 +2 +3	(min^{-1})	
Native	A E Y	C F N	0.63 ± 0.07	66 s
E2	G W N	C M L	0.74 ± 0.3	56 s
E4	G L T	C M Y	0.72 ± 0.12	58 s
E6	R G K	C W E	1.6 ± 0.2	28 s
E12	G A L	C W N	0.41 ± 0.01	101 s
E16	A E K	C W L	0.9 ± 0.3	46 s
EP	G D L	C K Q	0.06 ± 0.004	11.5 min
Inactive control	A E Y	A F N	<0.01	>>1 h*

* No *trans*-splicing product was detected after 1 hour.

insertion sites where we have confidence that splicing should occur. It appears that efficient splicing requires the native C-extein Cys-Phe-Asn or the genetically selected C-extein Cys-Trp and Cys-Met plus one of several acceptable +3 residues. Splicing experiments of the Npu DnaE intein within luciferase in principle validates the use of this predictive approach, albeit the targeted extein sequence was artificially inserted into the enzyme. However, the specificity data alone do not solve all problems when inserting inteins in heterologous proteins. Intein insertion is a complicated balance between expression, stability, and folding in addition to the requirements for splicing. These criteria must still be experimentally probed, but the specificity data provides more opportunities to find sites that permit splicing.

The Preferred Substrate Is Not the Native Extein—All orthologs of the DnaE intein have the same flanking extein sequence (1, 20). Previous studies suggested that DnaE inteins have minimal specificity requirements for the N-terminal extein, but the Cys-Phe-Asn C-terminal extein sequence is required for efficient catalysis (4). The Npu DnaE intein can also support splicing when the +2 position was mutated to Gly, Glu, or Arg, although with significantly lower efficiencies (20). When Iwai *et al.* (17) examined the cross-reactivity of Npu

DnaE I^N with Ssp DnaE I^C where the +2 position was mutated to all 20 amino acids, *trans*-splicing was observed with most uncharged residues after a 6-h induction. Our study is consistent with these findings while also defining alternative C-extein sequences that permit very efficient splicing.

To aid in comparison of selective pressure at each position, we calculated a selectivity ratio (Table 4) that equals the observed frequency of an amino acid divided by the theoretical frequency based on codon degeneracy. A selectivity ratio of ~1.0 indicates an unbiased representation, whereas higher values indicate greater selective pressure for the given amino acid, and lower values indicate selective pressure against that residue or overwhelming selective pressure for another amino acid. The most pronounced substrate specificity requirements for the Npu DnaE intein were found at the +2 and +3 positions. To our surprise, the specificity was for Trp+2 rather than the naturally occurring Phe+2, although all residues were present in the non-selected sample (data not shown). The selectivity ratio for Trp+2 was 29.4, whereas Phe+2 was 0.07 (Table 4). A lower stringency pilot selection yielded a relaxed specificity at +2 and +3 (Fig. 5) that is consistent with the findings of Iwai *et al.* (17).

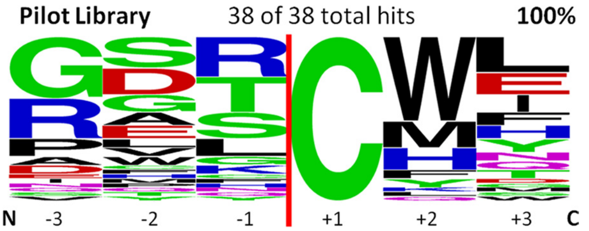

FIGURE 5. **Logo diagram of flanking extein sequences identified in the low stringency pilot study with the native Kan^R RBS reveals expanded specificity at the +2 and +3 positions.** This figure was made using the Weblogo program (28) and shows 38 randomly chosen positive hits selected for splicing of the Npu DnaE intein in Kan^R site C. This construct results in expression of more Kan^R protein than needed for resistance and, therefore, allows selection of precursors that splice slowly or with a low percentage of spliced product (data not shown). The *vertical red line* indicates the position of the intein. The +1 position was fixed as Cys.

The +3 position shows significant selective pressure for Asn, Glu, and Tyr reflected in selectivity ratios of 8.61, 5.5, and 5.23 respectively (Table 4). Glu is similar to the native Asn+3 residue, whereas Tyr may reflect a distinct specificity. The charge or polarity of the +3 amino acid does not have a significant effect on the identity of any position in the N-terminal extein (Fig. 3, *E* and *F*), suggesting that there is no direct amino acid side chain cross-talk between the two halves of the extein or that such cross-talk is highly context-dependent. The only pattern that emerges is the abundance of Gly in the N-terminal extein (Fig. 3*D* and Table 4), with 44% of the selected population containing Gly at either −3 or −2. These two positions have an unbiased frequency for Gly in the unselected control library. Presumably, the intein active site needs some flexibility in the adjacent N-extein to facilitate splicing.

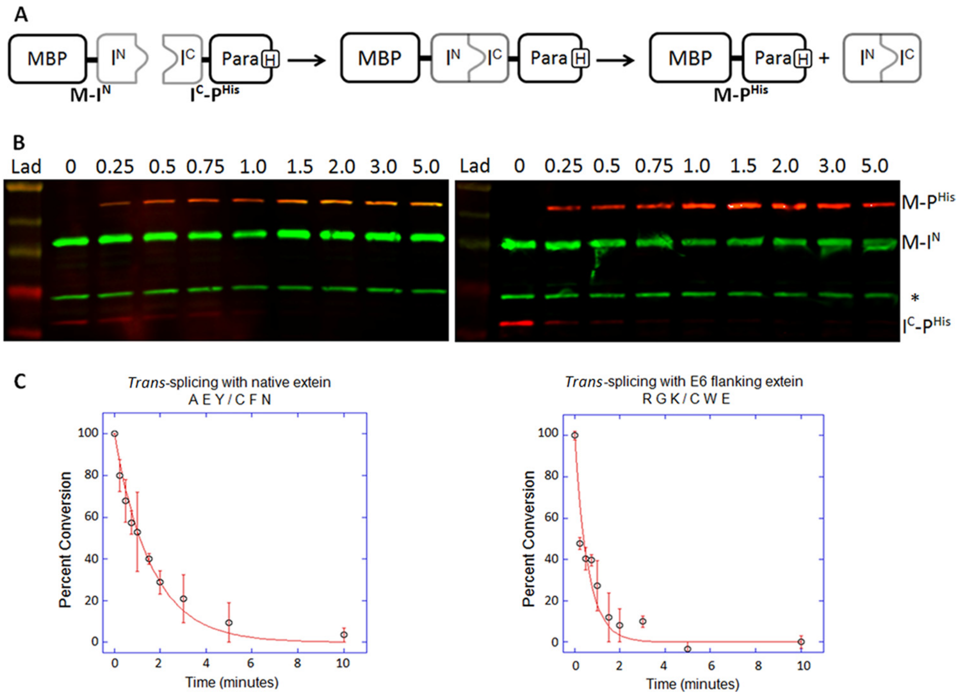

FIGURE 6. **Kinetic analysis of *trans*-splicing inteins with selected flanking extein sequences.** *A*, shown is a schematic representation of the *trans*-splicing reaction. The N-terminal precursor fragment (*M-I^N*) consists of the maltose binding protein (M or MBP) followed by a short linker, three native or selected extein residues, and the N-terminal half of the Npu DnaE intein (*I^N*: residues 1–102). The C-terminal precursor fragment (*I^C-P^His*) consists of the C-terminal half of the Npu DnaE intein (*I^C*; residues 103–138), three native or selected extein residues, a short linker, the ΔSal fragment of paramyosin (*P*) and a C-terminal His tag. M-I^N and I^C-P^His spontaneously assemble when mixed resulting in intein activation and protein splicing. Any unreacted precursor complex and the I^N-I^C product complex dissociate during SDS-PAGE sample preparation. *B*, Western blots show the time course of *trans*-splicing with native (Ala-Glu-Tyr/Cys-Phe-Asn, *left panel*) or E6 (Arg-Gly-Lys/Cys-Trp-Glu, *right panel*) flanking extein residues. Time points from 0 to 5 min are listed across the *top* of each blot. The nitrocellulose membranes were simultaneously probed with mouse IgG anti-His tag antibody and rabbit anti-MBP antiserum followed by detection with LI-COR IRDye 680 goat anti-mouse (*red*) and IRDye 800 goat anti-rabbit secondary antibodies (*green*). The band labeled *M-P^His* was quantified to measure the rate of spliced product formation. The *asterisk* marks endogenous *E. coli* MBP. Lane *Lad* contains the NEB prestained broad range (10–230 kDa) ladder. *C*, shown are plots of the percent conversion of precursor to product during the *trans*-splicing time courses with native (*left panel*) or E6 (*right panel*) flanking extein residues. The *red line* represents the fit of the data to a first order decay reaction, and the S.D. is indicated by the *error bars*.

TABLE 4

Selectivity ratio for the plasmid population that survived kanamycin selection

Data for the 483 randomly chosen clones in Table 1 are shown. The selectivity ratio is defined as the observed frequency divided by the theoretical frequency. Selectivity ratio values near or equal to 1 indicate no selection. Higher selectivity ratios indicate stronger selective pressure for that amino acid, whereas values under 1 indicate negative selective pressure for that amino acid or high selective pressure for another amino acid. The native flanking Npu DnaE extein residues are labeled with asterisks. The +1 position was fixed to the native Cys residue and was thus not included in this table. The intein is present between the −1 and +1 position.

Amino Acid	N-terminal Extein			C-terminal Extein	
	(-3)	(-2)	(-1)	(+2)	(+3)
D	1.39	1.52	0.00	0.00	1.26
E	1.26	*3.51	0.07	0.00	5.50
N	0.93	1.26	2.19	0.00	*8.61
Q	1.39	0.86	0.33	0.00	0.27
H	1.06	1.26	2.25	0.00	1.59
K	1.59	0.93	4.44	0.07	0.00
R	1.81	0.20	0.86	0.00	0.00
S	0.99	1.24	1.37	0.00	0.07
C	0.46	0.40	0.80	0.07	0.07
T	0.63	0.89	1.59	0.00	0.96
P	0.80	1.52	0.00	0.00	0.00
G	4.67	2.68	0.17	0.00	0.03
A	*0.63	0.83	1.92	0.00	0.23
V	0.56	0.50	0.23	0.03	0.10
I	0.07	0.53	0.00	0.00	0.40
L	0.11	0.57	0.75	0.00	0.82
M	0.13	1.13	0.86	2.19	1.46
F	0.13	0.73	1.26	*0.07	2.25
Y	0.20	0.60	*2.05	0.13	5.23
W	0.07	0.40	0.99	29.42	0.07
STOP	0.00	0.00	0.00	0.00	0.00

A second specificity group representing only ~7% of the selected clones has Met at +2. The Met+2 specificity group has a different requirement than the Trp+2 group, preferring a large hydrophobic residue at +3 and a very strong preference for Gly, but only at −3 (Fig. 3*C*).

The relatively broad specificity observed in the Npu DnaE intein is not unprecedented (20). Selective substrate sequestration is critical for enhancing catalytic efficiency and preventing undesirable side reactions in multiple turnover enzymes. In contrast, such sophisticated molecular recognition strategies are not necessary in single turnover enzymes where the substrate is covalently linked to the enzyme, allowing inteins more potential variability in substrate specificity. Furthermore, selective pressure on mobile elements such as inteins may require a higher degree of substrate promiscuity to enable invasion of new sites.

A Paradigm Shift; the Native Extein Is Not Always the Best—In most enzyme systems the natural substrate is the most efficient substrate for the enzyme. However, selection with the Npu DnaE intein did not pull out the native extein sequence. Even more surprising is that we identified extein sequences that surpass the rate of splicing measured with the naturally occurring N- and C- terminal extein. Clearly the paradigm that the optimal splicing site for an intein is its natural context does not hold true in this case. Because natural intein insertion sites are an evolutionary balance between the specificity requirements of the intein, the need to retain a functional host protein, and the specificity requirement of the homing endonuclease, the natural splice site is narrowly defined. However, this narrow specificity does not necessarily reflect the optimal splicing condition for the intein. Our study decouples the specificity requirements of the intein from effects caused by the host protein or homing endonuclease and so provides a more accurate picture of this intein preferred substrates. These data expand the range of known splicable sequences for the Npu DnaE intein and should prove valuable for the reproducible design of heterologous splicing experiments.

Although our studies focused on only one intein, the conclusion that the natural intein insertion sequence only reflects a subset of efficiently splicable extein sequences can likely be generalized to other inteins. As such, it is important to reexamine the substrate specificity of commonly used inteins to understand their true catalytic potential.

Conclusion—We examined the substrate specificity of the Npu DnaE intein under highly stringent conditions and demonstrated that the fastest splicing rates were obtained with extein sequences that are significantly different from the native extein. Whereas previous studies suggested that a C-terminal Cys-Phe-Asn sequence was necessary to achieve highly efficient and rapid splicing (4, 20), our data show that sequences containing Trp and Met at +2 can splice as efficiently as the native extein. These data indicate that the selective pressure to maintain the natural extein sequence overrides the pressure to optimize catalytic efficiency of this intein. These data expand the range of sequences known to allow splicing of this important intein and provide a rubric for evaluating potential intein insertion sites for protein splicing technologies without having to mutate the intein to splice at a desired insertion site sequence.

Acknowledgments—We thank Prof. Philip Farabaugh (University of Maryland, Baltimore County) for generously donating the pEK4 plasmid. We thank Richard Roberts, Gregory Lohman, Paul Balbo, and Andy Gardner (all from NEB) for reading the manuscript and helpful discussion. We thank Don Comb (NEB) for support.

REFERENCES

1. Perler, F. B. (2002) InBase. The intein database. *Nucleic Acids Res.* **30,** 383–384
2. Tori, K., Dassa, B., Johnson, M. A., Southworth, M. W., Brace, L. E., Ishino, Y., Pietrokovski, S., and Perler, F. B. (2010) Splicing of the mycobacteriophage Bethlehem DnaB intein. Identification of a new mechanistic class of inteins that contain an obligate block F nucleophile. *J. Biol. Chem.* **285,** 2515–2526
3. Gangopadhyay, J. P., Jiang, S. Q., van Berkel, P., and Paulus, H. (2003) *In vitro* splicing of erythropoietin by the *Mycobacterium tuberculosis* RecA intein without substituting amino acids at the splice junctions. *Biochim. Biophys. Acta* **1619,** 193–200
4. Lockless, S. W., and Muir, T. W. (2009) Traceless protein splicing utilizing evolved split inteins. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 10999–11004
5. Dalgaard, J. Z., Moser, M. J., Hughey, R., and Mian, I. S. (1997) Statistical modeling, phylogenetic analysis, and structure prediction of a protein splicing domain common to inteins and hedgehog proteins. *J. Comput.*

Biol. **4,** 193–214

6. Flavell, R. R., and Muir, T. W. (2009) Expressed protein ligation (EPL) in the study of signal transduction, ion conduction, and chromatin biology. *Acc. Chem. Res.* **42,** 107–116
7. Fong, B. A., Wu, W. Y., and Wood, D. W. (2010) The potential role of self-cleaving purification tags in commercial-scale processes. *Trends Biotechnol.* **28,** 272–279
8. Volkmann, G., and Iwaï, H. (2010) Protein *trans*-splicing and its use in structural biology. Opportunities and limitations. *Mol. Biosyst.* **6,** 2110–2121
9. Mootz, H. D. (2009) Split inteins as versatile tools for protein semisynthesis. *Chembiochem* **10,** 2579–2589
10. Amitai, G., Callahan, B. P., Stanger, M. J., Belfort, G., and Belfort, M. (2009) Modulation of intein activity by its neighboring extein substrates. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 11005–11010
11. Lew, B. M., and Paulus, H. (2002) An *in vivo* screening system against protein splicing useful for the isolation of non-splicing mutants or inhibitors of the RecA intein of *Mycobacterium tuberculosis*. *Gene* **282,** 169–177
12. Oeemig, J. S., Zhou, D., Kajander, T., Wlodawer, A., and Iwaï, H. (2012) NMR and crystal structures of the *Pyrococcus horikoshii* RadA intein guide a strategy for engineering a highly efficient and promiscuous intein. *J. Mol. Biol.* **421,** 85–99
13. Shemella, P. T., Topilina, N. I., Soga, I., Pereira, B., Belfort, G., Belfort, M., and Nayak, S. K. (2011) Electronic structure of neighboring extein residue modulates intein C-terminal cleavage activity. *Biophys. J.* **100,** 2217–2225
14. Chong, S., Montello, G. E., Zhang, A., Cantor, E. J., Liao, W., Xu, M. Q., and Benner, J. (1998) Utilizing the C-terminal cleavage activity of a protein splicing element to purify recombinant proteins in a single chromatographic step. *Nucleic Acids Res.* **26,** 5109–5115
15. Appleby-Tagoe, J. H., Thiel, I. V., Wang, Y., Wang, Y., Mootz, H. D., and Liu, X. Q. (2011) Highly efficient and more general cis- and trans-splicing inteins through sequential directed evolution. *J. Biol. Chem.* **286,** 34440–34447
16. Oeemig, J. S., Aranko, A. S., Djupsjöbacka, J., Heinämäki, K., and Iwaï, H. (2009) Solution structure of DnaE intein from *Nostoc punctiforme*. Structural basis for the design of a new split intein suitable for site-specific chemical modification. *FEBS Lett.* **583,** 1451–1456
17. Iwai, H., Züger, S., Jin, J., and Tam, P. H. (2006) Highly efficient protein *trans*-splicing by a naturally split DnaE intein from *Nostoc punctiforme*. *FEBS Lett.* **580,** 1853–1858
18. Zettler, J., Schütz, V., and Mootz, H. D. (2009) The naturally split Npu DnaE intein exhibits an extraordinarily high rate in the protein *trans*-splicing reaction. *FEBS Lett.* **583,** 909–914
19. Carvajal-Vallejos, P., Pallissé, R., Mootz, H. D., and Schmidt, S. R. (2012) Unprecedented rates and efficiencies revealed for new natural split inteins from metagenomic sources. *J. Biol. Chem.* **287,** 28686–28696
20. Shah, N. H., Dann, G. P., Vila-Perelló, M., Liu, Z., and Muir, T. W. (2012) Ultrafast protein splicing is common among cyanobacterial split inteins. Implications for protein engineering. *J. Am. Chem. Soc.* **134,** 11338–11341
21. Nurizzo, D., Shewry, S. C., Perlin, M. H., Brown, S. A., Dholakia, J. N., Fuchs, R. L., Deva, T., Baker, E. N., and Smith, C. A. (2003) The crystal structure of aminoglycoside-3′-phosphotransferase-IIa, an enzyme responsible for antibiotic resistance. *J. Mol. Biol.* **327,** 491–506
22. Fürste, J. P., Pansegrau, W., Frank, R., Blöcker, H., Scholz, P., Bagdasarian, M., and Lanka, E. (1986) Molecular cloning of the plasmid RP4 primase region in a multi-host-range tacP expression vector. *Gene* **48,** 119–131
23. Kramer, E. B., and Farabaugh, P. J. (2007) The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. *RNA* **13,** 87–96
24. Daugelat, S., and Jacobs, W. R., Jr. (1999) The *Mycobacterium tuberculosis* recA intein can be used in an ORFTRAP to select for open reading frames. *Protein Sci.* **8,** 644–653
25. Tori, K., Cheriyan, M., Pedamallu, C. S., Contreras, M. A., and Perler, F. B. (2012) The *Thermococcus kodakaraensis* Tko CDC21–1 intein activates its N-terminal splice junction in the absence of a conserved histidine by a compensatory mechanism. *Biochemistry* **51,** 2496–2505
26. Shi, J., and Muir, T. W. (2005) Development of a tandem protein trans-splicing system based on native and engineered split inteins. *J. Am. Chem. Soc.* **127,** 6198–6206
27. Martin, D. D., Xu, M. Q., and Evans, T. C., Jr. (2001) Characterization of a naturally occurring trans-splicing intein from *Synechocystis* sp. PCC6803. *Biochemistry* **40,** 1393–1402
28. Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004) WebLogo. A sequence logo generator. *Genome Res.* **14,** 1188–1190