

The crystal structure of the naturally split gp41-1 intein guides the engineering of orthogonal split inteins from *cis*-splicing inteins

Hannes M. Beyer¹, Kornelia M. Mikula¹, Mi Li^{2,3} , Alexander Wlodawer²  and Hideo Iwai¹ 

¹ Research Program in Structural Biology and Biophysics, Helsinki Life Science Institute-Institute of Biotechnology, University of Helsinki, Finland

² Macromolecular Crystallography Laboratory, National Cancer Institute, Frederick, MD, USA

³ Basic Science Program, Leidos Biomedical Research, Frederick National Laboratory for Cancer Research, MD, USA

Keywords

crystal structure; gp41-1 intein; orthogonal split intein; protein engineering; protein splicing

Correspondence

H. Iwai, Research Program in Structural Biology and Biophysics, Institute of Biotechnology, University of Helsinki, 00014 Helsinki, Finland
Tel: +358-2941 59752
E-mail: hideo.iwai@helsinki.fi

(Received 22 July 2019, revised 1 October 2019, accepted 29 October 2019)

doi:10.1111/febs.15113

Protein *trans*-splicing catalyzed by split inteins has increasingly become useful as a protein engineering tool. We solved the 1.0 Å-resolution crystal structure of a fused variant from the naturally split gp41-1 intein, previously identified from environmental metagenomic sequence data. The structure of the 125-residue gp41-1 intein revealed a compact pseudo-C2-symmetry commonly found in the Hedgehog/Intein superfamily with extensive charge-charge interactions between the split N- and C-terminal intein fragments that are common among naturally occurring split inteins. We successfully created orthogonal split inteins by engineering a similar charge network into the same region of a *cis*-splicing intein. This strategy could be applicable for creating novel natural-like split inteins from other, more prevalent *cis*-splicing inteins.

Database

Structural data are available in the RCSB Protein Data Bank under the accession number [6QAZ](#).

Introduction

Protein splicing is a post-translational modification where an intervening protein (intein) residing within an unrelated host protein excises itself, while covalently ligating the N- and C-terminally flanking sequences (exteins) with a standard peptide bond [1–4]. As a result, the ligated product is scar-less and devoid of any indication of its previous merged existence, while the function of the host protein is generally restored (Fig. 1). Inteins are commonly regarded as selfish parasitic elements, albeit some evidence attributes a regulatory role in controlling the activity of host proteins in response to environmental cues triggering the splicing reaction [3,5].

During recent years, inteins have become increasingly popular for diverse applications in biotechnology, chemical biology, and synthetic biology due to the following properties. First, intein-mediated protein splicing tolerates the deliberate exchange of extein sequences [3,6]. Second, the existence of naturally occurring split inteins that are capable of reconstituting a functional protein from two polypeptide chains, as well as the possibility of splitting *cis*-splicing inteins, generates ample possibilities for applications with protein *trans*-splicing (PTS) [3,7–9] (Fig. 1B). Ever since the discovery of protein splicing by inteins, engineering of inteins toward high performance, high tolerance of

Abbreviations

CI, charge-introduced; CS, charge-swapped; DnaB, bacterial helicase; DnaE, catalytic α subunit of DNA polymerase III; GB1, B1 domain of the *Streptococcus* sp. IgG-binding protein G; HINT, Hedgehog/INTEin; Int_C, C-terminal intein split fragment; Int_N, N-terminal intein split fragment; IPTG, isopropyl β -D-1-thiogalactopyranoside; *Npu*, *Nostoc punctiforme*; NTA, nitrilotriacetic acid; Oth, orthogonal; PDB, protein data bank; PEG, polyethylene glycol; PTS, protein *trans*-splicing; SUMO, yeast small ubiquitin-like modifier domain.

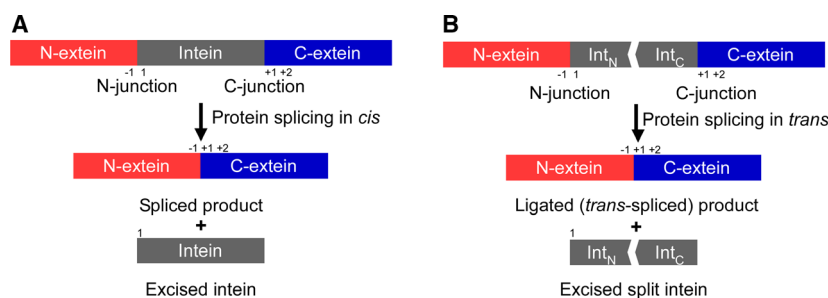


Fig. 1. Schematic representation of protein splicing in *cis* and *trans*. (A) *Cis*-splicing inteins excise themselves from a precursor where the N- and C-exteins flank the intein on the same polypeptide. (B) PTS ligates N- and C-exteins, each originating from an independent polypeptide, with a covalent peptide bond resulting in a *trans*-spliced product. Interaction of the N- (Int_N) and C-terminal (Int_C) split intein halves initiates the protein-splicing reaction. The N-terminal junction sequence at the front of an intein is termed as the -1 position. The $+1$ position after the intein sequence usually has a Cys, Ser, or Thr residue. The second residue after an intein is numbered as the $+2$ position.

junction sequences, and smaller variants has been an ongoing quest [9,10]. Successfully engineered inteins arose from the accumulation of beneficial mutations upon directed evolution [11–13], propagation of consensus sequence [14], and as a result of rational design [15,16].

Although more than 1500 putative inteins have been identified, only a few naturally occurring split inteins, such as the cyanobacterial catalytic α subunit of DNA polymerase III (DnaE) inteins, have been reported [17]. The naturally fragmented gp41-1 intein was found as a result of metagenomic sequencing [18]. It is one of the smallest reported split inteins with very fast *trans*-splicing activity [19], consisting of 88-residue N-terminal (Int_N) and 37-residue C-terminal (Int_C) fragments. Its small size and robust protein-splicing activity make it an attractive template for protein engineering [19]. Also, gp41-1 intein makes use of Ser as the catalytic residue at the $+1$ position, the first extein residue following the intein sequence. Given the much higher frequency of Ser over Cys within prokaryotic and eukaryotic proteins, inteins with $+1\text{Ser}$ allow a broader spectrum of possible insertion sites for scar-less protein ligation than naturally split inteins with Cys at the $+1$ position, thereby expanding potential applications.

Despite increasing interest in the utilization of various split inteins for protein engineering purposes, the repertoire of split inteins with both robust protein-splicing activity and high sequence tolerance at the splice junctions is still very small. Particularly, pairs of orthogonal (Oth) split inteins are desirable for one-pot multiple-fragment protein ligation and biorthogonal conjugation by PTS requiring two or more orthogonal split inteins [20–23]. Previous engineering attempts to derive novel split inteins from naturally occurring *cis*-splicing inteins did not result in highly robust split inteins, indicating that *cis*-splicing inteins are not

optimized for *trans*-splicing, unlike naturally occurring split inteins [15,24,25].

Here we report the 1.0 Å-resolution crystal structure of a variant of the naturally split gp41-1 intein. Based on the crystal structure, we grafted the structural features of naturally split inteins onto a *cis*-splicing intein to develop novel natural-like split inteins and demonstrate the engineering of orthogonal split intein fragments from a *cis*-splicing intein.

Results

Crystal structure of the gp41-1 intein

As the first step to engineer inteins based on the gp41-1 intein, we created a *cis*-splicing gp41-1 intein variant by genetically fusing the gp41-1_N and gp41-1_C split fragments. We found that the *cis*-splicing gp41-1 intein retained high protein-splicing activity when the three native extein residues were kept (Fig. 2). Next, for structure determination, we crystallized an inactive mutant of the *cis*-splicing gp41-1 intein bearing an alanine mutation as the first residue (C1A), to prevent N-cleavage due to the N-S acyl shift. Because the covalent fusion of gp41-1_N and gp41-1_C forces the association between the split intein fragments, we could solve the crystal structure of that inactive variant of the gp41-1 intein at the resolution of 1.0 Å. We used the crystal structure of the *Npu*DnaE intein as a search model for molecular replacement (Table 1). The structure of gp41-1 has the canonical intein horseshoe shape, termed Hedgehog/INTEin (HINT) fold (Fig. 2A). The structure confirmed that the loop engineering of naturally split gp41-1 intein fragments by introducing a peptide bond did not induce domain-swapping as it previously occurred with a short connection linker [26]. The active site region surrounding the terminal residues C1A and Asn125 is depicted in

Fig. 2B and involves His63, Asp107, Thr123, and His124. The side chain of Ala1 is pointing away from the last Asn125 residue, similar to other intein structures with ‘open’ rather than ‘closed’ conformation [15,16,27]. A Dali server search identified the engineered bacterial helicase (DnaB) mini-intein (PDB ID: 4or1) from *Nostoc punctiforme* (*NpuDnaB*^{A290}) as the closest structure to the gp41-1 intein, with a Z-score of 20.1 and an rmsd of 1.4 Å for C^α atoms of 127 residues [28]. *NpuDnaB*^{A290} intein is composed of 139 residues, which is 14 residues larger than the gp41-1 intein [15]. The main differences in the length between the two structures are found in two distinct regions (Fig. 2C,D). One is in the split fragment-connecting loop where canonical inteins harbor a homing endonuclease domain insertion (C35 site) [25], while the other is a loop at the pseudo-C2-symmetry related site (N35 site) [25]. These regions account for 11 residues of the size difference.

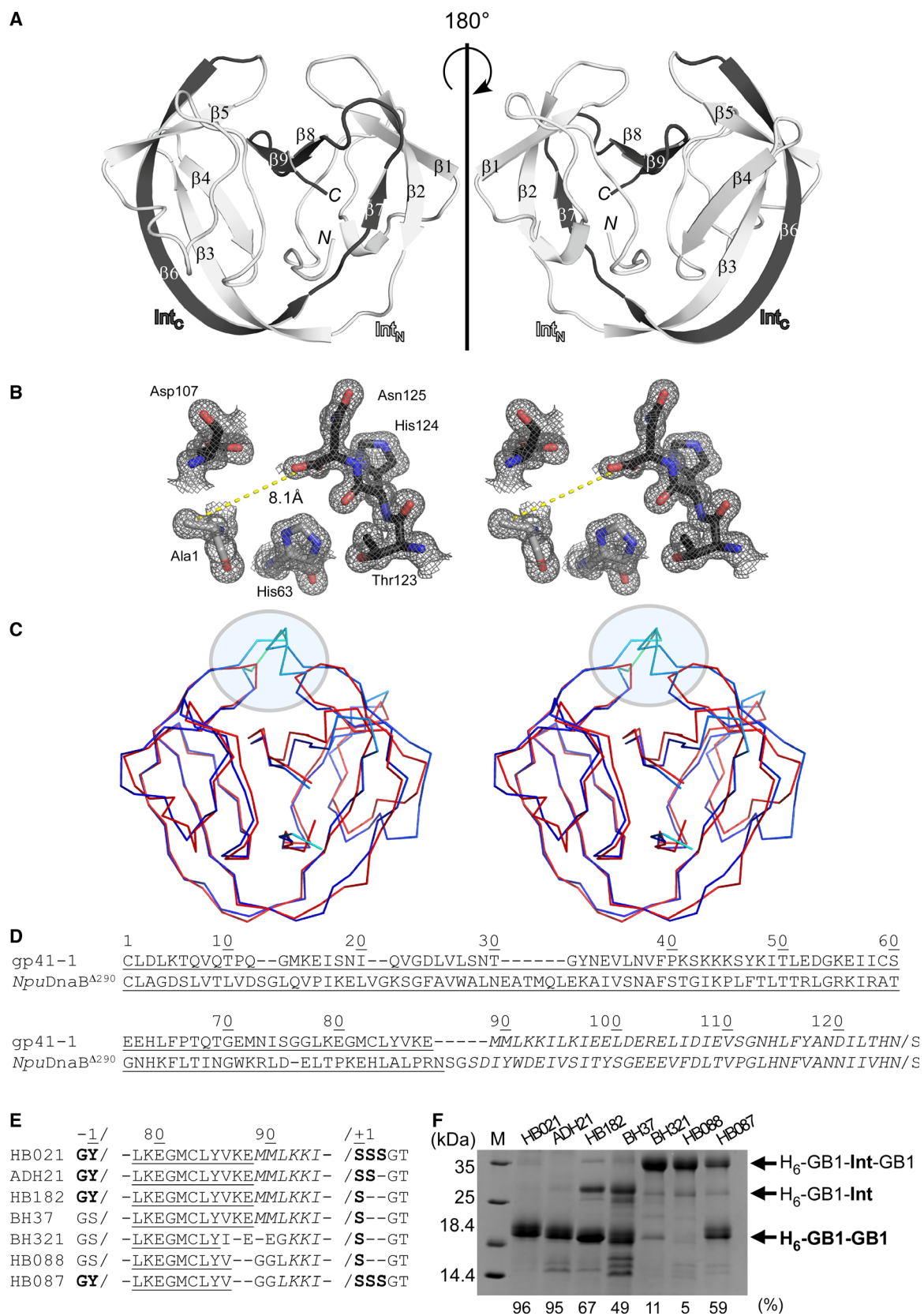
The smaller size of the gp41-1 intein led to a compact pseudo-C2-symmetric structure found in the HINT fold with shorter loops due to the shortened insertions [29] (Fig. 2). The two C2-symmetry-related regions (residues 3–52 and residues 60–110) can be well superimposed [25,29] (Fig. 3A). The gp41-1 intein structure can thus be dissected into four distinct units: the first C2-symmetry-related unit, β-strand 4 (β4), the second C2-symmetry-related unit, and two β-strands (β8, β9) (Fig. 3B). The C2-symmetry-related unit can further be divided into a globular region and two β-strands (β2 and β3, or β6 and β7). The natural split site of the gp41-1 intein is located within the second C2-symmetry-related unit, separating the C2-symmetry unit into a globular region and two β-strands (Fig. 3B).

Minimizing the gp41-1 intein

The first question we asked was whether it is possible to minimize the *cis*-splicing gp41-1 intein to a size even

smaller than 125 residues. The gp41-1 intein is among the smallest inteins identified to date, indicating that it underwent a natural selection for the size reduction. Because many intein-based applications would favor minimal interference, one would generally wish to use small rather than large inteins. The conserved insertion site for a homing endonuclease found in canonical inteins also overlaps with the split sites most commonly found for many split inteins [25]. We removed two residues from the linker where Int_N and Int_C were connected, that is, at the natural split site of the gp41-1 intein. This deletion reduced the protein-splicing efficiency by about 40% (Fig. 2D–E), indicating that the naturally reduced size of the gp41-1 intein has presumably reached a functional minimum. Our attempt at optimizing the linker sequence to rescue the robust splicing activity of the gp41-1 intein did not succeed (Fig. 2D–E). While *NpuDnaB*^{A290} intein, the closest crystal structure, shows higher B-factors for the backbone atoms of the corresponding linker region ($43.4 \pm 2.7 \text{ \AA}^2$), the B-factors for the corresponding regions in the gp41-1 intein crystal structure are much lower ($22.8 \pm 3.1 \text{ \AA}^2$), even though the linkers are not involved in crystal contacts. The lower B-factors thus suggest that this region might eventually be less flexible, exhibiting a higher degree of order in the gp41-1 intein. This region also contains an unusual *cis* peptide bond between Lys87 and Glu88, presumably caused by the covalent bond to fuse the gp41-1 intein split fragments without any linker insertion, corroborating that further linker shortening likely affects the activity of the intein. The presence of the *cis* peptide bond is unambiguously supported by the excellent electron density, although part of the side chain of Glu88 appears to be disordered. Because the gp41-1 intein does not seem to easily tolerate any deletion, the linker length at this site could play an essential role during productive folding of some HINT superfamily members [30]. Moreover, we observed deteriorating effects

Fig. 2. Crystal structure of the C1A variant of the engineered *cis*-splicing gp41-1 intein. (A) Ribbon representations of the gp41-1 intein structure. The region corresponding to the C-terminal split fragment (Int_C) is colored in dark gray. N and C indicate N and C termini, respectively. (B) Stereoview of the active site residues of the gp41-1 intein structure. The distance between the C^β atom of Ala1 and the carbonyl carbon of Asn125 is shown. (C) Stereoview of an overlay of the crystal structures of the gp41-1 intein (red) and the closest related structure, the *NpuDnaB* mini-intein (blue) (PDB: 4or1). A circle indicates the C35- and C2-symmetry-related N35 sites where additional residues are inserted in the *NpuDnaB* mini-intein. (A–C) Structures were drawn using PyMOL. (D) Sequence alignment between the *cis*-splicing gp41-1 intein and *NpuDnaB* mini-intein (*NpuDnaB*^{A290}). Underlined and italicized letters indicate sequence corresponding to the N- and C-terminal split fragments, respectively. (E) Engineering of the gp41-1 intein in the loop and splicing junction regions and their effects on protein splicing in *cis*. HB021, ADH21, HB182, BH37, BH321, HB088, and HB087 indicate the short names for different constructs with the sequence variations shown in the sequence alignment. (D, E) Sequence alignments were generated using Clustal Omega. (F) SDS/PAGE analysis of the *cis*-splicing activity of the engineered gp41-1 intein variants. M stands for molecular weight markers. H₆-GB1-Int-GB1 indicates unspliced precursor proteins (GB1, B1 domain of the *Streptococcus* sp. IgG-binding protein G). H₆-GB1-GB1 indicates *cis*-spliced products with various junction sequences causing minor variations in the migration profile. H₆-GB1-Int indicates an off-pathway C-cleavage product. The splicing efficiency in percent quantified from the gel is shown at the bottom. The gel summarizes representative results of one to four individually performed experiments per construct.



on the protein-splicing activity when deviating the extein sequences from the native sequences ('Y' as N-extein and 'SSS' as C-extein; Fig. 2D–E). Thus, gp41-1 intein might not be suitable for further engineering compared with the more robust DnaE intein from *N. punctiforme* (*NpuDnaE*), the latter tolerating various amino acid types at the splicing junctions [31].

The charge network in the gp41-1 intein

Previously, it has been suggested that local charge distributions between naturally split intein halves are important for their association [18,23,32]. We observed extensive charge–charge interactions in the crystal structure of the gp41-1 intein, as also identified among other naturally split inteins [32] (Fig. 3C,D). Particularly, they are located in the interacting regions within

the β -strands of the two C2-symmetry-related units. Recently, a 'capture and collapse' model has been proposed as a folding mechanism for the naturally split *NpuDnaE* intein, in which the first step of the interaction between the split fragments is initiated by electrostatic interactions on the extended β -strands [33]. A dissociation constant of 1.2 nM reported for the split *NpuDnaE* intein suggests strong binding between naturally split intein fragments, although the binding constant could be extein-dependent [33,34]. We identified similar electrostatic networks between β_6 at the beginning of Int_C and β_3 at the C-end region of Int_N in the structure of the gp41-1 intein, but with inverted charges compared to the *NpuDnaE* intein (Fig. 3C,D). These two anti-parallel β -strands appeared to form a charge zipper, reminiscent of leucine zipper structures but embedded in extended strands rather than helices [35]. We also compared the charge patterns in the same regions with the naturally split *NpuDnaE* intein and its closest structural homolog, the *cis*-splicing *NpuDnaB* mini-intein (Fig. 3D). Although the *NpuDnaB* mini-intein does not contain such an extensive charge network in the corresponding region, the gp41-1 intein encompasses more prominent charge interactions than the *NpuDnaE* intein (Fig. 3D). This observation might support the notion that the 'capture and collapse' model suggested for the naturally split *NpuDnaE* intein might also be valid for the naturally split gp41-1 intein [33].

Interestingly, the Int_C region of the gp41-1 intein is dominated by negative charges, as opposed to more positive charges found in the same region of the *NpuDnaE* intein (Fig. 3D). The charge distributions in β_3 and β_6 are thus opposite between the naturally split *NpuDnaE* and gp41-1 inteins (Fig. 3D). Therefore, we decided to test if it is possible to swap the charge distribution in β_3 and β_6 by mimicking the charge pattern of the gp41-1 intein onto the *NpuDnaE* intein. We introduced three lysines in Int_N and three glutamates in the Int_C of the *NpuDnaE* intein (Fig. 4A). This Charge-Swapped *NpuDnaE* intein (CS-*NpuDnaE*) could efficiently splice in *cis* (96% splicing efficiency), confirming that swapping these charges does not influence protein splicing in *cis* (Fig. 4B). This result is in line with the previous report in which the charge-swapping of *NpuDnaE* intein was successfully introduced into the entire *NpuDnaE_N* and *NpuDnaE_C* fragments to suppress the cross-reactivity [23].

Orthogonality of charge-swapped split inteins

The naturally split gp41-1 intein seems to be more sensitive to sequence changes at the splice and loop junctions than we anticipated (Fig. 2E). This poorer

Table 1. Data collection and refinement statistics.

Data collection	Diamond i02
Wavelength	0.9795
Space group	I222
Molecules/a.u.	1
Unit cell <i>a</i> , <i>b</i> , <i>c</i> (Å); $\alpha = \beta = \gamma$ (°)	48.81, 69.99, 71.24 90
Resolution (Å) ^a	49.92–1.02 (1.04–1.02)
<i>R</i> _{merge} (%) ^b	3.8 (28.8)
<i>R</i> _{pim} (%) ^c	1.9 (23.1)
No. of reflections (measured/unique)	242 564/55 708
$\langle I/\sigma I \rangle$	28.3 (2.8)
Completeness (%)	89.4 (37)
Redundancy	4.35
Refinement	
Resolution (Å)	49.92–1.02
No. of reflections (refinement/ <i>R</i> _{free})	53 080/2642
<i>R</i> / <i>R</i> _{free} ^d	12.12/14.98
No. atoms	
Protein	1093
Ligand/ion	51
Water	220
Rmsd from ideal	
Bond lengths (Å)	0.019
Bond angles (°)	2.086
Ramachandran plot	
Favored (%)	98.4
Allowed (%)	1.6
PDB code	6QAZ

^aThe highest resolution shell is shown in parentheses; ^b $R_{\text{merge}} = \sum_h \sum_i |I_i - \langle I \rangle| / \sum_h \sum_i I_i$, where I_i is the observed intensity of the i -th measurement of reflection h , and $\langle I \rangle$ is the average intensity of that reflection obtained from multiple observations; ^c Defined in Weiss *et al.*, 2001 [59]; ^d $R = \sum \|F_o - |F_c| \| / \sum |F_o|$, where F_o and F_c are the observed and calculated structure factors, respectively, calculated for all data. *R*_{free} was defined in Brünger, 1992 [60].

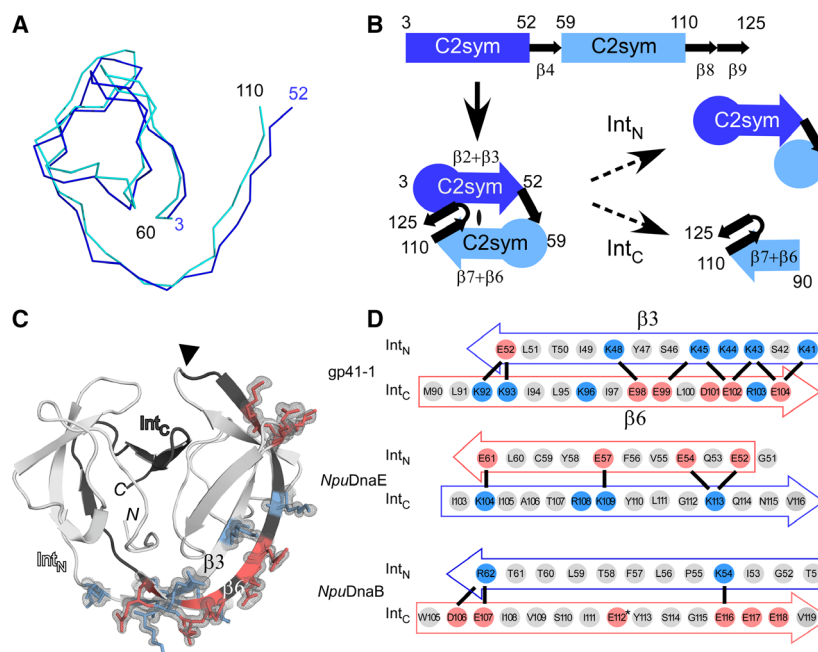


Fig. 3. The modular architecture of the gp41-1 intein and the charge network. (A) An overlay of the backbone atoms of the two C2-symmetry-related units (residues 3-52 and 60-110) observed in the gp41-1 intein structure. (B) The arrangement of the C2-symmetry-related units and connections of the secondary structures. The natural split site of the gp41-1 intein locates within the second C2-symmetry-related part and at the front of strand $\beta 6$. (C) The charged network found in the gp41-1 intein structure. The side-chains of the charged residues in the $\beta 3$ and $\beta 6$ strands are shown together with the electron density map. Residues with negative and positive charges are highlighted in red and blue, respectively. A filled triangle indicates the natural split site. *N* and *C* indicate the termini. (D) Comparison of the charged residues in the $\beta 3$ and $\beta 6$ strands between the gp41-1, *NpuDnaE*, and *NpuDnaB* inteins. Thick lines indicate possible favored charge interactions. An asterisk indicates Glu112 modeled as Val112 in the coordinate of the *NpuDnaB* mini-intein structure (PDB: 4o1r). (A, C) Figures were produced by PyMOL.

tolerance could constrain its practical applications by PTS. In contrast, the naturally split *NpuDnaE* intein and its homologs are more tolerant of amino acid changes at the splice junctions, making them better suited for protein engineering applications than the gp41-1 intein [14,31,36–38]. However, naturally split DnaE inteins from cyanobacteria are cross-reactive to each other [18,31]. This cross-activity limits their application in, for example, one-pot three-fragment ligation by PTS requiring two split inteins. For such multi-fragment applications, two or more non-cross-active (Oth) split inteins are needed to suppress undesired cross-activity. Several approaches have been used to circumvent the cross-reactivity, such as utilizing different split sites of the *NpuDnaE* intein, or kinetic control of two split inteins [21–23]. The three-dimensional structure of the gp41-1 intein revealed charge distributions different from the *NpuDnaE* intein in the corresponding $\beta 3$ and $\beta 6$ strands. Next, we asked if the charge network found in $\beta 3$ and $\beta 6$ strands of naturally split inteins can be responsible for the orthogonality of split inteins. We created a split intein from the CS *NpuDnaE* intein (CS-*NpuDnaE*) and tested the

cross-activity of the split fragments. CS-*NpuDnaE_N* and CS-*NpuDnaE_C* could still efficiently splice with the wild-type Int_C (*NpuDnaE_C*, 54% yield) and Int_N (*NpuDnaE_N*, 92% yield) fragments of the *NpuDnaE* intein, respectively, compared to the reference sample where both fragments were CS (100% yield). This high cross-activity of CS-*NpuDnaE* suggests that the charge network in the $\beta 3$ and $\beta 6$ regions alone cannot account for the cross-activity among the naturally split DnaE inteins (Fig. 4C). Nevertheless, the charges in this region may play an important role, for example, for keeping split intein fragments soluble. This observation is consistent with the previous report that the C-terminal 16-residue fragment of *NpuDnaE* intein is sufficient for efficient *trans*-splicing of the *NpuDnaE* intein [34,39,40].

Engineering of orthogonal split inteins

In contrast to naturally split inteins, *cis*-splicing inteins generally possess a less pronounced charge network within the regions corresponding to $\beta 3$ and $\beta 6$ in the gp41-1 intein structure [32]. Artificially split inteins

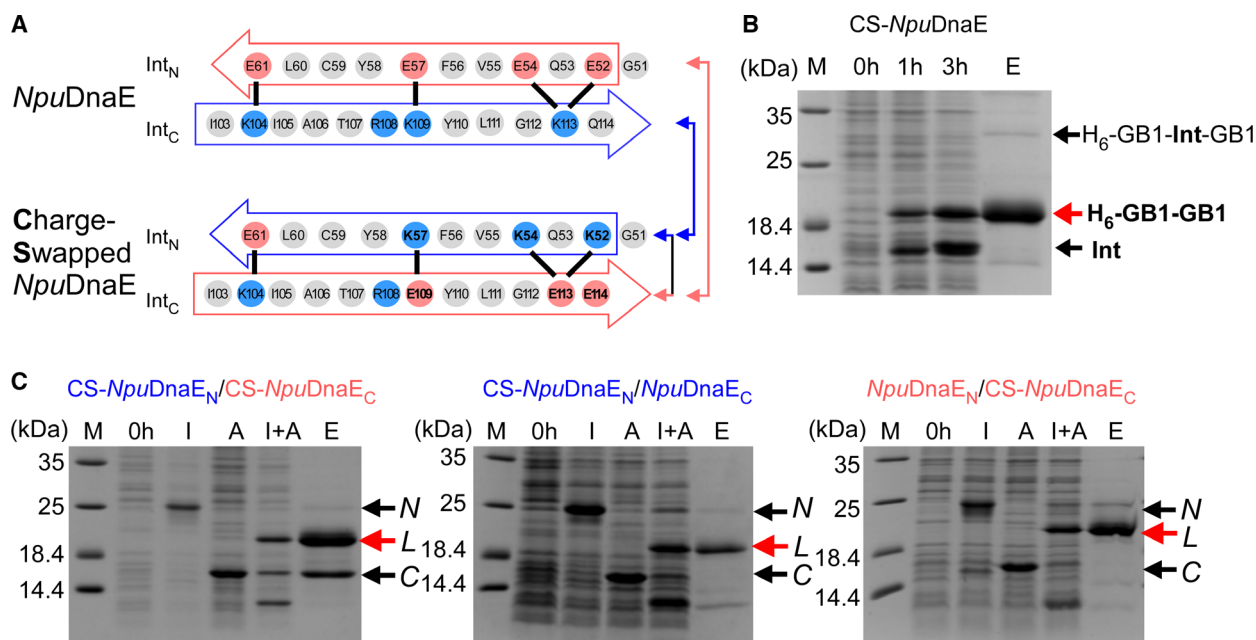
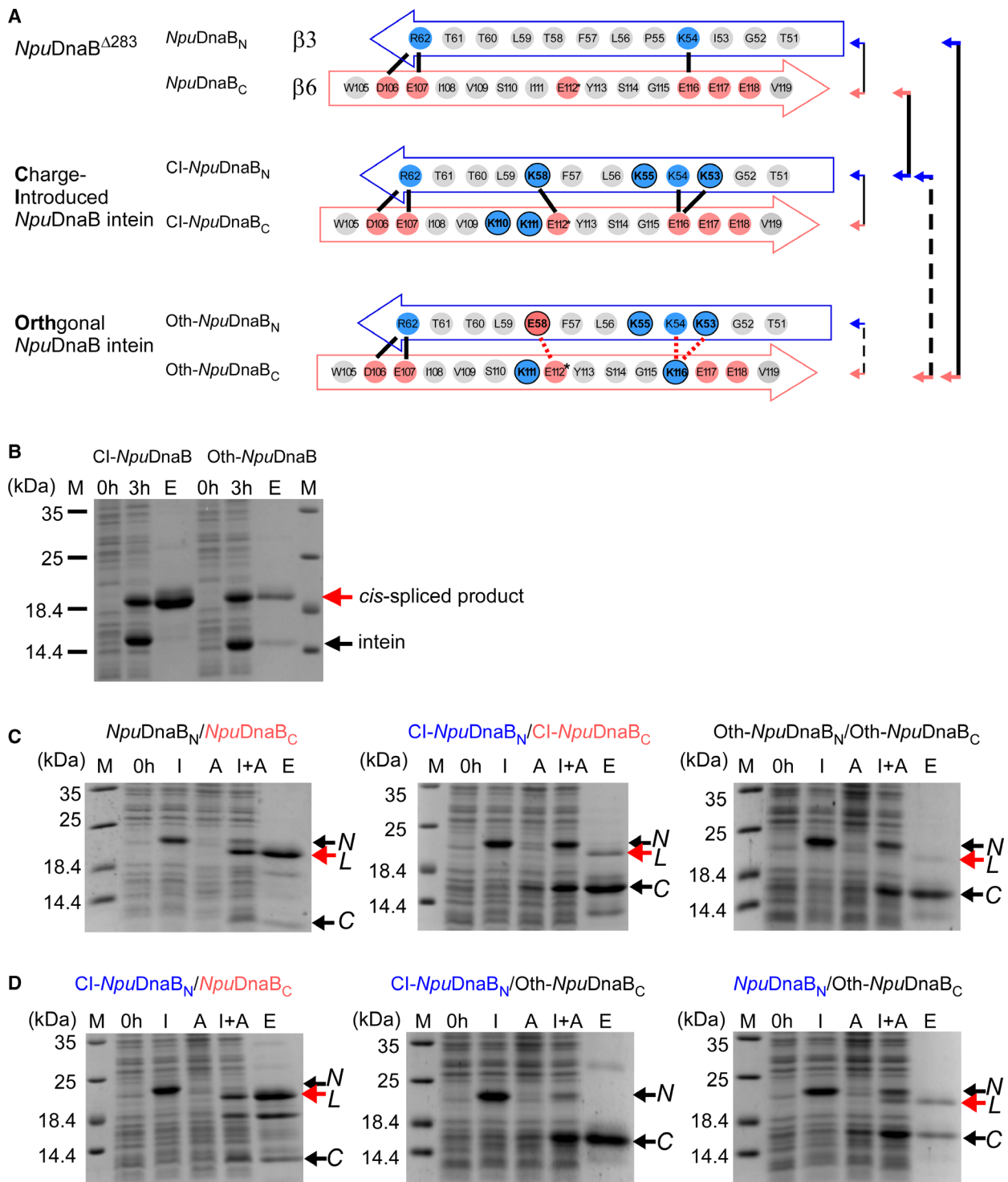


Fig. 4. Charge engineering of the *NpuDnaE* intein. (A) Charge distributions of the *NpuDnaE* and the CS-*NpuDnaE* inteins corresponding to strands $\beta 3$ and $\beta 6$ in the gp41-1 intein structure. Color-coded arrows indicate the pairs of split inteins tested for *trans*-splicing in panel C with lines. (B) *Cis*-splicing analysis of the CS-*NpuDnaE* intein by SDS/PAGE. M, 0 h, 3 h, and E stand for molecular markers, 0 h, 3 h after induction, and elution from Ni-NTA spin columns. A red arrow indicates the band corresponding to the *cis*-spliced H₆-GB1-GB1 product. A representative gel of three individual experiments is shown. (C) Cross-activity between split *NpuDnaE* and CS-*NpuDnaE* inteins. SDS/PAGE analysis of *trans*-splicing for three different pairs of split inteins, CS-*NpuDnaE*_N/CS-*NpuDnaE*_C (left), CS-*NpuDnaE*_N/*NpuDnaE*_C (middle), and *NpuDnaE*_N/CS-*NpuDnaE*_C (right), which are illustrated by arrows and lines in panel A. N, L, and C denote the N-terminal fragment with Int_N, the ligated product, and the C-terminal fragment with Int_C, respectively. I, A, I+A, and E stand for IPTG induction, arabinose induction, both IPTG and arabinose induction, and elution from Ni-NTA spin columns. IPTG induction produces the N-terminal precursor fragment, N. Arabinose induces the expression of the C-terminal precursor, C. Only dual induction by IPTG and arabinose (I+A) is expected to produce the ligated product L by PTS. A representative gel of two individual experiments for each combination is shown.

derived from *cis*-splicing inteins are often poorly soluble and might not be suitable for protein ligation because they would require unfolding/refolding processes to initiate PTS [41,42]. Hence, it would be of particular interest if one could convert *cis*-splicing inteins into natural-like split inteins by grafting the charge network similar to the one observed in

naturally split inteins. We chose the *cis*-splicing *NpuDnaB* mini-intein for the grafting approach. This intein is the closest structural homolog to the gp41-1 intein and is superior in tolerating sequence alterations at the splice junctions while retaining high splicing activity [15,43,44]. We introduced five lysine residues (three in $\beta 3$ and two in $\beta 6$) (Fig. 5A). The

Fig. 5. Engineering of the *cis*-splicing *NpuDnaB* mini-intein toward an orthogonal split intein pair. (A) Schematic comparison between the original and engineered *NpuDnaB* mini-inteins in the regions corresponding to $\beta 3$ and $\beta 6$ in the gp41-1 intein structure. Solid lines indicate possible favored charge interactions. Dashed red lines indicate possible unfavored charge interactions. Pairs of the split inteins tested in panels B and C are indicated by arrows and lines, where broken lines indicate poor or no *trans*-splicing. (B) *Cis*-splicing of the Cl-*NpuDnaB* and the designed Oth-*NpuDnaB* mini-inteins by SDS/PAGE analysis. *Cis*-spliced products and excised inteins are indicated by red and black arrows, respectively. M, 0 h, 3 h, and E above the lanes indicate molecular markers, before induction, 3 h after induction, and elution fraction from Ni-NTA columns. The gel of a single experiment is shown. (C) *Trans*-splicing of split inteins derived from the *NpuDnaB* intein (left), Cl-*NpuDnaB* intein (middle), and the designed Oth-*NpuDnaB* intein (right) by splitting at the canonical split site. The results of one to three individual experiments for each combination are shown. (D) Test for orthogonality between split inteins derived from the *NpuDnaB* intein. SDS/PAGE analysis of *trans*-splicing for the pairs of Cl-*NpuDnaB*_N/*NpuDnaB*_C (left), Cl-*NpuDnaB*_N/Oth-*NpuDnaB*_C (middle), and *NpuDnaB*_N/Oth-*NpuDnaB*_C (right) indicated by arrows with solid and broken lines in panel A. The results of one or two individual experiments per combination are shown. For panels C and D, N and C with arrows indicate the bands for the N- and C-terminal precursors, respectively. L indicates the ligated product by *trans*-splicing. M, 0 h, I, A, I+A, and E stand for molecular markers, before induction, IPTG induction, arabinose induction, both IPTG and arabinose induction, and elution fraction from Ni-NTA spin columns, respectively.



Charge-Introduced *NpuDnaB* mini-intein (CI-*NpuDnaB*) was still able to splice efficiently in *cis* (Fig. 5B, 98% splicing efficiency). As the introduced charged residues did not impair *cis*-splicing, we derived a split

intein pair from the CI-*NpuDnaB* mini-intein (CI-*NpuDnaB_N*/CI-*NpuDnaB_C*) by splitting at the conserved insertion site of the homing endonuclease domain [15,25]. *Trans*-splicing of the split CI-

NpuDnaB mini-intein became less efficient than that of the split intein derived from the *NpuDnaB* mini-intein, as increased amounts of unreacted precursors appeared (Fig. 5C, yield 25% vs. 77%). This observation suggests that the charge interactions in the $\beta 3$ and $\beta 6$ could play a critical role in the association of the two split intein fragments derived from the *NpuDnaB* mini-intein. To confirm this hypothesis, we introduced unfavorable interactions by mutating Lys58 to Glu in $\beta 3$ and Glu116 to Lys in $\beta 6$. This orthogonal design of *NpuDnaB* mini-intein (Oth-*NpuDnaB*) was still able to efficiently splice in *cis* as no precursor had been left due to spontaneous splicing when expressed in *Escherichia coli* (Fig. 5B, 97% splicing efficiency). The efficient *cis*-splicing of Oth-*NpuDnaB* intein verifies that the introduced mutations were not detrimental to the protein-splicing reaction, which is an important prerequisite for engineering split inteins. We thus split Oth-*NpuDnaB* intein into a pair of two fragments of Oth-*NpuDnaB_N* and Oth-*NpuDnaB_C* at the canonical split site and tested *trans*-splicing activity (Fig. 5C). Unlike the covalently connected *cis*-splicing Oth-*NpuDnaB* intein, *trans*-splicing between Oth-*NpuDnaB_N* and Oth-*NpuDnaB_C* derived from Oth-*NpuDnaB* intein was drastically impaired (Fig. 5C, 5% yield). While the combination of CI-*NpuDnaB_N* and Oth-*NpuDnaB_C* did not yield any *trans*-spliced product (0% yield), the pair of CI-*NpuDnaB_N*/*NpuDnaB_C* could produce the ligated product (Fig. 5D, 100% yield (reference sample)). Oth-*NpuDnaB_C* could still react further with the wild-type *NpuDnaB_N* fragment (23% yield), confirming that it is functional (Fig. 5D). These observations confirm that *NpuDnaB_C* and Oth-*NpuDnaB_C* fragments have become orthogonal with CI-*NpuDnaB_N*. Engineering of the charge network in the region corresponding to $\beta 3$ and $\beta 6$ in the gp41-1 intein was indeed sufficient to create orthogonal split inteins from a *cis*-splicing intein, at least with the test case of the *NpuDnaB* mini-intein.

Discussion

Intein-based technologies gave rise to a broad range of widely applied methods in both biotechnology and basic research. Many of these methods utilize naturally occurring split inteins and their homologs because artificially split inteins derived from *cis*-splicing inteins are usually less efficient and/or require denaturing/refolding due to poor solubility [15,25,41,42,45]. The poor solubility of precursor fragments particularly limits their application *in vitro*. Previous attempts to derive artificially split inteins from several *cis*-splicing inteins have not been hugely successful, resulting in less productive *trans*-splicing [15,25]. The poor solubility of artificially split inteins

has recently been alleviated in part by using a salt-inducible split intein. The highly soluble split intein was derived from inteins from extreme halophilic archaea, but requires salt addition to induce the protein-splicing reaction [46]. Having two or more robust and orthogonal split inteins could widen the applications of PTS, for example, in multiple-fragment ligation for segmental isotopic labeling, bioorthogonal protein conjugations by semi-synthesis (e.g., antibody–drug conjugate), and protein-based logic gates [20,22,23,45,47]. Naturally split inteins, such as the cyanobacterial DnaE inteins, are not only rare but also often cross-active to each other; however, they exhibit a robust splicing activity and high tolerance of variations at the splice junctions [31,32,48]. The gp41-1 intein is another fragmented split intein exhibiting robust splicing activity [19] and could be used as an orthogonal intein with other split DnaE inteins. However, the splicing activity of the gp41-1 intein turned out to be rather sensitive to variations at the splice junctions (Fig. 2). We determined the crystal structure of the *cis*-splicing gp41-1 intein at 1.0 Å, which is hitherto the highest resolution available for intein structures. The structure sheds light on the features common between the two naturally occurring split inteins, providing the structural basis to guide the engineering of natural-like split inteins from prevalent *cis*-splicing inteins. The three-dimensional structures of both the gp41-1 and *NpuDnaE* inteins highlight an extended charge network on the strands corresponding to $\beta 3$ and $\beta 6$ in the gp41-1 intein. This network is absent in most *cis*-splicing inteins and could play an essential role in efficient *trans*-splicing. Charge-swapping between the corresponding $\beta 3$ and $\beta 6$ regions in the *NpuDnaE* intein did not affect *cis*- nor *trans*-splicing, suggesting that the charge network in the $\beta 3$ and $\beta 6$ regions alone cannot sufficiently account for orthogonality within the naturally split *NpuDnaE* intein fragments.

In contrast, we successfully demonstrated an orthogonal design by charge engineering of the split *NpuDnaB* mini-intein, derived from a *cis*-splicing intein, in the same $\beta 3$ and $\beta 6$ regions. *Trans*-splicing of the *NpuDnaB* mini-intein *in vivo* can be as efficient as the *NpuDnaE* intein and has a high tolerance of variations at the splicing junction [15,43,44]. Split inteins engineered from *NpuDnaB* mini-intein are new additions to the protein engineering toolbox using PTS and contribute to overcoming the junction sequence and extein dependencies that currently complicate PTS applications. More than 1500 inteins or intein-like domains have been identified from the sequence databases [49]. Not all *cis*-splicing inteins can be converted into active mini-inteins by deleting the inserted homing endonuclease regions due to a mutualism developed between HINT

and homing endonuclease domains [30,50]. However, a few hundred mini-inteins in the intein database that carry various junction sequences remain experimentally untested and unexplored. The common structural features found among naturally split inteins could be exploited to convert many other naturally occurring *cis*-splicing inteins into natural-like split inteins with robust *trans*-splicing activity. This process would yield new orthogonal split inteins with desired features such as optimal junction sequences and a high tolerance of foreign extein sequences. The resulting engineered inteins expand the applicability of PTS in protein engineering, chemical biology, and synthetic biology, in particular when applications require scar-less protein ligation.

Materials and methods

Plasmid constructions

All plasmids used and designed in this study are listed and summarized in Table S1, including the oligonucleotide sequences used. The gp41-1 intein variant for crystallization was cloned in plasmid pBHRSF38 as a yeast small ubiquitin-like modifier domain (SUMO) fusion protein with an inactivating C1A substitution and a stop codon after the last residue of Asn125 for purification [51]. *Cis*-splicing gp41-1 intein variants with loop or junction variations are encoded in plasmids pADHDuet21, pBHDuet37, pBHDuet321, pHBDuet021, pHBDuet087, pHBDuet088, and pHBDuet182. pHBDuet093 is a *cis*-splicing vector with the CS *NpuDnaE* intein. *Cis*-splicing vectors containing the CI and Oth *NpuDnaB* mini-intein variants are pHBDuet139 and pHBDuet140, respectively. A dual vector system using the pair of pHBDuet095 and pHBBAD106, derived from pHBDuet093, was used for testing *trans*-splicing of the CS-*NpuDnaE* intein, in which the N- and C-terminal fragments can be induced with isopropyl β -D-1-thiogalactopyranoside (IPTG) and arabinose, respectively [52]. Plasmids pSKDuet01 (Addgene #12172) and pSKBAD2 (Addgene #15335) encoding the natural *NpuDnaE_N* and *NpuDnaE_C* intein fragments were used to test the orthogonality of CS-*NpuDnaE_C* and CS-*NpuDnaE_N*, respectively [31]. Two previously described plasmids pSADuet259 (Addgene #121910) and pSABAD250 (Addgenem, Watertown, MA, USA #45612) encoding *NpuDnaB^{A283} _{Δ C39}* and *NpuDnaB_{C39}*, respectively, were used as a reference for *trans*-splicing of the *NpuDnaB* mini-intein [15]. A pair of two precursor fragments with CI-*NpuDnaB^{A290} _{Δ C39}* (pHBDuet148, Addgene #121911) and CI-*NpuDnaB_{C39}* (pHBBAD113, Addgene #121912) were derived from plasmid pHBDuet139 (Addgene #121913). Split intein fragments derived from Oth-*NpuDnaB*, that is, Oth-*NpuDnaB^{A290} _{Δ C39}* and Oth-*NpuDnaB_{C39}*, were encoded in pHBDuet116 (Addgene #121915) and pHBBAD168 (Addgene #121916), respectively, which were derived from pHBDuet140

(Addgene #121914). The plasmids with Addgene numbers are available from www.addgene.org (Addgene).

Protein production and purification

All recombinant proteins were produced in *E. coli* T7 Express (New England Biolabs, Ipswich, MA, USA). For small-scale expression and purification of amounts sufficient to analyze protein splicing in *cis* and *trans*, 5-mL LB medium cultures supplemented with 25 $\mu\text{g mL}^{-1}$ kanamycin, 100 $\mu\text{g mL}^{-1}$ ampicillin, or both were grown at 37 °C until an OD₆₀₀ of 0.6 was reached. Cultures to express a precursor protein containing a *cis*-splicing intein were then induced with a final concentration of 1 mM IPTG for 3 h. For coexpression of N- and C-terminal precursors for *trans*-splicing, 0.04% arabinose induction was followed by IPTG addition with a delay of 30 min. The coexpression lasted for a total time of 4 h at 30 °C. The cell cultures were harvested by centrifugation at 4700 *g* for 10 min at 4 °C and lysed using 400 μL of B-PER bacterial protein extraction reagent (Thermo Fisher Scientific, Waltham, MA, USA) according to the instructions of the manufacturer. Elution fractions from IMAC purification using Ni²⁺-nitrilotriacetic acid (NTA) spin columns (Qiagen, Hilden, Germany) were analyzed by 16.5% polyacrylamide SDS/PAGE gels stained with Coomassie Blue. *Cis*-splicing efficiency in percent was determined by gel band intensity quantification of the IMAC elution fractions from the depicted gels using IMAGE J 2.0.0 [53] and estimated from $100 \times [\text{Spliced product}] / ([\text{Precursor}] + [\text{C-cleavage products}])$. *Trans*-splicing was evaluated from the gel pictures by determination of the relative IMAC purification yield from each combination of split inteins, that is, the band intensity for the eluted *trans*-spliced product normalized to the highest yielding combination of split pairs (reference sample), as previously described [15]. In order to compare *trans*-splicing results among different gels, the intensities were normalized to the sum of three marker bands (14.4, 18.4, and 25 kDa). Inactive gp41-1 intein with C1A mutation utilized in structural studies was produced in 2 L of LB medium supplemented with 25 $\mu\text{g mL}^{-1}$ kanamycin by induction with a final concentration of 1 mM IPTG at an OD₆₀₀ of 0.6 for 3 h. The cells were harvested by centrifugation as described above and lysed in Buffer A (50 mM sodium phosphate pH 8.0, 300 mM NaCl) by continuous passing through an EmulsiFlex-C3 homogenizer (Avestin, Ottawa, ON, Canada) at 15 000 psi for 10 min, 4 °C. The cell lysate was cleared by centrifugation at 38 000 *g* for 60 min at 4 °C, and loaded on a 5 mL HisTrap column (GE Healthcare, Chicago, IL, USA) for purification. The protein was purified by following the previously described two-step protocol, including the removal of the hexahistidine tag and SUMO fusion domain [51]. The purified protein contained an additional

sequence 'SGG' as the N-terminal extein sequence of the gp41-1 intein. For crystallization, the protein was dialyzed against deionized water and concentrated to a final concentration of 45 mg·mL⁻¹ using an ultrafiltration device.

Crystallization, data collection, and structure solution

Diffraction crystals of the fusion protein comprising the N- and C-terminal gp41-1 intein fragments were obtained at room temperature by mixing 100 nL of concentrated protein with 100 nL of mother liquor (100 mM citric acid, pH 3.5, 100 mM magnesium sulfate, and 30% w/v polyethylene glycol 3350); 30% polyethylene glycol 3350 was sufficient as a cryoprotectant. Data were collected at beamline i02 at Diamond Light Source (Didcot, UK) equipped with a Pilatus 6MF detector. Data were processed using HKL3000 [54] at the nominal resolution of 1.02 Å (Table 1). The structure was solved by molecular replacement with Phaser [55] using the *NpuDnaE* intein (PDB ID: 4kl5) [26] as the starting model. The structure was rebuilt with Coot [56] and refined with REFMAC5 [57]. Although data completeness in the outermost shell (1.04–1.02 Å) was only 37%, the $\langle I/\sigma I \rangle$ ratio was quite significant at 2.8. Since the completeness in the 1.08–1.06 Å shell was 74% and $\langle I/\sigma I \rangle$ 3.6, we could safely claim the effective resolution of at least 1.06 Å. However, all data were used in refinement, with almost 2800 reflections present beyond this effective resolution limit.

The protein chain could be traced in the electron density without breaks for all 128 residues (125 intein residues and three residues of the amino acid sequence 'SGG' preceding the first intein residue). Alternate conformations were modeled for 23 protein residues, extending to the main chain for 18 of them. A noncanonical *cis* peptide bond was modeled between Lys87 and Glu88 based on unambiguous electron density for this part of the main chain (the electron density is also unambiguous for the side chain of Lys87, whereas the side chain of Glu88 appears to be partially disordered). Final validation was performed with MolProbity [58], showing an acceptable quality of the model (score 1.74, 40th percentile). The coordinates and structure factors were deposited in the Protein Data Bank with the accession code 6qaz.

Structures were drawn using PyMOL (The PyMOL Molecular Graphics System, Version 2.2.0, Schrödinger, LLC, New York, NY, USA). Sequence alignments were generated using Clustal Omega 1.2.4 (Conway Institute, University College Dublin, Ireland).

Acknowledgements

We thank B. Haas, S. Jääskeläinen, and AD. Hietikko for their technical help in the preparation of proteins and plasmids. We thank K. Kogan for his assistance at the

crystallization facility. This work was supported in part by the Academy of Finland (137995, 277335), Novo Nordisk Foundation (NNF17OC0025402 to HMB., NNF17OC0027550 to HI), Sigrid Jusélius Foundation, and by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research, as well as with Federal funds from the National Cancer Institute, NIH, under Contract No. HHSN261200800001E (to ML). The crystallization and NMR facilities have been supported by Biocenter Finland and HiLIFE-INFRA. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views or policies of the US Department of Health and Human Services, nor does the mention of trade names, commercial products, or organizations imply endorsement by the US Government.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

HI designed and supervised the project; HMB, KMM, and HI performed the experiments and analyzed the data; ML and AW participated in the crystallographic studies. All authors contributed to writing the manuscript.

References

- 1 Hirata R, Ohsumk Y, Nakano A, Kawasaki H, Suzuki K & Anraku Y (1990) Molecular structure of a gene, VMA1, encoding the catalytic subunit of H(+)-translocating adenosine triphosphatase from vacuolar membranes of *Saccharomyces cerevisiae*. *J Biol Chem* **265**, 6726–6733.
- 2 Kane PM, Yamashiro CT, Wolczyk DF, Neff N, Goebel M & Stevens TH (1990) Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase. *Science* **250**, 651–657.
- 3 Paulus H (2000) Protein splicing and related forms of protein autoprocessing. *Annu Rev Biochem* **69**, 447–496.
- 4 Perler FB, Davis EO, Dean GE, Gimble FS, Jack WE, Neff N, Noren CJ, Thorner J & Belfort M (1994) Protein splicing elements: inteins and exteins—a definition of terms and recommended nomenclature. *Nucleic Acids Res* **22**, 1125–1127.
- 5 Belfort M (2017) Mobile self-splicing introns and inteins as environmental sensors. *Curr Opin Microbiol* **38**, 51–58.
- 6 Cooper AA, Chen YJ, Lindorfer MA & Stevens TH (1993) Protein splicing of the yeast TFP1 intervening

- protein sequence: a model for self-excision. *EMBO J* **12**, 2575–2583.
- 7 Topilina NI & Mills KV (2014) Recent advances in *in vivo* applications of intein-mediated protein splicing. *Mob DNA* **5**, 5.
 - 8 Sarmiento C & Camarero JA (2019) Biotechnological applications of protein splicing. *Curr Protein Pept Sci* **20**, 408–424.
 - 9 Volkmann G & Iwai H (2010) Protein trans-splicing and its use in structural biology: opportunities and limitations. *Mol BioSyst* **6**, 2110.
 - 10 Mills KV, Johnson MA & Perler FB (2014) Protein splicing: how inteins escape from precursor proteins. *J Biol Chem* **289**, 14498–14505.
 - 11 Buskirk AR, Ong Y-C, Gartner ZJ & Liu DR (2004) Directed evolution of ligand dependence: small-molecule-activated protein splicing. *Proc Natl Acad Sci USA* **101**, 10505–10510.
 - 12 Peck SH, Chen I & Liu DR (2011) Directed evolution of a small-molecule-triggered intein with improved splicing properties in mammalian cells. *Chem Biol* **18**, 619–630.
 - 13 Thiel IV, Volkmann G, Pietrokovski S & Mootz HD (2014) An atypical naturally split intein engineered for highly efficient protein labeling. *Angew Chem Int Ed Engl* **53**, 1306–1310.
 - 14 Stevens AJ, Sekar G, Shah NH, Mostafavi AZ, Cowburn D & Muir TW (2017) A promiscuous split intein with expanded protein engineering applications. *Proc Natl Acad Sci USA* **114**, 8538–8543.
 - 15 Aranko AS, Oemig JS, Zhou D, Kajander T, Wlodawer A & Iwai H (2014) Structure-based engineering and comparison of novel split inteins for protein ligation. *Mol BioSyst* **10**, 1023–1034.
 - 16 Oemig JS, Zhou D, Kajander T, Wlodawer A & Iwai 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.
 - 17 Wu H, Hu Z & Liu XQ (1998) Protein trans-splicing by a split intein encoded in a split DnaE gene of *Synechocystis* sp. PCC6803. *Proc Natl Acad Sci USA* **95**, 9226–9231.
 - 18 Dassa B, London N, Stoddard BL, Schueler-Furman O & Pietrokovski S (2009) Fractured genes: a novel genomic arrangement involving new split inteins and a new homing endonuclease family. *Nucleic Acids Res* **37**, 2560–2573.
 - 19 Carvajal-Vallejos P, Pallissé R, Mootz HD & Schmidt SR (2012) Unprecedented rates and efficiencies revealed for new natural split inteins from metagenomic sources. *J Biol Chem* **287**, 28686–28696.
 - 20 Otomo T, Ito N, Kyogoku Y & Yamazaki T (1999) NMR observation of selected segments in a larger protein: central-segment isotope labeling through intein-mediated ligation. *Biochemistry* **38**, 16040–16044.
 - 21 Shi J & Muir TW (2005) Development of a tandem protein trans-splicing system based on native and engineered split inteins. *J Am Chem Soc* **127**, 6198–6206.
 - 22 Busche AEL, Aranko AS, Talebzadeh-Farooji M, Bernhard F, Dötsch V & Iwai H (2009) Segmental isotopic labeling of a central domain in a multidomain protein by protein trans-splicing using only one robust DnaE intein. *Angew Chem Int Ed Engl* **48**, 6128–6131.
 - 23 Shah NH, Vila-Perelló M & Muir TW (2011) Kinetic control of one-pot trans-splicing reactions by using a wild-type and designed split intein. *Angew Chem Int Ed Engl* **50**, 6511–6515.
 - 24 Sun W, Yang J & Liu X-Q (2004) Synthetic two-piece and three-piece split inteins for protein trans-splicing. *J Biol Chem* **279**, 35281–35286.
 - 25 Aranko AS, Wlodawer A & Iwai H (2014) Nature's recipe for splitting inteins. *Protein Eng Des Sel* **27**, 263–271.
 - 26 Aranko AS, Oemig JS, Kajander T & Iwai H (2013) Intermolecular domain swapping induces intein-mediated protein alternative splicing. *Nat Chem Biol* **9**, 616–622.
 - 27 Mizutani R, Nogami S, Kawasaki M, Ohya Y, Anraku Y, Satow Y (2002) Protein-splicing reaction via a thiazolidine intermediate: crystal structure of the VMA1-derived endonuclease bearing the N and C-terminal propeptides. *J Mol Biol* **316**, 919–929.
 - 28 Holm L & Laakso LM (2016) Dali server update. *Nucleic Acids Res* **44**, W351–W355.
 - 29 Hall TM, Porter JA, Young KE, Koonin EV, Beachy PA & Leahy DJ (1997) Crystal structure of a Hedgehog autoprocessing domain: homology between Hedgehog and self-splicing proteins. *Cell* **91**, 85–97.
 - 30 Iwai H, Mikula KM, Oemig JS, Zhou D, Li M & Wlodawer A (2017) Structural basis for the persistence of homing Endonucleases in transcription factor IIB inteins. *J Mol Biol* **429**, 3942–3956.
 - 31 Iwai H, Züger S, Jin J & Tam P-H (2006) Highly efficient protein trans-splicing by a naturally split DnaE intein from *Nostoc punctiforme*. *FEBS Lett* **580**, 1853–1858.
 - 32 Dassa B, Amitai G, Caspi J, Schueler-Furman O & Pietrokovski S (2007) Trans protein splicing of cyanobacterial split inteins in endogenous and exogenous combinations. *Biochemistry* **46**, 322–330.
 - 33 Shah NH, Eryilmaz E, Cowburn D & Muir TW (2013) Naturally split inteins assemble through a “capture and collapse” mechanism. *J Am Chem Soc* **135**, 18673–18681.
 - 34 Aranko AS, Züger S, Buchinger E & Iwai H (2009) *In vivo* and *in vitro* protein ligation by naturally occurring and engineered split DnaE inteins. *PLoS ONE* **4**, e5185.
 - 35 Walther TH, Gottselig C, Grage SL, Wolf M, Vargiu AV, Klein MJ, Vollmer S, Prock S, Hartmann M, Afonin S *et al.* (2013) Folding and self-assembly of the TatA translocation pore based on a charge zipper mechanism. *Cell* **152**, 316–326.

- 36 Li Y, Aboye T, Breindel L, Shekhtman A & Camarero JA (2016) Efficient recombinant expression of SFTI-1 in bacterial cells using intein-mediated protein trans-splicing. *Biopolymers* **106**, 818–824.
- 37 Jagadish K, Borra R, Lacey V, Majumder S, Shekhtman A, Wang L & Camarero JA (2013) Expression of fluorescent cyclotides using protein trans-splicing for easy monitoring of cyclotide-protein interactions. *Angew Chem Int Ed Engl* **52**, 3126–3131.
- 38 Bi T, Li Y, Shekhtman A & Camarero JA (2018) In-cell production of a genetically-encoded library based on the θ -defensin RTD-1 using a bacterial expression system. *Bioorg Med Chem* **26**, 1212–1219.
- 39 Muona M, Aranko AS & Iwai H (2008) Segmental isotopic labelling of a multidomain protein by protein ligation by protein trans-splicing. *ChemBioChem* **9**, 2958–2961.
- 40 Oeemig JS, Aranko AS, Djupsjöbacka J, Heinämäki K & Iwai 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.
- 41 Southworth MW, Adam E, Panne D, Byer R, Kautz R & Perler FB (1998) Control of protein splicing by intein fragment reassembly. *EMBO J* **17**, 918–926.
- 42 Otomo T, Teruya K, Uegaki K, Yamazaki T & Kyogoku Y (1999) Improved segmental isotope labeling of proteins and application to a larger protein. *J Biomol NMR* **14**, 105–114.
- 43 Iwai H & Aranko AS (2017) Protein ligation by HINT domains. *Chemical Ligation* (D'Andrea LD & Romanelli A, eds), pp. 421–445. John Wiley & Sons Inc, Hoboken, NJ, USA.
- 44 Ellilä S, Jurvansuu JM & Iwai H (2011) Evaluation and comparison of protein splicing by exogenous inteins with foreign exteins in *Escherichia coli*. *FEBS Lett* **585**, 3471–3477.
- 45 Lohmueller JJ, Armel TZ & Silver PA (2012) A tunable zinc finger-based framework for Boolean logic computation in mammalian cells. *Nucleic Acids Res* **40**, 5180–5187.
- 46 Ciragan A, Aranko AS, Tascon I & Iwai H (2016) Salt-inducible Protein Splicing in cis and trans by Inteins from extremely Halophilic Archaea as a novel protein-engineering tool. *J Mol Biol* **428**, 4573–4588.
- 47 Möhlmann S *et al.* (2011) Site-specific modification of ED-B-targeting antibody using intein-fusion technology. *BMC Biotechnol* **11**, 76.
- 48 Cheriyan M, Peadarallu CS, Tori K & Perler F (2013) Faster protein splicing with the *Nostoc punctiforme* DnaE intein using non-native extein residues. *J Biol Chem* **288**, 6202–6211.
- 49 Novikova O, Jayachandran P, Kelley DS, Morton Z, Merwin S, Topilina NI & Belfort M (2016) Intein clustering suggests functional importance in different domains of life. *Mol Biol Evol* **33**, 783–799.
- 50 Hiraga K, Derbyshire V, Dansereau JT, Van Roey P & Belfort M (2005) Minimization and stabilization of the *Mycobacterium tuberculosis* recA intein. *J Mol Biol* **354**, 916–926.
- 51 Guerrero F, Ciragan A & Iwai H (2015) Tandem SUMO fusion vectors for improving soluble protein expression and purification. *Protein Expr Purif* **116**, 42–49.
- 52 Züger S & Iwai H (2005) Intein-based biosynthetic incorporation of unlabeled protein tags into isotopically labeled proteins for NMR studies. *Nat Biotechnol* **23**, 736–740.
- 53 Rueden CT, Schindelin J, Hiner MC, DeZonia BE, Walter AE, Arena ET & Eliceiri KW (2017) ImageJ2: imageJ for the next generation of scientific image data. *BMC Bioinformatics* **18**, 529.
- 54 Minor W, Cymborowski M, Otwinowski Z & Chruszcz M (2006) HKL-3000: the integration of data reduction and structure solution—from diffraction images to an initial model in minutes. *Acta Crystallogr D Biol Crystallogr* **62**, 859–866.
- 55 McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC & Read RJ (2007) Phaser crystallographic software. *J Appl Crystallogr* **40**, 658–674.
- 56 Emsley P, Lohkamp B, Scott WG & Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486–501.
- 57 Murshudov GN, Vagin AA & Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* **53**, 240–255.
- 58 Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS & Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12–21.
- 59 Weiss MS (2001) Global indicators of X-ray data quality. *J Appl Crystallogr* **34**, 130–135.
- 60 Brünger AT (1992) Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–475.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. Plasmids and oligonucleotides used in this study.