FEBS Letters

FEBS PRESS science publishing by scientists

# Substrate specificities of inteins investigated by QuickDrop-cassette mutagenesis

Jesper S. Oeemig, Hannes M. Beyer*, A. Sesilja Aranko§, Justus Mutanen and Hideo Iwaï

Institute of Biotechnology, University of Helsinki, Helsinki, Finland

**Correspondence**
H. Iwaï, Institute of Biotechnology, University of Helsinki, P.O. Box 65, Helsinki FIN-00014, Finland
Tel: +358 2941 59752
E-mail: hideo.iwai@helsinki.fi

**Present address**
* Institute of Synthetic Biology and CEPLAS, University of Düsseldorf, Düsseldorf, Germany
§ Department of Bioproducts and Biosystems, School of Chemical Engineering, Aalto University, Espoo, Finland

(Received 12 July 2020, revised 5 August 2020, accepted 10 August 2020, available online 26 September 2020)

doi:10.1002/1873-3468.13909

Edited by Miguel De la Rosa

**Inteins catalyze self-excision from host precursor proteins while concomitantly ligating the flanking substrates (exteins) with a peptide bond. Noncatalytic extein residues near the splice junctions, such as the residues at the −1 and +2 positions, often strongly influence the protein-splicing efficiency. The substrate specificities of inteins have not been studied for many inteins. We developed a convenient mutagenesis platform termed "QuickDrop"-cassette mutagenesis for investigating the influences of 20 amino acid types at the −1 and +2 positions of different inteins. We elucidated 17 different profiles of the 20 amino acid dependencies across different inteins. The substrate specificities will accelerate our understanding of the structure–function relationship at the splicing junctions for broader applications of inteins in biotechnology and molecular biosciences.**

**Keywords:** intein; mutagenesis; protein ligation; protein splicing; substrate specificity

Enzymes are protein catalysts that accurately recognize the substrates and convert the chemical structure with high accuracy. Intervening protein sequences termed inteins are unique single-turnover enzymes that catalyze their self-excision and concomitant peptide ligation of flanking protein sequences termed exteins (Fig. 1) [1]. This process is called protein splicing. Inteins catalyzing protein splicing have their substrates covalently connected with peptide bonds at the flanking N and C termini. Since the discovery of the protein-splicing phenomenon, the concerted chemical reaction catalyzed by inteins has stimulated broad biotechnological applications such as protein ligation tools by split inteins in biotechnology, biosensors, protein purification, synthetic biology, chemical biology, and protein engineering [2–7]. However, the use of inteins has been relatively limited, despite the original high expectations [8–11]. Inteins seemingly adapted to the specific insertion sites within their host proteins, where they are inserted [12]. Thus, protein-splicing efficiency, a critical consideration when using inteins, is often affected by the splicing junction sequences as well as extein sequences in the foreign contexts [13,14]. In some cases, mutations of these junction positions may considerably reduce or even entirely abolish the splicing activity, resulting in side reactions of cleavages or inactivity. The junction dependence of inteins has been limiting the potential of inteins when used with artificial exteins [14,15]. The situation is even more complicated when split inteins are used for protein

**Abbreviations**
AA, amino acid; GB1, the B1 domain of IgG-binding protein G; HEN, homing endonuclease; IMAC, immobilized metal affinity chromatography; PTS, protein *trans*-splicing; TCA, trichloroacetic acid; TCEP, tris(2-carboxyethyl)phosphine.



FEBS Letters **594** (2020) 3338–3355 © 2020 The Authors. *FEBS Letters* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies

*trans*-splicing (PTS) to ligate separate foreign polypeptides (Fig. 1B). This complication arises due to the limited solubility of precursors, extein dependency, and association of the two split intein fragments [16]. Consequently, engineering of inteins is not always straightforward for various biotechnological applications in non-native contexts, thereby restricting more extensive uses of various inteins.

In addition to the protein-splicing catalysis, many canonical inteins are bifunctional enzymes containing homing endonuclease (HEN) domains that cleave the DNA sequence near the intein insertion site and are involved in horizontal gene transfer of intein genes [1,17]. While some inteins are proficient of protein splicing and can splice without the HEN, other inteins have developed a mutualism between the protein-splicing domain and the nested endonuclease domain. The requirement of the HEN domain for protein splicing intricates the engineering of inteins and split inteins [18].

The number of identified inteins and intein-related proteins is continuously growing as genomic and metagenomic sequence data are increasingly becoming available at a rapid rate. However, the splicing

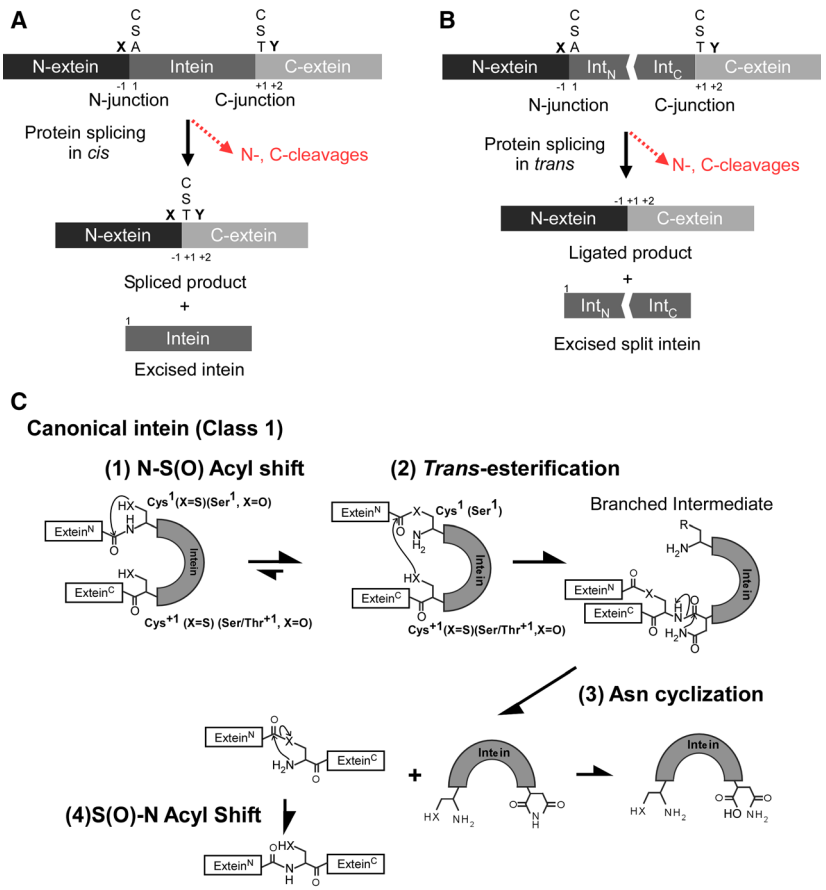



**Fig. 1.** Protein splicing and canonical protein-splicing reaction steps. (A) Protein splicing in *cis*. The intein is flanked by N- and C-terminal exteins. The intein is autocatalytically excised during protein *cis*-splicing, thereby covalently ligating the flanking exteins. The −1, +1, and +2 positions of exteins are indicated together with the first residue of inteins (indicated by '1'). Positions subjected to QuickDrop mutagenesis are depicted in bold and indicated with an **X** for the −1 position and **Y** for the +2 position. The +1 residue is shown with C, S, and T for Cys, Ser, and Thr, respectively. A red arrow indicates nonproductive N and C cleavages (A, B). (B) PTS by a split intein. Two precursor fragments split within the intein ligate two exteins sequences with a peptide bond by PTS. Mutations at the splicing junction (e.g., **X** and **Y**) might increase the by-products caused by N and/or C cleavage (red arrows). (C) Generally accepted protein-splicing steps of canonical class 1 inteins: (1) N-S(O) acyl shift by the first intein residue (Cys or Ser), (2) *trans*-esterification from Cys, Ser, or Thr at the +1 position of the first C-extein residue, (3) Asn cyclization by the last inteins residue, releasing the branched intermediate consisting of an N-extein and C-extein ester, (4) S(O)-N acyl shift forming an energetically favorable peptide bond.

activity, the splicing junction dependence, and the extein dependence of the majority of identified inteins are mostly unknown, hindering various potential biotechnological applications of many naturally occurring inteins [12,14,15,19,20]. Due to a limited understanding of the substrate specificities (junction sequence dependencies), inteins have often been used with the assumption that the native N- and C-extein sequences were the best sequences at the junctions for optimal protein-splicing activity [12]. Therefore, one to five natural flanking extein sequences were often kept at both junctions [12,19,21,22]. These flanking sequences remain in the primary structure as a 'scar' of flanking extein residues after protein splicing, which could potentially lower the usefulness of protein splicing in many applications (Fig. 1B).

To overcome the bottleneck imposed by the junction problem of inteins, several groups attempted to engineer more promiscuous inteins by directed evolution or to identify more robust and promiscuous inteins from natural sources [14,21–26]. However, the substrate specificities of various inteins, namely the junction sequence dependencies, have not been thoroughly investigated except for only a very few inteins [8,14,20]. Thus, limited information about the substrate specificity for inteins is currently available, presumably because the analysis of a complete set of 20 amino acids (AA) variants at the so-called −1 and/or +2 position even for a limited set of inteins can be time-consuming and tedious (Fig. 1A). Such quantitative analysis of the substrate specificity at both the −1 and +2 positions of one intein would require a complete set of 20 × 20 (400) vectors to cover all combinations for the two junction positions for each intein. Thus, a limited quantitative analysis of the substrate specificity exists only for a few inteins [8,14,20]. As a consequence, for each new intein or target sequence one must empirically test whether the intein could tolerate artificial exteins with the desired junction sequences.

In this work, we elucidated the substrate specificities of different inteins using the newly developed QuickDrop mutagenesis. The QuickDrop mutagenesis is a cost- and labor-effective mutagenesis platform to introduce 20 amino acid (AA) types at one residue in the sequence by taking advantage of the conserved sequence among the family members for the comparison. We used this approach to introduce 20 different AA-types at the −1 and +2 positions of class 1 inteins by making use of the conserved first residue of class 1 inteins and also the highly conserved first residue of the C-terminal extein. As little is hitherto known for the structural basis of the substrate specificity of individual inteins, we elucidated the AA-type dependency of different inteins at the splicing junctions. We demonstrated that the junction sequence specificity of many different inteins could not only potentially guide the application of different inteins, but also assist in understanding the structure–function relationship of different inteins.

## Materials and methods

### Construction of the QuickDrop −1 libraries

First, to create a set of standard vector libraries with 20 AA-variants at the −1 position, we used the *Npu*DnaE intein. A *Bse*RI restriction site was introduced at the N-terminal splicing junction of the plasmid pSKDuet16 containing the *cis*-splicing *Npu*DnaE intein gene with two B1 domains of IgG-binding protein G (GB1) as N- and C-exteins [21]. The *Bse*RI site was introduced by amplifying the N-terminal GB1 by PCR with the oligonucleotides HK683 and HK894 (Table S1). The PCR product bearing the *Bse*RI site at the 3′-terminus was digested with *Nco*I and *Bam*HI and ligated into predigested pSKDuet16 to replace the N-terminal His-tagged GB1, resulting in pLLSDuet1 with Ser at the −1 position. The other 19 plasmids were constructed following the QuickChange™ (Stratagene, San Diego, CA, USA) protocol using pairs of oligonucleotides listed in Table S2 [27]. These plasmids bearing 20 AA-types at the −1 position and a *Bse*RI site upstream of the N-splicing junction constitute the QuickDrop −1 library with Cys1.

To cover all class 1 inteins, we constructed another set of the QuickDrop −1 library with Ser at the 1 position for inteins containing Ser as the first intein residue, by mutating the first residue of *Npu*DnaE intein in pLLSDuet1 to Ser. We introduced the Cys1 to Ser1 mutation and simultaneously a *Nde*I site within the *Npu*DnaE intein gene by PCR using the oligonucleotides I212 and I213, resulting in pJMDuet20. The remaining 19 mutations were introduced by amplifying the N-terminal GB1 gene from pJMDuet20 using the oligonucleotide Duet-MCS1-fw (Table S1) and a corresponding oligonucleotide as listed in Table S2. These PCR products were cloned into pJMDuet20 using *Nco*I and *Nde*I, resulting in the QuickDrop −1 library with Ser1.

### Construction of the QuickDrop +2 libraries

A set of 20 AA-variants bearing 20 different AA-types at +2 the position of inteins was first derived from a plasmid encoding the *Pho*RadA intein with Thr at the +1 position (Fig. 1A). The *Pho*RadA intein gene was amplified from pCADuet99 using the oligonucleotides I101 and SZ015 and inserted into pSKDuet16 using *Bam*HI and *Kpn*I, resulting in pSCFDuet64 bearing *Kpn*I and *Bse*RI sites downstream

of the +2 position (Fig. 2C) [20,21]. The entire set of 20 AA-variants at the +2 position was constructed by PCR amplification of the *Pho*RadA intein gene from pCADuet99 using the oligonucleotides HK375 and an oligonucleotide encoding one of the 20 AA-types substituting the +2 position as listed in Table S3. The 20 PCR products were digested with *Kpn*I and *Bam*HI and subsequently cloned into pSCFDuet64, resulting in the QuickDrop +2 library with Thr/Ser+1. This QuickDrop +2 library is compatible with both Ser and Thr at the +1 position (Fig. 4).

To completely cover Cys, Ser, and Thr at the +1 position of inteins, we created another set of 20 AA-variants at the +2 position of the *Npu*DnaE intein. The *Npu*DnaE intein gene with an AA mutation at the +2 position and Cys at the +1 position was amplified from pSKDuet16 using the oligonucleotide SK092 and an oligonucleotide encoding one of 20 AA-types at the +2 position as listed in Table S3. The PCR product was cloned into pSCFDuet64 using *Bam*HI and *Kpn*I, resulting in the QuickDrop +2 library with Cys+1, which is also compatible with Ser+1. Therefore, we call it the QuickDrop +2 library with Cys/Ser+1.

## Cloning of different inteins using the QuickDrop libraries

During constructing the 20 AA-type variants at the −1 and +2 positions of *Npu*DnaE and *Pho*RadA inteins, in total, we developed four sets of the QuickDrop libraries, namely, QuickDrop −1 library with Ser1, QuickDrop −1 library with Cys1, QuickDrop +2 library with Thr/Ser+1, and QuickDrop +2 library with Cys/Ser+1. These libraries should suffice to cover all class 1 inteins for the 20-AA mutagenesis by the

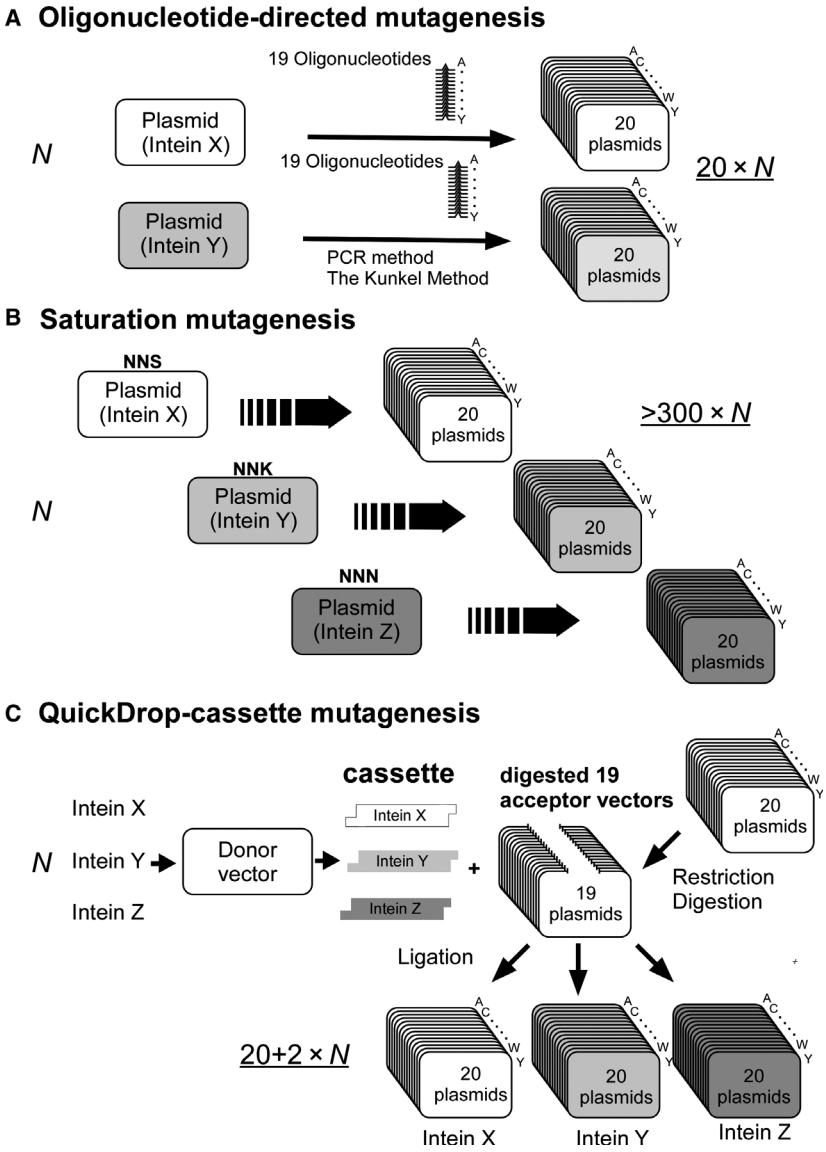



**Fig. 2.** Comparison between oligonucleotide-directed mutagenesis, saturation mutagenesis, and QuickDrop-cassette mutagenesis. (A) Oligonucleotide-directed mutagenesis for several inteins using additional 19 oligonucleotides for each intein to cover the 20 AA-types at one position. (B) Saturation mutagenesis using a degenerated oligonucleotide with 'NNS', 'NNK', or 'NNN' for each intein. (C) Cassette mutagenesis termed QuickDrop mutagenesis for many inteins to cover the 20 AA-types at the −1 or +2 position.

cassette-mutagenesis approach. We applied the QuickDrop strategy for the analysis of the substrate specificities of the *Npu*DnaB$^{\Delta 290}$, *Pho*RadA, *Pho*RadA(E71T), gp41-1, *Hwa*rPolA″, and *Hut*MCM2 inteins.

### *Pho*RadA intein

The RadA intein from *Pyrococcus furiosus* (*Pho*RadA) was amplified from pHYDuet183 by PCR with the oligonucleotides HK979 and HK376 and inserted into pLLSDuet1, resulting in pSCFDuet2 [20]. Two internal *Bse*RI sites within the *Pho*RadA intein gene were deleted by two cycles of inverse PCR mutagenesis using oligonucleotides I38, I39, I40, and I41, resulting in pCADuet99 as donor vector. The *Pho*RadA intein gene was excised by digestion of 1.5–2.0 μg pCADuet99 with 10 U of *Hind*III (Fermentas, Waltham, MA, USA) at 37 °C for 3 h followed by the addition of *Bse*RI and incubation for additional 60 min at 37 °C in 20 μL (NEB, Ipswich, MA, USA) buffer 2. The gene fragment was subsequently purified on a 1.2% agarose gel and extracted using the GeneJET™ Gel Extraction Kit (Fermentas). The purified gene fragment was ligated into the remaining 19 plasmids of the QuickDrop −1 library with Cys1. The 19 acceptor vectors were predigested with *Bse*RI/*Hind*III and treated with alkaline phosphatase (FastAP; Fermentas) before separation on a 1.2% agarose gel and gel extraction. 30–50 ng of the digested plasmid was ligated with 10–15 ng of the excised intein gene in the total volume of 5 μL in the presence of 2.5 U of T4 DNA ligase in T4 DNA ligation buffer (Fermentas). The reaction mixture was incubated at 20 °C for 1 h before transformation of 2 μL of the ligation reaction into *Escherichia coli* DH5α. 1–2 colonies were typically selected for the plasmid isolation and screened for positive clones using *Bam*HI prior to DNA sequencing. For *Pho*RadA(E71T) at the −1 position, the *Pho*RadA intein carrying an E71T mutation was created by inverse PCR using the plasmid encoding Asp at the −1 position (pSCFDuet22), resulting in pJODuet27 [20]. The remaining plasmids encoding the additional 19 AA-variants at the −1 position were created by digesting pJODuet27 with *Bse*RI and *Hind*III and ligating the excised gene cassette into the 19 predigested plasmids as prepared above.

### *Npu*DnaB$^{\Delta 290}$ intein

The 20 AA-variants at the −1 position of the DnaB mini-intein from *Nostoc punctiforme* (*Npu*DnaB$^{\Delta 290}$) were created in the same way as for the *Pho*RadA intein. The intein was amplified from the template plasmid pALBDuet28 using the oligonucleotides HK978 and HK212 [16]. The amplified gene was inserted into pLLSDuet1 between *Kpn*I and *Bam*HI, resulting in pLLSDuet4. The QuickDrop intein cassette was excised from pLLSDuet4 with *Bse*RI and *Hind*III and cloned into the predigested 19 plasmids of the QuickDrop −1 library with Cys1 using *Bse*RI and *Hind*III. For the +2 position of the *Npu*DnaB$^{\Delta 290}$ intein, the gene was PCR-amplified from pALBDuet28 with the oligonucleotides HK151 and HK212. We cloned the PCR product into pSCFDuet80 using *Kpn*I and *Bam*HI, resulting in pJMDuet21 with Ile at the +2 position [16]. The resulting donor plasmid (pJMDuet21) was digested with *Nco*I and *Bse*RI to make the QuickDrop cassette of the *Npu*DnaB$^{\Delta 290}$ intein and cloned into the remaining predigested 19 plasmids of the QuickDrop +2 library with Ser+1. Two variants of Met and His at the +2 position were cloned using *Bam*HI and *Bse*RI sites instead of using the *Nco*I site due to the presence of an additional *Nco*I site.

### gp41-1 intein

The gene of the gp41-1 intein was amplified from pBHDuet37 by PCR with two oligonucleotides I785 and I786 [28]. The gp41-1 gene was digested with *Bam*HI and *Kpn*I and cloned into pLLSDuet1, resulting in the donor plasmid pLKDuet7. The QuickDrop cassette was purified after digesting with *Bse*RI and *Hind*III and ligated into the predigested QuickDrop −1 library with Cys1. The amplified gene was also cloned into pLVDuet9, resulting in the donor plasmid pLKDuet9. The QuickDrop cassette was prepared by digestion with *Bse*RI and *Bam*HI and ligated into the QuickDrop +2 library with Ser+1.

### *Hut*MCM2 intein

For the −1 position, the gene was amplified from pSADuet616 with two primers SZ015 and I142 [29]. The PCR product was ligated into pLLSDuet1 with *Bam*HI and *Kpn*I, resulting in the donor plasmid pSADuet646. The QuickDrop cassette was isolated from pSADuet646 by digestion with *Bse*RI and *Hind*III and ligated into the predigested QuickDrop −1 library with Cys1. For the +2 position, the gene was amplified from pSADuet616 by PCR with two oligonucleotides I119 and I360. The PCR product was ligated into pLVDuet19, resulting in the donor vector pSADuet794. The donor plasmid was digested with *Bam*HI and *Bse*RI and ligated into the QuickDrop +2 library with Ser+1.

### *Hwa*rPolA″ intein

The gene was amplified from pSADuet685 with two primers I527 and I179 and ligated into pLLSDuet1 with *Bam*HI and *Kpn*I, resulting in the donor plasmid pJODuet143 [29]. The QuickDrop cassette was isolated from pJODuet143 by digestion with *Bse*RI and *Hind*III and ligated into the predigested QuickDrop −1 library with Cys1 for analysis at the −1 position. For the +2 position

with the wild-type Thr at the +1 position, the gene was PCR-amplified with two primers I386 and I178 and ligated into pSCRDuet76, resulting in the donor vector pJMDuet106. The digested QuickDrop cassette created by digesting pJMDuet106 with *Bam*HI and *Bse*RI was ligated into QuickDrop +2 library with Thr+1. The gene was also amplified with I178 and I548 by PCR and ligated into pSCRDuet76, resulting in the donor vector pJODuet163 with Ser+1. The QuickDrop cassette from pJODuet163 digested with *Bam*HI and *Bse*RI was ligated into the QuickDrop +2 library with Ser+1. The list of all the plasmids used in this study is summarized in Tables S4–S6.

### Protein expression and purification for comparison of the 20 AA-variants

Protein-splicing efficiencies were assayed by transformation of each plasmid into *E. coli* T7 Express (NEB #C2566H) and protein expression in 5 mL LB medium supplemented with 25 $\mu g \cdot mL^{-1}$ kanamycin. The cells harboring the corresponding plasmid were grown at 37 °C and induced at $OD_{600}$ = 0.4–0.6 with 1 mM IPTG. After a 4-h induction, the cells were harvested by centrifugation at 4500 ***g*** at 4 °C for 10 min. The cell pellets were lysed using B-PER® Bacterial Protein Extraction Reagent (Thermo Scientific, Waltham, MA, USA). After centrifugation, the soluble fraction was purified with Ni-NTA spin columns (Qiagen, Hilden, NRW, Germany) or His MultiTrapTM HP 96-well plate (GE Healthcare, Chicago, IL, USA) according to the manufacturer's purification protocol with a Whatman® UniVac3 vacuum manifold collection device (Whatman, Little Chalfont, UK). The samples from the elution fractions were analyzed on home-made 18% SDS-PAGE gels or precast gels (10–20% Criterion TGX with 26 wells 1.0 mm #5671115; Bio-Rad, Hercules, CA, USA) using Coomassie blue staining. For the protein splicing of halophilic inteins, the elution fraction was incubated with a final concentration of 3 M NaCl, 0.5 mM TCEP at 37 °C overnight [29]. The reaction mixture was precipitated by trichloroacetic acid (TCA) to remove the salt before the SDS-PAGE analysis. IMAGEJ (NIH, Bethesda, MD, USA) was used for the quantification of bands [30]. The error bars are the standard deviation estimated from the intensity quantifications from at least three independent experiments analyzed on more than three SDS-PAGE gels.

## Results

### Strategy for constructing 20 AA-type variants at the −1 and +2 positions for many inteins

One of the widely used methods to introduce a mutation at one residue in a protein is oligonucleotide-directed mutagenesis, in which the synthetic oligonucleotides encode an amino acid mutation for the conventional PCR method or Kunkel method [27,31]. Oligonucleotide-directed mutagenesis requires the same number of oligonucleotides and DNA sequencing reactions as the required mutations. To cover the 20 amino acid (AA) types for $N$ inteins, one needs $20 \times N \times 2$ oligonucleotides for both the −1 and +2 positions of inteins. The procedure of performing site-directed mutagenesis by PCR is not challenging but rather cumbersome. It gets quickly too labor-intensive to evaluate many substrate specificities at the −1 and +2 positions for many different inteins and many variants of an intein. The other commonly used method for introducing 20 AA-types at one position is site-directed saturation mutagenesis using a degenerated oligonucleotide with the sequence 'NNS', 'NNK', or 'NNN' encoding the amino acid mutation [32]. The saturation mutagenesis method involves only two oligonucleotides for the −1 and +2 positions to construct a library of plasmids by PCR covering all 20 AA-types at one residue. However, saturation mutagenesis entails much higher screening efforts of the library to isolate 20 individual plasmids with all 20 AA-types. It was estimated that this screening might require more than 300 clones for DNA sequencing to cover all 20 AA-types with > 99% probability [33,34]. Relying on saturation mutagenesis, one might have to screen $> 300 \times N \times 2$ clones for $N$ inteins in the worst-case scenario. However, trinucleotide cassettes offer an alternative to oligonucleotide randomization and facilitate the 20-AA selection by avoiding stop codons and codon bias [35]. Thus, the saturation mutagenesis has significant advantages when the mutation is directly coupled to a phenotype for directed evolution or screening to identify active clones without isolating all 20 AA-type variants from the library for individual characterization [31]. Therefore, we thought of developing a more straightforward method for the investigation of the −1 and +2 positions of inteins. Class 1 inteins share semi-conserved residues (Cys or Ser at the first residue and Cys, Ser, or Thr at the +1 position) that can be used to develop a general mutagenesis strategy. We devised a novel cassette mutagenesis strategy, termed QuickDrop, by making use of the conserved positions and a type IIS restriction endonuclease (Fig. 2A) [36]. Canonical inteins catalyze protein splicing with the four concerted steps: (1) N-S(O) acyl shift; (2) *trans*-esterification step; (3) Asn cyclization; (4) S(O)-N acyl shift (Fig. 2B) [2]. The first step of the N-S(O) acyl shift is induced by the first residue of inteins that is typically a Cys or Ser residue at the N terminus of inteins.

Therefore, Cys or Ser of the first residue of inteins is a highly conserved position, particularly for class 1 inteins [36]. The *trans*-esterification step (2) requires a thiol or hydroxyl group in Cys, Ser, or Thr as a nucleophile. Thus, Cys, Ser, or Thr at the so-called +1 position is indispensable for protein splicing [2]. As the underlying core principle of the QuickDrop method, we use these conserved residues of inteins as a shared DNA cleavage site by a restriction enzyme among different inteins.

## QuickDrop plasmid libraries designed for the −1 position

Type IIS restriction enzymes comprise a group of restriction enzymes that recognize asymmetric DNA sequences and cleave at a defined distance outside of their recognition sequence [36]. These enzymes lately became highly popular in a DNA multifragment assembly method called 'Golden Gate Assembly', as type IIS restriction enzymes allow the simultaneous utilization with DNA ligases in one pot because the ligation product is protected from digestion [37]. We selected *Bse*RI, which recognizes the 6-base DNA sequence and cleaves 8-base pairs downstream of the recognition sequence with a 2-base 3′-overhang [38]. The 8-base pairs are sufficient to accommodate another 6-base pair recognition site of a commonly used restriction enzyme such as *Bam*HI and *Kpn*I or a codon encoding the 20 amino acid types. We located the DNA cleavage site after the −1 position of an intein so that the 2-base 3′-overhang is located within the highly conserved first residue of the intein (Fig. 3). Additionally, we utilized the 8-base pair space to introduce another restriction site (*Bam*HI), facilitating the insertion of an intein of choice for creating the −1-donor vector (Fig. 3, Step I). QuickDrop mutagenesis can be performed in a single step by simultaneously transferring the QuickDrop cassette from the donor vector into the QuickDrop library of predigested acceptor vectors using *Bse*RI and *Kpn*I (or *Hind*III) to obtain the remaining 19 plasmids encoding different AA at the −1 position. We thus obtained the entire set of 20AA-variants at the −1 position, flanked by two B1 domains of the IgG-binding protein G (GB1) as model exteins. We used the GB1s for flanking N- and C-extein because we previously used them for analyzing the protein-splicing efficiency by SDS-PAGE from samples produced in *E. coli* [14,21]. Moreover, we constructed another set of the 20 vectors for inteins starting with Ser instead of Cys at the first residue (Fig. 3).

## The QuickDrop plasmid library designed for the +2 position

Not only the −1 position preceding inteins but also the second residue following inteins (called the +2 position) has been shown to largely influence protein-splicing efficiency [14] (Fig. 1A). The amino acid dependency at the +2 position is also critical for protein splicing in foreign contexts. Thus, we constructed the QuickDrop +2 library in a similar way as the QuickDrop −1 library. Unlike the −1 position, we used the +1 residue of the C-extein, which is usually Ser, Cys, or Thr, and used the *Bse*RI recognition site on the complementary strand. A *Kpn*I site was introduced between the recognition and cleavage sites of *Bse*RI to facilitate the +2 mutations as well as creating a donor vector (Fig. 4).

## QuickDrop mutagenesis workflow

QuickDrop mutagenesis can conveniently generate vector libraries encoding all 20 possible amino acid variants at the −1 or +2 position around the splice junctions of inteins. Utilizing the QuickDrop platform (available from www.addgene.org/Hideo_Iwai) allows generating the target library in three simple and straightforward steps (Figs 3 and 4):

**I.** Cloning of the target intein into the −1-donor vector using PCR and *Bam*HI/*Hind*III (or *Kpn*I) restriction sites. This donor vector has Ser at the −1 position due to the *Bam*HI site.

**II.** Excision of the target intein (QuickDrop) cassette from the −1 donor vector by digestion with the two restriction enzymes *Bse*RI and *Hind*III (or *Kpn*I).

**III.** Parallel ligation of the QuickDrop cassette into the −1 QuickDrop library digested by *Bse*RI and *Hind*III (or *Kpn*I), resulting in the remaining 19 vectors bearing the 19 different amino acid types at the −1 position. The QuickDrop library can be stored as a predigested library, facilitating rapid construction for other intein libraries. Isolated clones can be subjected to screening by *Bam*HI digestion, followed by DNA sequencing.

Similarly, the target intein can be cloned into the +2-donor vector using *Bam*HI/*Kpn*I, resulting in the +2-donor vector. The donor vector could have any amino acid at the +2 position. The QuickDrop cassette is excised from the donor vector using *Bse*RI and *Nco*I (or *Bam*HI) followed by ligation into the predigested +2 QuickDrop library by using *Bse*RI and *Nco*I (or *Bam*HI), resulting in the remaining 19 vectors. All the

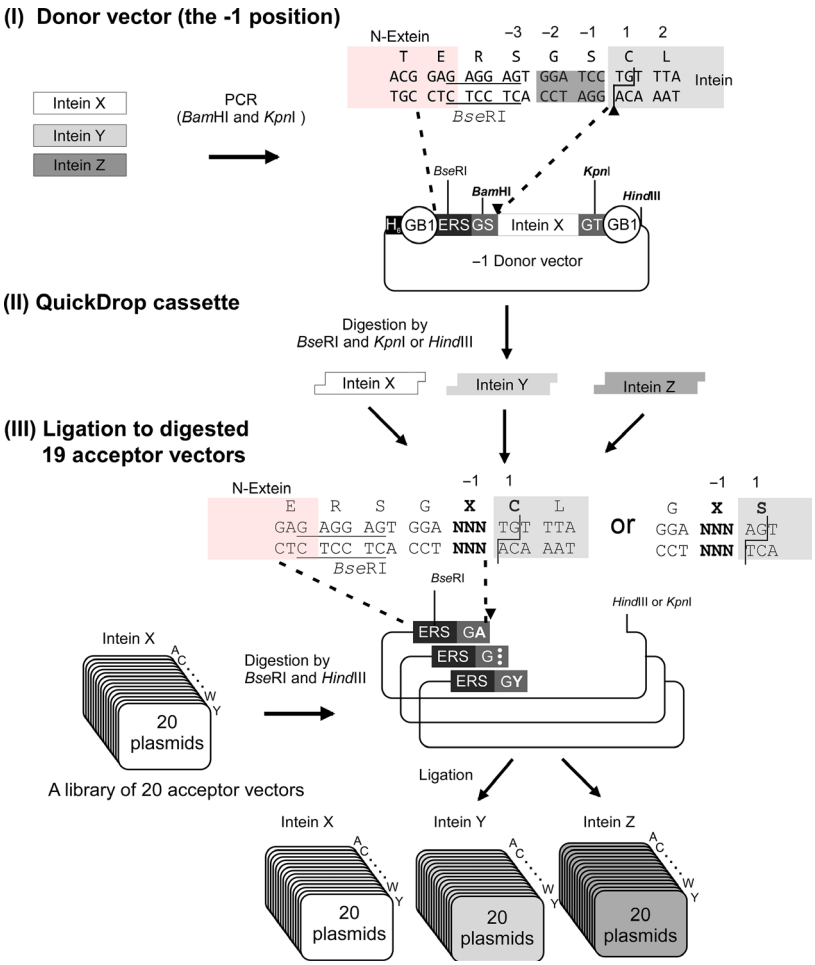



**Fig. 3.** Design of the QuickDrop −1 libraries. The QuickDrop mutagenesis procedure is depicted by steps (**I**), (**II**), and (**III**). (**I**) Cloning of the donor vector for the QuickDrop −1 libraries. The DNA sequences at the N-junction of the donor vector are shown, including the *Bam*HI and *Bse*RI cloning sites. (**II**) Excision of the QuickDrop cassette from the donor vector from Step I by *Bse*RI and *Hind*III (or *Kpn*I). (**III**) Ligation into the predigested remaining 19 acceptor vectors of the QuickDrop −1 libraries to cover all 20 AA-types. The DNA sequences of the cloning site near the −1 position are shown, representing the 20 AA-types as 'NNN', which are mutated from the donor vector by oligonucleotide-directed mutagenesis to create the QuickDrop −1 libraries. The DNA sequence for the part of the intein and N-extein are highlighted in gray and light red shadows, respectively. The recognition sequences of *Bam*HI and *Bse*RI are indicated by dark gray and underlines, respectively. The two DNA sequences for both inteins with Cys1 and Ser1 are designed separately, namely, the QuickDrop −1 library with Cys1 and the QuickDrop −1 library with Ser1.

predigested vectors can be reused with different inteins so that the initial library preparation must only be performed once. Thus, QuickDrop cassette mutagenesis does not require any additional oligonucleotides except those for cloning the intein into the donor vector. However, it requires the first library with the 20 plasmids bearing the 20 different AA-types at the −1 or +2 position, which will be available at addgene.org (Tables S4 and S5).

## QuickDrop mutagenesis reveals the substrate-specificity profiles of different inteins

We applied QuickDrop mutagenesis to profile the substrate specificity (or junction sequence dependency) at the −1 and +2 positions of selected inteins. In total, we investigated 17 profiles of different inteins, including their variants at the first residue and the +1 position (Table 1, Figs 5, 6, and 8). Whereas oligonucleotide-directed mutagenesis requires $20 \times N$ oligonucleotides and PCR reactions for $N$ different inteins, the QuickDrop approach requires $20 + 2 \times N$ oligonucleotides and PCRs in theory, reducing the effort and cost to $(1/N + 1/10)$ of the conventional oligonucleotide-directed mutagenesis. For example, QuickDrop requires approximately 1/5 of the PCRs and oligonucleotides for one position to investigate ten inteins, compared with oligonucleotide-directed mutagenesis.

Inteins are self- and auto-catalytic enzymes. Thus, efficient inteins usually start protein splicing during the precursor expression in *E. coli*. To quantify the splicing efficiency, we introduced a His-tag at the N terminus of the precursor containing two GB1s as the N- and C-exteins flanking the inserted intein. Efficient protein splicing should result in only $H_6$-GB1-GB1 and an intein after *cis*-splicing with 100% efficiency. We purified His-tagged proteins from *E. coli* after the 3-hour precursor expression by immobilized metal affinity chromatography (IMAC). When *cis*-splicing had 100% efficiency, we would expect only the spliced product of $H_6$-GB1-GB1 in the IMAC elution fraction. Side products of protein splicing could originate from N-cleavage at the

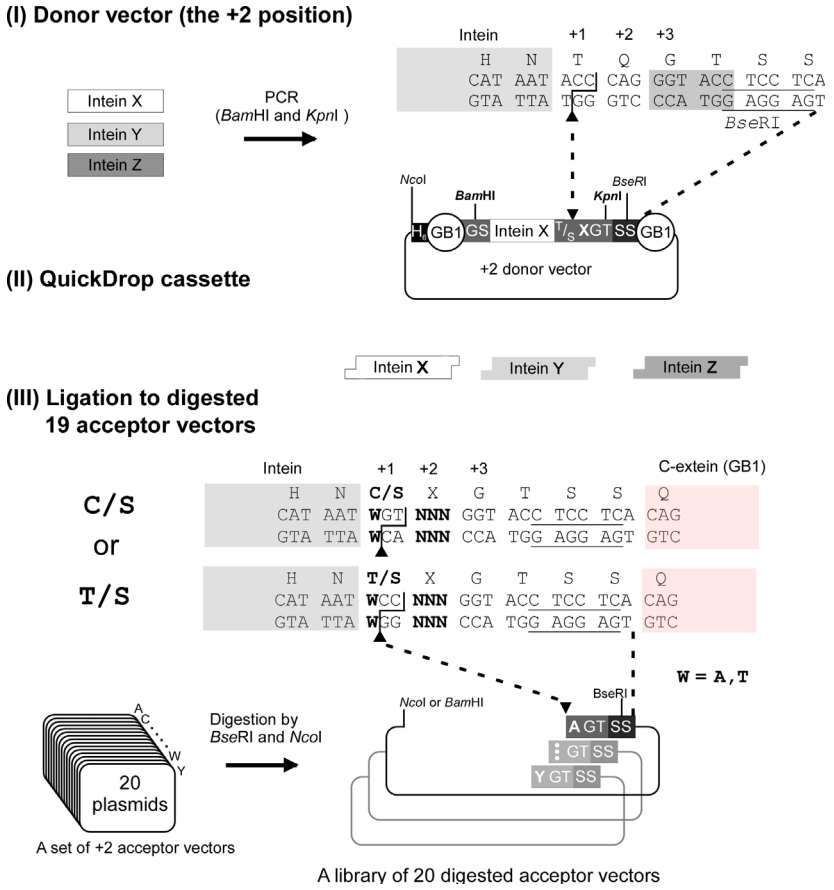



**Fig. 4.** Design of the QuickDrop +2 libraries. The QuickDrop mutagenesis procedure for the +2 position is indicated by steps (**I**), (**II**), and (**III**). (**I**) Cloning of the donor vector of the QuickDrop +2 libraries. The DNA sequences at the +2 position of the donor vector are shown, including the *Kpn*I and *Bse*RI sites. The recognition sequences of *Kpn*I and *Bse*RI are indicated by dark gray and underlines, respectively. (**II**) Excision of the QuickDrop cassette from the donor vector of Step I with *Bse*RI and *Nco*I (or *Bam*HI). (**III**) Ligation to the predigested remaining 19 acceptor vectors of the QuickDrop +2 libraries to cover 20 AA-types. The DNA sequence at the C-junction is depicted, representing the 20 AA-types as 'NNN', which are mutated from the donor vector by oligonucleotide-directed mutagenesis to create the QuickDrop +2 libraries. The DNA sequences for inteins with Cys+1, Ser+1, and Thr+1(Ser+1) are designed, namely, the QuickDrop +2 library with Cys/Ser+1 and the QuickDrop +2 library with Thr/Ser+1. The DNA sequences for the part of the intein and C-extein are highlighted in gray and light red shadow, respectively. The *Bse*RI recognition sequence is underlined.

junction between N-extein and the intein and C-cleavage at the junction between the intein and C-extein. We quantified the spliced product with respect to the total proteins purified, which might include the His-tagged precursor, N- and C-cleaved products, and the spliced product. We assumed that Coomassie blue dye binds equally to each protein according to their molecular weight for the quantification (Fig. 5). Figure 6 presents seven profiles of the −1 position for different inteins and five profiles of the +2 position for different inteins from the SDS-PAGE analysis.

## The effect of the conserved residues of *Npu*DnaE intein on protein splicing

*Npu*DnaE intein has been widely used as one of the most efficient split inteins because of its broad tolerance for various AA-types at the +2 position and its fast splicing activity [14,39]. As both Cys and Ser as the first residue of inteins are common among class 1 inteins, we tested the Cys1 to Ser1 mutation at the first residue of *Npu*DnaE intein [2,36]. *Npu*DnaE(C1S) did not splice any more with all types of amino acids at the −1 position (Fig. 6B). In the case of *Npu*DnaE, the nucleophilicity of Ser does not seem to be high enough to induce the first step of the N-O acyl shift (Fig. 1B). Whereas Cys to Ser mutation at the −1 position abolished protein splicing completely, *Npu*DnaE bearing the Cys+1Ser mutation at the first residue of the C-extein is still capable of splicing with a reduced efficiency for many amino acid types (Fig. 6C) [40]. While the splicing efficiency of *Npu*DnaE was largely affected at the −1 position by the Cys+1Ser mutation, the same Cys+1Ser mutation reduced the efficiency for only a few amino acid types at the +2 position such as Phe, Tyr, His, and Trp showing high splicing efficiency with Cys+1. Some residues did not show a reduced efficiency by the C+1S mutation. For applications such as segmental isotopic labeling by PTS, we think that *cis*-splicing efficiency should preferably reach at least >80-90% to be of practical use (pink lines in Fig. 6) [41]. Thus, *Npu*DnaE(C+1S) probably requires further improvement for the practical use, for example by directed evolution or rational design [22,23,25].

**Table 1.** List of inteins and their wild-type and tested junction sequences. A mutation at the catalytic residues is colored in red. The position for 20-AA mutations is indicated by "X" in bold.

| Intein name | WT N-extein/the 1st residue of intein (−3, −2, −1/1) | WT C-extein (/+1,+2,+3) | N-junction at the −1 position | C-junction at the +2 position |
|---|---|---|---|---|
| *Npu*DnaE | AEY/C | /CFN | SG**X**/C—N/CFN | EGS/C—N/C**X**G |
| *Npu*DnaE(C1S) | AEY/C | /CFN | SG**X**/**S**—N/CFN | – |
| *Npu*DnaE(C+1S) | AEY/C | /CFN | SG**X**/C—N/**S**FG | EGS/C—N/**SX**G |
| *Npu*DnaB$^{\Delta 290}$ | ESG/C | /SIE | SG**X**/C—N/SIG | GSG/C—N/S**X**G |
| *Pho*RadA | SGK/C | /TQL | SG**X**/C—N/TQL | SGK/C—N/T**X**G |
| *Pho*RadA(E71T) | SGK/C | /TQL | SG**X**/C—N/TQL | – |
| gp41-1 | SGY/C | /SSS | SG**X**/C—N/SGG | EGS/C—N/S**X**G |
| *Hut*MCM2 | KMR/C | /SED | SG**X**/C—N/SED | SMR/C—N/S**X**G |
| *Hwar*PolA″ | TQM/S | /TMN | SG**X**/S—N/TMN | GSM/S—N/T**X**G |
| *Hwar*PolA″(T+1S) | TQM/S | /TMN | – | GSM/S—N/**SX**G |

## Structure-based engineering of the specificity of *Pho*RadA intein

Understanding the structure–function relationship of inteins at the splicing junctions might enable us to rationally design inteins with designed features, thereby widening potential applications. The crystal structure of the *Pho*RadA intein previously revealed the interaction between Met at the −1 position and Glu71 of *Pho*RadA intein [20]. We previously found that Glu at the −1 position (E-1) lowered the splicing activity by analyzing the substrate specificity of *Pho*RadA intein [20]. We hypothesized that the negative charge of Glu at the −1 position might have unfavorable interactions with E71 of *Pho*RadA. Therefore, we introduced E71T mutation in *Pho*RadA to remove the negative charge. Indeed, the E71T mutation improved the splicing efficiency of the E-1 variant and also retained the splicing efficiency of the wild-type Lys residue at the −1 position (K-1) [20]. Here, we characterized the full AA-type dependency profile at the −1 position of *Pho*RadA(E71T) for the comparison with the wild-type *Pho*RadA and evaluated the impact of a point mutation on the dependency profile at the −1 position (Fig. 6D,E). While the splicing efficiency of the E-1 variant has improved due to the E71T mutation, several AA-types such as Trp, Leu, Met, and Asn lowered the *cis*-splicing efficiency by 30–50%. The other AA-types did not change the splicing efficiency of *Pho*RadA severely. It would be of particular interest to analyze how other point mutations would influence the substrate specificity or junction dependency to understand the structure–function relationships of inteins. We also obtained the junction residue profile for the −1 position of *Npu*DnaB$^{\Delta 290}$, which allowed us to compare it with *Pho*RadA (Fig. 6F). *Npu*DnaB$^{\Delta 290}$ could also tolerate Glu at the −1 position without a significant reduction of the splicing activity. When we compared the structures of *Npu*DnaB$^{\Delta 290}$ and *Pho*RadA, we found that E71 of *Pho*RadA locates at the position of T51 in the structure of *Npu*DnaB$^{\Delta 290}$ (Fig. 7) [16,20]. Although the wild-type Gly at the −1 position of *Npu*DnaB$^{\Delta 290}$ does not show any interaction with T51 in the structure, T51 in *Npu*DnaB$^{\Delta 290}$ might be responsible for accepting E-1 similar to *Pho*RadA.

Furthermore, gp41-1 intein found by metagenomic sequencing could also tolerate many amino acid types with our model system, including E-1 (Fig. 6G) [42,43]. The comparison between *Pho*RadA and gp41-1 indicated that E71 in *Pho*RadA could correspond to S42 in the structure of gp41-1 (Fig. 7) [20,28]. The inspection of the substrate-specificity profiles and three-dimensional structures implies that it might be feasible to identify the common structure–function relationship among different inteins, even though the junction sequence dependency profile seems to be unique to individual inteins.

## The junction dependencies of obligate halophilic inteins

Saturation mutagenesis with degenerated oligonucleotides is a very convenient way to create a library of plasmids with various mutations at desired positions. It has been used for monitoring the phenotypes without isolating individual mutants for directed evolution [23,24,32,33]. In the case of obligate halophilic inteins, the selection of mutants based on the phenotype is not feasible with nonhalophilic organisms such as *E. coli*. Obligate halophilic inteins require a high salinity condition for protein splicing, such as 3–4 M NaCl [29]. Therefore, the QuickDrop mutagenesis is a more suitable method of choice for the specificity analysis of obligate halophilic inteins. The vectors with 20 variants at the −1 and +2 positions were constructed by

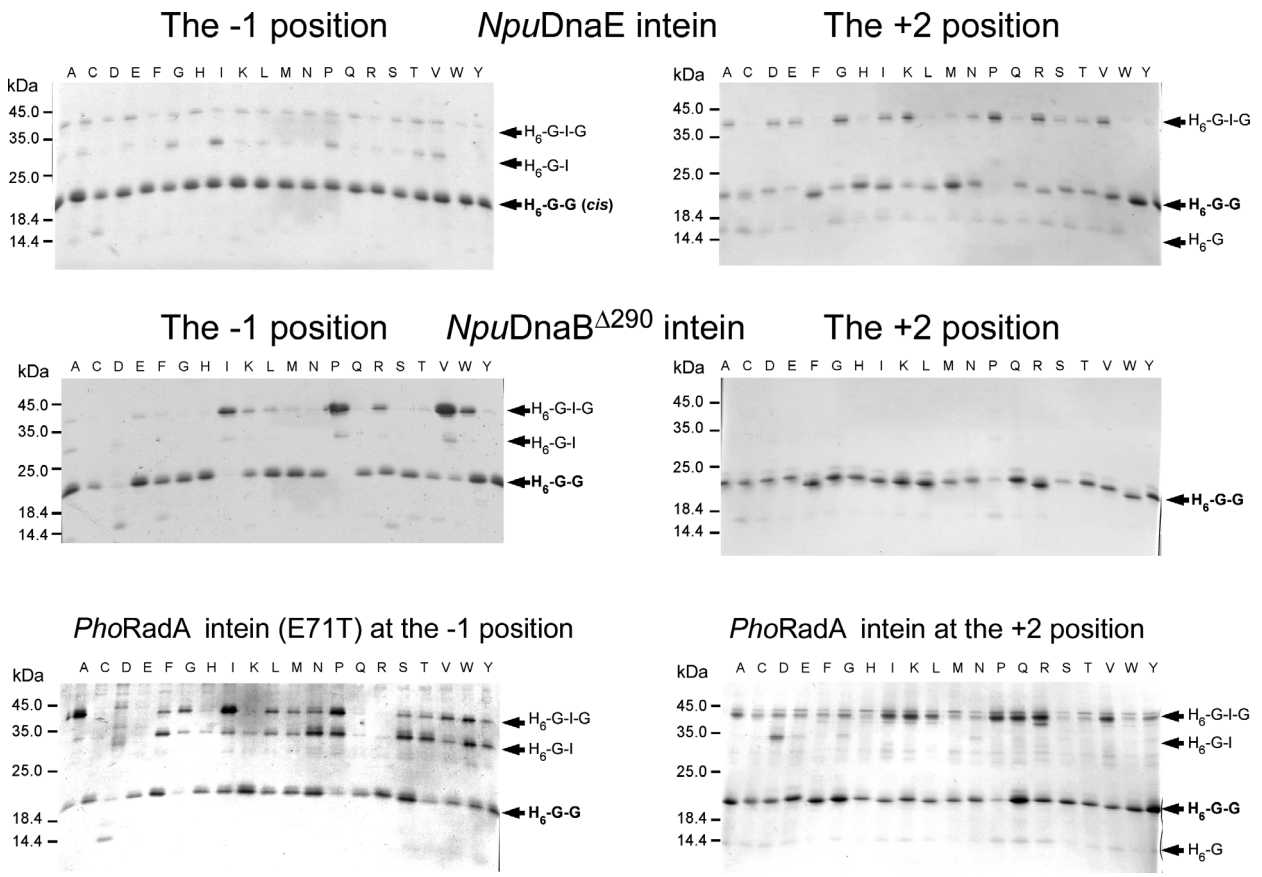
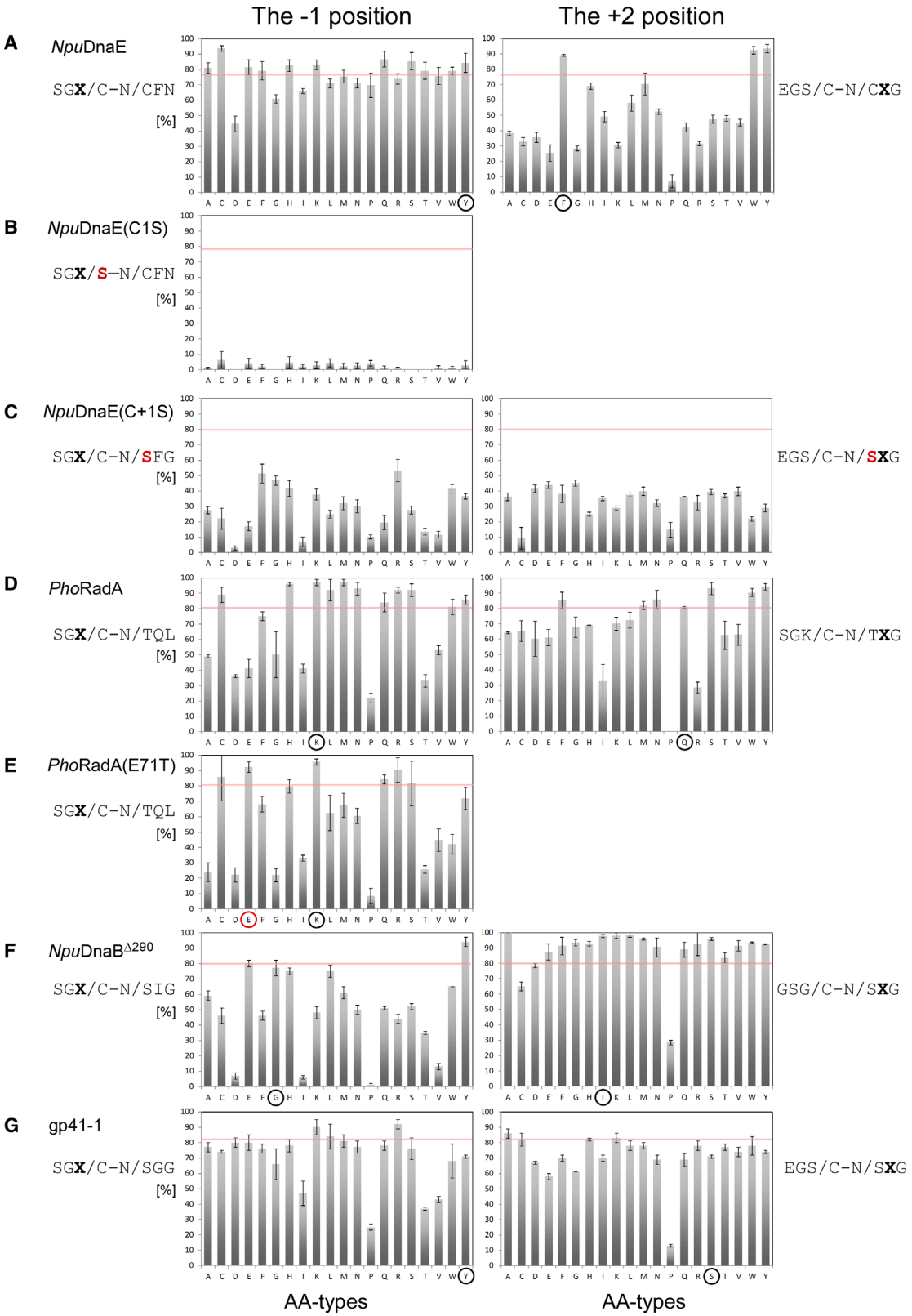



**Fig. 5.** Example of SDS-PAGE analysis of protein-splicing efficiencies for 20 AA-variants for selected inteins obtained by QuickDrop mutagenesis. Examples of SDS gels are shown, which were used for the *cis*-splicing analysis for *Npu*DnaE intein (top), *Npu*DnaB$^{\Delta 290}$ (middle), and *Pho*RadA(E71T) and *Pho*RadA (bottom row). Precursor proteins with the 20 AA-variants at the −1 and +2 positions were expressed in *Escherichia coli*. His-tagged proteins were purified by IMAC and analyzed by SDS-PAGE (18%). Arrowheads indicate bands corresponding to unspliced precursors ($H_6$-G-I-G), C-cleaved products ($H_6$-G-I), and *cis*-spliced products ($H_6$-G-G).

QuickDrop mutagenesis and expressed and purified by IMAC for analysis. The 20 purified precursors were incubated in the presence of 3 M NaCl at room temperature overnight. The salts in the reaction mixture were removed by precipitating the proteins using TCA. The precipitated proteins were then analyzed by SDS-PAGE and quantified (Fig. 8A and Figs S1 and S2). The residue type at the −1 position of MCM2 intein from *Halorhabdus utahensis* (*Hut*MCM2) affected the splicing activity of *Hut*MCM2 as previously reported and the best residue type was Arg at the −1 position [29,44]. Whereas the −1 position of *Hut*MCM2 does not generally tolerate any other amino acid type other than the wild-type Arg, the +2 position widely accepts any AA-types, including Pro, which does not splice with most inteins.

We also analyzed the junction sequence dependencies of the DNA-directed RNA polymerase subunit A″ intein from in *Haloquadratum walsbyi* (*Hwa*rPolA″ intein). *Hwa*rPolA″ intein was selected as one of the

**Fig. 6.** Summary of the substrate specificities of different inteins presented as *cis*-splicing efficiencies. Bar graphs present quantified *cis*-splicing efficiencies in % from SDS gels *versus* 20-AA-types at the −1 and +2 positions for tested nonhalophilic inteins. The data for (A) *Npu*DnaE, (B) *Npu*DnaE(S1), (C) *Npu*DnaE(S+1), (D) *Pho*RadA, (E) *Pho*RadA(E71T), (F) *Npu*DnaB$^{\Delta 290}$, and (G) gp41-1 are presented. Black circles indicate the wild-type residue types. Three-residue N- and C-extein sequences and the first and last residues of each intein are shown next to the graph with **X** presenting 20 AA-types. A red circle indicates E-1 in *Pho*RadA(E71T) used for improving efficiency. Pink lines indicate 80%-efficiency lines. The graphs for *Npu*DnaE and *Npu*DnaB$^{\Delta 290}$ are the reproduction of the data in ref. [51]. The data for *Pho*RadA at the −1 position was the reproduction of the data from ref. [20]. Error bars indicate standard deviations derived from at least three independent experiments and quantifications.


The -1 position
The +2 position
A
NpuDnaE
SGX/C-N/CFN
[%]
EGS/C-N/CXG
B
NpuDnaE(C1S)
SGX/S-N/CFN
[%]
C
NpuDnaE(C+1S)
SGX/C-N/SFG
[%]
EGS/C-N/SXG
D
PhoRadA
SGX/C-N/TQL
[%]
SGK/C-N/TXG
E
PhoRadA(E71T)
SGX/C-N/TQL
[%]
F
NpuDnaB^Δ290
SGX/C-N/SIG
[%]
GSG/C-N/SXG
G
gp41-1
SGX/C-N/SGG
[%]
EGS/C-N/SXG
AA-types
AA-types

most robust inteins among all tested inteins from *Haloquadratum walsbyi* (data not shown). *Hwa*rPolA″ intein seems to accept many AA-types at both the −1 and +2 positions except for Pro. Interestingly, the *Hwa*rPolA″ (T+1S) variant also tolerates many AA-types at the +2 position, although with a generally reduced splicing efficiency. Only a few examples of junction dependencies are available for obligate halophilic inteins. There is currently no three-dimensional structure of halophilic inteins, which makes it difficult to extrapolate any tendency from the current data. Characterization of obligate halophilic inteins remains to be further investigated for salt-dependent protein splicing [29].

## Discussion

Elucidation of the substrate specificities of the single-turnover intein enzymes is the first step to understand the molecular and structural basis of the protein-splicing reaction. Profiling the AA-type dependency at the splicing junctions is essential, particularly for the rational design of inteins as well as for the selection of inteins for biotechnological applications. We described the QuickDrop mutagenesis approach using a type IIS restriction enzyme and conserved residues among inteins for studying the junction dependency at the −1 and +2 positions. This method could be useful for the 20 AA-type mutageneses with many enzymes possessing conserved residues because it could reduce the required number of oligonucleotides and PCR steps down to on-tenth in comparison with conventional oligonucleotide-directed mutagenesis. The QuickDrop mutagenesis is also not limited to inteins as the target but could also be used with other proteins for testing 20 AA-types among homologs with conserved residues [45,46]. However, one of the limitations of this strategy are possible internal restriction sites for cloning, that are, *Bse*RI, *Bam*HI, and *Kpn*I sites, requiring removal or a change of the restriction site. This problem could also be easily solved by synthetic genes for small inteins together with codon optimizations. Another limiting factor is that the linker sequences connecting the intein and exteins are defined by the recognition sequence of the type IIS restriction enzyme, like *Bse*RI [38]. However, the recognition sequence from *Bse*RI does not particularly disturb the comparison between the 20 AA-variants at the critical −1 and +2 positions of inteins as it is used as an artificial amino acid linker in the model system (Table 1). The full investigation on the −1 and +2 positions of an intein requires the analysis of all the 20 × 20 (400) combinations. However, we tested only the 20 + 20 (40) combinations by fixing the AA-type at one of the two junctions, assuming multiplicative effects from both N- and C-junction residues (Table 1). We believe that the characterized substrate specificities of inteins by the 20 + 20 (40) combination approach still guides a choice of AA-types for each junction for various applications, yet keeps the experimental number relatively small. Further similar analysis of many more inteins having efficient *cis*-splicing and *trans*-splicing activities could help our understanding of the structure–function relationships of each intein for broader applications. These constructed 20 + 20 (40) vectors could also be easily converted into a dual vector system by splitting within the intein for testing *trans*-splicing using two compatible inducible vectors as previously used [7,13,14].

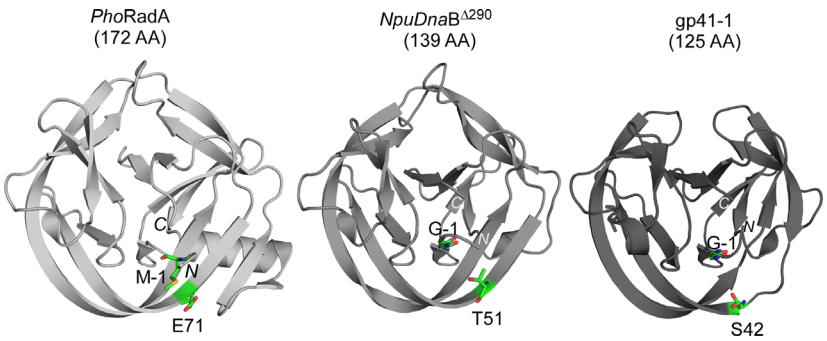



**Fig. 7.** Three-dimensional structures of *Pho*RadA, *Npu*DnaB$^{\Delta 290}$, and gp41-1. E71 of *Pho*RadA is shown in the stick model on the cartoon representation of the *Pho*RadA structure (PDB: 4E2U) [20]. T51 of *Npu*DnaB$^{\Delta 290}$ is presented in a stick model on the cartoon of the *Npu*DnaB$^{\Delta 290}$ structure (PDB: 4O1R) [16], which corresponds to E71 in the structure of *Pho*RadA intein. Similarly, S42 of gp41-1 intein structurally corresponds to E71 in the structure of *Pho*RadA intein, presented in a stick model (PDB: 6QAZ) [28]. These three residues are colored in green. N and C indicate the N and C termini, respectively. The M-1 residue interacting with E71 of *Pho*RadA is also shown in the stick model. The corresponding residues of G-1 for *Npu*DnaB$^{\Delta 290}$ and gp41-1 inteins are illustrated with green stick models.

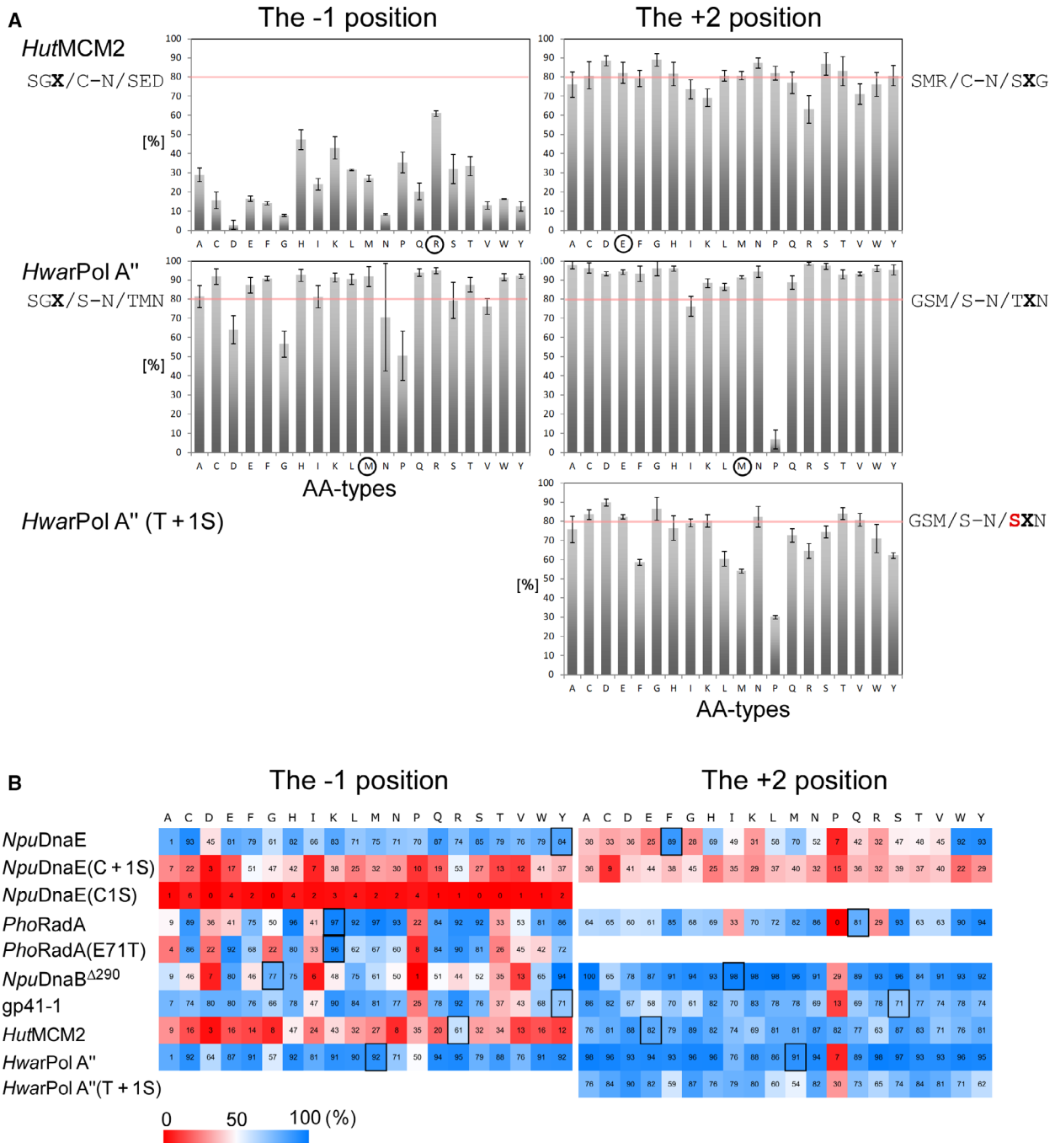



**Fig. 8.** Substrate-specificity profiles of halophilic inteins and the combined summary. (A) Summary of the substrate-specificity profiles for three halophilic *Hut*MCM2 (top), *Hwar*PolA″ (middle), and *Hwar*PolA″(T+1S) inteins. The left and right panels are *cis*-splicing efficiencies in % *versus* 20 AA-types at the −1 and +2 positions, respectively. Black circles indicate the wild-type residue types. Three-residue N- and C-extein sequences and the first and last residues of each intein are shown next to the graph with **X** for 20 AA-types. Pink lines indicate 80%-efficiency lines. Error bars indicate standard deviations derived from at least three independent experiments and quantifications. (B) Summary of the combined substrate-specificity profiles by the heatmap presentation for the ten inteins, including their variants. The color scale is represented with 0%, 50%, and 100% for red, white, and blue, respectively. Rectangles highlight the wild-type amino acid types.

Generally, we could detect some trends from the substrate profiles from the tested inteins and their variants in this study. For example, branched amino acids such as Ile, Val, and Thr are generally unfavorable at the −1 position, comprising a group of the lower splicing efficiency in many inteins in line with previous

reports [14]. A similar trend for branched amino acids observed with native chemical ligation might suggest that these amino acid types slow down the *trans*-esterification step [47]. Negatively charged residues like Asp tend to have lower efficiencies, presumably because the side-chain carboxyl group of Asp could react with the backbone peptide, resulting in cleavages [48]. Pro residue seems to be the worst AA-type for both the −1 and +2 positions for most inteins, suggesting that the cyclic imino acid restricts the backbone conformation or/and dynamics, which are required for protein splicing. However, these observations are only tendencies with several exceptions like in the case of *Hut*MCM2 intein. Therefore, these trends cannot be generalized for all inteins. Indeed, inteins with Pro at the −1 or +2 position or Asp at the −1 position are prevalent [49,50]. Such naturally occurring inteins suggest that some inteins adapted to the unfavorable AA-types at the −1 or +2 position in their host proteins. Therefore, it is safe to assume that the junction dependency (substrate specificity) is unique to individual inteins like conventional enzymes with multiple turnovers. It is still unknown how some inteins have adapted to the junction sequences by mutations at the atomic level. The primary structures of inteins alone do not indicate strong correlations with the small number of available substrate-specificity profiles. However, the combination with the three-dimensional structures might provide some insights into the structure–function relationship of inteins. We previously found that Glu71(E71) of *Pho*RadA can be mutated to Thr to improve the splicing efficiency of the E-1 variant based on the interactions found in the structure of *Pho*RadA [20]. We examined the structures of *Npu*DnaB$^{\Delta 290}$ and gp41-1, showing better efficiency with Glu at the −1 position [16,28]. Both *Npu*DnaB and gp41-1 structures have Thr (T51) and Ser (S42), respectively, at the corresponding residue of E71 in *Pho*RadA. This observation might indicate that the negative charge at the position corresponding to E71 in *Pho*RadA could be unfavorable for the splicing efficiency in many inteins. It is also noteworthy that not only the −1 and +2 positions but also the −2 and +3 positions could synergistically influence protein-splicing efficiency, presumably by defining the local conformation due to the interactions between them [28,40]. We also observed the same effect from other residues than at the −1 and +2 positions on protein splicing by the gp41-1 intein, pointing out that the −1 and +2 positions are not the only determinants for the splicing efficiency [28]. Not only the AA-type but also local structures in the presence of exteins and conformational dynamics such as flexibility could play essential roles in protein splicing [9,40]. The splicing efficiencies derived in this experiment using model exteins are mere reference values. The amino acid type dependency of *cis*-splicing studied with the model proteins should not be generalized for all contexts.

Nevertheless, an increasing number of intein structures, along with their quantitative substrate-specificity analysis, would disclose further valuable and detailed structure–function relationships of the junction dependencies. Such information would directly assist in the design of various applications utilizing different inteins. It has become evident in the past that not all naturally occurring inteins could be easily engineered for their applications, such as protein ligation requiring splitting of the inteins [7,16,18]. A selection of an optimal intein from naturally occurring inteins for a specific junction sequence might not always be feasible, with the assumption that a natural junction sequence is an optimal sequence for protein splicing. Directed evolution of an intein for adapting a particular junction sequence can be laborious and time-consuming and cannot generally be used for each application [7,20,21,23,24]. The newly developed QuickDrop mutagenesis approach could facilitate the quantitative analysis of the substrate specificities of many different inteins for overcoming one of the bottlenecks in their broader applications without optimizations, for example, by directed evolution. We anticipate that further biochemical and structural studies on many different inteins will improve our understanding of critical aspects of the protein-splicing mechanism, including the substrate specificities at the atomic resolution. A better understanding of the structure–function relationship of many different inteins would not only facilitate the rational design of promiscuous and robust inteins with desired features for wider applications as protein ligases but also shed light on the evolutionary origins of different inteins.

## Acknowledgements

We thank Cathrin Albert, Sandra Jääskeläinen, Britta Haas, Laura Knaapi, Lynn Schlierkamp, and Lauri Vaija for contributing to the preparation of various plasmids used in this study.

## Author contributions

HI conceived and supervised the study; HI, JSO, and ASA designed experiments; JSO, HMB, JM, and ASA performed experiments; JSO, ASA, HMB, and HI analysed data; HI, JSO, and HMB prepared the manuscript; HI and HMB made manuscript revisions with inputs from others.

## Funding

This work was supported by the Academy of Finland [decision number 137995, 277335]; Sigrid Juselius Foundation. HMB was supported by the Novo Nordisk Foundation [NNF17OC0025402]. JSO was, in part, supported by the National Doctoral Programme in Informational and Structural Biology.

## References

1 Paulus H (2000) Protein splicing and related forms of protein autoprocessing. *Annu Rev Biochem* **69**, 447–496.
2 Noren CJ, Wang J and Perler FB (2000) Dissecting the chemistry of protein splicing and its applications. *Angew Chem Int Ed Engl* **39**, 450–466.
3 Truong D-JJ, Kühner K, Kühn R, Werfel S, Engelhardt S, Wurst W and Ortiz O (2015) Development of an intein-mediated split–Cas9 system for gene therapy. *Nucleic Acids Res* **43**, 6450–6458.
4 Wood DW and Camarero JA (2014) Intein applications: from protein purification and labeling to metabolic control methods. *J Biol Chem* **289**, 14512–14519.
5 Volkmann G and Iwaï H (2010) Protein trans-splicing and its use in structural biology: opportunities and limitations. *Mol Biosyst* **6**, 2110–2121.
6 Topilina NI and Mills KV (2014) Recent advances in in vivo applications of intein-mediated protein splicing. *Mob DNA* **5**, 5.
7 Pinto F, Thornton EL and Wang B (2020) An expanded library of orthogonal split inteins enables modular multi-peptide assemblies. *Nat Commun* **11**, 1529.
8 Chong S, Montello GE, Zhang A, Cantor EJ, 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.
9 Yamazaki T, Otomo T, Oda N, Kyogoku Y, Uegaki K, Ito N, Ishino Y and Nakamura H (1998) Segmental isotope labeling for protein NMR using peptide splicing. *J Am Chem Soc* **120**, 5591–5592.
10 Mootz HD and Muir TW (2002) Protein splicing triggered by a small molecule. *J Am Chem Soc* **124**, 9044–9045.
11 Severinov K and Muir TW (1998) Expressed protein ligation, a novel method for studying protein-protein interactions in transcription. *J Biol Chem* **273**, 16205–16209.
12 Gangopadhyay JP, Jiang SQ, 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.
13 Aranko AS, Züger S, Buchinger E and Iwaï H (2009) In vivo and in vitro protein ligation by naturally occurring and engineered split DnaE inteins. *PLoS One* **4**, e5185.
14 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.
15 Southworth MW, Amaya K, Evans TC, Xu MQ and Perler FB (1999) Purification of proteins fused to either the amino or carboxy terminus of the *Mycobacterium xenopi* gyrase A intein. *Biotechniques* **27**, 110–120.
16 Aranko AS, Oeemig JS, Zhou D, Kajander T, Wlodawer A and Iwaï H (2014) Structure-based engineering and comparison of novel split inteins for protein ligation. *Mol Biosyst* **10**, 1023–1034.
17 Nogami S, Fukuda T, Nagai Y, Yabe S, Sugiura M, Mizutani R, Satow Y, Anraku Y and Ohya Y (2002) Homing at an extragenic locus mediated by VDE (PI-SceI) of *Saccharomyces cerevisiae*. *Yeast* **19**, 773.
18 Iwaï H, Mikula KM, Oeemig JS, Zhou D, Li M and Wlodawer A (2017) Structural basis for the persistence of homing endonucleases in transcription factor IIB inteins. *J Mol Biol* **429**, 3942–3956.
19 Neugebauer M, Böcker JK, Matern JC, Pietrokovski S and Mootz HD (2017) Development of a screening system for inteins active in protein splicing based on intein insertion into the LacZα-peptide. *Biol Chem* **398**, 57–67.
20 Oeemig JS, 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.
21 Ellilä S, Jurvansuu JM and Iwaï H (2011) Evaluation and comparison of protein splicing by exogenous inteins with foreign exteins in *Escherichia coli*. *FEBS Lett* **585**, 3471–3477.
22 Cheriyan M, Pedamallu CS, Tori K and Perler F (2013) Faster protein splicing with the *Nostoc punctiforme* DnaE intein using non-native extein residues. *J Biol Chem* **288**, 6202–6211.
23 Lockless SW and Muir TW (2009) Traceless protein splicing utilizing evolved split inteins. *Proc Natl Acad Sci USA* **106**, 10999–11004.
24 Appleby-Tagoe JH, Thiel IV, Wang Y, Wang Y, Mootz HD and Liu X-Q (2011) Highly efficient and more general cis- and trans-splicing inteins through sequential directed evolution. *J Biol Chem* **286**, 34440.
25 Beyer HM and Iwaï H (2019) Off-pathway-sensitive protein splicing screening based on a toxin/antitoxin system. *Chembiochem* **20**, 1933–1938.
26 Stevens AJ, Sekar G, Shah NH, Mostafavi AZ, Cowburn D and Muir TW (2017) A promiscuous split intein with expanded protein engineering applications. *Proc Natl Acad Sci USA* **114**, 8538–8543.

27 Braman J, Papworth C and Greener A (1996) Site-directed mutagenesis using double-stranded plasmid DNA templates. *Methods Mol Biol* **57**, 31–44.
28 Beyer HM, Mikula KM, Li M, Wlodawer A and Iwaï H (2020) The crystal structure of the naturally split gp41-1 intein guides the engineering of orthogonal split inteins from a cis-splicing intein. *FEBS J* **287**, 1886–1898.
29 Ciragan A, Aranko AS, Tascon I and Iwaï 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.
30 Rueden CT, Schindelin J, Hiner MC, DeZonia BE, Walter AE, Arena ET and Eliceiri KW (2017) ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics* **18**, 529.
31 Kunkel TA (1985) Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proc Natl Acad Sci USA* **82**, 488–492.
32 Miyazaki K and Arnold FH (1999) Exploring nonnatural evolutionary pathways by saturation mutagenesis: rapid improvement of protein function. *J Mol Evol* **49**, 716–720.
33 Reetz MT, Kahakeaw D and Lohmer R (2008) Addressing the numbers problem in directed evolution. *Chembiochem* **9**, 1797–1804.
34 Nov Y (2012) When second best is good enough: another probabilistic look at saturation mutagenesis. *Appl Environ Microbiol* **78**, 258–262.
35 Virnekäs B, Ge L, Plückthun A, Schneider KC, Wellnhofer G and Moroney SE (1994) Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. *Nucleic Acids Res* **22**, 5600–5607.
36 Pingoud A, Fuxreiter M, Pingoud V and Wende W (2005) Type II restriction endonucleases: structure and mechanism. *Cell Mol Life Sci* **62**, 685–707.
37 Engler C, Kandzia R and Marillonnet SA (2008) One pot, one step, precision cloning method with high throughput capability. *PLoS One* **3**, e3647.
38 Mushtaq R, Naeem S, Sohail A and Riazuddin S (1993) BseRI a novel restriction endonuclease from a Bacillus species which recognizes the sequence 5'..GAGGAG..3'. *Nucleic Acids Res* **21**, 3585.
39 Aranko AS, Wlodawer A and Iwaï H (2014) Nature's recipe for splitting inteins. *Protein Eng Des Sel* **27**, 263–271.
40 Cheriyan M, Chan SH and Perler F (2014) Traceless splicing enabled by substrate-induced activation of the *Nostoc punctiforme* Npu DnaE intein after mutation of a catalytic cysteine to serine. *J Mol Biol* **426**, 4018–4029.
41 Muona M, Aranko AS, Raulinaitis V and Iwaï H (2010) Segmental isotopic labelling of multi-domain and fusion proteins by protein trans-splicing in vivo and in vitro. *Nat Protoc* **5**, 574–587.
42 Dassa B, London N, Stoddard BL, Schueler-Furman O and 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.
43 Carvajal-Vallejos P, Pallissé R, Mootz HD and Schmidt SR (2012) Unprecedented rates and efficiencies revealed for new natural split inteins from metagenomic sources. *J Biol Chem* **287**, 28686–28696.
44 Ciragan A, Backlund SM, Mikula KM, Beyer HM, Ollila OHS and Iwaï H (2020) NMR structure and dynamics of TonB investigated by scar-less segmental isotopic labeling using a salt-inducible split intein. *Front Chem* **8**, 136.
45 Mikula KM, Tascón I, Tommila JJ and Iwaï H (2017) Segmental isotopic labeling of a single-domain globular protein without any refolding step by an asparaginyl endopeptidase. *FEBS Lett* **591**, 1285–1294.
46 Mikula KM, Krumwiede L, Plückthun A and Iwaï H (2018) Segmental isotopic labeling by asparaginyl endopeptidase-mediated protein ligation. *J Biomol NMR* **71**, 225–235.
47 Hackeng TM, Griffin JH and Dawson PE (1999) Protein synthesis by native chemical ligation: expanded scope by using straightforward methodology. *Proc Natl Acad Sci USA* **96**, 10068–10073.
48 Stephenson RC and Clarke S (1989) Succinimide formation from aspartyl and asparaginyl peptides as a model for the spontaneous degradation of proteins. *J Biol Chem* **264**, 6164–6170.
49 Perler FB (1999) InBase, the New England Biolabs intein database. *Nucleic Acids Res* **27**, 346–347.
50 Novikova O, Jayachandran P, Kelley DS, Morton Z, Merwin S, Topilina NI and Belfort M (2016) Intein clustering suggests functional importance in different domains of life. *Mol Biol Evol* **33**, 783–799.
51 Aranko AS and Iwaï H (2017) Protein ligation by HINT domains. In Chemical Ligation: Tools for Biomolecule Synthesis and Modification (D'Andrea LD and Romanelli A, eds), pp. 421–445. Wiley, Hoboken, NJ.

## Supporting information

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

**Fig. S1.** Substrate specificty profile for *Hut*MCM2 intein at the +2 position.

**Fig. S2.** Substrate specificty profile for *Hut*MCM2 intein at the −1 position.

**Table S1.** Summary of oligonucleotides used for cloning genes and modifications of plasmids.
**Table S2.** Summary of oligonucleotides used for the introduction of 20-AA types at the −1 position of the *Npu*DnaE intein with C1/S1 at the first residue (QuickDrop −1 libraries with Cys1 and Ser1).
**Table S3.** Summary of oligonucleotides used for the introduction of 20-AA types at the +2 position for the construction of the QuickDrop +2 libraries with Thr/Ser+1 or Cys/Ser+1.
**Table S4.** Summary of plasmids for non-halophilic inteins with 20 AA mutations at the −1 position.
**Table S5.** Summary of plasmids for non-halophilic inteins with 20 AA mutations at the +2 position.
**Table S6.** Summary of plasmids for halophilic inteins with 20 AA mutations at the −1 and +2 positions.