Pervasive Suicidal Integrases in Deep-Sea Archaea

Catherine Badel,[1] Violette Da Cunha,[1] Patrick Forterre,[1,2] and Jacques Oberto *,[1]

[1]Microbiology Department, Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, University of Paris-Sud, Université Paris-Saclay, Gif-sur-Yvette cedex, France

[2]Unité de Biologie Moléculaire du Gène chez les Extrêmophiles, Département de Microbiologie, Institut Pasteur, Paris, France

*Corresponding author: E-mail: jacques.oberto@i2bc.paris-saclay.fr.

Associate editor: Miriam Barlow

Abstract

Mobile genetic elements (MGEs) often encode integrases which catalyze the site-specific insertion of their genetic information into the host genome and the reverse reaction of excision. Hyperthermophilic archaea harbor integrases belonging to the SSV-family which carry the MGE recombination site within their open reading frame. Upon integration into the host genome, SSV integrases disrupt their own gene into two inactive pseudogenes and are termed suicidal for this reason. The evolutionary maintenance of suicidal integrases, concurring with the high prevalence and multiples recruitments of these recombinases by archaeal MGEs, is highly paradoxical. To elucidate this phenomenon, we analyzed the wide phylogenomic distribution of a prominent class of suicidal integrases which revealed a highly variable integration site specificity. Our results highlighted the remarkable hybrid nature of these enzymes encoded from the assembly of inactive pseudogenes of different origins. The characterization of the biological properties of one of these integrases, $Int^{pT26\text{-}2}$ showed that this enzyme was active over a wide range of temperatures up to 99 °C and displayed a less-stringent site specificity requirement than comparable integrases. These observations concurred in explaining the pervasiveness of these suicidal integrases in the most hyperthermophilic organisms. The biochemical and phylogenomic data presented here revealed a target site switching system operating on highly thermostable integrases and suggested a new model for split gene reconstitution. By generating fast-evolving pseudogenes at high frequency, suicidal integrases constitute a powerful model to approach the molecular mechanisms involved in the generation of active genes variants by the recombination of proto-genes.

Key words: **tyrosine recombinase, evolution, pseudogenes.**

Introduction

The maintenance and propagation of mobile genetic elements (MGEs) such as plasmids and viruses impose the infection of a suitable cellular host and the deployment of appropriate strategies (Hulter et al. 2017). Brute force mechanisms such as high copy number grant MGE inheritance into daughter cells after cell division (Million-Weaver and Camps 2014). More refined toxin–antitoxin systems ensure MGE maintenance by relentlessly killing hosts trying to eliminate them (Harms et al. 2018). These mechanisms prove a burden for the host cells which develop effective countermeasures such as CRISPRs or restriction modification systems (Arber and Dussoix 1962; Hille et al. 2018). Alternatively, to favor their maintenance, MGEs alleviate their physiological cost for the host (Carroll and Wong 2018). For example, an efficient MGE partitioning allows propagation with a low copy number (Nordstrom 2006; Gerdes et al. 2010). Some MGEs even carry functions that present an advantage for the host such as resistance genes that increase the fitness of the symbiont in the presence of antibiotics (Carroll and Wong 2018). Contrastingly, particular MGEs have adopted a different potent survival strategy. They have acquired the capacity to integrate their DNA at a particular location of the cellular chromosome without overly altering both genetic programs, using a mechanism known as site-specific recombination (Landy 1989, 2015; Grindley et al. 2006). By disguising their genome as part of the host chromosome, these MGEs succeed in lowering their negative impact on the host metabolism and in bypassing defense mechanisms. This improved cellular acceptance ensures MGE maintenance and vertical propagation. The reverse reaction of excision regenerates the MGE in its independently replicating form (Gandon 2016) which can infect other host cells. The bistable mechanism of site-specific integration/excision is orchestrated by MGE-encoded enzymes belonging to serine- or tyrosine recombinases (Grindley et al. 2006). Tyrosine recombinases constitute the most widespread site-specific recombinases and their enzymatic properties have been investigated for decades (Chen et al. 1992; Guo et al. 1999; Landy 2015). They typically recognize short identical DNA sequences present simultaneously on the MGE DNA and on its host chromosome. According to the phage Lambda/*Escherichia coli* paradigm, these sequences are termed attB (for attachment Bacteria) and attP (for attachment Phage) (Landy 2015). Integrases catalyze site-specific recombination between these sequences using a timely orchestrated mechanism consisting of two sequentially integrase-generated single-stranded cuts in the two att sequences followed by strand-migration and

religation (Grindley et al. 2006). As a result, the exact MGE DNA is integrated into the host chromosome and bordered by attL (for attachment Left) and attR (for attachment Right) sequences which are hybrids of attB and attP. The site-specific recombination between attL and attR, known as excision, regenerates perfectly intact MGE and host chromosomes. The recombination reaction requires in some cases additional protein partners called recombination directionality factors (RDFs) that regulate the orientation of the reaction (Lewis and Hatfull 2001). Interestingly, integrases sharing very similar enzymatic properties have been identified in all domains of life. Bacterial and eukaryotic tyrosine recombinases have been extensively studied (Jayaram et al. 2015; Landy 2015; Van Duyne 2015; Dorman and Bogue 2016). In contrast, very few archaeal integrases have been fully characterized (Zhan et al. 2015; Cossu et al. 2017; Wang et al. 2018). Archaeal integrases belong to two distinctive types: the SSV type (type I) and the pNOB8 type (type II) (She et al. 2004). The site-specific recombination promoted by type II enzymes follows the Lambda Int paradigm with separate attP site and integrase gene. Type I is so far restricted to archaea and consists of peculiar suicidal integrases whose attP site resides within the integrase-coding gene (She et al. 2001). Upon integration, the integrase-coding gene is split into two inactive pseudogenes, int(N) and int(C), on each side of the integrated MGE. Suicidal integrases have been encountered in geothermal environments (She et al. 2001; Cossu et al. 2017), the natural habitat of the Euryarchaeal *Thermococcales* comprising one of the most hyperthermophilic organisms (Schut et al. 2007; Callac et al. 2016; Adam et al. 2017). Plasmid pTN3 (Gaudin et al. 2014) from *Thermococcus nautili* 30-1 (Gorlas et al. 2014; Oberto et al. 2014) encodes Int^{pTN3}, the only suicidal integrase that has been characterized in *Thermococcales* so far (Cossu et al. 2017). This integrase is present in a narrow range of *Thermococcales* and promotes massive genomic inversions in addition to bona fide site-specific recombination properties (Cossu et al. 2017).

The evolutionary maintenance of suicidal integrases, which destroy their own gene, is highly paradoxical and has not been studied so far. In order to elucidate this phenomenon using phylogenomic analyses, it was important to identify a larger data set than the one available for the Int^{pTN3}-like integrases. We recently uncovered a wide geographical distribution among hyperthermophilic archaea of pT26-2 type plasmids encoding suicidal integrases (Badel et al. 2019). This large plasmid family allowed us to perform robust phylogenies and comparative genomics. The characterization of the enzymatic properties of one of these integrases and the reconstitution of the evolution history of the entire family provided a strong rationale to explain the maintenance and widespread distribution of suicidal integrases in deep-sea archaea using a mechanism involving pseudogenes. If pseudogenes are often described as "junk DNA," they also provide a source of genetic diversity (Vihinen 2014). In contrast, de novo gene birth via transitory proto-genes remains poorly understood (Siepel 2009; Carvunis et al. 2012). By generating pseudogenes at high frequency, suicidal integrases could constitute a powerful model to investigate the generation of active genes variants by the recombination of proto-genes.

Results

A Cluster of Related Integrases Is Prevalent in Hyperthermophilic Euryarchaea

Plasmid pT26-2 from *Thermococcus* sp.26-2 was the first plasmid isolated from hyperthermophilic Euryarchaea shown to encode a putative integrase, Int^{pT26-2} (Soler et al. 2010). We wondered whether we would detect Int^{pT26-2} homologs in other hyperthermophilic organisms. A similarity search first performed on the pT26-2 plasmid family (Badel et al. 2019) and further extended as indicated in Materials and Methods identified 73 integrases constituting data set 1 (table 1). This data set comprises hyperthermophilic integrases, 54 from *Thermococcales* and 14 from *Archaeoglobales*, and 5 mesophilic integrases from *Methanosarcinales* (table 1 and supplementary table S1, Supplementary Material online). To decipher the evolutionary relationship between these integrases, we built similarity networks including data set 1 and all previously known suicidal integrases, using two levels of sensitivity and random walk as detailed in Materials and Methods (fig. 1). The lower similarity threshold assigned all the integrases from data set 1 to the same cluster while excluding pTN3-like integrases and SSV-like integrases (fig. 1*A*). A more stringent similarity threshold applied to data set 1 clustered into data set 2 most hyperthermophilic integrase from *Thermococcales* and *Archaeoglobales*, constituting data sets 3 and 4, respectively (fig. 1*B*). The integrases of data set 2 were present in 30% of closed *Thermococcales* chromosomes (15/51) and in 50% of closed *Archaeoglobales* chromosomes (4/8). Several genomes even contained several copies of these integrases, up to five for *Archaeoglobus profundus* DSM5631. Remarkably, we did not detect any multiple or tandem

Table 1. Integrase Data Sets Used in This Study.

Integrase Data Set	Number of Integrases	Tandem Integration Number	Host Phyla	Data Set Description
1	73	0	*Thermococcales*, *Archaeoglobales*, and *Methanosarcinales*	Larger data set. All suicide integrases clustering with Int^{pT26-2} in network A (fig. 1*A*)
2	62	0	*Thermococcales* and *Archaeoglobales*	Subset of data set 1. Clustering integrases in network B (fig. 1*B*)
3	54	0	*Thermococcales*	Subset of data set 2. *Thermococcales* integrases
4	8	0	*Archaeoglobales*	Subset of data set 3. *Archaeoglobales* integrases

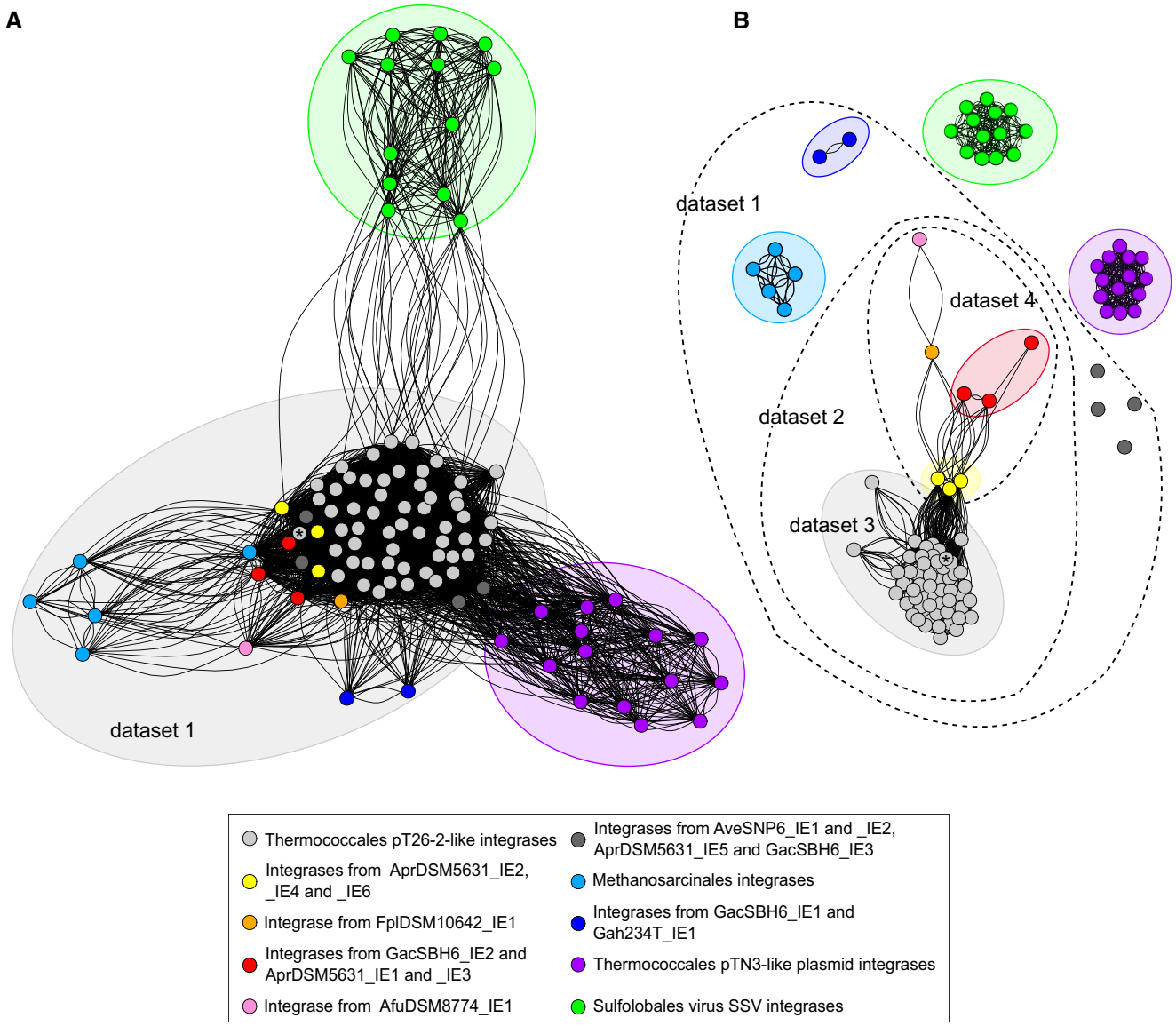

FIG. 1. Archaeal suicidal integrases similarity network. All available archaeal suicidal integrases identified as described in Materials and Methods were analyzed through a similarity network. Each dot corresponds to a protein. A random walk algorithm was used for protein clustering. For both networks, proteins are colored depending on their clustering as indicated in the boxed legend. The star points to Int^{pT26-2}. The data sets defined in table 1 are indicated. (*A*) Links between two proteins refer to a BlastP pairwise similarity >25% over 60% of the protein. (*B*) Pairwise similarity >35% over 60% of the protein.

integration events at the same chromosomal site (supplementary table S1, Supplementary Material online). Additionally, the integrases from data set 2 originate from organisms with an optimal growth temperature >75 °C (supplementary table S1, Supplementary Material online). The very high prevalence rate of the members of data set 2 denoted the extreme pervasiveness of these integrases. This observation prompted us to investigate whether particular biochemical properties of these enzymes could explain their pervasiveness.

Selection of a Suicidal Integrase and Its Target Sites for Biochemical Analysis

The majority of the suicidal integrase-coding genes contained in data set 2 consists of pseudogenes generated by the insertion of MGE sequences into host chromosomes. Due to the fact that pseudogenes rapidly accumulate deleterious mutations (Liu et al. 2004), we selected an intact integrase gene encoded by the replicative form of plasmid pT26-2 in *Thermococcus* sp. 26-2 (Soler et al. 2010). The genomic analysis of the host chromosome revealed the additional presence of an integrated copy of pT26-2. DNA sequence comparison between the plasmid sequence and the extremities of the integrated copy identified the attachment sites of plasmid pT26-2 (fig. 2*A*). The attP site corresponded to a portion of the integrase-coding gene as expected for suicidal integrases. The chromosomal attachment site (attB) was found in a gene coding for a $tRNA^{Arg}$(TCT). The identification of these sequences allowed us to reconstitute the molecular integration scenario of plasmid pT26-2 into its host chromosome (fig. 2*B*).

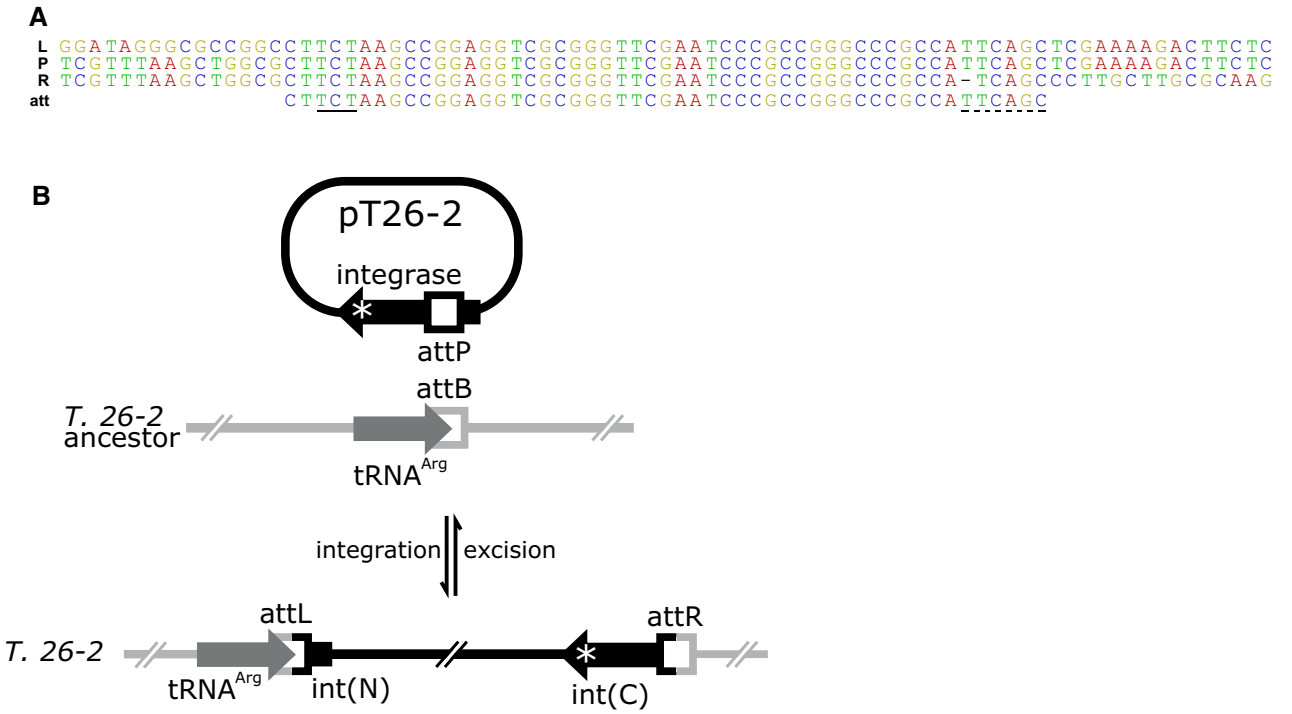

FIG. 2. Plasmid pT26-2 integration model. (*A*) Alignment of the pT26-2 attP (P) sequence with the attL (L) and attR (R) sequences from *Thermococcus* sp. 26-2. The conserved sequence is the attachment site (att) that corresponds to the 3′ end of a $tRNA^{Arg}$(TCT) gene. The anticodon sequence is underlined. The attB site starts two nucleotides upstream of the anticodon sequence and extends six nucleotides downstream of the tRNA gene (dotted line). This sequence identity between attP and attB extends over 51 bp. The sequences of the integrase and tRNA genes are antiparallel. (*B*) Plasmid pT26-2 was present as a freely replicating element and integrated in the chromosome of *Thermococcus* sp. 26-2. The chromosomal attachment site (attB) corresponds to a the $tRNA^{Arg}$(CTC) gene (in gray). Upon integration, the integrase gene is split in two parts named int(C) and int(N). The catalytic tyrosine residue (*) is located in the int(C) part.

Int^{pT26-2} Can Catalyze All Three Canonical Site-Specific Recombination Activities

To investigate the enzymatic properties of this integrase, we overproduced in *E. coli* strep-tagged versions of Int^{pT26-2} and of the Int^{pT26-2}Y327F variant, where the catalytic tyrosine is substituted by a phenylalanine. We purified these enzymes and tested their in vitro recombination activities using synthetic DNA substrates. We designed the synthetic Int^{pT26-2} recombination site carried by plasmid pCB568 as the entire sequence of the $tRNA^{Arg}$(TCT) gene followed by six nucleotides downstream (fig. 2*B*). The most straightforward assay to rapidly assert the activity of purified Int^{pT26-2} was an in vitro integration reaction. The integrase-catalyzed recombination between identical supercoiled plasmids carrying a single att site was monitored as described previously (Cossu et al. 2017) through the formation of plasmid dimers and higher order multimers (fig. 3*A*). Int^{pT26-2} was capable of efficiently catalyzing site-specific integration in vitro. The capacity of Int^{pT26-2} to promote excision was assayed in a recombination reaction using supercoiled plasmid pCB596 carrying two att sites in direct orientation followed by endonuclease restriction. This excision reaction effectively produced two smaller circular DNA molecules each containing a single att site (fig. 3*B*). The substrate and the excised products were easily discriminated via their respective restriction pattern (fig. 3*B*). The comparable efficiency of the IntpT26-2-promoted integration and excision suggested that these reactions do not require additional helper proteins contrarily to phage lambda excision and virus SNJ2 integration (Abremski and Gottesman 1981; Wang et al. 2018). Intramolecular DNA inversion constituted the third canonical site-specific recombination reaction. We assayed the Int^{pT26-2} inversion activity on supercoiled plasmid pCB598 containing two att sites in opposite orientations in a recombination reaction followed by endonuclease restriction. In the presence of Int^{pT26-2}, the assay produced the inversion of the sequence delimited by the att sites readily identified by restriction pattern analysis (fig. 3*C*). The Int^{pT26-2}Y327F variant was unable to produce detectable site-specific recombination for all three canonical reaction (fig. 3*A*–*C*). The reported activity of several integrases demonstrated a high or mandatory requirement for negatively supercoiled templates (Mizuuchi et al. 1978; Reed 1981). In hyperthermophilic archaeal cells, the topological state of DNA is still conjectural even if some reports favor a relaxed chromosome (Lopez-Garcia and Forterre 1997). The inversion and integration activities of Int^{pT26-2} were therefore tested on supercoiled, linear, and relaxed DNA templates but no marked preference was observed for a particular topological state (supplementary figs. S1 and S2, Supplementary Material online). All three positive recombination assays demonstrated that Int^{pT26-2} is a fully functional tyrosine recombinase able to catalyze efficient DNA topology-independent site-specific integration, excision, and inversion in vitro in the absence of additional cofactors. We then explored other biochemical properties of Int^{pT26-2}

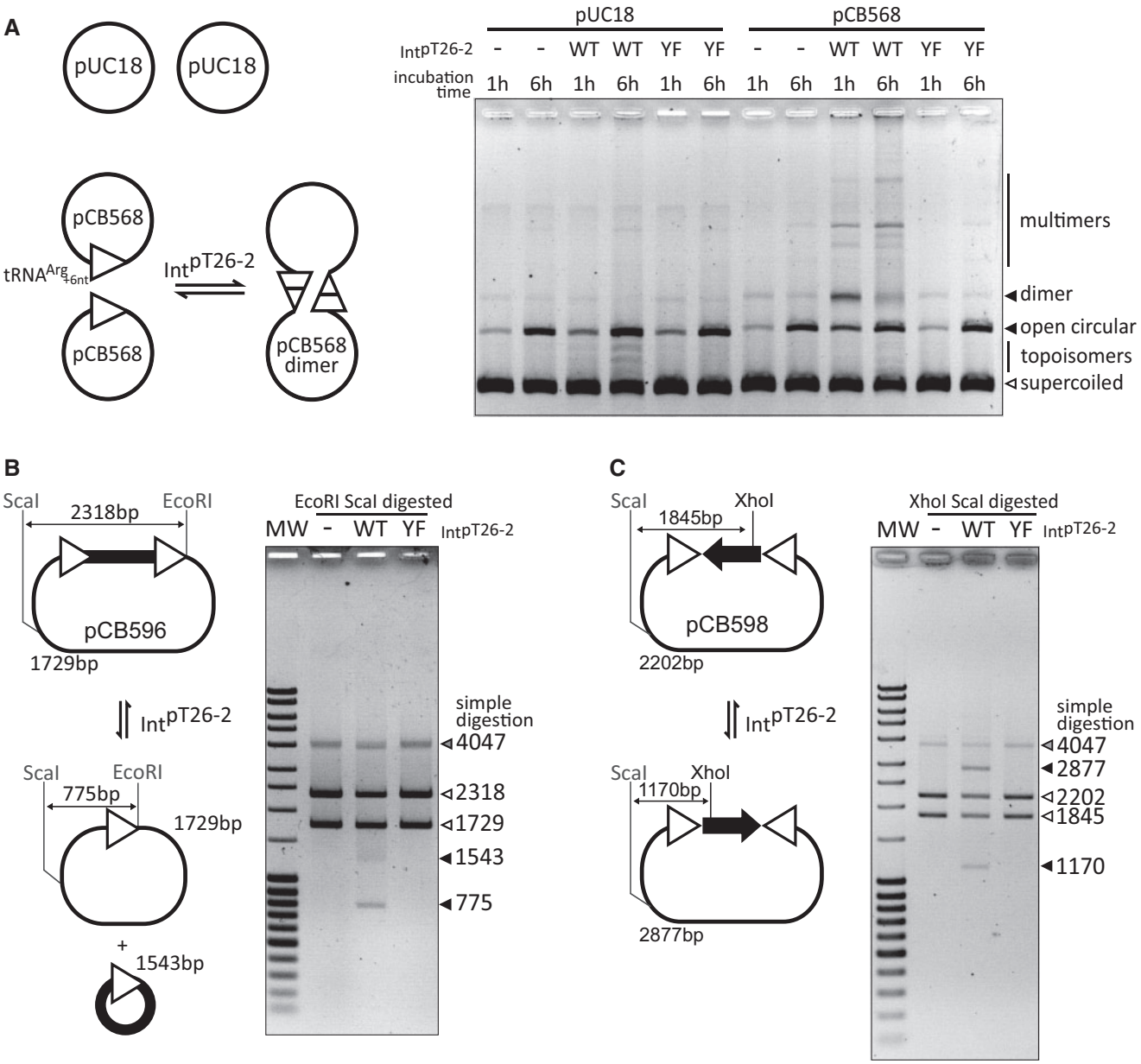

FIG. 3. Int^{pT26-2} site-specific recombination assays for the three canonical activities: integration, excision, and inversion. The recombination model is presented for each activity assay. (*A*) Integration. Recombination between two att sites (triangles) carried by two identical plasmids pCB568 producing plasmid dimers. Plasmid pUC18 without att site cannot undergo site-specific recombination. Plasmids containing zero or one att site were incubated with purified Int^{pT26-2} (WT) or variant Int^{pT26-2}Y327F (YF) at 75 °C during 1 or 6 h. Samples were treated with proteinase K and separated on agarose gel. The Y recombinase and att site are necessary and sufficient for the integration activity. Wild-type Int^{pT26-2} introduces topoisomers in supercoiled templates devoid of att site such as pUC18. This indicates that IntpT26 2 can perform the first step of recombination, that is, nonspecific single-stranded cleavage, followed by religation of the nonspecific substrate, leading to the formation of a topoisomer ladder. (*B*) Excision. Intramolecular recombination between two att sites in direct orientation leading to the formation of two plasmids (excision) with one att site each. Different ScaI-EcoRI restriction identify substrate and products. Plasmid pCB596 was incubated with WT or YF at 75 °C during 2 h and digested with ScaI and EcoRI. (*C*) Inversion. Intramolecular recombination between two att sites in inverted orientation leads to the inversion of the intervening segment. The substrate and product have different ScaI-XhoI restriction patterns. Plasmid pCB598 was incubated with WT or YF at 75 °C during 2 h and digested with ScaI and XhoI.

to explain their pervasiveness among hyperthermophilic archaea.

Int^{pT26-2} att Site Extremities Are Not Highly Stringent

In order to identify the attachment site requirements for positive Int^{pT26-2} recombination, we produced a set of nested deletions starting from a full-length tRNA^{Arg} gene followed by six additional nucleotides used in the above in vitro reactions (fig. 4). Surprisingly, this test did not provide clear-cut limits for the Int^{pT26-2} att site. At the 5′ end, we observed a progressive reduction in recombination efficiency: the segments L56, L55, and L54 are positive for recombination, whereas

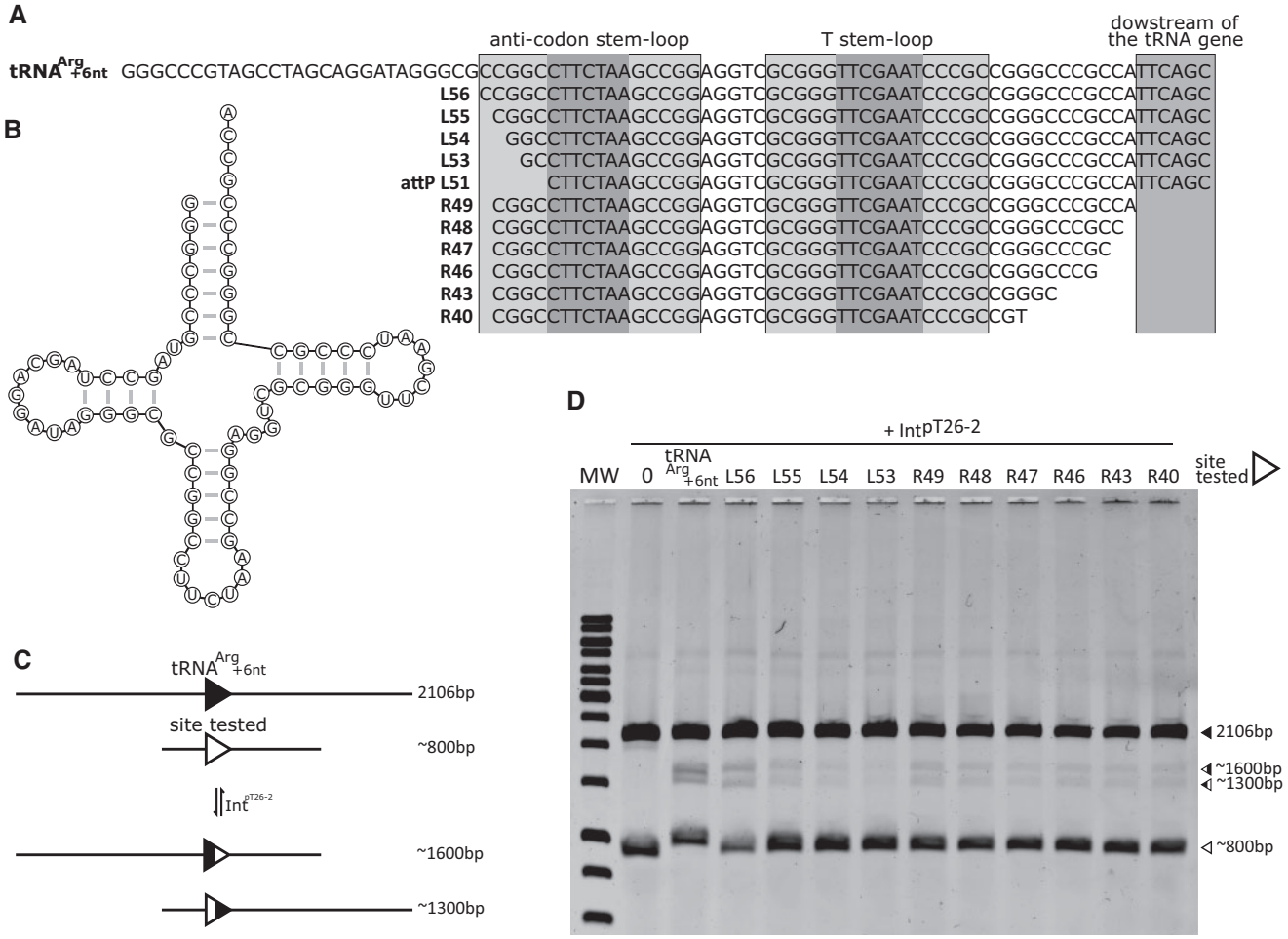

FIG. 4. Minimal att recombination site. (*A*) A nested deletion set of tRNA^{Arg}(TCT) sequences were tested as substrates for Int^{pT26-2} recombination. (*B*) The leaf-like structure of *T. 26-2* tRNA^{Arg}(TCT) is presented. (*C*) The set of nested sequences was tested for recombination against a full-length tRNA^{Arg}(TCT) gene plus 6 nt downstream. When recombination occurs, two chimeric linear substrates of intermediate size are produced. (*D*) The two linear substrates were incubated with purified Int^{pT26-2} for 2 h at 75 °C, treated with proteinase K, and run on an agarose gel.

sequence L53 is weakly active (fig. 4*D*). At the 3′ end, a wide range of sequences exhibited a barely detectable gradient of reduced recombination. The observed trend also strongly suggested that the attP site detected in silico (L51) is not recombination-proficient in vitro. Previously characterized archaeal integrases such as Int^{pTN3} and Int^{SSV2} recognize att sites of different lengths and sequence but always show an abrupt loss of activity when few nucleotides are removed from the shortest recombination-proficient DNA substrate (Zhan et al. 2015; Cossu et al. 2017). Contrarily to what was observed previously, it appeared that Int^{pT26-2} retained partial activity over a remarkably wide range of recombination site deletions.

IntpT26-2 Is Active at Near-Boiling Water Temperature

The natural hosts of the pT26-2 plasmid belong to *Thermococcales* which actively grow up to 95 °C therefore constituting some of the most hyperthermophilic organisms known to date. The particular distribution of plasmid pT26-2 raised the question whether the recombinase activity of Int^{pT26-2} was optimized for, and restricted to, high temperatures. All in vitro Int^{pT26-2} activity assays described above were performed at a near optimal 75 °C which constituted the highest documented temperature for in vitro site-specific recombination. The optimal reported temperature for other hyperthermophilic recombinases never exceeded 65 °C (Cortez et al. 2010; Zhan et al. 2015; Cossu et al. 2017). It was therefore of great interest to test whether the Int^{pT26-2} integrase would catalyze recombination reactions at yet higher temperatures. We performed the inversion assay described in figure 3C across a wide range of incubation temperatures, from 60 to 99 °C (fig. 5). Interestingly, Int^{pT26-2} was able to efficiently catalyze site-specific recombination over the whole temperature range, whereas the maximal amount of recombination product was obtained between 75 and 80 °C (fig. 5*A*). Temperature-dependent DNA degradation was accounted for in the reaction (fig. 5*B*). Remarkably, the inversion product was still observed at 99 °C, which was the highest temperature we could assay at atmospheric pressure and constituted the highest reported temperature for the activity of a tyrosine recombinase.

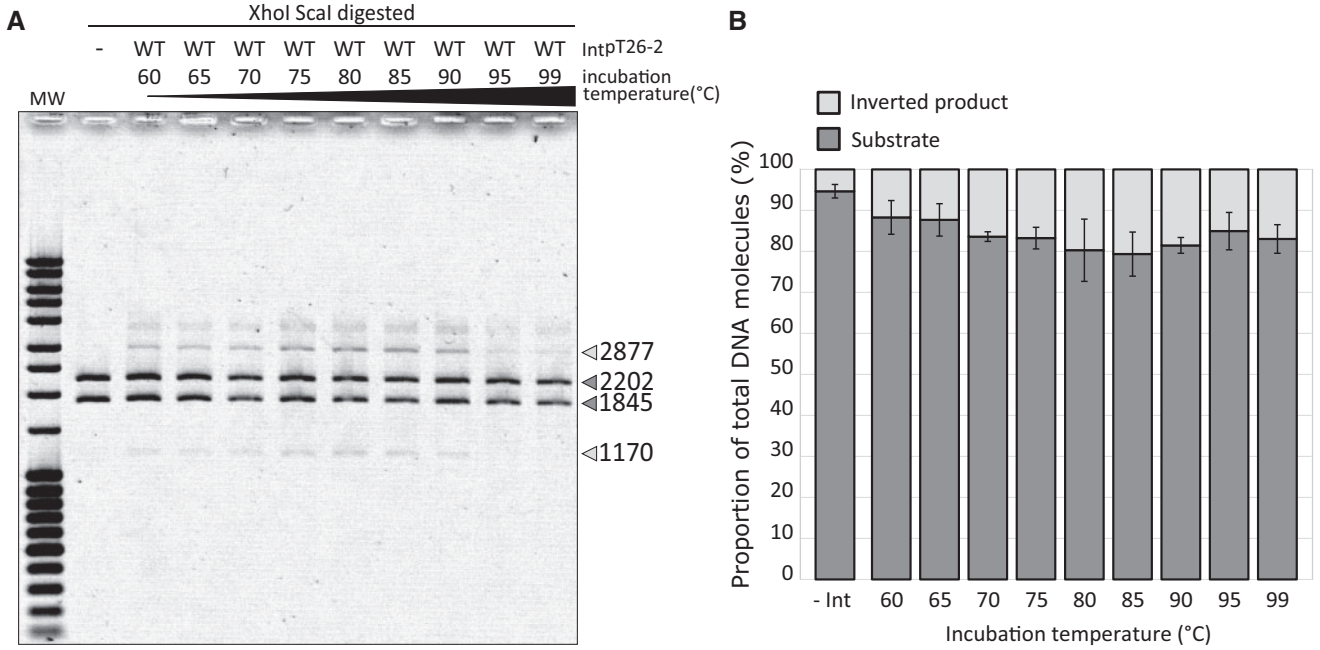

FIG. 5. Temperature activity range of Int^{pT26-2}. The inversion assay presented in figure 3C was used to test the temperature activity range of Int^{pT26-2}. (*A*) Plasmid pCB598 was incubated with purified Int^{pT26-2} at different temperatures during 0.5 h and digested with ScaI and XhoI. (*B*) Template DNA was decaying probably due to thermal degradation. To take degradation into account, we quantified the substrate/product ratio in three replicate experiments which demonstrated an optimal inversion rate between 80 and 85 °C. Relative amounts of substrate and product were calculated for each lane, in triplicate. The error bar represents a 95% confidence interval. The difference between apparent and real in vitro IntpT26 2 optimal recombination temperatures was therefore due to DNA degradation at the highest temperatures.

Limited Choice of Integrases among Highly Variable Hyperthermophilic MGEs

In order to correlate the particular properties of Int^{pT26-2} with their widespread presence among hyperthermophilic archaea, we first analyzed the MGEs from which they originate by comparative genomics. The more stringent network analysis clearly restricted the distribution of Int^{pT26-2} homologs from data set 2 to hyperthermophilic Euryarchaeota. We uncovered a total of 47 different MGEs, most of them being integrated elements except virus TPV1 (Gorlas et al. 2012), plasmids pT26-2 (Soler et al. 2010), pGE2 (Badel et al. 2019), and pIRI06c from *Thermococcus* IRI06c (Courtine D, personal communication). These MGEs were highly variable in size, from 8 to 38 kb and in genetic content. Based on their genetic content (see Materials and Methods), these MGEs could be ranked in four families related to plasmid pT26-2 (25/47), plasmid pAMT11 or TKV1 (8/47) (Gonnet et al. 2011), fuselloviruses encoding the same major capsid protein (8/47) (Krupovic et al. 2014) or without known relative (6/47) (supplementary table S1 and fig. S3, Supplementary Material online). Overall, a wide range of different elements, plasmids and viruses, recruited integrases from data set 2 further underlining the pervasiveness of these recombinases.

Assessing the Diversity of Int^{pT26-2}-Related Hyperthermophilic Suicidal Integrases and Their Targets

To investigate the relationships between the more closely related integrases of data set 2, we built a phylogenetic tree (fig. 6). Based on the network analysis results presented in figure 1, we rooted the tree between the *Thermococcales* integrases (data set 3) and the eight *Archaeoglobales* integrases from data set 4. The distal branches of *Thermococcales* were well resolved even if *Archaeoglobales* and basal *Thermococcales* were poorly supported. *Archaeoglobales* integrases displayed long branches in the phylogenetic tree which did not permit us to infer evolutionary relationships (fig. 6). The *Thermococcales* integrases of data set 3 presented a mixture of closely related and divergent enzymes providing the opportunity to study integrase evolution at different scales. To assess more precisely the evolution history of these enzymes, we superimposed over the integrase phylogeny their respective chromosomal integration site and their MGE family as defined above (fig. 6).

The sequences of the attB and attP target sites were identified as the direct DNA repeats bordering integrated elements (attL and attR) or by comparing episomal MGEs with their host chromosome. The attachment sites of all integrases from data set 1 consisted of the 3′ end of various tRNA genes without supplementary loop. As a notable exception, the integrases from the *Archaeoglobales* elements AprDSM5631_IE2 and AveSNP6_IE2 recombined att sites with a supplementary loop, whereas element AveSNP6_IE1 recombined at the 5′ end of its tRNA gene target (supplementary fig. S4, Supplementary Material online). The 54 *Thermococcales* integrases from data set 3 used 14 different tRNA genes for integration, whereas the 14 *Archaeoglobales* integrases from data set 4 used 9 different tRNA genes reflecting a flexible integration specificity (supplementary table S1,

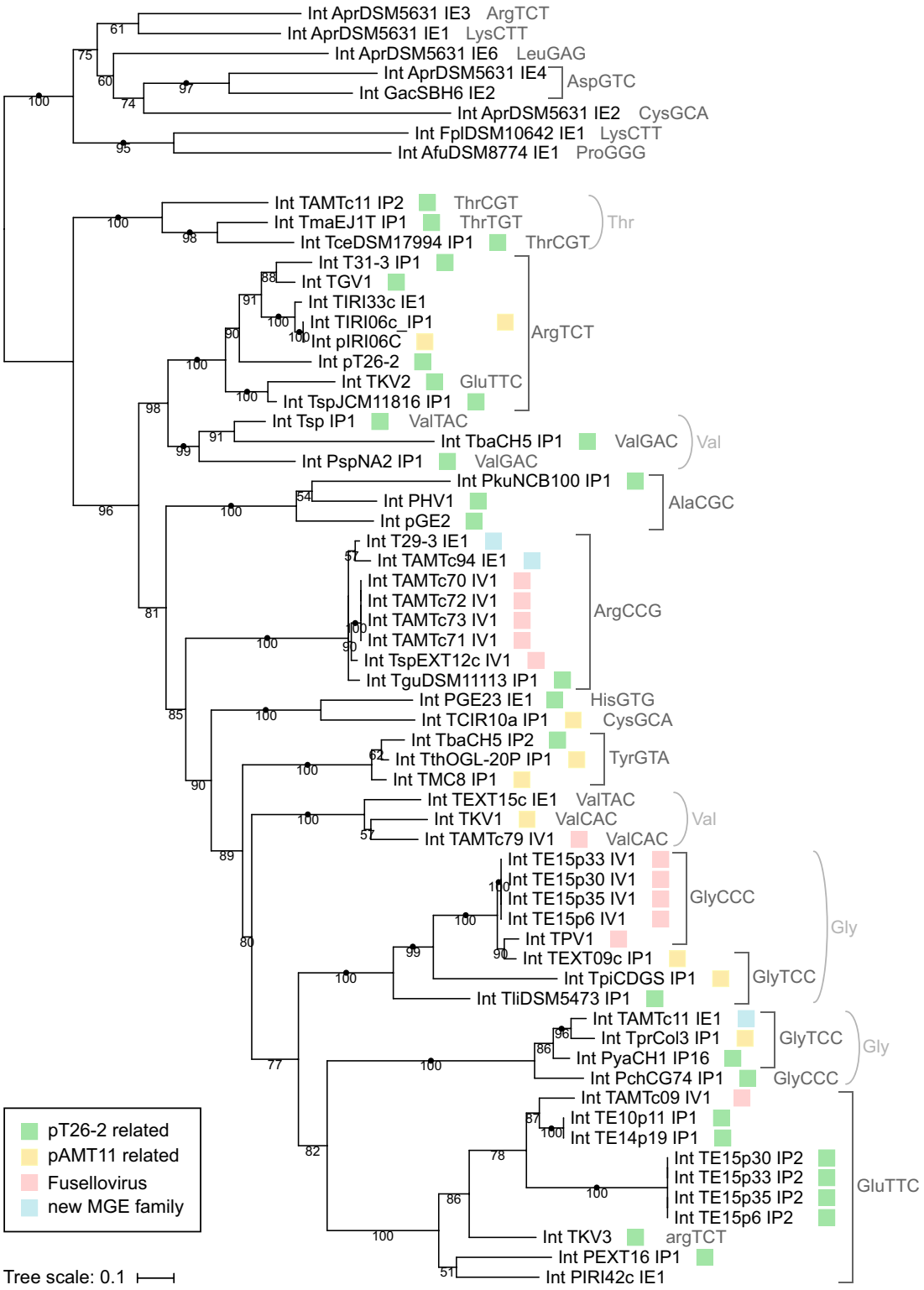

FIG. 6. Maximum likelihood phylogenic tree of the integrases from data set 2. Branch values represent the posterior probability. Branches supported by both the posterior probability and ultrafast bootstrap (>95%) are indicated by a black dot. The integrated element classification is color-indicated when known, see also supplementary table S1, Supplementary Material online. The individual tRNA genes used for integration are indicated as well as their anticodon sequence. The scale bar represents the average number of substitutions per site.

Supplementary Material online). All tRNA without supplementary loop of a given organism such as *T. kodakarensis*, displayed a more conserved sequence downstream of the anticodon (73% mean pairwise similarity) than upstream (63%) (supplementary data file S2*A* and *B*, Supplementary Material online). All att sites of the *Thermococcales* integrases of data set 3 shared 75% mean pairwise identity (supplementary data file S2C, Supplementary Material online), whereas a

portion of the 3′ region, the T stem-loop, was even more conserved, at 90% (supplementary data file S2*D*, Supplementary Material online). In *T. kodakarensis*, all T stem-loops shared 85% similarity (supplementary data file S2*G*, Supplementary Material online).

It is to be noted that for a given integrated element, the attL and attR sequences might differ. We evidenced one such case of nonspecific integration with the *Thermococcales* element TspEXT12c_IV1 integrating in a $tRNA^{Arg}$(CGC) gene (supplementary fig. S5, Supplementary Material online). The attL and attR sequences of this element presented a single A–G nucleotide mismatch at the tip of the tRNA T loop (supplementary fig. S6*A*–*C*, Supplementary Material online). Both the A and the G alleles were found for $tRNA^{Arg}$(CGC) in *Thermococcales* (supplementary data file S3, Supplementary Material online) therefore ruling out sequencing errors or random mutations. Strikingly, the sequences corresponding to attL and attR were also present in the $tRNA^{Arg}$(TCG) gene of *Thermococcales* (supplementary data file S3, Supplementary Material online).

Differential Evolution History of the N- and C-Terminal of Suicidal Integrases and Their Targets

Suicidal integrases share as common characteristic to use part of their own gene as attP integration site. The integrase protein sequence at the junction between the Int(N) and Int(C) moieties therefore corresponds to the translation of the att site (fig. 1). The phylogenetic analysis presented in figure 6 showed that, as a general rule, closely related integrases targeted the same tRNA gene for integration. In a number of cases, however, cognate integrases appeared to have switched specificity resulting in a substantial modification of their amino-acid sequence. In order to understand this phenomenon, we subdivided each integrase gene from data set 3 into three parts: 5′ end, attP, and 3′ end and compared them and their respective translation to the other members of the data set. The results revealed five different patterns of unusual sequence conservation both at the DNA and protein level (fig. 7 and supplementary fig. S6, Supplementary Material online). The first case demonstrated the acquisition of different att sites, the tRNAVal(CAC) and tRNATyr(GTA) genes, respectively, by the distantly related integrases of TKV1 and TthOGL-20P_IP1. These target sites were very likely acquired via two independent events (fig. 7*A* and supplementary fig. S6*A* and *B*, Supplementary Material online). The second case illustrated the recruitment of an identical tRNAVal(TAC) att site by the two phylogenetically distant integrases from Tsp_IP1 and TEXT15c_IE1 (figs. 7 and 8*B*). The two att sites exhibited different lengths and three nucleotide mismatches giving rise to a different protein sequence in the corresponding segment (supplementary fig. S6*C* and *D*, Supplementary Material online). The att site similarity presumably constituted a convergence due to the limited pool size of the possible tRNA genes for integration rather than a character inherited from their common ancestor, explaining the variation in att site size and translation. In the third case, the closely related integrases of pT26-2 and TGV1 shared the same specificity for a tRNAArg(TCT) gene (fig. 7*C* and supplementary fig. S6*E* and *F*, Supplementary Material online). These proteins exhibited high amino-acid similarity (>70%) (fig. 7*C*) as reflected by their proximity in the phylogenetic analysis (fig. 6). On the other hand, the amino-acid sequences corresponding to their respective att site were strikingly different. This difference was caused by two translation frameshifts occurring immediately upstream and downstream the att site, accounting also for a slight difference in site length (supplementary fig. S6*E* and *F*, Supplementary Material online). Surprisingly, in its phylogenetic clade, the IntpT26-2 integrase was the only one exhibiting these frameshifts therefore suggesting a single att site acquisition for all clade members followed by a unique shifting event for one member. A similar situation of frameshifting was observed in a fourth case for the integrases of PIRI42c_IE1 and TE10P11_IP1 even if it resulted in similar glycine and proline-rich sequences due to the high GC content of the att site (fig. 7*D* and supplementary fig. S6*G* and *H*, Supplementary Material online). Notably, these proteins and their respective gene exhibited differential sequence conservation upstream and downstream the att region, suggesting a hybrid origin for the two moieties. The fifth case also illustrated the recombinant nature of these enzymes. Integrases originating from two different phylogenetic clades and carried by TE15P30_IV1 and TpiCDGS_IP1 opted for att sites in the related $tRNA^{Gly}$(CCC) and $tRNA^{Gly}$(TTC) genes (fig. 7*E* and supplementary fig. S6*I* and *J*, Supplementary Material online). Contrarily to other integration events, the in silico reconstituted integrase genes of TpiCDGS_IP1 carried a frameshift mutation due to a missing nucleotide in the attachment site. The presence of this mutation was confirmed by sequence read mapping of the *T. piezophilus* CDGS genome (kindly provided by the original authors) (Dalmasso et al. 2016). This situation constituted the exact converse of the differential sequence conservation upstream and downstream the att region observed in the fourth case. The integration of TpiCDGS_IP1 in *T. piezophilus* CDGS further exposed the recombination mechanism involved in the evolution of suicidal integrases.

Taken together with our in vitro data demonstrating $Int^{pT26\text{-}2}$ relaxed target recognition, the succession of cases presented here suggested the presence of an efficient mechanisms for the evolution and specificity switch of suicidal integrases.

Discussion

Suicidal integrases carry their attP DNA recombination site within their own coding sequence. The site-specific recombination reaction with a compatible attP target on the host chromosome causes the disruption—or suicide—of the integrase gene into two inactive stumps. These pseudogenes cannot produce active integrase and therefore prevent MGE excision. Intuitively, episomal MGEs encoding such suicidal integrases would become irreversibly bound to their host genome, incapable of producing further rounds of infection and eventually disappear. Strikingly, we observed recently that the pT26-2 plasmid family encoding such integrases was worldwide distributed and pervaded archaeal populations

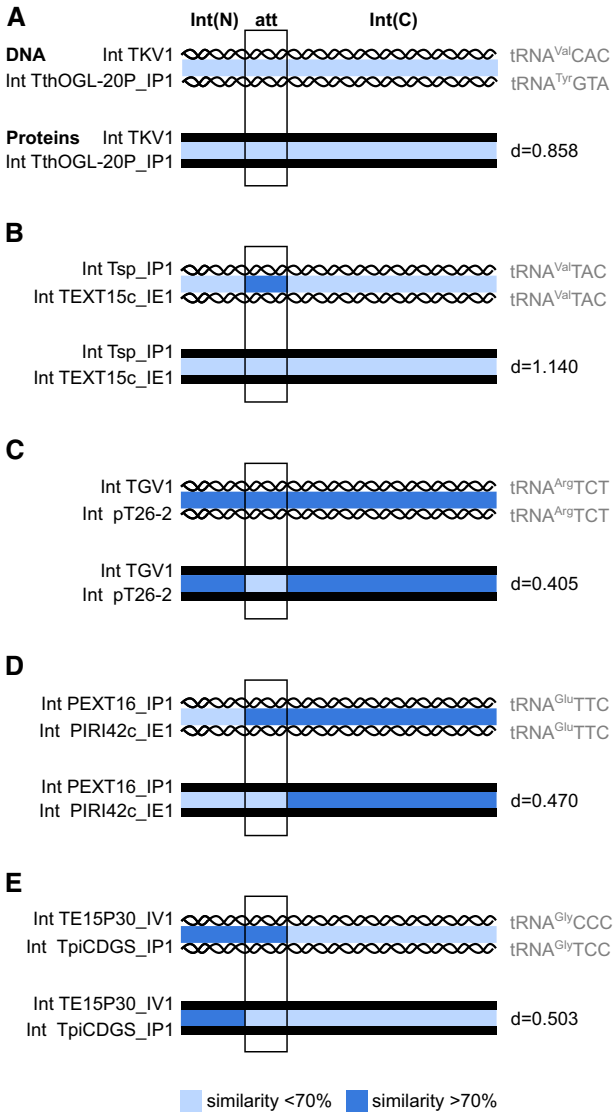

FIG. 7. Independent evolution of integrases and their target sites. For suicidal integrases, the att site is located inside the gene coding for the integrase and is therefore translated along with the integrase. Different cases illustrating the independent evolution of the integrases of data set 3 and their respective target sites are summarized here. Gene sequences (DNA) or integrase protein sequences (proteins) were aligned. Mean pairwise similarity over the Int(N), att, or Int(C) regions is indicated by a color scale. High similarities (>70%) are indicated in dark blue. Lower similarities (<70%) are indicated in light blue. The 70% cutoff was selected because it corresponds to the similarity between the closely related integrases from elements TGV1 and pT26-2. The phylogenetic distance (d) between proteins is calculated in the same units as in figure 6. (*A*) General case: completely divergent integrases at the DNA and protein levels. (*B*) Two divergent integrases sharing the same att site but translated in different frames. (*C*) The integrases are closely related as indicated by their similar gene and protein sequences. The same att sequence is translated in a different frame. (*D*) The two integrases are closely related at their Int(C) as indicated by their similar gene and protein sequences but with divergent Int(N) segments. Similar to (*C*), the att sequence translation is different between the two proteins, due to a frameshift. (*E*) The two integrases are closely related at their Int(N) as indicated by their similar gene and protein sequences but with divergent Int(C) segments. The att site is translated in a different frame. Complete att site and protein alignments are available in supplementary figure S7, Supplementary Material online.

both as freely replicative plasmids or integrated elements (Badel et al. 2019). We were interested in solving this apparent paradox using complementary approaches using phylogenomics and in vitro enzyme characterization. We examined if the particular in vitro recombination properties of one of these enzymes could convey some form of selective advantage to the MGE or its host and provide clues for the evolutionary success of these suicidal integrases.

Suicidal Integrases Are Active at Boiling Water Temperature and Present Relaxed Integration Site Specificities

We have selected Int$^{pT26\text{-}2}$, the prominent integrase from data set 2 to conduct a series of in vitro recombination tests. Our results indicated that this enzyme could outstandingly catalyze all the canonical reactions involved in site-specific recombination over a wide range of temperatures up to 99 °C

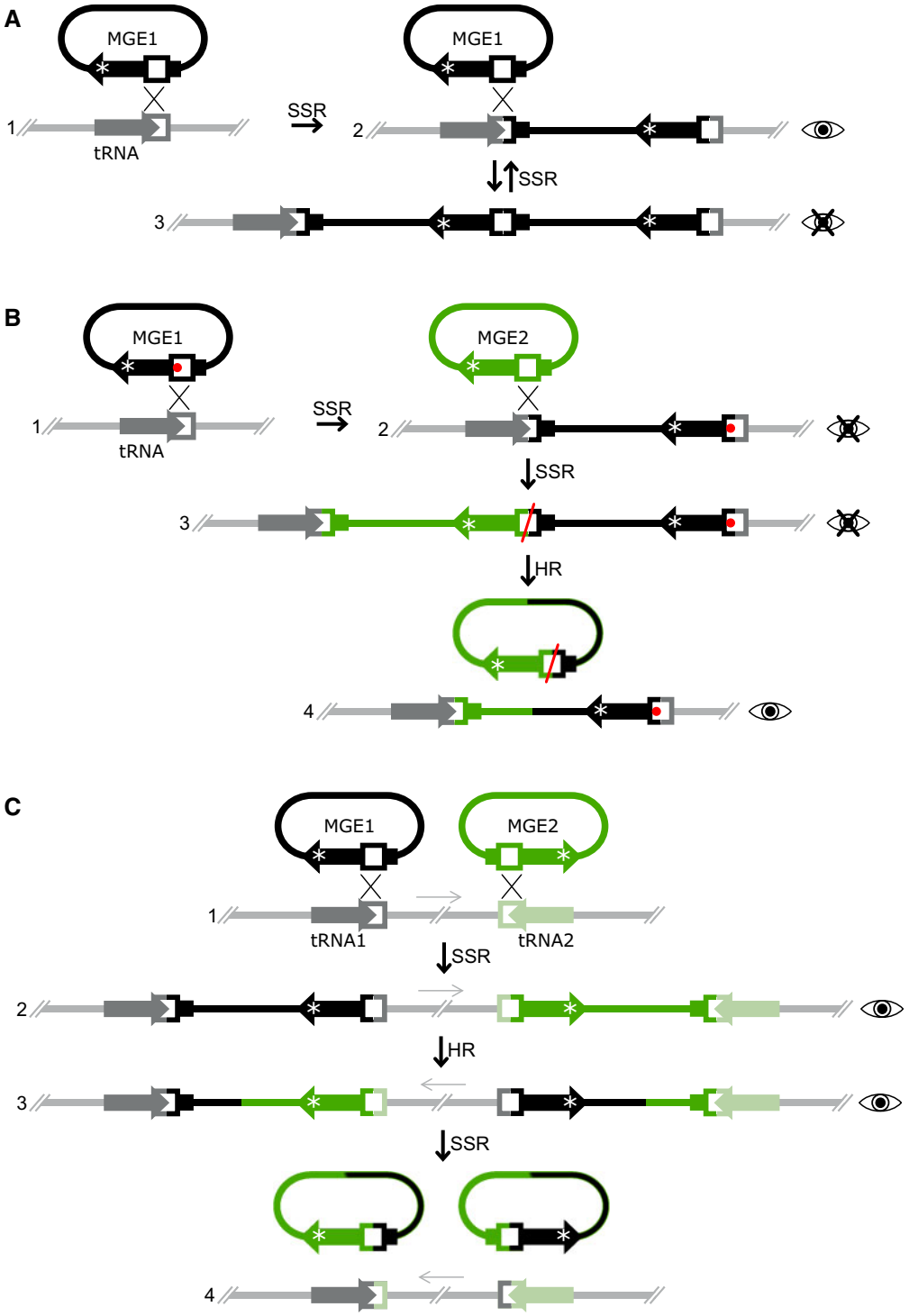

FIG. 8. Model for the formation of hybrid integrases. (*A*) Tandem insertion of the same MGE in the same tRNA gene target reconstituting a functional integrase gene able to excise the element. Identical tandem insertions have never been observed. (*B*) A first MGE integration event generated an attR site with a single-nucleotide deletion as compared with the original $tRNA^{Gly}$ gene (red dot) (supplementary fig. S6*I* and *J*, Supplementary Material online). The second integration event involved a related MGE but with a more distant integrase. This integration generates an inactive integrase gene (red bar) due to frameshifting. Homologous recombination between related MGE backbones could have excised a hybrid plasmid leading to the situation observed for the integrated TpiCDGS_IP1. The Int(N) and Int(C) segments of its integrase have a different evolution history and cannot be assembled due to a mirrored frameshift in the att region. (*C*) Multiple MGE integration events at separate chromosomal locations and in inverted orientation can give rise to a large genomic inversion by homologous recombination between related MGE backbones as reported (Gehring et al. 2017). This inversion generates hybrid MGEs which could excise by the means of a compatible integrase provided in trans via superinfection. The asterisk refers to the codon of the catalytic tyrosine. The eye icon indicates whether particular MGE forms were observed and described.

(figs. 3 and 5). We reported previously the in vitro characterization of Int^{pTN3}, the first integrase isolated from *Thermococcales* and capable of catalyzing site-specific recombinations as well as low-sequence specificity recombination reactions with the same outcome as homologous recombination events (Cossu et al. 2017). Here, we showed that Int$^{pT26\text{-}2}$ did not carry the additional subdomains found in Int^{pTN3} and performed exclusively site-specific recombination reactions. Additional hyperthermophilic site-specific recombinases have been characterized and their activity was assayed in vitro at a maximal temperature never exceeding 65 °C (Serre et al. 2002; Zhan et al. 2015; Cossu et al. 2017; Jo et al. 2017). Additionally, these enzymes encoded by self-replicating mobile elements infect hosts with optimal growth temperatures of 85 °C at the most. On the other end, the integrases from the data set 2 are encoded by MGEs infecting hosts with much higher optimal growth temperatures, up to 105 °C as reported for *Pyrococcus kukulkanii* NCB100 (Callac et al. 2016). Integrases such as Int$^{pT26\text{-}2}$ are therefore particularly well suited to efficiently catalyze integration and spread in environments with a wide range of temperature, including extreme hyperthermophilic conditions.

The integration module conveyed by suicidal integrases is much simpler than what is found in most MGEs. The integration module of bacteriophage Lambda is composed of the integrase gene, a separate att site, and additional genes encoding RDFs to avoid spontaneous MGE excision (Landy 2015). In contrast, the suicidal Int^{pTN3} and Int^{SSV2} integrases were shown to promote in vitro excision without RDFs (Zhan et al. 2015; Cossu et al. 2017), the disruption of their gene upon integration acting as directionality regulator. This property was also confirmed for Int$^{pT26\text{-}2}$, which was able to perform both in vitro integration and excision reactions with comparable efficiencies (fig. 3). The compactness of an integration module not requiring RDFs and carrying attP imbedded in the integrase gene constituted very likely a strong advantage to explain the pervasiveness of suicidal integrases among related organisms.

The in vitro characterization of Int$^{pT26\text{-}2}$ revealed an additional peculiar property of this integrase regarding its target site. The reported recombination activity of other archaeal integrases such as Int^{pTN3} and Int^{SSV2} was impaired as soon as very few nucleotides were removed from their target substrate (Zhan et al. 2015; Cossu et al. 2017). By assaying the recombination activity of Int$^{pT26\text{-}2}$ on nested deletions of attB, we observed that the requirements for a specific site were far less stringent. This integrase was active over a wide range of site deletions as long as the core site was present. In these experiments, the last ten nucleotides present in both attB and attP and corresponding to the arm sequence were not crucial to allow site-specific recombination (fig. 4). These observations suggested that recombinases promiscuous in site selection could target various tRNA gene locations of the same genome or even different related hosts. The high occurrence of this type of integrase in data set 1 suggested a selective advantage of mobile elements carrying such a promiscuous integrase. A pertinent phylogenomic analysis confirmed these observations for the entire Int$^{pT26\text{-}2}$ integrase data set.

Integrases from data set 1 showed the capacity to target a high variety of sites on the host chromosome namely 18 out of the 46 possible tRNA genes, either at the 5′ or at the 3′ end (supplementary fig. S5, Supplementary Material online). These attachment sites consisted for the vast majority of the 3′ end of these tRNA genes, comprising the T stem-loop which is significantly more conserved than the rest of the tRNA with a mean pairwise identity of 90% (supplementary fig. S4 and data file S2C, Supplementary Material online). We surmised that this conserved T stem-loop constitutes the core attachment site carrying the cleavage and strand exchange positions of the attB × attP recombination reaction. Outside of this conserved core, the various target sites were more variable both in sequence and length (supplementary data file S3 and fig. S5, Supplementary Material online). The combination of our phylogenomic analyses with the in vitro activity data presented above strongly suggested that all integrases from data set 1 share the intrinsic propensity to easily switch between different att targets with similar core sites.

Thermococcales Integrases Are Not Species-Specific and Are Frequently Exchanged between MGEs

Our phylogenomic analysis investigated the evolution history of hyperthermophilic suicidal integrases composing data set 3 at four different levels. On top of the integrase sequence phylogeny, we superposed their particular target sites and the mobile element of origin (fig. 6). In addition, we correlated integrases and host species (supplementary table S1, Supplementary Material online). The wide distribution of *Thermococcales* integrases we observed among the various types of elements such as fuselloviruses, pT26-2-, or pAMT11-like plasmids and unidentified MGEs can be explained by two evolution histories: 1) the congruence of the phylogenies of the MGES and their associated integrase indicating that theses enzymes diverged from a single common ancestor and coevolved with the mobile element or 2) the exchange of integrases between the different MGE types. Strikingly, very similar integrases (94% mean pairwise similarity) were found in the genomes of very distinct mobiles elements: in fuselloviruses (TspEXT12C_IV1 and TAMTc70_IV1), in a pT26-2-like integrated plasmid (TguDSM11113_IP1), and in unidentified integrated elements (T29-3_IE1 and TAMTc94_IE1) (supplementary fig. S7, Supplementary Material online). Such high similarity values indicated a recent exchange of integrase genes between these integrated elements. However, we could not trace the directionality of the transfer due to the lack of bootstrap support. In a similar process, the pAMT11-related plasmid family presumably captured integrases from data set 3 at least twice independently, in TpIRI06c_IP1 and TprCol3_IP1 (fig. 6). Interestingly, the pAMT11 plasmid described originally did not encode an integrase (Gonnet et al. 2011), suggesting either a corresponding gene loss in this particular plasmid or independent integrase gene acquisitions in the pAMT11-related elements identified in this study. Module exchange between related MGEs is a well-known process (Oberto et al. 1994; Hendrix et al. 2000; Iranzo et al. 2016). In the case of this integrase family, the frequency of genetic exchange or acquisition highlighted the selective

advantage provided by Int^{pT26-2}-related integrases to their respective MGE. Additionally, our phylogenetic analysis indicated clearly that the phylogenies of integrases and host chromosomes are not congruent (fig. 6 and supplementary table S1, Supplementary Material online). *Thermococcales* from distinct genera such as *T. barophilus* CH5 and *Pyrococcus* sp. NA2 harbored very closely related integrases, whereas the distant integrases of elements TKV1, 2, and 3 were found in the same *T. kodakarensis* KOD1 isolate. On the other hand, we observed a limited *Pyrococcus* genus specificity for the integrases of pGE2, PkuNCB100_IP1, and PHV1. Overall, the integrases from data set 3 seemed to be capable of pervading all *Thermococcales*, without species specificity.

Molecular Model for Suicidal Integrase Evolution and Target Site Switching

The Int^{pT26-2} integrase family allows MGE integration in a variety of chromosomal sites and in a wide range of archaeal organisms belonging to three distinct taxonomic orders. These enzymes are uniquely resilient by efficiently switching target specificity. The comparison of all chromosomal attachment sites demonstrated that these integrases target the 3′ end of various tRNA genes which corresponds to their most conserved region. In addition, the in vitro activity analysis of Int^{pT26-2}, the most prominent integrase of this data set, clearly showed a relaxed requirement for specific att site extremities. These two properties certainly contributed to the evolution of these enzymes but were not sufficient to explain the extensive target site exchange among closely related integrases (fig. 6). One would expect that any abrupt att switching would lead to drastic changes in the protein sequence in the att site segment and that these alterations could also extend further downstream due to frameshifting. It can be intuited that in both cases the resulting protein would lose its integrase function. Unexpectedly in data set 2, integrase sequences corresponding to the att site diverged either due to different att sequences or to identical att sites translated in alternate frames. In the latter case, we observed frequent site size variation and the presence of indels bordering the att site. These changes, allowing the restoration of a sense reading frame for the C-terminal end of the protein were often found among closely related integrases and were compatible with our biochemical evidence of relaxed sequence requirement at att borders. In addition, the variability of protein sequence encompassing the att site was somewhat constrained by the extensive conservation of the 3′ end of the target tRNA genes and its high GC content giving rise to proline- or glycine-rich protein segments. Overall, it appeared clearly that protein sequence changes corresponding to the att site did not affect protein function, making specificity switching easier than anticipated.

The aforementioned results and the thorough genomic comparison of 54 chromosomal integration events from data set 1 permitted us to propose a molecular model explaining the prevalence and pervasiveness of suicidal integrases in hyperthermophilic organisms. This model describes the mechanism used for att target switching and is based on successive MGE integrations in the same cellular host. Any integration episode would generate identical attL and attR sequences at its borders while disrupting the suicidal integrase gene (fig. 1). Each of these att sites can be targeted by the same MGE in a second event of integration to produce a tandem integration reconstituting an intact copy of the integrase gene (fig. 8A). This particular situation is prone to efficient excision catalyzed by the intact integrase and has not been observed even in the larger data set 1 nor for other MGEs carrying suicidal enzymes (Redder et al. 2009; Cossu et al. 2017). On the other hand, tandem integration has been observed for MGEs carrying type II nonsuicidal integrases (Krupovic et al. 2010, 2019) as their excision might be regulated by RDFs. The integration instability of suicidal tandem MGEs could also be used to generate new hybrid suicidal integrases as observed in the case of *T. piezophilus* TpiCDGS_IP1 (fig. 7E and supplementary fig. S6I and J, Supplementary Material online).

In our model, the tandem integration of two related MGEs carrying divergent integrases followed by homologous recombination releases a hybrid plasmid carrying a potential frameshift in the reconstituted hybrid integrase gene. This event would leave behind a conversely hybrid MGE integrated in the chromosome and presenting two integrase gene moieties of different origin and in different reading frames. We observed this exact situation for the TpiCDGS_IP1 element (fig. 8B). We have documented additional cases of hybrid integrases displaying separate evolution histories in their Int(N) and Int(C) moieties (fig. 7D and E and supplementary fig. S6G–J, Supplementary Material online). Efficient genomic homologous recombination between cognate integrated copies of MGEs was proposed as a mechanisms for the evolution of fuselloviruses in *Sulfolobales* (Redder et al. 2009), demonstrated more recently by direct sequence analysis in *T. kodakarensis* as discussed below (Gehring et al. 2017) and fully supports this model.

An alternative scenario could also account for the generation of hybrid suicidal integrases. Cognate MGEs carrying various integrases from data set 2 are often found inserted in different locations of the same host chromosome (supplementary table S1, Supplementary Material online) as shown for other MGEs encoding suicidal integrases (She et al. 2004; Wang et al. 2007). Two cognate MGEs integrated in opposite orientations and sharing enough DNA similarity could undergo homologous recombination and generate chromosomal inversions events as reported for the *T. kodakarensis* TKV2 and TKV3 elements (Gehring et al. 2017) (fig. 8C). Such an inversion would bring heterologous attL and attR sites and heterologous integrase moieties into the correct register. A new incoming MGE with a relaxed integrase specificity could excise these recombinant MGEs and generate hybrid integrases with modified target specificities (fig. 8C).

The simple site-specific recombination scheme of suicidal integrase shown in figure 1 seemed to imply that these enzymes which destroy their own gene would be doomed to disappear by leaving only inactive genes relics. This work demonstrated on the contrary that by integrating, these enzymes generated a fertile bed of fast-evolving pseudogenes whose combinations created a wide array of new integrases

able to efficiently target 18 different tRNA genes. Our data showed that this variability, a somewhat relaxed target specificity, a very compact integration module devoid of RDFs and an extreme thermostability very likely accounted for the prevalence and unique pervasiveness of this integrase family in hyperthermophilic archaea. It is well accepted that pseudogenes increase the genetic diversity through recombination and gene conversion (Vihinen 2014). In contrast, the emergence in all organisms of new genes via pseudogenes and transitory proto-genes remains poorly understood (Siepel 2009; Carvunis et al. 2012). By generating pseudogenes and at high frequency, pervasive suicidal integrases could constitute an efficient model to approach the molecular mechanisms involved in the generation of active genes variants by the recombination of proto-genes.

Materials and Methods

Detection of Mobile Elements and Integrase Homologs in Euryarchaea

A classical similarity search in the protein databases to detect proteins closely related to Int$^{pT26\text{-}2}$ could not be implemented because SSV-type integrase genes are often mis-annotated due to their fragmentation after integration. Instead, we used TBlastN with already known and subsequently detected Int(N) and Int(C) moieties as query. As subject sequence, we used the nr/nt nucleotide collection and our own collection of sequenced *Thermococcales* genomes (to be published elsewhere).

The detection of genomic integrated elements is a two-step process. In the first step, disrupted integrases and their adjacent att site are located by sequence comparison. We selected hits with an e-value lower than 1e-30 and then reconstituted the complete integrase-coding gene. In the second step, the surroundings (<30–40 kb) of these locations are scanned for the cognate att direct repeat. This arrangement is unequivocal as tandemly inserted MGEs were never observed. The sequences of integrated MGEs were obtained by extracting from GenBank files DNA segments comprised between attL and attR pairs (supplementary data file S1, Supplementary Material online). About 73 integrases were detected, 20 were already published, and 53 were newly identified, including 34 in our genome collection (to be published elsewhere). Mobile elements were assigned to a MGE family based on the presence of marker genes: core genes for the pT26-2 plasmid family (supplementary fig. S3, Supplementary Material online) and the major capsid protein gene for the fuselloviruses (Krupovic et al. 2014). For the pAMT11 plasmid family, no marker gene was previously proposed. We used the three longer genes (ORF1 to 3) conserved between the two previously known members of the family pAMT11 and TKV1 (Gonnet et al. 2011).

Assessment of Integrases Relatedness Using Similarity Networks

All-against-all BlastP analyses were performed on all the integrases comprises in data set 1, a set of *Sulfolobales* integrases identified in free Fuselloviridae, and all available pTN3 integrases. The all-against-all integrases BlastP results were grouped using the SiLiX (for Single Linkage Clustering of Sequences) package v1.2.8 (http://lbbe.univ-lyon1.fr/SiLiX) (Miele et al. 2011). This approach for the clustering of homologous sequences is based on single transitive links with alignment coverage constraints. Several different criteria can be used separately or in combination to infer homology separately (percentage of identity, alignment score or E-value, alignment coverage). For this integrase data set, the results of the all-against-all BlastP analyses were filtered with the additional thresholds of BlastP pairwise similarity >25% or >35% over 60% for the protein (fig. 1). The network was visualized using the igraph package from R (https://igraph.org/). In order to find densely connected communities in a graph via random walks, we used the cluster_walktrap function of the igraph package.

Thermococcales Isolation and Sequencing

Thermococcales strains were isolated during the Starmer (1989), Amistad (1999), CIR (2001), Extreme (2001), and Iris (2001) Ifremer campaigns and originate from the Indian Ocean, the Oriental Pacific Ridge, and the Mid-Atlantic Ridge (Badel et al. 2019). DNA sequencing was performed by Genoscope (Centre National de Séquençage, France), using Illumina MiSeq. Reads were assembled with Newbler (release 2.9) and gap closure was performed by PCR, Sanger sequencing, and Oxford Nanopore MinION. The sequences of all the integrated elements detected in these isolates are publicly available (supplementary data file S1, Supplementary Material online).

Recombinant Protein Production and Purification

The gene coding for the integrase of plasmid pT26-2 (Int$^{pT26\text{-}2}$, NCBI protein accession YP_003603594.1) was PCR amplified from pT26-2 plasmid DNA with primers pT26-2_F and _R (supplementary table S2, Supplementary Material online). The forward primer added a sequence coding for the Strep-tag at the 5′ end of the gene. The PCR product was then assembled with the linearized expression vector pET-26b(+) by Gibson assembly (NEB) and transformed into *E. coli* strain XL1-Blue. The resulting plasmid pCB558 was verified by DNA sequencing. Plasmid pCB616 encoding the variant Int$^{pT26\text{-}2}$Y327F was constructed with the Q5 site-directed mutagenesis kit (NEB) using the primers pT26-2_Y327F_F and _R (supplementary table S2, Supplementary Material online). *Escherichia coli* Rosetta BL21 (DE3) carrying pCB558 or pCB616 was grown in LB medium to OD 0.5 and recombinant protein production was induced with 250 µM IPTG. Int$^{pT26\text{-}2}$ overproduction in *E. coli* was somewhat toxic. After 1.5-h induction, cells were harvested and resuspended in the purification buffer (1 M KCl, 40 mM Tris–HCl, pH 8, 10% glycerol, and 5 mM β-mercaptoethanol) supplemented with a protease inhibitor cocktail (cOmplete ULTRA Tablets, EDTA-free, Roche). Cells were lysed by a pressure shock with a one-shot cell disruptor (Constant Systems Ltd) and centrifuged at 4 °C for 30 min at 18,000×g. The supernatant was recovered, heated at 65 °C for 10 min, centrifuged at 5,000×g for 15 min and filtered. The solution was then loaded on a

1-ml StrepTrap HP column (GE Healthcare). The STREP-tagged Int$^{pT26\text{-}2}$ and Int$^{pT26\text{-}2}$Y327F were eluted by the purification buffer supplemented with 2.5 mM d-desthiobiotin. The buffer of Int$^{pT26\text{-}2}$Y327F was depleted in d-desthiobiotin by buffer exchange with a Vivaspin Centrifugal Concentrators (Sartorius). Int$^{pT26\text{-}2}$ was subsequently loaded on a HiLoad 16/600 75 prep grade column (GE Healthcare) for size exclusion chromatography and fractions containing the protein were concentrated with a Vivaspin Centrifugal Concentrators (Sartorius). Protein solutions harvested at different steps of the purification were analyzed by SDS–PAGE (supplementary fig. S8, Supplementary Material online). The purified concentrated proteins contained the N-ter strep-tag and their concentration was determined by spectrophotometry.

Integrase Substrates Construction

To construct plasmid pCB568, we annealed oligonucleotides BamHI-tRNAarg + 6-EcoRI_A and _B (supplementary table S2, Supplementary Material online) corresponding to the *T. sp. 26-2* tRNA^{arg} gene and including six nucleotides downstream of the gene. The annealing product was digested by EcoRI and BamHI, ligated into a similarly digested pUC18 and transformed into *E. coli* XL1-Blue. The same method was applied for plasmids pCB590, pCB588, and pCB584 with the oligonucleotides BamHI-L56-coRI_A and _B, BamHI-L55-coRI_A and _B, and BamHI-L53-coRI_A and _B, respectively. Plasmid pCB596 was obtained by Gibson assembly of the following three fragments: 1) pCB568 digested by NdeI, 2) a PCR product amplified from pUC4K with the primers KanR-ex1 and 2 (supplementary table S2, Supplementary Material online) and corresponding to the KmR gene, and 3) a PCR product amplified from pCB568 with the primers tRNAarg + 6-ex1 and 2 (supplementary table S2, Supplementary Material online) and corresponding to tRNA^{arg} gene and additional six nucleotides downstream. The assembled product was transformed into *E. coli* XL1-Blue. The same strategy was used to obtain plasmid pCB598 but with the primers KanR-inv1 and 2 and tRNAarg + 6-inv1 and 2 (supplementary table S2, Supplementary Material online) that lead to the assembly of the tRNA^{arg} in the opposite orientation. To obtain plasmids pCB586, pCB602, pCB604, pCB630, pCB632, pCB636, and pCB638, pUC18 was PCR amplified with the forward primer pUC18-H_FOR and the reverse primer L54-pUC18_REV or R49-pUC18_REV or R48-pUC18_REV or R47-pUC18_REV or R46-pUC18_REV or R43-pUC18_REV or R40-pUC18_REV, respectively. PCR product was digested by NcoI and HindIII, ligated, and transformed into *E. coli* XL1-Blue. All plasmids were verified by DNA sequencing. The plasmids used in this work are listed in supplementary table S3, Supplementary Material online.

Integrase Substrates Production

Supercoiled plasmids were extracted from *E. coli* XL1-Blue using NucleoSpin Plasmid (Macherey-Nagel) or NucleoBond Xtra Midi (Macherey-Nagel) accordingly to the manufacturer instructions. Relaxed pCB568 and pCB598 were obtained by Nt.BspQI digestion (NEB) followed by a column purification (NucleoSpin Gel and PCR clean-up, Macherey-Nagel). ScaI and PvuII pCB568 fragments were obtained by ScaI and PvuII digestion (FastDigest, ThermoFisher) followed by a gel purification of the desired fragment (NucleoSpin Gel and PCR clean-up, Macherey-Nagel). Linear pCB598 was obtained by ScaI digestion (FastDigest, ThermoFisher) followed by a gel purification (NucleoSpin Gel and PCR clean-up, Macherey-Nagel). A 2,106-bp fragment of pCB568 was amplified by Phusion Polymerase (ThermoFisher) with the primers pUC1481-1503 and P30-REV followed by column purification (NucleoSpin Gel and PCR clean-up, Macherey-Nagel). The various fragments of 800 bp were amplified from the appropriate plasmid by Phusion Polymerase (ThermoFisher) with the primers pUC195-217 and pZE21_rev followed by column purification (NucleoSpin Gel and PCR clean-up, Macherey-Nagel).

In Vitro Integrase Enzymatic Assay

For in vitro enzymatic assays, 500 µg substrate DNA and 200 ng (240 nM) integrase were incubated for 1 h at 75 °C in 300 mM KCl, 7 mM Tris–HCl, pH 8, 0.4% glycerol, and 825 µM β-mercaptoethanol in a total volume of 20 µl unless otherwise indicated. In certain cases, two different substrates were mixed in an equimolar ratio for a total mass of 500 µg. For integration assays, reaction product were treated with proteinase K, separated by agarose gel electrophoresis at 50 V and subsequently stained with ethidium bromide for visualization. For inversion and excision assays, reaction products were purified with the NucleoSpin gel and PCR clean-up kit (Macherey-Nagel), digested with the appropriate restriction enzymes (FastDigest, ThermoFisher), and separated by gel electrophoresis. Band intensity was quantified with ImageJ (Schneider et al. 2012) on nonsaturated gel pictures using three repetitions of the activity assay.

Protein Alignment, Trimming, and Phylogenetic Analysis

The alignment used for phylogenetic analyses was performed using MAFFT v7 with default settings (Katoh and Standley 2013) and trimmed with BMGE (Criscuolo and Gribaldo 2010) with a BLOSUM30 matrix, and the -b 1 parameter. IQ-TREE v1.6 (http://www.iqtree.org/) (Nguyen et al. 2015) was used to calculate maximum likelihood trees with the best model as suggested by the best model selection option (Kalyaanamoorthy et al. 2017). Branch robustness was estimated with the nonparametric bootstrap procedure (100 replicates) or with the SH-like approximate likelihood ratio test (Guindon et al. 2010) and the ultrafast bootstrap approximation (1,000 replicates) (Hoang et al. 2018). The integrases phylogenetic tree shown in figure 6 corresponds to the tree obtained with the VT+F + I+G4 model on a matrix of 318 positions and with ultrafast bootstrap to indicate the tree robustness. The phylogenetic tree was shaped with the iTOL webtool (Letunic and Bork 2019).

Other Bioinformatics Analyses

Synteny maps were created using EasyFig (Sullivan et al. 2011). Pairwise alignments and att site alignments were performed with MUSCLE (Edgar 2004). *Thermococcus kodakarensis* tRNA genes were extracted with GtRNAdb (Chan and Lowe 2016).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors wish to thank Dr Philippe Oger for kindly providing *T. piezophilus* sequencing reads and Dr Damien Courtine for communicating the sequence of *Thermococcus* IRI06c. This work was funded by CNRS, the European Research Council under the European Union's Seventh Framework Program (FP/2007-2013)/Project EVOMOBIL—ERC (Grant No. 340440) (P.F.) and the Agence Nationale de la Recherche ANR-19-CE11-0007. C.B. is supported by "Ecole Normale Supérieure de Lyon."

References

Abremski K, Gottesman S. 1981. Site-specific recombination Xis-independent excisive recombination of bacteriophage lambda. *J Mol Biol*. 153(1):67–78.

Adam PS, Borrel G, Brochier-Armanet C, Gribaldo S. 2017. The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. *ISME J*. 11(11):2407–2425.

Arber W, Dussoix D. 1962. Host specificity of DNA produced by *Escherichia coli*. I. Host controlled modification of bacteriophage lambda. *J Mol Biol*. 5(1):18–36.

Badel C, Erauso G, Gomez AL, Catchpole R, Gonnet M, Oberto J, Forterre P, Da Cunha V. 2019. The global distribution and evolutionary history of the pT26-2 archaeal plasmid family. *Environ Microbiol*. 21(12):4685–4705.

Callac N, Oger P, Lesongeur F, Rattray JE, Vannier P, Michoud G, Beauverger M, Gayet N, Rouxel O, Jebbar M, et al. 2016. *Pyrococcus kukulkanii* sp. nov., a hyperthermophilic, piezophilic archaeon isolated from a deep-sea hydrothermal vent. *Int J Syst Evol Microbiol*. 66(8):3142–3149.

Carroll AC, Wong A. 2018. Plasmid persistence: costs, benefits, and the plasmid paradox. *Can J Microbiol*. 64(5):293–304.

Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and de novo gene birth. *Nature* 487(7407):370–374.

Chan PP, Lowe TM. 2016. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res*. 44(D1):D184–D189.

Chen JW, Lee J, Jayaram M. 1992. DNA cleavage in trans by the active-site tyrosine during Flp recombination – switching protein partners before exchanging strands. *Cell* 69(4):647–658.

Cortez D, Quevillon-Cheruel S, Gribaldo S, Desnoues N, Sezonov G, Forterre P, Serre MC. 2010. Evidence for a Xer/dif system for chromosome resolution in archaea. *PLoS Genet*. 6(10):e1001166.

Cossu M, Badel C, Catchpole R, Gadelle D, Marguet E, Barbe V, Forterre P, Oberto J. 2017. Flipping chromosomes in deep-sea archaea. *PLoS Genet*. 13(6):e1006847.

Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol*. 10(1):210.

Dalmasso C, Oger P, Courtine D, Georges M, Takai K, Maignien L, Alain K. 2016. Complete genome sequence of the hyperthermophilic and piezophilic archeon *Thermococcus piezophilus* CDGST, able to grow under extreme hydrostatic pressures. *Genome Announc*. 4(5):e01018-16.

Dorman CJ, Bogue MM. 2016. The interplay between DNA topology and accessory factors in site-specific recombination in bacteria and their bacteriophages. *Sci Prog*. 99(4):420–437.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32(5):1792–1797.

Gandon S. 2016. Why be temperate: lessons from bacteriophage lambda. *Trends Microbiol*. 24(5):356–365.

Gaudin M, Krupovic M, Marguet E, Gauliard E, Cvirkaite-Krupovic V, Le Cam E, Oberto J, Forterre P. 2014. Extracellular membrane vesicles harbouring viral genomes. *Environ Microbiol*. 16(4):1167–1175.

Gehring AM, Astling DP, Matsumi R, Burkhart BW, Kelman Z, Reeve JN, Jones KL, Santangelo TJ. 2017. Genome replication in *Thermococcus kodakarensis* independent of Cdc6 and an origin of replication. *Front Microbiol*. 8:2084.

Gerdes K, Howard M, Szardenings F. 2010. Pushing and pulling in prokaryotic DNA segregation. *Cell* 141(6):927–942.

Gonnet M, Erauso G, Prieur D, Le Romancer M. 2011. pAMT11, a novel plasmid isolated from a *Thermococcus* sp. strain closely related to the virus-like integrated element TKV1 of the *Thermococcus kodakaraensis* genome. *Res Microbiol*. 162(2):132–143.

Gorlas A, Croce O, Oberto J, Gauliard E, Forterre P, Marguet E. 2014. *Thermococcus nautili* sp. nov., a hyperthermophilic archaeon isolated from a hydrothermal deep sea vent (East Pacific Ridge). *Int J Syst Evol Microbiol*. 64(Pt 5):1802–1810.

Gorlas A, Koonin EV, Bienvenu N, Prieur D, Geslin C. 2012. TPV1, the first virus isolated from the hyperthermophilic genus *Thermococcus*. *Environ Microbiol*. 14(2):503–516.

Grindley NDF, Whiteson KL, Rice PA. 2006. Mechanisms of site-specific recombination. *Annu Rev Biochem*. 75(1):567–605.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 59(3):307–321.

Guo F, Gopaul DN, Van Duyne GD. 1999. Asymmetric DNA bending in the Cre-loxP site-specific recombination synapse. *Proc Natl Acad Sci U S A*. 96(13):7143–7148.

Harms A, Brodersen DE, Mitarai N, Gerdes K. 2018. Toxins, targets, and triggers: an overview of toxin-antitoxin biology. *Mol Cell*. 70(5):768–784.

Hendrix RW, Lawrence JG, Hatfull GF, Casjens S. 2000. The origins and ongoing evolution of viruses. *Trends Microbiol*. 8(11):504–508.

Hille F, Richter H, Wong SP, Bratovic M, Ressel S, Charpentier E. 2018. The biology of CRISPR-Cas: backward and Forward. *Cell* 172(6):1239–1259.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*. 35(2):518–522.

Hulter N, Ilhan J, Wein T, Kadibalban AS, Hammerschmidt K, Dagan T. 2017. An evolutionary perspective on plasmid lifestyle modes. *Curr Opin Microbiol*. 38:74–80.

Iranzo J, Koonin EV, Prangishvili D, Krupovic M. 2016. Bipartite network analysis of the archaeal virosphere: evolutionary connections between viruses and capsidless mobile elements. *J Virol*. 90(24):11043–11055.

Jayaram M, Ma CH, Kachroo AH, Rowley PA, Guga P, Fan HF, Voziyanov Y. 2015. An overview of tyrosine site-specific recombination: from an Flp perspective. *Microbiol Spectr*. 3(4), doi:10.1128/microbiolspec.MDNA3-0021-2014.

Jo M, Murayama Y, Tsutsui Y, Iwasaki H. 2017. In vitro site-specific recombination mediated by the tyrosine recombinase XerA of *Thermoplasma acidophilum*. *Genes Cells* 22(7):646–661.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 14(6):587–589.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4):772–780.

Krupovic M, Forterre P, Bamford DH. 2010. Comparative analysis of the mosaic genomes of tailed archaeal viruses and proviruses suggests common themes for virion architecture and assembly with tailed viruses of bacteria. *J Mol Biol*. 397:144–160.

Krupovic M, Makarova KS, Wolf YI, Medvedeva S, Prangishvili D, Forterre P, Koonin EV. 2019. Integrated mobile genetic elements in Thaumarchaeota. *Environ Microbiol*. 21(6):2056–2078.

Krupovic M, Quemin ER, Bamford DH, Forterre P, Prangishvili D. 2014. Unification of the globally distributed spindle-shaped viruses of the archaea. *J Virol*. 88(4):2354–2358.

Landy A. 1989. Dynamic, structural, and regulatory aspects of lambda site-specific recombination. *Annu Rev Biochem*. 58(1):913–949.

Landy A. 2015. The lambda integrase site-specific recombination pathway. *Microbiol Spectr*. 3(2):MDNA3-0051-2014.

Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*. 47(W1): W256–W259.

Lewis JA, Hatfull GF. 2001. Control of directionality in integrase-mediated recombination: examination of recombination directionality factors (RDFs) including Xis and Cox proteins. *Nucleic Acids Res*. 29(11): 2205–2216.

Liu Y, Harrison PM, Kunin V, Gerstein M. 2004. Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome Biol*. 5(9):R64.

Lopez-Garcia P, Forterre P. 1997. DNA topology in hyperthermophilic archaea: reference states and their variation with growth phase, growth temperature, and temperature stresses. *Mol Microbiol*. 23(6):1267–1279.

Miele V, Penel S, Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* 12(1):116.

Million-Weaver S, Camps M. 2014. Mechanisms of plasmid segregation: have multicopy plasmids been overlooked? *Plasmid* 75:27–36.

Mizuuchi K, Gellert M, Nash HA. 1978. Involement of supertwisted DNA in integrative recombination of bacteriophage lambda. *J Mol Biol*. 121(3):375–392.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32(1):268–274.

Nordstrom K. 2006. Plasmid R1–replication and its control. *Plasmid* 55:1–26.

Oberto J, Gaudin M, Cossu M, Gorlas A, Slesarev A, Marguet E, Forterre P. 2014. Genome sequence of a hyperthermophilic archaeon, *Thermococcus nautili* 30-1, that produces viral vesicles. *Genome Announc*. 2:e00243–00214.

Oberto J, Sloan SB, Weisberg RA. 1994. A segment of the phage HK022 chromosome is a mosaic of other lambdoid chromosomes. *Nucleic Acids Res*. 22(3):354–356.

Redder P, Peng X, Brugger K, Shah SA, Roesch F, Greve B, She Q, Schleper C, Forterre P, Garrett RA, et al. 2009. Four newly isolated fuselloviruses from extreme geothermal environments reveal unusual morphologies and a possible interviral recombination mechanism. *Environ Microbiol*. 11(11):2849–2862.

Reed RR. 1981. Transposon-mediated site-specific recombination: a defined in vitro system. *Cell* 25(3):713–719.

Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods*. 9(7):671–675.

Schut GJ, Bridger SL, Adams MW. 2007. Insights into the metabolism of elemental sulfur by the hyperthermophilic archaeon *Pyrococcus furiosus*: characterization of a coenzyme A-dependent NAD(P)H sulfur oxidoreductase. *J Bacteriol*. 189(12):4431–4441.

Serre MC, Letzelter C, Garel JR, Duguet M. 2002. Cleavage properties of an archaeal site-specific recombinase, the SSV1 integrase. *J Biol Chem*. 277(19):16758–16767.

She Q, Chen B, Chen L. 2004. Archaeal integrases and mechanisms of gene capture. *Biochem Soc Trans*. 32(2):222–226.

She Q, Peng X, Zillig W, Garrett RA. 2001. Gene capture in archaeal chromosomes. *Nature* 409(6819):478–478.

Siepel A. 2009. Darwinian alchemy: human genes from noncoding DNA. *Genome Res*. 19(10):1693–1695.

Soler N, Marguet E, Cortez D, Desnoues N, Keller J, van Tilbeurgh H, Sezonov G, Forterre P. 2010. Two novel families of plasmids from hyperthermophilic archaea encoding new families of replication proteins. *Nucleic Acids Res*. 38(15):5088–5104.

Sullivan MJ, Petty NK, Beatson SA. 2011. Easyfig: a genome comparison visualizer. *Bioinformatics* 27(7):1009–1010.

Van Duyne GD. 2015. Cre recombinase. *Microbiol Spectr*. 3(1):MDNA3-0014-2014.

Vihinen M. 2014. Contribution of pseudogenes to sequence diversity. *Methods Mol Biol*. 1167:15–24.

Wang J, Liu Y, Liu Y, Du K, Xu S, Wang Y, Krupovic M, Chen X. 2018. A novel family of tyrosine integrases encoded by the temperate pleolipovirus SNJ2. *Nucleic Acids Res*. 46(5):2521–2536.

Wang Y, Duan Z, Zhu H, Guo X, Wang Z, Zhou J, She Q, Huang L. 2007. A novel *Sulfolobus* non-conjugative extrachromosomal genetic element capable of integration into the host genome and spreading in the presence of a fusellovirus. *Virology* 363(1): 124–133.

Zhan ZY, Zhou J, Huang L. 2015. Site-specific recombination by SSV2 integrase: substrate requirement and domain functions. *J Virol*. 89(21):10934–10944.