Archaeal tyrosine recombinases

Catherine Badel, Violette da Cunha, J. Oberto

FEMS Microbiology Reviews, fuab004, 1–27

doi: 10.1093/femsre/fuab004
Advance Access Publication Date: 1 February 2021
Review Article

REVIEW ARTICLE

Archaeal tyrosine recombinases

Catherine Badel[†,‡], Violette Da Cunha[†,#] and Jacques Oberto[*,§]

Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198 Gif-sur-Yvette, France

[*]**Corresponding author:** Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198 Gif-sur-Yvette, France. E-mail: jacques.oberto@i2bc.paris-saclay.fr

One sentence summary: A number of new archaeal tyrosine recombinases have been reported to catalyze canonical DNA integration, excision, inversion and also reactions beyond site-specific recombination, prompting us to review their properties and analyze their phylogenomic relationships with other known recombinases.

[†]Both authors contributed equally to this manuscript.

Editor: Sonja-Verena Albers

[‡]Catherine Badel, http://orcid.org/0000-0002-1723-0350
[#]Violette Da Cunha, http://orcid.org/0000-0002-9035-7825
[§]Jacques Oberto, http://orcid.org/0000-0003-1680-636X

ABSTRACT

The integration of mobile genetic elements into their host chromosome influences the immediate fate of cellular organisms and gradually shapes their evolution. Site-specific recombinases catalyzing this integration have been extensively characterized both in bacteria and eukarya. More recently, a number of reports provided the in-depth characterization of archaeal tyrosine recombinases and highlighted new particular features not observed in the other two domains. In addition to being active in extreme environments, archaeal integrases catalyze reactions beyond site-specific recombination. Some of these integrases can catalyze low-sequence specificity recombination reactions with the same outcome as homologous recombination events generating deep rearrangements of their host genome. A large proportion of archaeal integrases are termed suicidal due to the presence of a specific recombination target within their own gene. The paradoxical maintenance of integrases that disrupt their gene upon integration implies novel mechanisms for their evolution. In this review, we assess the diversity of the archaeal tyrosine recombinases using a phylogenomic analysis based on an exhaustive similarity network. We outline the biochemical, ecological and evolutionary properties of these enzymes in the context of the families we identified and emphasize similarities and differences between archaeal recombinases and their bacterial and eukaryal counterparts.

Keywords: tyrosine recombinase; Archaea; mobile genetic element; horizontal transfer; genome evolution; site-specific recombination

INTRODUCTION

Recombination of DNA is an essential mechanism ensuring the maintenance, propagation and evolution of genetic information in all living organisms. Homologous recombination is complex, requires energy, involves a number of protein complexes and operates over large regions sharing extensive sequence identity (Sung and Klein 2006; Sun *et al.* 2020). The exchange point can occur anywhere between these regions. Site-specific recombination is an energy-independent process catalyzed by DNA transaction proteins whose primary function is to specifically recognize and recombine two short DNA duplexes sharing some degree of sequence identity (Craig 1988; Dorman and Bogue 2016). In this case, the breakage and joining of DNA requires a particular catalytic amino acid forming a transient covalent bond between the protein and the DNA substrate (Pargellis

Received: 28 May 2020; **Accepted:** 13 January 2021

et al. 1988; Smith *et al.* 2004). Conservative site-specific recombinases are classified into two unrelated families, serine and tyrosine recombinases, referring to this catalytic residue (Grindley, Whiteson and Rice 2006; Stark 2014). The recombination promoted by these enzymes is conservative since strand exchange occurs in a precise location of short sequence identity without loss or addition of DNA. Non-conservative site-specific recombinases such as DDE transposases do not depend on sequence identity in the target sequences and require DNA synthesis. Site-specific recombinases are encountered in the three domains of life and have been characterized extensively in both bacteria and eukarya where they perform a number of biological functions such as integration and excision of viral DNA into the host chromosome (Landy 2015), phage resolution (Van Duyne 2015; Meinke *et al.* 2016), plasmid flipping (Jayaram *et al.* 2015), resolution of chromosome dimers (Castillo, Benmohamed and Szatmari 2017), integron shuffling (Escudero *et al.* 2015; Engelstadter, Harms and Johnsen 2016), insertion sequence (IS) element transposition (Siguier *et al.* 2015; Arinkin, Smyshlyaev and Barabas 2019) and induction of gene expression by phase variation (Bayliss 2009). The usefulness of tyrosine recombinases as DNA transaction tools has emerged in a wide range of biotechnological and medical applications (Jayaram *et al.* 2015; Van Duyne 2015; Meinke *et al.* 2016).

In archaea, serine recombinases and DDE recombinases have been observed in transposons but were never fully investigated (Filee, Siguier and Chandler 2007; Krupovic *et al.* 2019). On the other hand, the activities of several archaeal tyrosine recombinases have been analyzed, reviewed and ranked into two classes (She, Brugger and Chen 2002; She, Chen and Chen 2004). Class I corresponds to the SSV-like integrases found in *Sulfolobales* viruses and whose gene is fragmented upon integration. Class II groups the pNOB8 plasmid-like integrases that follow the phage λ integration paradigm. Additionally, XerA resolvases are encoded by archaeal chromosomes. In the recent years, the study of archaeal tyrosine recombinases has generated considerable experimental data.

In the present review, we discuss archaeal integrases from biochemical, ecological and evolutionary points of view. After a brief overview of the historical context of the discovery of archaeal integrases, we present the insights gained from their sequences and the current knowledge about their mechanisms and recombination target. These aspects are put into the perspective of a systematic phylogenomic analysis. Because the recombination reaction catalyzed by integrases is central to mobile genetic element (MGE) lifestyles, we then highlight some ecological consideration related to archaeal integrases. Finally, we describe recent findings on integrase evolution and genome evolution mediated by integrases. The deep comparison of all available archaeal tyrosine recombinases presented here allows the ranking of these enzymes in defined families and to underline functional similarities and differences with known bacterial and eukaryal recombinases.

FIRST ACCOUNTS OF INTEGRASE-PROMOTED SITE-SPECIFIC RECOMBINATION

Early investigations reported the discovery of an ultramicroscopic virus capable of either modifying bacteria or destroying them (Twort 1915). A similar virus able to lyse *Shigella* was later isolated and called bacteriophage (d'Hérelle 1917). Bordet and Ciuca defined these bacteria modified by bacteriophages as lysogens since they carried the potential to lyse other cells (Bordet and Ciuca 1920). With the advent of phage genetics, the mechanisms of lytic cycle and lysogeny later became understandable in molecular terms. Building upon his hypothesis of a small region of homology between phage and host chromosome where recombination would take place, Campbell elaborated the foundations of the site-specific recombination pathway and illustrated how bacteriophage λ and other episomes can integrate into, or excise from, bacterial chromosomes (Campbell 1963) (Fig. 1). Zissler isolated the first phage λ mutants unable to lysogenize and interpreted them as lacking a region necessary for integration in the host chromosome that he called *int* (for integration deficient) (Zissler 1967). Simultaneously, the requirement of an enzymatic activity for site-specific recombination was reported for bacteriophage φ80 (Signer and Beckwith 1966). The product of the *int* gene along with the newly identified attachment sites attϕ (attachment φ80) and attB (attachment bacteria) was used to demonstrate site-specific integration by formal genetics (Weil and Signer 1968). Ausubel performed the first radiochemical purification of this enzyme that he called integrase, using phage λ-infected *Escherichia coli* cells (Ausubel 1974). Concomitantly, Nash also purified the Int protein (Nash 1974) that he later used to demonstrate integrative recombination activity *in vitro* (Nash 1975). Remarkably, this relentless experimentation extending over six decades succeeded in identifying all components involved of site-specific recombination at the advent of molecular biology and before DNA sequencing.

Three types of archaeal tyrosine recombinases

The archaea, which form the third domain of life, often carry extrachromosomal elements some of which were found integrated into the genome. The first report described the presence of a freely replicating and chromosome-integrated element of 15 kb, later called *Sulfolobus* spindle-shaped virus 1 or SSV1 (Schleper, Kubo and Zillig 1992), in the hyperthermophilic *Sulfolobus shibatae* (Grogan, Palm and Zillig 1990). Soon other elements were discovered such as SAV1 (Martin *et al.* 1984), pQX1 (Peng *et al.* 2000), XQ2 (She *et al.* 2001a) and SSV2 (Stedman *et al.* 2003) that could integrate at specific chromosomal loci in various *Sulfolobus* species of the *Crenarchaeota* phylum. SSV1 and SAV1 were shown to form virus-like particles upon UV induction (Martin *et al.* 1984; Frols *et al.* 2007) and turned out to contain identical genomes (Stedman *et al.* 2003). Interestingly, the SSV-type integrases encoded by SSV1, SSV2, pQX1 and XQ2 carry the DNA recombination site within their own gene that would become fragmented upon chromosomal integration, therefore precluding integrase-mediated excision in the absence of an intact gene (She *et al.* 2001a) (Fig. 2A). More recently, SSV-type integrases encoded by *Thermococcus nautili* plasmid pTN3 (Oberto *et al.* 2014; Cossu *et al.* 2017) and by *Thermococcus* sp. 26-2 plasmid pT26-2 (Badel *et al.* 2020) were uncovered in the *Euryarchaeota* phylum as well. Remarkably, recombinases that disrupt their own gene are found exclusively in the archaeal domain and were named suicidal integrases (Badel *et al.* 2020). It would be logical to assume such MGEs would remain irreversibly integrated into their host genome. However, this is not the case. This aspect and the evolutionary implications of suicidal integrases will be discussed below.

The second type of archaeal integrases follows the more classical bacteriophage λ paradigm and maintains an intact gene after integration (Fig. 2B). *Sulfolobus* conjugative plasmid pNOB8

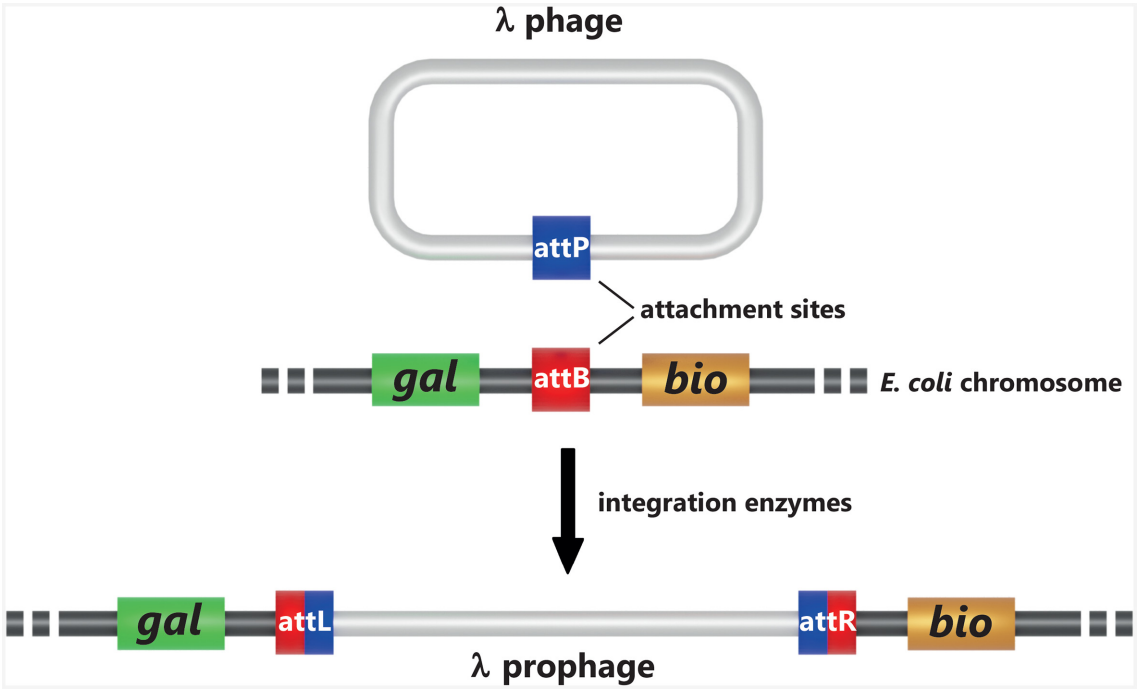

Figure 1. Campbell model for phage λ site-specific recombination. The model presented by Campbell (1963) suggested for the first time the breaking and rejoining of DNA sequences by integration enzymes in order to allow phage λ lysogenization in *Escherichia coli*.

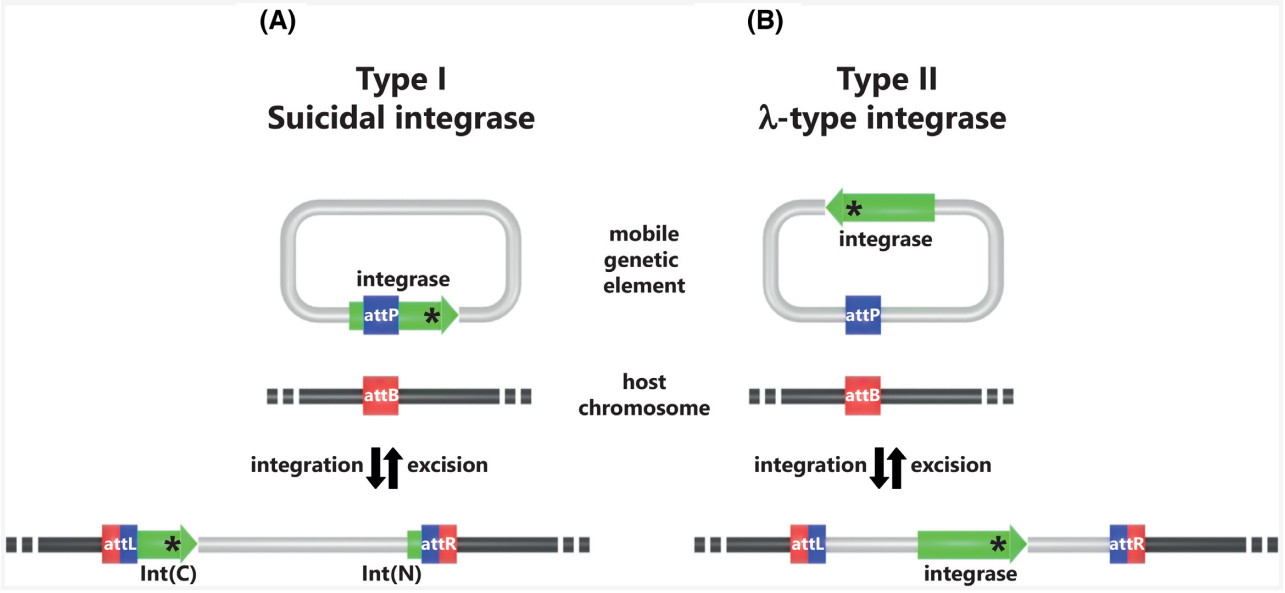

Figure 2. Classes of archaeal integrases. She, Brugger and Chen (2002) proposed the ranking of archaeal integrases into two distinct types: Type I **(A)** and Type II **(B)**. The blue and red squares correspond to the specific recombination sites.

(She *et al.* 1998) encoded such a recombinase allowing to integrate into the genome of *Sulfolobus tokodaii* and several other species (She, Chen and Chen 2004). A systematic genomic search for sequences encoding the more conserved catalytic domain identified a number of archaeal integrases of this type that could be ranked in five families (She, Brugger and Chen 2002).

In addition to the first two types of tyrosine recombinases that correspond to *bona fide* integrases, the majority archaeal genomes possess the XerA (also named XerC) site-specific recombinase that, similarly as the bacterial XerC/D homologs, resolves chromosome dimers occurring during DNA replication (Cortez *et al.* 2010). The major difference between bacteria and archaea is that the latter lack an FtsK homolog, which is essential in bacterial systems to displace the reaction equilibrium toward resolution (Bigot *et al.* 2004). This observation suggests that the regulation of archaeal chromosome dimer resolution operates with a different mechanism than in bacteria.

DEFINITION OF THE FAMILY OF TYROSINE RECOMBINASES

Insights from primary sequences

The development of sequence databases leads to the detection of novel enzymes related to phage λ integrase and the identification of their functional domains by comparative analysis. Despite distant relationships in protein primary sequences, all bacteriophage site-specific recombinase C-terminal ends could be aligned with the yeast 2μ plasmid FLP protein (Argos *et al.* 1986). In particular, the perfect conservation of the three residues (HXXR...Y) highlighted a potential active site. The presence of a tyrosine in that region suggested a catalytic function for this residue in DNA cleavage (Argos *et al.* 1986). A subsequent analysis revealed yet another conserved arginine residue upstream leading to the consensus tetrad R...HXXR...Y (Abremski and Hoess 1992). These enzymes defined a new family of proteins, the tyrosine recombinases. Afterward, the alignment of all the tyrosine recombinases available in the databases underlined an important sequence diversity among these enzymes and the conservation of important residues composing the catalytic domain that spans ∼180 residues as confirmed by mutational analysis (Esposito and Scocca 1997; Nunes-Duby *et al.* 1998).

The study of archaeal organisms also revealed integrase-encoding genes from *Sulfolobus shibatae* spindle-shaped virus (SSV1) (Palm *et al.* 1991) that split upon integration (Fig. 3A) and from *Sulfolobus* NOB8-H2 pNOB8 plasmid (She *et al.* 1998). On the basis of a slightly divergent consensus, two subfamilies were identified: the SSV1-type integrases (R...KXXR...Y) and the pNOB8-type integrases (R...YXXR...Y) (She, Chen and Chen 2004) (Fig. 3B). IntSSV1 and XerA from *Pyrococcus abyssi* (PaXerA) were shown to form a covalent intermediate with the substrate DNA and the implication of the tyrosine Y314 was evidenced for IntSSV1 (Serre *et al.* 2002, 2013; Zhan *et al.* 2012). A substitution of this tyrosine abolished IntSSV1 substrate cleavage activity confirming its importance for catalysis in archaea (Letzelter, Duguet and Serre 2004). Similarly to previously characterized tyrosine recombinases, the active residues are localized at the C-terminal end of the protein and are involved in DNA cleavage and ligation catalysis (Zhan, Zhou and Huang 2015) (Fig. 3B).

The highly divergent N-terminal regions of tyrosine recombinases suggested early on their involvement in features unique to each system such as specific sequence recognition (Argos *et al.* 1986). This diversity also reflected the fact that some recombinases such as phage integrases would recognize two distinct DNA sites instead of a single one (Esposito and Scocca 1997). It can be assumed that archaeal recombinases use their N-terminal regions to recognize and bind to their specific site. Gel retardation experiments with truncated forms of IntSSV1 indicated that in addition to the full-length protein, both the first half (N175) and the second half (C174) would bind to the specific DNA target (Zhan *et al.* 2012). A similar approach using *Sulfolobus islandicus* IntSSV2 demonstrated that the N-terminal extremity controls multimerization whereas the middle portion officiates in the specific DNA interaction (Zhan, Zhou and Huang 2015).

Extending the archaeal tyrosine recombinase diversity

Archaeal tyrosine recombinases belong to the NCBI DNA_BRE_C superfamily (cl00213 or cd00397) that groups the DNA breaking–rejoining enzymes with a catalytic domain in C-terminal position. In addition to tyrosine site-specific recombinases such as integrases, this superfamily also includes Type IB topoisomerases, as they share conserved active site residues and the same fold in their catalytic domain (Cheng *et al.* 1998). This DNA_BRE_C superfamily is composed of five major Pfam domains: Phage_integrase (PF00589), Phage_integr_3 (PF16795), Topoisom_I (PF01028), DUF3504 (PF12012) and Integrase_1 (PF12835). The DUF3504 and Integrase_1 domains are specific to eukarya and bacteria, respectively.

In order to propose a comprehensive overview of archaeal integrase diversity, we undertook a phylogenomic analysis based on an exhaustive similarity network to assign all these enzymes to their respective families. This classification also illustrates the phylogenetic relationships between the families and enlightens the evolutionary links between the three domains of life. For this deep investigation, we extracted from the conserved domain database of the NCBI all archaeal protein sequences belonging to the cl00213 superfamily. To this dataset, we added the previously reported integrases encoded by archaeal viruses (Pauly *et al.* 2019) and the reconstructed suicidal integrases from the pTN3 and pT26-2 plasmid families (Cossu *et al.* 2017; Badel *et al.* 2019, 2020). Partial sequences <150 residues were eliminated from the initial dataset and redundancy was reduced using the UCLUST program (Edgar 2010) to eliminate sequences sharing >90% identity (Fig. S1, Supporting Information). Due to the reported high divergence in tyrosine recombinase sequences especially in the N-terminal region, molecular phylogeny tools are ill- adapted to infer the evolutionary relationships between the members of this superfamily (Esposito and Scocca 1997). In a previous report, we successfully used a different methodology based on protein similarity networks to establish relationships between different archaeal integrase families (Badel *et al.* 2020). A similar approach was conducted on the present archaeal integrase dataset using the SiLiX program (Miele, Penel and Duret 2011) in order to define families sharing 25% identity covering at least 60% of the protein.

Among the 4341 archaeal sequences of the cl00213 superfamily, 93.6% were distributed into three major groups (Fig. 4; Table S1, Supporting Information). The first group, tentatively named pR1SE1, contains integrases encoded by *Haloarchaea* and their MGEs such as the pR1SE plasmid. The second group TopoIB corresponds to the eukaryotic-like topoisomerase I (TopoIB) encoded mostly by *Thaumarchaea*, *Bathyarchaea* and *Aigarchaea*. Our results indicated that no archaeal integrases belong to the group of eukaryotic-like topoisomerases IB, even though a conserved structure between eukaryotic topoisomerase I and the catalytic domain bacterial tyrosine recombinases was reported (Cheng *et al.* 1998). The absence of overall amino acid conservation between these enzymes (Cheng *et al.* 1998) further highlights the divergence we observed. The third group Supfam25-02 is the largest comprising over 89% of the total sequences. It encompasses the characterized archaeal XerA recombinases and integrases identified in several MGE families such as pNOB8, pT26-2, Met26-2, SNJ2, pTN3 and SSV (She, Brugger and Chen 2002; Liu *et al.* 2015; Cossu *et al.* 2017; Wang *et al.* 2018; Badel *et al.* 2020). A similar large clade of archaeal integrases from pNOB8, pT26-2, SNJ2, pTN3 and SSV was previously observed by phylogenetic analysis (Wang *et al.* 2018). Using a slightly more stringent criterion (30% identity covering 60% of the protein), we could rank these sequences into the previously well-characterized archaeal integrases families named after their respective representative SSV, pTN3, pT26-2, SNJ2 and pNOB8. This criterion was instrumental in ranking all archaeal tyrosine recombinases into 17 major families, including newly

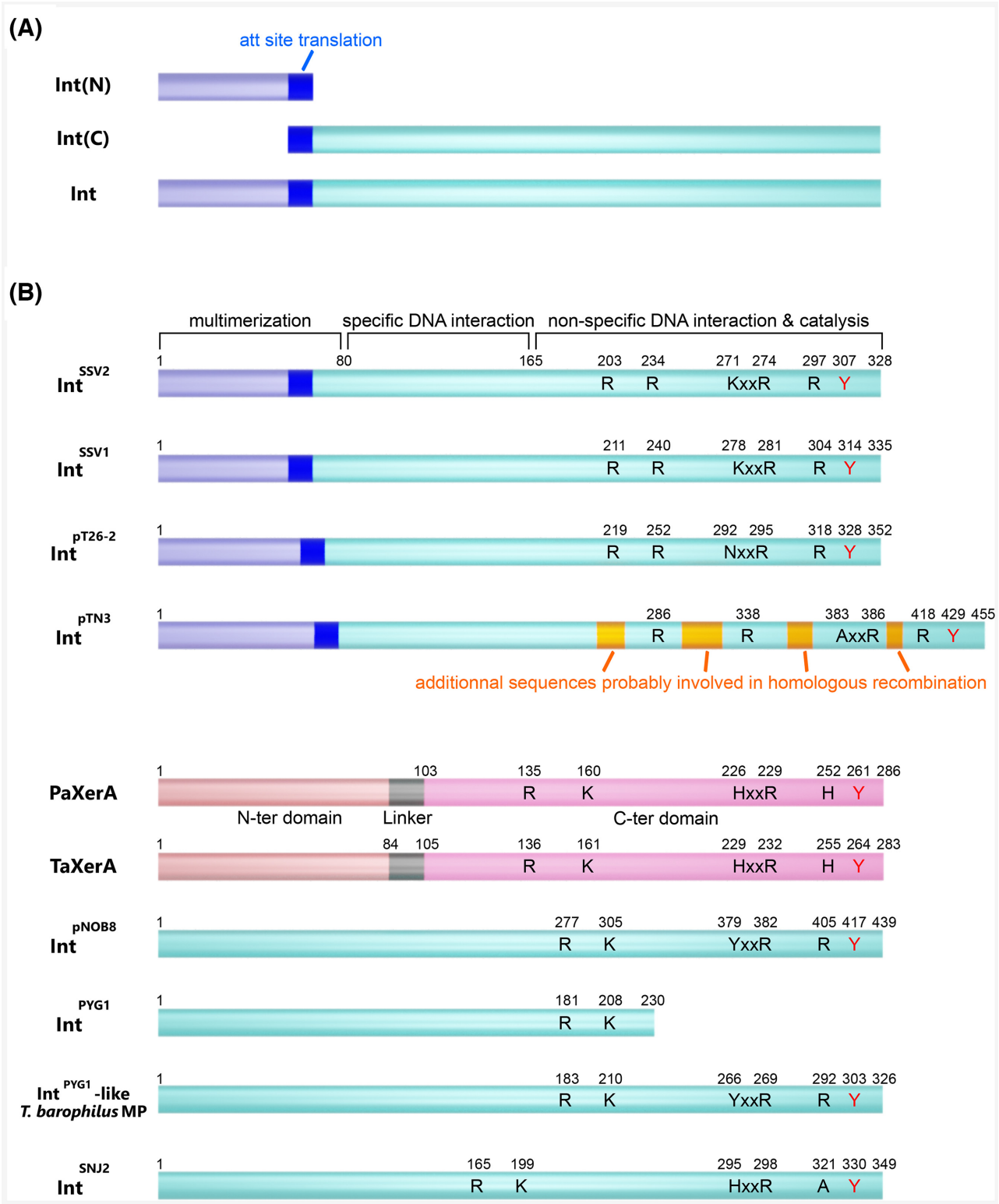

Figure 3. Archaeal integrases sequence domains and conserved residues. **(A)** Suicidal integrases are either encoded by their intact gene or the regions corresponding to the N-terminal portion or Int(N), and C-terminal region or Int(C) are separated upon MGE integration. **(B)** The conserved catalytic residues are indicated for all characterized archaeal tyrosine integrases from Table 1. The domains of particular interest are indicated. Functional domains were dissected for IntSSV2 (Zhan, Zhou and Huang 2015). IntpTN3 presents additional loop that may be responsible for its unprecedented dual catalytic activity (Cossu *et al.* 2017). PaXerA and TaXerA structure was resolved and corresponds to two domains separated by a linker (Serre *et al.* 2013; Jo *et al.* 2016).

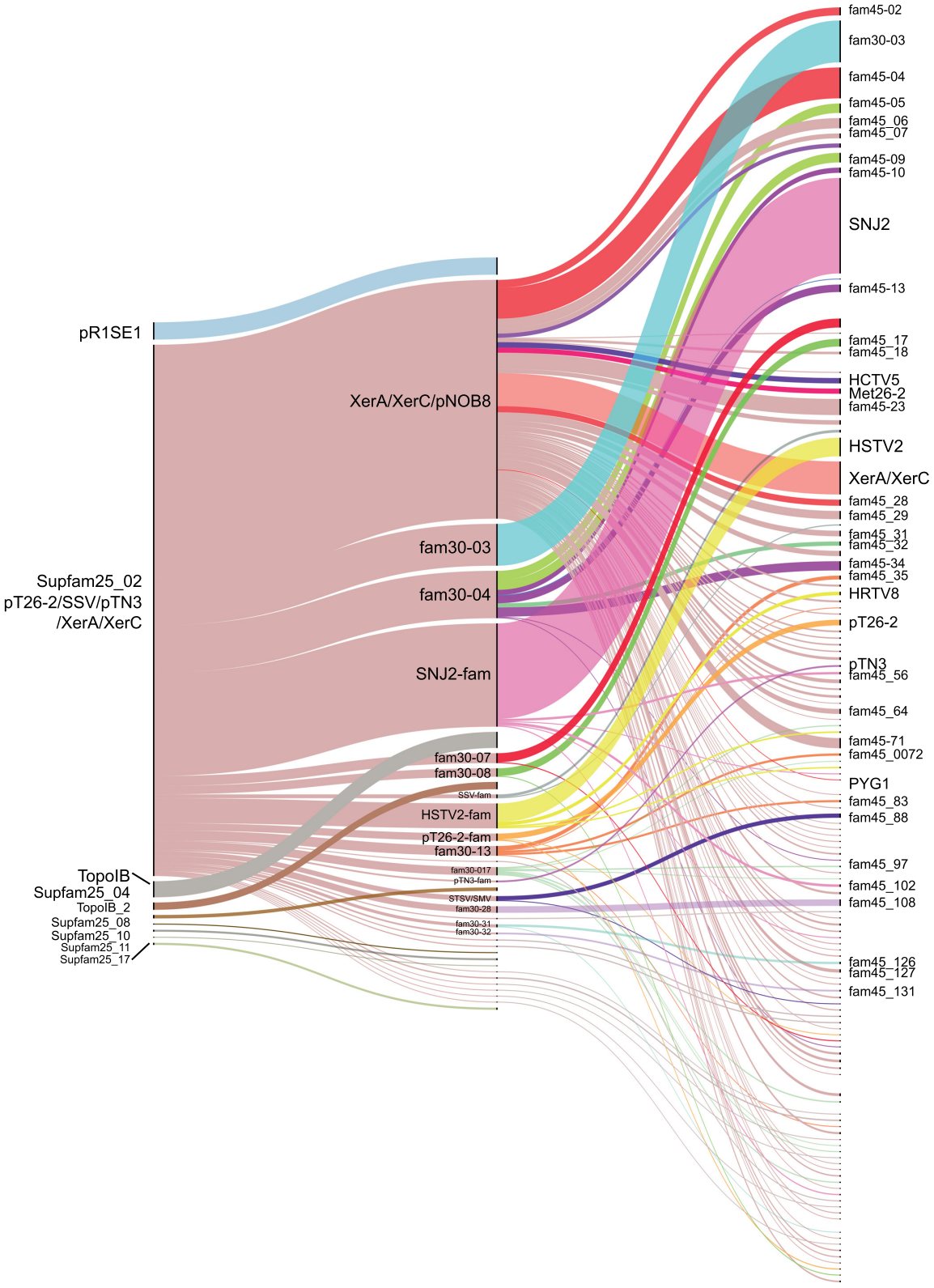

Figure 4. Classification of archaeal tyrosine recombinases. Alluvial diagram showing the integrase distribution across the different archaeal tyrosine recombinase superfamilies (left), families (center) and subfamilies (right), predicted based on the conservation identity threshold of 25%, 30% and 45%, respectively, using SiLiX (Miele, Penel and Duret 2011). In this diagram, blocks represent clusters of proteins and stream fields between the blocks represent changes in clustering attribution of these proteins to superfamily, family and subfamily over the selected identity threshold. Block height is proportional to the size of the protein cluster and the height of a stream field is proportional to the number of proteins contained within the blocks connected by the stream field. Our classification retained the 17 families containing >13 members. The graph was drawn using RawGraphs (https://app.rawgraphs.io) to map data dimensions onto visual variables. The raw data are available in Table S1 (Supporting Information).

identified integrases families such as HSTV2, STSV/SMV, fam30-07 and fam30-04 (Fig. 4; Table S1, Supporting Information).

In order to test whether these connections reflected the conservation of existing domains, we scanned each protein for Pfam domains using the Conserved Domain CD-Search with an *e*-value of 1e-05 (Marchler-Bauer and Bryant 2004) (Fig. S1 and Table S1, Supporting Information). We observed that the different families corresponded mainly to different combinations of Pfam domains, or yet unknown domains. The pTN3like, pT26-2like and SSV corresponded to suicidal integrases with the Phage_integr_3 domain (PF16795). Our classification suggests that suicidal integrases correspond to a recent adaptation originating from a single event within the monophyletic Superfamily 25_02 (Figs 4 and 5) in agreement with previous phylogenetic and network analyses (Wang *et al.* 2018; Badel *et al.* 2019). The TopoIb and TopoIb_2 families both contained the Topoisom_I central domain (PF01028). The TopoIb_2 family is related to viral/bacterial origin whereas the TopoIb family comprises the additional eukaryotic domains Topoisom_I_N (PF02919) Topo_C_assoc (PF14370) (Brochier-Armanet, Gribaldo and Forterre 2008). The major families XerA/XerC/pNOB8 or SNJ2 corresponded to proteins with both Phage_int_SAM (PF02899, PF13495) associated to the Phage_integrase domain (PF00589). In the families HSTV2like, fam30-07, fam30-07, fam30-17 and fam30-28 we detected the single Phage_integrase domain (PF00589). A particular family, fam30-04, composed of larger proteins carried the Phage_integrase (PF00589) BAT (PF15915) and HTH_10 (PF04967) domains with some of them also containing the GAF_2 domain (PF13185). We did not detect known Pfam domains in the fam30-08 and STSV/SMV families. Overall, the diversity of archaeal tyrosine recombinases is explained by the highly variable domain composition of the 17 major families as they only share a single Pfam domain, Phage_integrase (PF00589) (Figs 5 and 6). Archaeal integrases are not equally distributed among archaeal species, as for example the pR1SE, HSTV2 and SNJ2 families are restricted to Halobacteria and their MGEs (Wang *et al.* 2018). Our analysis confirmed this limited distribution, while other integrase families are spread among different classes or phyla such as XerA/XerC/pNOB8 and fam30-08. Further phylogenomic studies will be required to understand how horizontal gene transfer generated the patchy distribution of tyrosine recombinases among archaeal phyla.

A first observation of our network map (Fig. 7) indicated that all archaeal tyrosine recombinases are not connected through a single network, therefore highlighting the inability to perform a robust phylogeny on this dataset. Early reports already discussed the difficulty to compare the primary sequence of archaeal, bacterial and eukaryotic tyrosine recombinases due to the presence of large portions devoid of detectable homology (Esposito and Scocca 1997; Nunes-Duby *et al.* 1998). A number of recombinases diverged greatly from the main group (Esposito and Scocca 1997) and many singleton integrases could not be ranked in any subfamily (Williams 2002). Our network analysis also confirms two previously reported evolutionary relationships, the first connecting the SSV, pTN3 and pT26-2 integrases families and the second connecting the XerA/C recombinases and the SNJ2, Met26-2 and pNOB8 integrase subfamilies (Wang *et al.* 2018; Badel *et al.* 2019) (Fig. 7).

In order to verify established relationships between tyrosine recombinases in the three domains of life, we extended our network analysis with well-studied bacterial and eukaryotic enzymes from previous reviews (Esposito and Scocca 1997; Nunes-Duby *et al.* 1998). We were able to underline a close relationship between most archaeal tyrosine recombinases and bacterial XerC/D chromosomal resolvases (Colloms *et al.* 1990; Blakely *et al.* 1993), Fim tyrosine-integrase (McCusker, Turner and Dorman 2008), TnpA from transposon Tn554 (Bastos and Murphy 1988), conjugative element ICEBs1 integrase (Suzuki *et al.* 2020), P2 phage integrase (Nilsson *et al.* 2011), virophage integrase (La Scola *et al.* 2008) and Integron_Int (Nunes-Duby *et al.* 1998; Demarre *et al.* 2007). In agreement with previous observations by She *et al.* (She, Brugger and Chen 2002), we confirmed that most archaeal integrases belong to the same superfamily, together with many families of bacterial integrases (Fig. S2 and Table S2, Supporting Information). On the other hand, no direct relationship was observed between archaeal tyrosine recombinases and P1 phage Cre, λ phage Int, yeast 2μ plasmid FLP, DusA-associated integrases (DAI) (Farrugia *et al.* 2015), plasmids R64 shufflon-specific DNA recombinase Rci (Kubo, Kusukawa and Komano 1988) and Tn916 tyrosine recombinases (Lu and Churchward 1994) (Fig. S2 and Table S2, Supporting Information). Interestingly, we did not observe the clade formed by the archaeal integrases from the BJ1 and phiCh1/HCTV-5 viruses (Atanasova *et al.* 2012), P1 Cre and 2μ FLP that was reported by Wang *et al.* (2018). Considering that FLP was previously reported as one the tyrosine recombinases that had greatly diverged (Esposito and Scocca 1997), this discrepancy might be attributed to long branch attraction issues in Wang *et al.* phylogenetic analysis, an artifact to which network analysis is immune. Remarkably, the pR1SE1 archaeal integrase family (Erdmann *et al.* 2017) could not be connected to any of the other tested tyrosine recombinase families.

We have observed many integrase families that do not share relationships with any other known family (Fig. 4; Table S1, Supporting Information). Among these, 27 families harbor <7 members and are encoded, for example, by well-studied viruses such as the *Sulfolobus* turreted icosahedral viruses 1 and 2 (STIV1 and STIV2) (Rice *et al.* 2004; Happonen *et al.* 2010) or by the uncharacterized *Methanocaldococcus* sp. FS406.22 plasmid pFS01 (METSF_2, Joint Genome Institute, unpublished). In addition, we witnessed up to 104 individual integrases not ranked in any family and constituting as many shadow areas remaining to be explored. With the increase in available sequences, we anticipate further analyses able to establish phylogenetic relationships between these tyrosine recombinases and identify new conserved residues conveying additional functions.

Insights from tertiary structures

The first full length tyrosine recombinase structure was obtained for the *E. coli* XerD resolvase and consisted of two domains separated by a disorganized linker (Subramanya *et al.* 1997). Subsequently, the resolution of the co-crystal structure of phage P1 Cre recombinase with its Lox site led to a structural model of the site-specific recombination reaction catalyzed by tyrosine recombinases (Guo, Gopaul and van Duyne 1997) (Fig. 8A). Both XerD and Cre crystal structures harbored an unfolded linker that separates the N-terminal domain from the C-terminal catalytic domain. The structure of the yeast 2μ plasmid Flp recombinase tetramer bound to an Holliday junction later revealed that the helix containing the nucleophilic tyrosine is swapped to cut the DNA in *trans* (Chen *et al.* 2000). When DNA cleavage occurs in *trans*, an integrase monomer activates the sessile phosphodiester bond while the adjacent monomer supplies the catalytic tyrosine. On the contrary, in the XerD and Cre structures, DNA cutting occurs in *cis* meaning that the same integrase monomer supplies the entire active site, as for the majority of bacterial integrases (Jayaram *et al.* 2015). The crystal structures of phage λ Int with its DNA substrates showed the simultaneous binding of two separate protein domains to

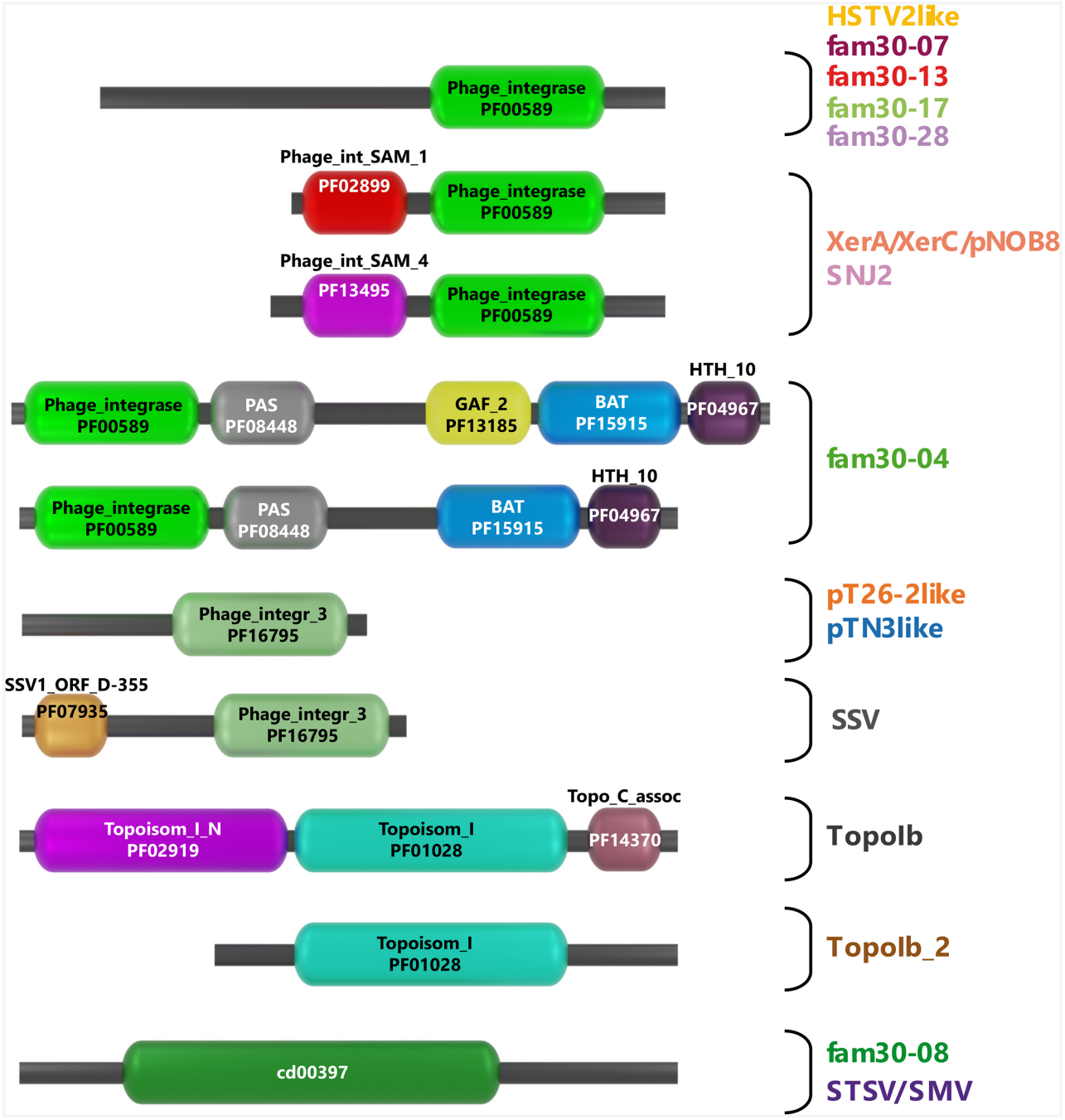

Figure 5. Pfam domain combinations in the 17 major tyrosine recombinase families. The most frequent combinations of Pfam domains for each major tyrosine recombinase are represented. No Pfam could be detected for the fam30-08 and STSV/SMV families even if both belong to the cd00397 superfamily. The result of the conserved domain search is available in Table S1 (Supporting Information).

DNA and suggested that the additional arm binding shifted the reaction equilibrium toward recombinant products (Biswas *et al.* 2005).

The only structural studies devoted to archaeal tyrosine recombinases concerned the integrase from the spindle-shaped virus SSV1 of *Sulfolobus shibatae* (suicidal integrase class) (Zhan *et al.* 2012), the resolvases PaXerA from *Pyrococcus abyssi* (Serre *et al.* 2013) and XerA from *Thermoplasma acidophilum* (TaXerA) (Jo *et al.* 2016). All three structures unsurprisingly revealed that archaeal tyrosine recombinases present a catalytic fold similar to bacterial and eukaryotic integrases (Eilers, Young and Lawrence 2012; Serre *et al.* 2013; Jo *et al.* 2017). Moreover, archaeal PaXerA and TaXerA proteins display the canonical structure of tyrosine recombinases comprising two domains surrounding the DNA substrate in a C-shape conformation (Serre *et al.* 2013) (Fig. 8A). The active sites of PaXerA and TaXerA assemble in *cis* (Jo *et al.* 2016) whereas IntSSV1 and *Thermococcus nautili* IntpTN3 catalyze DNA cleavage in *trans* (Letzelter, Duguet and Serre 2004; Eilers, Young and Lawrence 2012; Cossu *et al.* 2017). At this stage, only the structures of the catalytic C-terminal domain of archaeal recombinases have been resolved. The future resolution of the complete structure of archaeal tyrosine recombinases

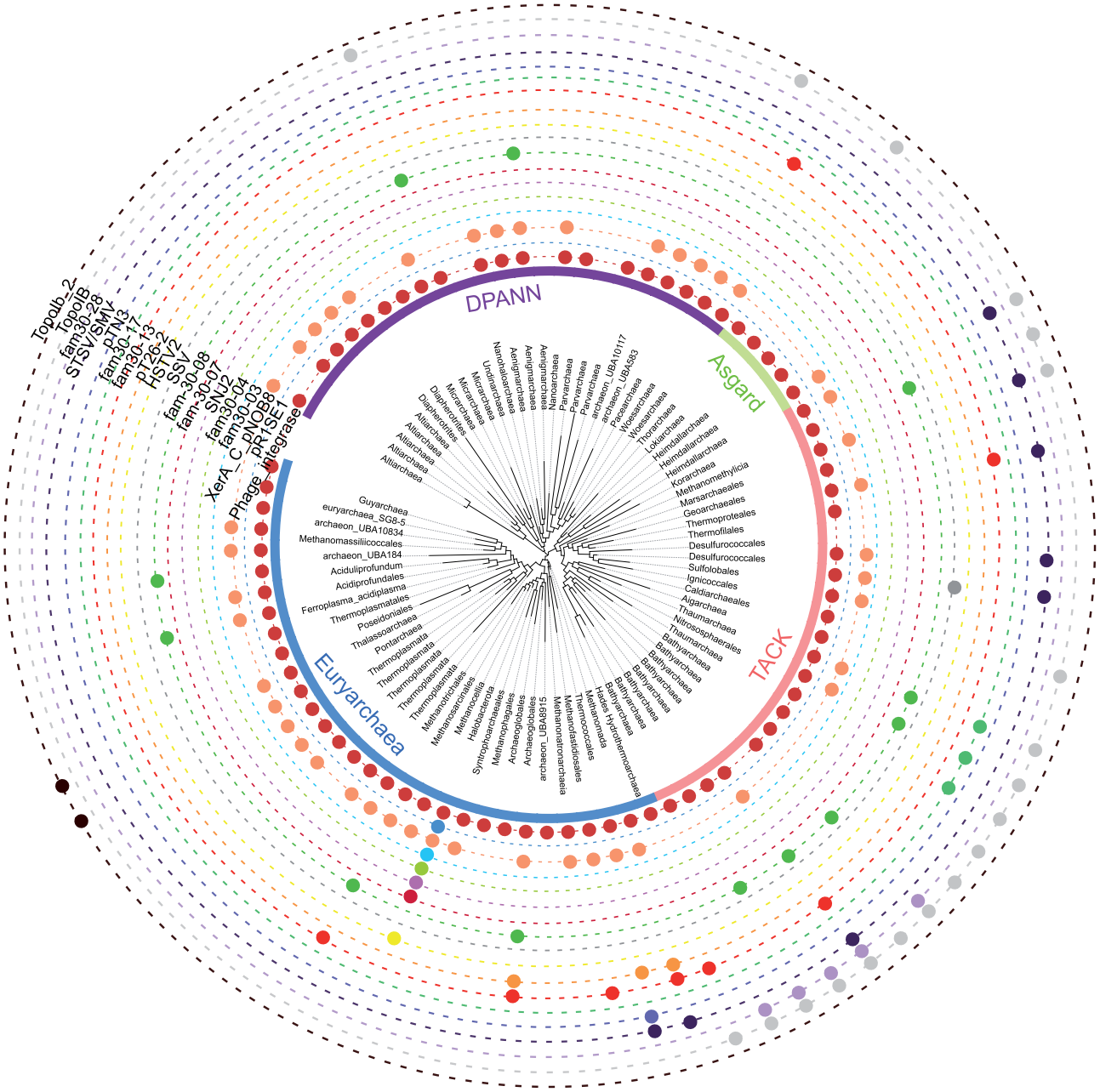

Figure 6. Distribution of the 17 major tyrosine recombinase families among the archaeal diversity. Each tyrosine recombinase family corresponds to a dot. In order to visualize the distribution of the major tyrosine recombinase families on the archaeal diversity, we used data from AnnoTree (Mendler *et al.* 2019) with additional manual curation. The archaeal tree was obtained from the Genome Taxonomy Database (GTDB Release 03-RS86) (Parks *et al.* 2020), generated from 122 core proteins and exported using taxonomic orders as resolution level. The presence/absence profiles for each family were visualized using iTOL (Letunic and Bork 2019).

especially while bound to their DNA substrate might highlight further archaeal peculiarities.

THE MECHANISM OF SITE-SPECIFIC RECOMBINATION

A common catalytic mechanism for different reaction directionalities

The standard site-specific recombination reaction requires a tetramer of recombinases and a pair of identical DNA sequences specific to the enzyme involved (Grindley, Whiteson and Rice 2006) (Fig. 8B). The identity constraint can however be relaxed for one of the two sequences (Rajeev, Malanowska and Gardner 2009). The first stage of the reaction corresponds to the recruitment of the integrases to the specific site and to their tetramerization, resulting in the formation of a synaptic complex. The recombinases then catalyze two coordinated staggered single strand DNA cuts through the formation of a transient 3′-phosphodiester covalent bond between the active tyrosine and DNA (Grindley, Whiteson and Rice 2006). Then, strand exchange occurs between the two DNA segments and a nucleophilic attack by the 5′-terminal hydroxyl group of the invading strand resolves the covalent complex. The process is repeated leading to a total of four recombination reactions.

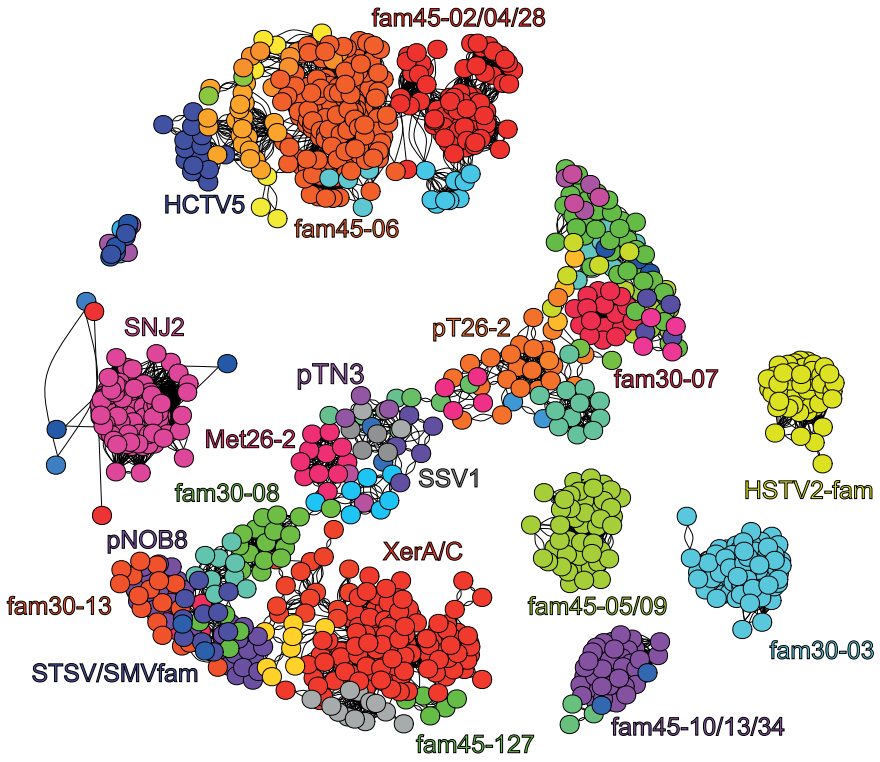

Figure 7. Similarity network of archaeal tyrosine recombinases. A similarity network performed with SiLiX (Miele, Penel and Duret 2011) and visualized with Igraph (https://igraph.org) was used to assign tyrosine recombinases from the superfamily Superfam25-02 to families and subfamilies. The similarities of all against all proteins of the dataset were assessed using BlastP (expect >0.001, with an identity threshold of 30% among 60% of the protein). Protein clustering was achieved by a random walk algorithm. Each circle corresponds to an individual protein colored according to the clustering.

Depending on the topology linkage of the two DNA sequences, the outcome of the recombination varies (Fig. 9A). Recombination between two sites carried by two independent circular molecules results in their integration. The newly formed chimeric circular molecule harbors the specific site in two copies in direct orientation. The inverse recombination reaction between these two sites produces an excision and the two initial circular molecules are restored. Recombination between two sites carried in opposite orientations by a single circular molecule produces an inversion (Fig. 9B). Integration corresponds to an intermolecular reaction while excision and inversions are intramolecular reactions. Site-specific recombinases can also catalyze recombination between two linear DNA molecules resulting in two chimeric linear DNA molecules (Fig. 9C).

Helper proteins and recombination directionality factors

The symmetry of the synaptic complex illustrates the reversible nature of site-specific recombination reactions (Fig. 8B). In order to control the directionality of the reaction, small accessory proteins are often involved in stimulating one reaction while suppressing the other. These proteins are referred as recombination directionality factors or RDFs (Lewis and Hatfull 2001). The best-studied RDF is the phage λ-encoded Xis protein that is necessary for excisive recombination and inhibits integrative recombination (Abremski and Gottesman 1982). Phage λ excision is also favored by the host-encoded Fis protein (Papagiannis *et al.* 2007). The bacterial helper protein IHF contributes to both λ integration and excision reactions although integration requires more IHF than does excision for optimal reaction (Bushman *et al.* 1985). The completion of the resolution reaction of dimeric bacterial chromosomes by the XerC/D recombinases required the activity of the cell division protein FtsK (Barre *et al.* 2000).

Unessential archaeal cofactors were identified *in vivo* concurring in IntSNJ2 activity in *Natrinema* sp. J7-1 (Wang *et al.* 2018). The gene orf1 coding for the integrase is transcribed in an operon with two the other genes orf2 and orf3. Orf2 and orf3 code for small proteins (111 and 140 residues, respectively) containing a coiled coil motif that could mediate protein-protein interactions and a MarR-like DNA binding domain, respectively. The presence of one or both proteins increased IntSNJ2 *in vivo* integration activity by 30% (Wang *et al.* 2018). For the inversion reaction, the recombination efficiency was increased 70-fold in the presence of a single protein and 180-fold in the presence of both. They cooperatively activated IntSNJ2 recombination activity through an undetermined mechanism. Nevertheless, IntSNJ2 is active in their absence and the operons of many SNJ2-like viruses do not encode these cofactors (Wang *et al.* 2018).

Site-specific recombination activity of unescorted archaeal integrases

The aforementioned molecular model can to all account be extended from bacterial and eukaryotic to archaeal tyrosine recombinases. In archaeal cells, site-specific recombination substrates are circular molecules. Several archaeal tyrosine recombinases were proven to be active *in vitro* and *in vivo* through various activity assays (Table 1; Fig. 10A). The first archaeal recombinase whose activity was tested *in vitro* is the *Sulfolobus* spindle-shape virus 1 integrase IntSSV1. The

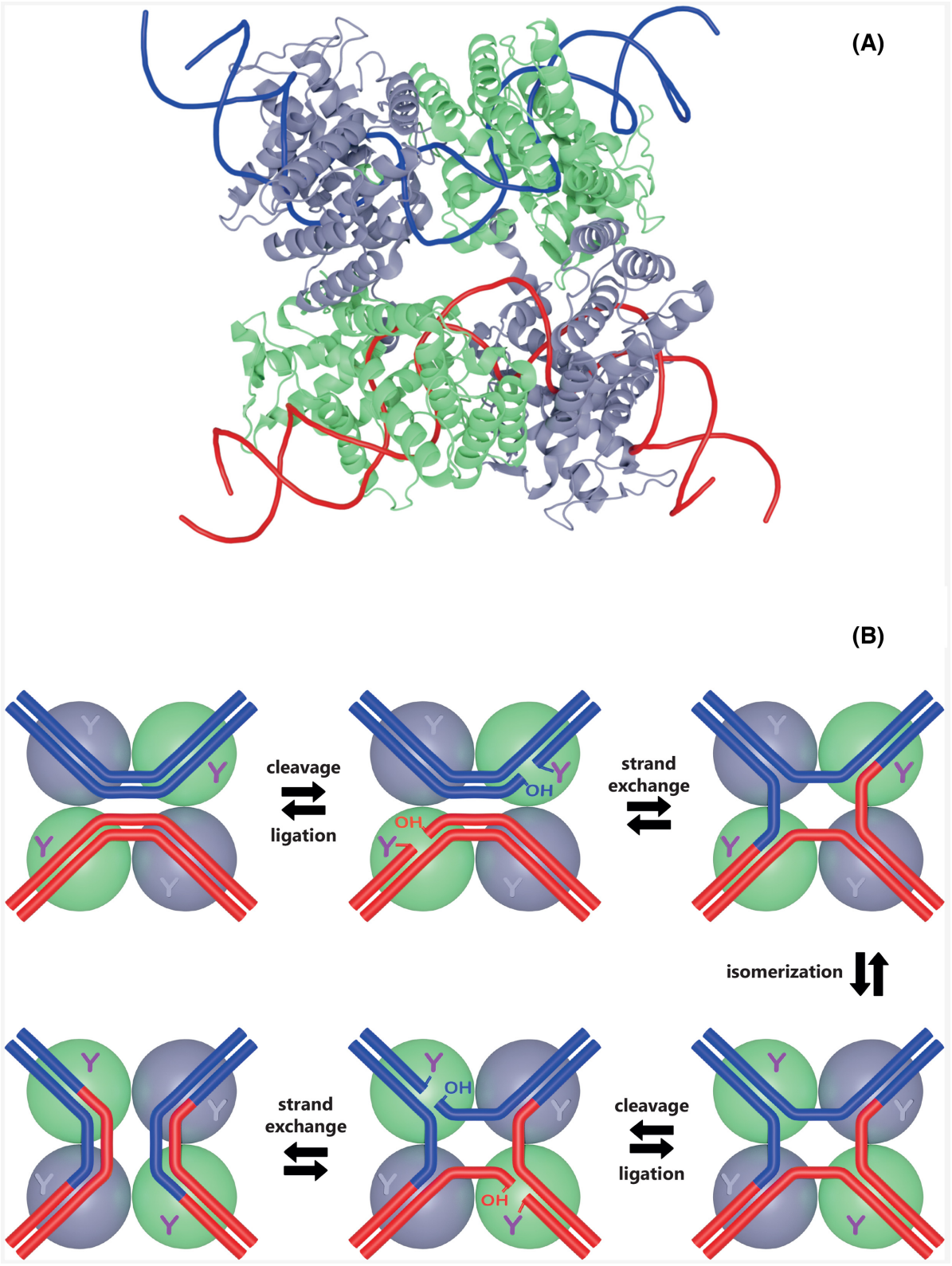

Figure 8. Tyrosine recombinase model and site-specific recombination reaction. **(A)** Structural model of the Cre-loxA pre-synapse complex. The tetrameric conformation of the Cre-loxA pre-synapse is clearly apparent with its active cleaving subunits in green color. The DNA components of the recombination complex are shown in blue and red backbone form. The tridimensional model is referenced as PDB 1NZB (Guo, Gopaul and van Duyne 1997). **(B)** Site-specific recombination model. Schematic representation of the consecutive reactions leading to the formation of recombinant DNA molecules by tyrosine recombinases. The color code is consistent with that of Panel A. The active tyrosine residues are shown in purple color.

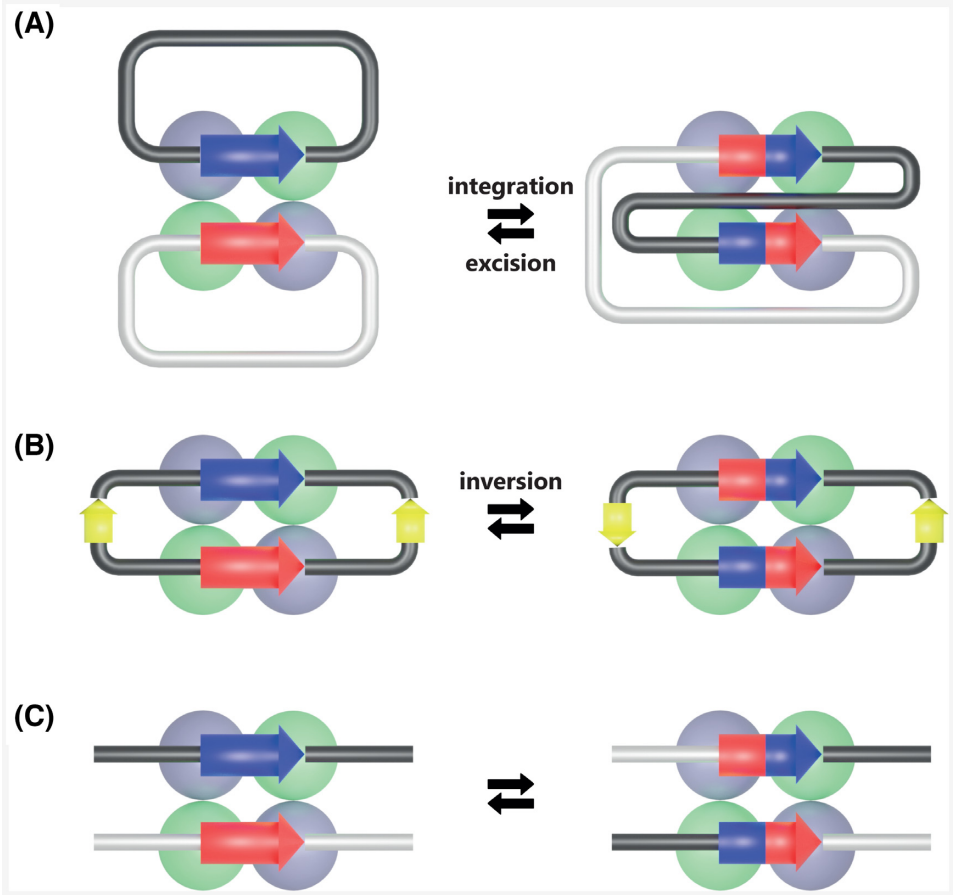

Figure 9. The different outcomes of site-specific recombination. The red and blue rectangles correspond to the specific recombination sites. With circular DNA molecules as substrates, the recombination outcome can be **(A)** integration and excision or **(B)** inversion, depending on the relative position of the two specific sites. **(C)** Recombination between two linear DNA molecules results in two chimeric linear DNA molecules.

Table 1. Published archaeal tyrosine recombinase groups.

Representative integrase	Integrase type	Host order	Activity demonstrated	Biochemical analysis	Structure resolution	Reference
XerA	Classical	All chromosomally encoded	Yes	Yes	Yes	(Cortez *et al.* 2010; Jo *et al.* 2017)
pNOB8 integrase	Classical	*Sulfolobales*	No	No	No	(She, Brugger and Chen 2002)
PYG1 integrase	Classical	*Thermococcales*	Yes	No	No	(Li *et al.* 2016)
SNJ2 integrase	Classical	*Halobacteriales*	Yes	No	No	(Wang *et al.* 2018)
SSV1 integrase	Suicidal	*Sulfolobales* (*Desulfurococcales*)	Yes	Yes	Yes	(Serre *et al.* 2002; Zhan, Zhou and Huang 2015)
pTN3 integrase	Suicidal	*Thermococcales*	Yes	Yes	No	(Cossu *et al.* 2017)
pT26-2 integrase	Suicidal	*Thermococcales*, *Archaeoglobales* (*Methanosarcinales*)	Yes	Yes	No	(Badel *et al.* 2020)

recruitment to a specific site and tetramerization was found to be rate limiting for IntSSV1 (Serre *et al.* 2002). Its recombinase activity observed *in vitro* by Muskhelishvili *et al.* (Muskhelishvili, Palm and Zillig 1993) could not be reproduced except for the first step of the recombination reaction, i.e. strand cleavage (Muskhelishvili, Palm and Zillig 1993; Serre *et al.* 2002; Letzelter, Duguet and Serre 2004; Zhan, Zhou and Huang 2015) (Fig. 10B). More recently, the tyrosine recombinases PaXerA from *Pyrococcus abyssi* and TaXerA from *Thermoplasma acidophilus*, typically resolving chromosome dimers *in vivo* were shown to catalyze integration reactions *in vitro* making of them *bona fide* integrases (Cortez *et al.* 2010; Serre *et al.* 2013; Jo *et al.* 2017). The two suicidal *Thermococcales* integrases IntpTN3 from plasmid pTN3 and IntpT26-2 from plasmid pT26-2, were also shown to catalyze site-specific recombination on circular substrates *in vitro*. (Cossu *et al.* 2017; Badel *et al.* 2020). All tested arrangements of specific sites allowed recombination in the absence of any additional cofactor. This suggests that, contrary to most bacterial

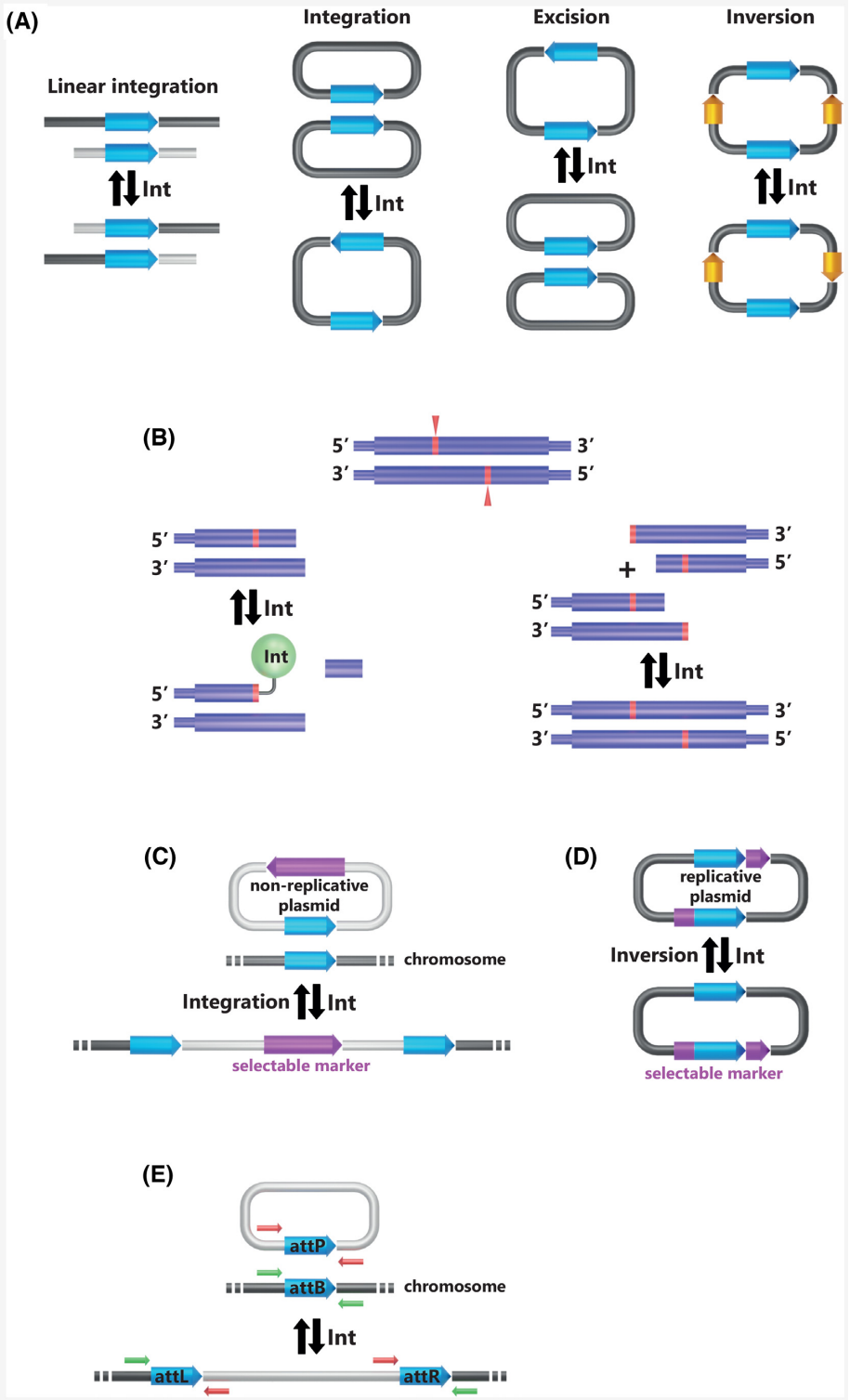

Figure 10. Assays to detect integrase activity *in vitro* and *in vivo*. **(A)** Different substrates harboring specific sites (light blue arrows) are incubated with the integrases *in vitro* and the products were monitored (Cortez *et al.* 2010; Zhan, Zhou and Huang 2015; Cossu *et al.* 2017; Jo *et al.* 2017; Badel *et al.* 2020). **(B)** The half-site strand transfer assay was first implemented *in vitro* by Serre *et al.* (2013). It allows the verification of the strand cleavage site (in red). The incubation of the integrase with half of the specific site results in a covalent DNA–protein complex that can be detected (left). The incubation of the integrase with two separated halves of the specific site results in the reconstruction of the entire specific site only if the substrate halves were designed accordingly to the cleavage site (right). **(C)** A non-replicative plasmid harboring the specific site and a selectable marker is introduced in a suitable host cell. Upon selection, only the cells where the integrase catalyzes plasmid integration can grow (Wang *et al.* 2018). **(D)** A replicative plasmid harboring two specific sites and a split selectable marker is introduced into the appropriate cell. The selectable marker is reconstituted only if the integrase catalyzes recombination between the two specific sites. The cell can then grow upon selection (Wang *et al.* 2018). **(E)** Different arrangements of the specific site result from integration or excision. They can be detected by polymerase chain reaction with different pairs of four primers (red and green arrows) (Li *et al.* 2016; Cossu *et al.* 2017; Wang *et al.* 2018).

integrases (Landy 2015), archaeal integrases do not require any RDF for efficient *in vitro* recombination. The integrases IntpTN3, IntSSV2, PaXerA and TaXerA could also catalyze site-specific recombination on linear substrates *in vitro* (Cortez *et al*. 2010; Serre *et al*. 2013; Landy 2015; Zhan, Zhou and Huang 2015; Cossu *et al*. 2017; Jo *et al*. 2017). Linear substrates are not their natural substrates but are useful to characterize some aspects of the integrase activity such as the strand cleavage site (Serre *et al*. 2013) (Fig. 10B). Additionally, three integrases (IntPYG1, IntpTN3, IntSNJ2) were shown to catalyze site-specific recombination *in vivo* (Li *et al*. 2016; Cossu *et al*. 2017; Wang *et al*. 2018)(Fig. 10C–E). Overall, the activity of several archaeal integrases from several families was characterized whose most remarkable aspect is the recurrent absence of necessary cofactor for catalysis, in stark contrast with most bacterial model integrases.

DNA relaxation activity of archaeal integrases

Despite the very limited overall sequence conservation observed in our network analysis, eukaryotic and bacterial tyrosine recombinases and topoisomerases IB share a common catalytic core that could have originated from an ancestral strand transferase (Cheng *et al*. 1998; Yang 2010). This relationship explains why tyrosine recombinases can often catalyze DNA relaxation (Abremski *et al*. 1986; Landy 2015). The archaeal recombinases IntpT26-2, IntSSV1 and TaXerA also presented non-specific DNA relaxation activities (Letzelter, Duguet and Serre 2004; Jo *et al*. 2017; Badel *et al*. 2020). This property underlines a similar relationship between archaeal tyrosine recombinases and topoisomerases IB. It is to be noted however that topoisomerases cleave and then join the same 5′ and 3′ termini, whereas site-specific recombinase transfer a 5′ hydroxyl to a new 3′ phosphate partner from a different strand.

ARCHAEAL DNA RECOMBINATION TARGETS

Attachment site characteristics

The attachment sites define the DNA segments containing the points of strand exchange and the binding site for site-specific recombinases. These enzymes catalyze the recombination between the attP (attachment phage) site on the episomal MGE and attB (attachment bacteria) on the chromosome (Landy 2015) (Fig. 2). This reaction generates the two hybrid attL (attachment Left) and attR (attachment Right) sites bordering the integrated MGE. In the canonical model derived from the structure of the Cre/LoxP synapse, the four att sites are strictly identical and correspond to the recombinase specific site (Fig. 8). The lambdoid phages constitute a notable exception to this rule: their ∼240 bp attP site carries multiple binding sites for Int, Xis and IHF (Hsu, Ross and Landy 1980) whereas the ∼20 bp attB carries only two Int binding sites (Mizuuchi and Mizuuchi 1980). These attB and attP share an identical DNA stretch of 15 bp called the core containing the two points of strand exchange. The 7bp interval between these exchanges on the two strands is the overlap region (Craig and Nash 1983) (Fig. 11A).

Archaeal attachment sites have been defined by the extent of exact DNA sequence shared by MGEs and their host chromosome. They usually extend over 40 to 50 bp for suicidal integrases (She, Brugger and Chen 2002; Cossu *et al*. 2017), 40 to 50 bp for *Sulfolobales* pNOB8 integrases (She, Brugger and Chen 2002; Erauso *et al*. 2006) and 50 to 60 bp for Methanococcales integrases (Badel *et al*. 2020). They can be as short as 8 bp for *Thaumarchaeota* integrated elements (Krupovic *et al*. 2019), 11 bp for the *Methanobacteriales* Msmi-Pro1 integrated virus (Krupovic, Forterre and Bamford 2010) or 13 bp for the *Halovivax* SNJ2 integrated virus (Liu *et al*. 2015) or they can be longer than 100 bp for some *Thermococcales* elements (102 bp for PkuNCB100_IP1 and 243 bp for TIRI33c_IE1) (Badel *et al*. 2019, 2020). Interestingly, the large att site of these two *Thermococcales* MGEs encompassed the shorter att site of the closely related integrases from the PHV1 and TGV1 viruses, respectively (Badel *et al*. 2020).

Several studies aimed at characterizing experimentally the extent of archaeal att sequences required for efficient integration and postulated the existence of a minimal and sufficient recombination site. It appears however that att sites defined by a strict sequence conservation between MGE and host chromosomes are inoperative for recombination. For IntSSV1 and IntSSV2, two sequences were suggested to be sufficient for recombination *in vitro*: the *stricto sensu* att site or the inverted-repeats separated by an overlap region, reminiscent of the phage λ core (Muskhelishvili, Palm and Zillig 1993; Serre *et al*. 2002; Zhan, Zhou and Huang 2015) (Fig. 11A). These two sequences did not completely overlap and the sequence at the intersection was not assayed for recombination. The minimal site remained therefore undefined. For IntSSV1, the attB strand exchanges were observed in a tRNA gene with the 5′ cuts bordering an overlap region that corresponded to the tRNA anticodon loop as for classical bacterial tyrosine recombinases (Serre *et al*. 2002; Grindley, Whiteson and Rice 2006). However, the *in vivo* observation of non-specific integration events suggested that the strand cleavage position could vary (Wiedenheft *et al*. 2004). For both IntpTN3 and IntpT26-2, the identical DNA stretch shared by the attL and attR sites is not sufficient for recombination *in vitro* (Cossu *et al*. 2017; Badel *et al*. 2019, 2020). Additional nucleotides are required in order to encompass the anticodon loop and the proximal stem extremity (Fig. 11A–C). It could be extrapolated that the cleavage sites for IntpTN3 and IntpT26-2 border the anticodon loop as reported for IntSSV1 or the D and T loops, respectively (Fig. 11B). A cleavage site at the extremities of a tRNA loop can also be considered for IntPYG1 but not for IntSNJ2 whose att site is very short (Fig. 11A–C). It would be interesting to determine whether the 14 nt long and stem loop-free att site from IntSNJ2 is sufficient for recombination. Finally, for IntpT26-2 recombination, the nucleotides of the acceptor stem are not necessary but their presence significantly increases recombination efficiency. The recombination site does not seem to be a precisely defined and finite sequence but rather a stretch of nucleotides that favor recombination. The effective recombination site of IntpT26-2 is not located at the center of the att site but shifted toward its 5′ end, similarly to IntSSV1 (Serre *et al*. 2002; Zhan, Zhou and Huang 2015) and numerous bacterial integrases (Campbell 1992). Overall, archaeal attachment sites are reminiscent of their bacterial counterparts but present the peculiarity to require additional nucleotides outside the conserved sequence between attB and attP.

If most integrases display a marked preference for a particular specific site on the host chromosome, they can also target slightly different sequences albeit with a reduced efficiency. Phage λ Int could recognize many such sites whose sequence deviated from the original att while retaining structural features such as the twist and roll angle between adjacent base pairs (Nussinov and Weisberg 1986). The presence of secondary attachment sites could play a determinant role in the specify switch and evolution of integrases (Rutkai *et al*. 2006). The existence of secondary attachment sites was investigated for

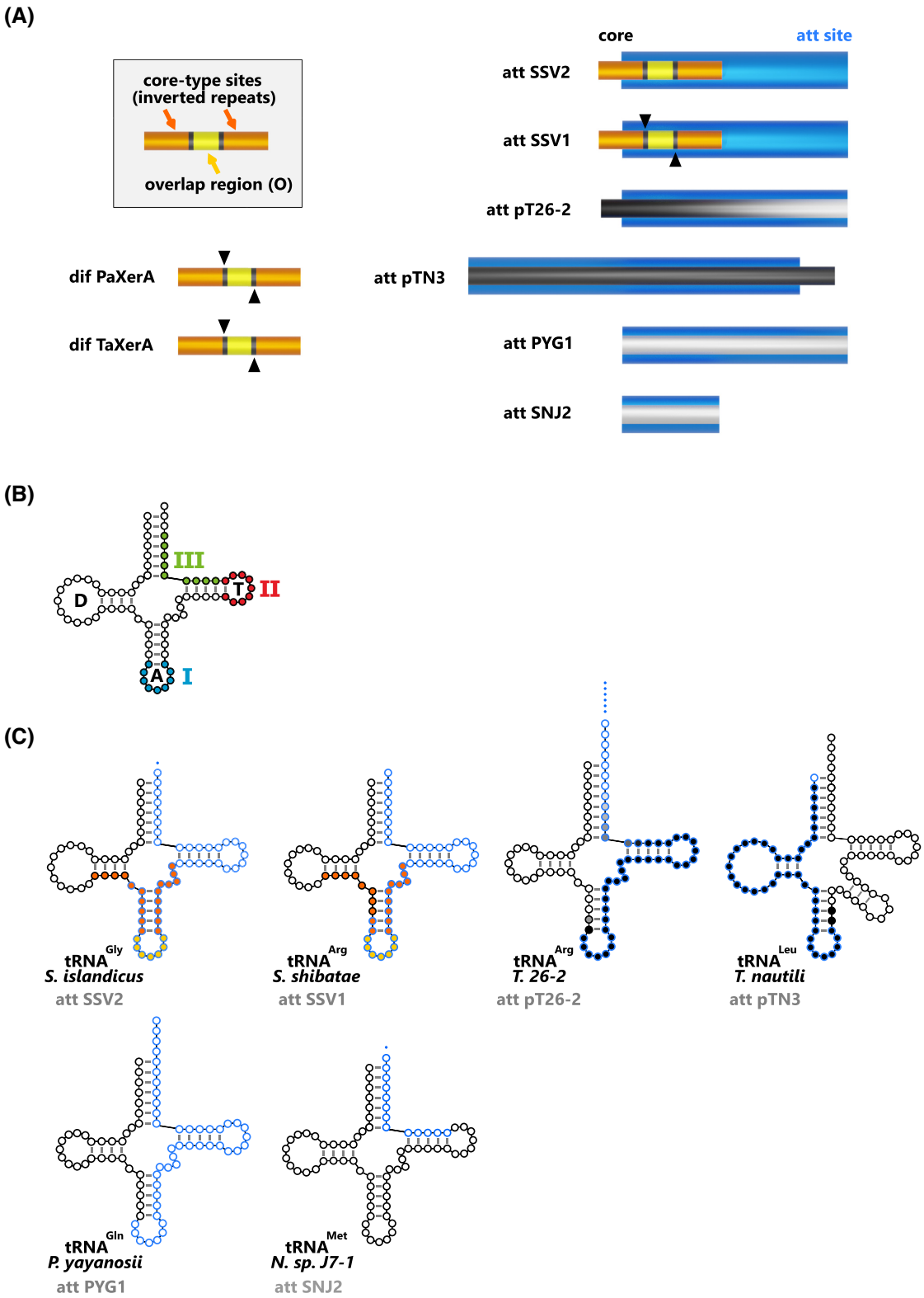

Figure 11. Archaeal tyrosine recombinase recombination sites. **(A)** Recombination sites are sketched for the characterized archaeal tyrosine recombinases. The orange and yellow boxes correspond to core-type sites as defined for bacteriophage λ (Landy 2015). The blue box corresponds to the att site. The black sequences are necessary and sufficient for recombination *in vitro* (Cossu *et al.* 2017; Badel *et al.* 2020). The black arrows indicate the cleavage site when experimentally determined (Serre *et al.* 2002, 2013; Jo *et al.* 2017). **(B)** General organization of a tRNA indicating the location of the T, A and D loops and the three preferred integration locations for bacterial integrases (I, II and III) (Williams 2002). **(C)** Att sites often correspond to tRNA sequences. The leaf-like structure of the targeted tRNA is indicated with att site nucleotides circled in blue.

integrases of archaeal fuselloviruses. The SSV2 virus was only found integrated in its cognate attB site and not in any other slightly divergent site (Contursi *et al.* 2006). The recombination site specificity was found to be more relaxed for the SSV1 virus that, in absence of its cognate attB site, could integrate in a sequence differing in two nucleotides (Schleper, Kubo and Zillig 1992; She *et al.* 2001a; Contursi *et al.* 2006).

Most integration sites reside in tRNA genes

The abundance of attachment sites located within tRNA genes has been extensively documented for bacterial integrases (Williams 2002). Due to the conservativeness of site-specific recombination, an intact tRNA gene sequence is reconstituted after MGE integration. With very few exceptions, the recombinant bacterial tRNA gene is expressed from its original promoter (Williams 2002). The preference for these targets is due in part to the remarkable structural similarities between tRNA genes features and the canonical attachment site defined by phage λ attB. DNA segments corresponding to the usually 7 bp-long anticodon loop and flanking palindromic stem match remarkably the consensus 7 bp overlap region and core site organization (Campbell 2003). This similarity suggests that tRNA genes are common integration sites because they were the target of a primordial tyrosine recombinase (Reiter, Palm and Yeats 1989; Campbell 1992). Additional reasons were invoked for the selection of tRNA genes as preferential attachment targets. First, all tRNA genes harbor several characteristic regions of dyad symmetry that could serve as binding sites for recombinases (Reiter, Palm and Yeats 1989). Second, tRNA genes are very stable through time (Williams 2002), it is therefore possible that ancestral tyrosine recombinases targeting other sequences disappeared when their target sequence changed. Finally, tRNA genes belong to a multigene family offering a multitude of potential target sites with only limited nucleotide changes (Winckler, Szafranski and Glockner 2005).

Archaeal tRNA genes constitute preferential integration targets as well (She, Brugger and Chen 2002; Cossu *et al.* 2017; Wang *et al.* 2018; Badel *et al.* 2019, 2020; Krupovic *et al.* 2019). Recently, a systematic survey of all integrated MGE in thaumarchaeal genomes showed that more than half of the attB sites were located in tRNA genes (Krupovic *et al.* 2019). Additionally, half of the tRNA genes present in *Thaumarchaeota* were used as integration site at least once, including tRNA genes with introns (Krupovic *et al.* 2019). The evolutionary stability of tRNA genes sequences is also exploited by archaeal integrases. This sequence conservation allowed the SSV2 virus to integrate in multiple genomes such as those of *Sulfolobus islandicus* and *S. solfataricus* (Contursi *et al.* 2006). Similarly, the closely related integrases form plasmid pXZ1 and SSVA fusellovirus targeted two separate $tRNA^{Glu}$ genes differing by a single nucleotide (Peng 2008).

In a more detailed approach, it appeared that three different regions of the tRNA gene can be used for bacterial integration (Williams 2002). Two of these regions, the anticodon stem loop and the T stem loop contain a dyad symmetry whereas the third region, located at the 3′ end, has no symmetry (Fig. 11B). While archaeal integrases have been found to target the same three regions, a preference emerged for the 3′ end of tRNA genes with various 5′ limits (Serre *et al.* 2002; Badel *et al.* 2019, 2020). The archaeal att site could be somewhat larger and overlapped both the anticodon loop and the T stem loops as for SSV2 integration (Contursi *et al.* 2006) or could be as short as the amino acid attachment site as for SNJ2 integration (Liu *et al.* 2015). The pTN3-like integrases were found to be unique in archaea in that their attB site corresponds to the 5′ half of the tRNA gene (Cossu *et al.* 2017). As it is for all integration events, pTN3-like integrases restore an uninterrupted copy of the original tRNA gene. A notable difference resides in the fact that these recombinant tRNA genes are expressed from the integrase promoter (Krupovic and Bamford 2008; Cossu *et al.* 2017). A similar situation has been seldomly encountered in bacteria (Williams 2002). In rare documented event, the non-specific integration of virus SSVK1 into a $tRNA^{Glu}$ gene sporadically generated a $tRNA^{Asp}$ gene (Wiedenheft *et al.* 2004).

We inventoried the tRNA genes used as integration sites in archaea and observed marked differences in the targeting frequencies of the various tRNA genes (Fig. 12). Out of the collection of 44 tRNA genes, only 7 were never targeted suggesting that their use as integration site would be deleterious. On the other hand, $tRNA^{GluTTG}$, $tRNA^{ArgTCT}$ and $tRNA^{ValCAC}$ were used more frequently than the others. By comparing preferred codon usage and integration frequency in all sequenced *Thermococcales*, it appeared that targeting occurs preferentially in genes encoding tRNAs that read rare codons (our unpublished observation). This result is consistent with the observation that the tRNA genes most frequently used as integration targets by *E. coli* phages were the least expressed and corresponded to the rarer codons (Bobay, Rocha and Touchon 2013).

In general, archaeal integrases select their tRNA gene targets following the same rules as bacterial integrases with some notable exceptions. Archaeal attB sites can be somewhat larger and extend over the tRNA D-loop and even outside the tRNA gene sequence.

Integration in other intragenic sequences and intergenic regions

Genes encoding tRNAs are not the sole targets of archaeal integrases. Several archaeal attB sites were reported in intergenic sequences or within protein coding genes (Luo *et al.* 2001; Krupovic, Forterre and Bamford 2010; Krupovic *et al.* 2019). The att sites of prophage ΨM100 of *Methanothermobacter wolfeii*, correspond to an intergenic region and prophage integration has no effect on adjacent gene transcription (Luo *et al.* 2001). It is noteworthy that prophage ΨM100 is a *Siphoviridae* like phage λ (Brussow and Desiere 2001) and targets an AT-rich att site in an intergenic region similarly to phage λ (Campbell 1992). Att sites are also found in coding regions (Krupovic, Forterre and Bamford 2010). When the function of the targeted gene is known, it can be as diverse as a gene coding for 3-hydroxy-3-methylglutaryl-coenzyme A reductase for the *Halorubrum* sp. virus BJ1 (Krupovic, Forterre and Bamford 2010), for heavy metal cation efflux system for the *Methanobrevibacter smithii* provirus Msmi-Pro1 (Krupovic, Forterre and Bamford 2010) or for AsnC family transcriptional regulator gene for the integrated element NitGar-E6 of Candidatus *Nitrososphaera gargensis* Ga9.2 (Wang *et al.* 2018). Recombination was not tested *in vivo* or *in vitro* with these inter- and intra-genic regions. It would be interesting to determine their efficiency for recombination and which positions are essential. Especially, since the presence of dyad symmetry in their DNA substrates is crucial for the activity of most tyrosine recombinases analyzed to date, we wonder whether those genes present such features.

Dimer resolution at dif sites

The integration of MGE sequences into their host genome does not constitute the sole function of site-specific recombinases.

Alanine (Ala)	AGC	**GGC**	CGC ■ ▥	**TGC**		
Arginine (Arg)	ACG	GCG ■ ● ▲	CCG ●	TCG ▥ ● ▲	**CCT**	TCT ■ ▥ ▨ ▲
Asparagine (Asn)	ATT	GTT ▲				
Aspartic acid (Asp)	ATC	GTC ▨ ● ▲				
Cysteine (Cys)	ACA	GCA ■ ▥ ▨				
Glutamic acid (Glu)	CTC ▨ ● ▲	TTC ■ ▨ ● ▲				
Glutamine	CTG ●	TTG ■				
Glycine (Gly)	ACC	GCC ▲	CCC ■ ●	TCC ■		
Histidine (His)	ATG	GTG ■ ■ ▲				
Isoleucine (Ile)	AAT	**GAT**	TAT			
Leucine (Leu)	AAG	GAG ●	CAG ▨	TAG ▲	CAA ■	TAA ■ ●
Lysine (Lys)	CTT ▨ ▲	TTT ● ▲				
Methionine (Met)	CAT ▥ ▲					
Phenylalanine (Phe)	AAA	GAA ▥ ▨ ▲				
Proline (Pro)	AGG	GGG ▥ ▨	**CGG**	TGG ▥ ▲		
Serine (Ser)	AGA	**GGA**	CGA ■ ▲	TGA ■ ▨ ▲	GCT ■ ▲	ACT
Threonine (Thr)	AGT	**GGT**	CGT ■ ▥ ▲	TGT ■		
Tryptophane (Trp)	CCA ▥					
Tyrosine (Tyr)	ATA	GTA ■ ▥				
Valine (Val)	AAC	GAC ■ ▲	CAC ■ ◘ ● ▲	TAC ■ ◘ ▲		

tRNA targeted in Thermococcales ■ Methanococcales ■ Sulfolobales ●
Halobacteriales ▥ Methanosarcinales ◘ Thaumarchaeota ▲
Archaeoglobales ▨

Figure 12. Wide tRNA genes targeting by archaeal tyrosine recombinases. All tRNA anticodon combinations are listed along with their corresponding amino acids. Anticodons that are not found in archaeal tRNAs are indicated in light gray. An array of symbols indicates the utilization of a tRNA gene with the specified anticodon as attB site for *Thermococcales* (Li *et al.* 2016; Cossu *et al.* 2017; Badel *et al.* 2019, 2020), *Halobacteriales* (Krupovic, Forterre and Bamford 2010; Liu *et al.* 2015), *Archaeoglobales* (Badel *et al.* 2020), *Methanococcales* (Krupovic, Forterre and Bamford 2010; Badel *et al.* 2019), *Methanosarcinales* (Badel *et al.* 2020), *Sulfolobales* (She, Brugger and Chen 2002; Wang *et al.* 2007; Peng 2008; Redder *et al.* 2009) and *Thaumarchaeota* (Krupovic *et al.* 2019). Anticodons corresponding to untargeted tRNA genes are highlighted in bold green.

As for bacteria, archaeal chromosomally-encoded Xer tyrosine recombinases exert an essential role in genome maintenance and integrity. The sites of Xer recombinases are named dif and are present in single copy on circular chromosomes of archaea and bacteria (Castillo, Benmohamed and Szatmari 2017; Cossu *et al.* 2017). The Xer/dif system improves cell viability by resolving concatenated chromosomes occurring by homologous recombination during DNA replication. When a chromosome dimer is formed, the dif sequence appears duplicated and the dimer is resolved through site-specific recombination between the two dif sites. In bacteria and archaea, the dif sequence is composed of two 11-nt inverted repeats separated

by a 6-nt spacer (Cortez *et al.* 2010; Cossu *et al.* 2017; Jo *et al.* 2017). The extremities of the spacer correspond to the position of the tyrosine-catalyzed cleavage (Serre *et al.* 2013; Castillo, Benmohamed and Szatmari 2017). The dif sequence is variable to a certain extent except for some conserved positions (Cortez *et al.* 2010). It seems that both the stem loop structure and the nucleotide sequence are important for the functionality of dif sites. Interestingly, PaXerA can bind its dif substrate with a high affinity and stem-looped structures of unrelated att sites with a lower affinity (Cortez *et al.* 2010). In the bacterial model of dif recombination, single XerC and XerD monomers each bind to an inverted repeat of the dif site (Castillo, Benmohamed and Szatmari 2017). Similarly, TaXerA and PaXerA can bind each dif inverted repeats (Serre *et al.* 2013; Jo *et al.* 2017). The activity of TaXerA was assayed on a series of dif site variants mutated in the inverted repeats (Jo *et al.* 2017). All variants allowed recombination even if to a lesser extent than the wild-type dif site. Depending on the variant, the reduced activity was due to reduced binding affinity or reduced strand exchange (Jo *et al.* 2017) indicating that key dif positions are involved either in sequence recognition or in the strand-transfer reaction. Due to the deleterious effect of chromosome dimers at cell division, it is vital that the equilibrium of the reversible Xer/dif reaction be displaced toward resolution. In bacteria, this function is officiated by the cell division protein FtsZ (Barre *et al.* 2000) whereas no equivalent system has been identified in archaea so far.

ECOLOGY OF ARCHAEAL INTEGRATION

Integrases are responsible for integration into and excision from a chromosome, the central switches in mobile genetic elements life history. As such, integrase catalytic activities should be considered in the light of the ecological consequences of the mobile genetic element lifestyle. In this section, we present the current knowledge about the host specificity of integrases and the type of mobile genetic elements that encode them. We also discuss the advantages of encoding an integrase for mobile genetic elements and the control of integrase directionally and temporality.

Integrase host specificity

Tyrosine integrases are detected in many archaeal phyla but closely related integrases are mostly found in a narrower range of organisms indicating a certain degree of host specificity (Fig. 6 and Table 1). In *Methanococcales* and *Thermococcales* that were thoroughly investigated, it was found that plasmids related to pT26-2 harbor two distinct integrase families (Badel *et al.* 2019). Furthermore, it was observed that similar integrases from pleolipoviruses were present in the chromosomes of 10 different *Halobacteriaceae* genera from various geographical locations (Liu *et al.* 2015). Related integrases are not restricted to a local area but are found all around the globe where their host is present.

Mobile element recruitment

In the archaeal domain, integrases are carried by a wide variety of MGEs: conjugative plasmids (She, Brugger and Chen 2002), cryptic plasmids (Erdmann *et al.* 2017; Badel *et al.* 2019), viruses from several viral families such as *Myoviridae* (Klein *et al.* 2002; Tang *et al.* 2002), *Pleolipoviridae* (Atanasova *et al.* 2018a), *Fuselloviridae* (Goodman and Stedman 2018) and unidentified MGEs (Li *et al.* 2016). The *Thermococcales* pT26-2 family of integrases is present in plasmids from the pT26-2 family, in plasmid from the pAMT11 family, in Fusellovirus and in unidentified MGEs (Badel *et al.* 2020). Similarly, two very similar integrases were identified in *Sulfolobus solfataricus*, on a plasmid and a virus (86% nucleotide similarity, 94% amino acid similarity) (Peng 2008) indicating that integrases from the same family can be recruited by several mobile elements. Additionally, some MGE families include integrative and non-integrative members as the *Thermococcales* pAMT11 family (Argos *et al.* 1986) or the haloarchaeal pleolipoviruses (Roine *et al.* 2010; Liu *et al.* 2015). Strikingly, the sequences of the two pleolipoviruses HHPV3 and HHPV4 are very syntenic and similar except for a 3 kb HHPV4-specific region carrying an integrase gene (Atanasova *et al.* 2018a). This situation could result either from integrase acquisition or loss for some MGE members. For suicidal integrases, the att site is included within the integrase gene resulting in a compact module that could favor exchange between mobile elements (Ausubel 1974). The pING1 plasmid was identified as encoding an integrase exhibiting all the conserved residues of its family of pNOB8-like integrases but no attP site could be determined (Erauso *et al.* 2006). It is possible that in that case, attP site loss would lead to integrase gene degeneration and/or loss. Finally, even when the integration module is conserved in a plasmid family, its evolutionary history can be complex. This is observed in all the conjugative *Sulfolobales* plasmids that exhibit conserved conjugation and integration modules. However, the phylogenetic trees of the two modules are not congruent suggesting intrafamily module exchanges (Erauso *et al.* 2006). On the whole, the frequent integrase exchange between mobile elements is featured in a network of all archaeal viruses where some integrases represent connector genes between virus clades (Iranzo *et al.* 2016). However, in the network, other integrases represent a signature gene of a clade evidencing their favored residence in those particular MGEs. Some archaeal integrases seem 'settled' (Iranzo *et al.* 2016) whereas the majority is frequently exchanged, gained or lost between MGEs.

Integration is a major lifestyle for archaeal mobile elements

The primary function of tyrosine integrases is to catalyze the integration of the MGE that encode them into the host chromosome or the reverse reaction of excision (Fig. 2). Such integrase-encoding MGEs are widely present in archaeal genomes (Soler *et al.* 2010; Gaudin *et al.* 2014; Wang *et al.* 2018; Krupovic *et al.* 2019). In *Thermococcales*, it was shown that >30% of the published genomes contain an integrated element encoding an integrase of the pT26-2 family (Iranzo *et al.* 2016). The proportion of genomes presenting any integrase-encoding MGE is most probably higher than that. In the phylum *Thaumarchaea*, integrated MGEs were systematically detected and found in 20 out of 21 analyzed genomes (Krupovic *et al.* 2019). In halophilic archaea, SNJ2-like integrases from integrated MGEs form a large, well-supported clade with the MGE-encoded hyperthermophilic integrases (Wang *et al.* 2018). In this systematic search, several integrated MGEs would not encode an integrase. This presumably results from the integrase gene loss after integration similarly to what was observed for integrated plasmids of *Methanococcales* (Badel *et al.* 2019). Several related or unrelated MGEs can be integrated in the same chromosome at different loci (Pauly *et al.* 2019; Badel *et al.* 2020) or integrated in tandem at the same locus (Krupovic, Forterre and Bamford 2010; Krupovic *et al.* 2019). No account for λ-type immunity system has been reported for archaeal MGEs therefore enabling co-infection or superinfection. Overall, integrated MGEs are widely present in archaeal

genomes suggesting strong evolutionary advantages for integration in this domain (Fig. 6).

Advantages of mobile element integration: why code for an integrase?

Advantages were uncovered for phage λ lysogenic state: integration increases long-term MGE maintenance (Echols 1972) and the integrated state cell provides a solution when chances of finding a new suitable host are low (Levin, Stewart and Chao 1977). The canonical integration model established for phage λ proposes that the integrative state would help the MGE to survive through adverse environmental conditions. During lysogeny, the MGE genome only exists in the integrated form and is silenced. When stressful conditions are encountered by the cell, the MGE excises and enters the lytic cycle and virions are released into the environment though cell lysis (Paul 2008; Gandon 2016). Depending on the environmental conditions, the MGE chooses to reproduce vertically (integration) or horizontally (infection) though highly controlled mechanisms. The same lifestyle was observed for *Acidianus* convivator bicaudavirus ATV (Prangishvili *et al.* 2006). Under optimal growth temperature conditions, it adopts a lysogenic lifestyle and integrates into the host chromosome. Inversely, under suboptimal growth temperature conditions, the virus adopts a lytic lifestyle resulting in host cell lysis.

For the archaeal fuselloviruses, which are the most studied archaeal MGEs encoding an integrase and exemplified by the model virus SSV1, the integration implications differ on several aspects from the lysis/lysogeny switch paradigm of phage λ (Prangishvili, Stedman and Zillig 2001). (i) SSV1 viral production is induced by a UV irradiation (Martin *et al.* 1984), mitomycin C treatment (Liu and Huang 2002) or by shaking the culture (Liu and Huang 2002) similarly to phage λ, but cells do not lyse massively after viral production and return to the lysogenic sate (Martin *et al.* 1984). Virus TPV1 replication is also induced by UV-treatment without any extensive cellular lysis (Gorlas *et al.* 2012). (ii) During the SSV1 integrative stage, a few circular copies of the viral genome remain in the cell (Yeats, McWilliam and Zillig 1982; Pauly *et al.* 2019). Similarly, a high copy number of TPV1 circular DNA is present in its host cells (Gorlas *et al.* 2012). (iii) During the SSV1 integrative stage, the majority of the viral ORFs are expressed, including the integrase gene and the structural proteins (Frols *et al.* 2007). It is not known whether the transcription template corresponds to the integrated or episomal copy of the viral genome. A transcriptional regulator was identified that is probably involved in lysogeny regulation (Fusco *et al.* 2015b) but it does not result in provirus silencing as it is the case for phage λ. Contrastingly, the SSV2 integrase is not basally expressed (Fusco *et al.* 2015b). (iv) For the virus SSV1, evidence point toward the replication of already present circular DNA independently of the integrated copy rather than an excision and subsequent replication of the circular DNA similarly to λ (Fusco *et al.* 2015b). Overall, and contrarily to the lambdoid paradigm, it seems that the integrase of lysogenic fuselloviruses is not involved in the regulation of virus replication and virion production. Nevertheless, most fuselloviruses encode a suicidal tyrosine integrase (Gorlas *et al.* 2012; Goodman and Stedman 2018) suggesting a probable evolutionary importance for virus survival. SSV1 viruses lacking the integrase gene were found to be outcompeted by wild-type viruses (Clore and Stedman 2007). However, mutant viruses were infectious and stably maintained in *Sulfolobus* and no clear benefit was associated with integrase activity. The exact evolutionary advantage of fusellovirus integrase still remains to be determined.

A number of archaeal plasmids were identified that are present in the cell both in the integrated and episomal states (Basta *et al.* 2009; Gaudin *et al.* 2014; Cossu *et al.* 2017). Contrarily to highly controlled lysis/lysogeny switch of temperate phages, archaeal plasmids such as pTN3 and pAH1 use a rudimentary safekeeping mechanism. Integration appears in this case as a simple and efficient solution to ensure the propagation of replicative plasmids when targeted by host defenses or other superinfecting MGEs. The initial isolate of *Thermococcus nautili* carried plasmid pTN3 in both replicative and integrated forms whereas the circular form was lots after successive subculturing (Cossu *et al.* 2017). This plasmid loss was caused presumably by the clustered regularly interspaced short palindromic repeats defense system (CRISPR-Cas9) (Oberto *et al.* 2014). Similarly, the *Acidianus hospitalis* pAH1 plasmid was evidenced to be stably maintained simultaneously in integrated and episomal states (Basta *et al.* 2009). When the host was co-infected with the virus AFV1, the episomal form disappeared rapidly while the integrated form persisted. These observations suggest that the integrated form can act as a safekeeping copy of the disappearing plasmid.

Integrating the host chromosome might force the cell into accepting the MGE and shutting down its defense systems. The targeting of integrated MGEs by the CRISPR system might induce an autoimmune response and death of the infected cell (Stern *et al.* 2010; Wimmer and Beisel 2019). It was postulated also that multiple integrations of related fuselloviruses and frequent recombinations among their highly similar genomes might provide a means to evade their hosts CRISPR system (Redder *et al.* 2009). Transcriptional activation of the CRISPR-Cas system was observed during SSV2 fusellovirus infections leading to a significant reduction in SSV2 copy number, its integration into the host chromosome and the deletion of several repeats-spacer units from the CRISPR array (Fusco *et al.* 2015b). As a result, all copies of the intact integrase gene were lost abolishing excision and effectively trapping the provirus in the chromosome. From a population genetics point of view, MGEs encoding suicidal integrases could be considered 'kamikazes' which role would be to defeat host defense mechanisms.

Integrated MGEs can provide functions that are beneficiary for the host (Schuch and Fischetti 2009; Wang *et al.* 2010) and therefore increase the probability of MGE retention. *Thermococcus kodakarensis* mutants lacking each of the four integrated TKV1 to TKV3 elements displayed impaired growth suggesting their importance for cellular metabolism at least in laboratory conditions (Tagashira *et al.* 2013). Similarly, the integrated element PYG1 was shown to increase its host resistance to temperature (Li *et al.* 2016).

Integration/excision temporality control in archaeal mobile elements

The control of MGE integration and excision was thoroughly investigated for the bacterial lambdoid phages evidencing a complex regulatory genetic network (Oppenheim *et al.* 2005). Two levels of regulation were observed: (i) reaction temporality control and (ii) reaction directionality control (integration or excision). It is interesting to investigate whether integration and excision are also tightly regulated in archaea and if similar regulatory networks are implemented. In *Pyrococcus abyssi*, it was proposed that the integrase of the genomic island PYG1 can spontaneously catalyze excision since PYG1 does not carry an

identified replication module and the element can be found in a circular state (Li *et al*. 2016). MGE excision seems in that case loosely controlled.

The first level of integration/excision temporal regulation consists in the regulation of integrase transcription. In some pNOB8-like integrases, the presence of a HTH domain was proposed to be involved in the transcriptional regulation of the integration/excision of the MGE (Erauso *et al*. 2006). For the *Sulfolobus* spindle-shaped viruses, transcription temporality was investigated by several studies. In SSV1, the integrase is under the control of an early promoter that allows a rapid expression after UV-induction (Frols *et al*. 2007) and the F55 repressor downregulates expression of the integrase operon in the absence of induction (Fusco *et al*. 2013, 2015b). Contrastingly, the integrase from virus SSV2 is expressed in the late infection phase consistently with the provirus detection >7 h after infection (Ren, She and Huang 2013). Moreover, SSV1 and SSV2 integrases are expressed from polycistronic operons while for other SSV viruses, the integrase is proposed to be translated from a monocistronic mRNA transcript (Goodman and Stedman 2018). The mechanisms of integrase expression regulation in the various SSV viruses appears to be diverse but still remains largely unexplored.

Some archaeal halophilic tailed viruses belong to the Caudovirales, which also include tailed bacteriophages (Sencilo *et al*. 2013; Krupovic *et al*. 2018). Among them, the archaeal Myovirus φCh1 can integrate into its host genome (Witte *et al*. 1997) and two potential tyrosine integrase sequences were identified (Klein *et al*. 2002). φCh1 regulatory network for the switch from the lysogenic to the lytic cycle was partially elucidated (Iro *et al*. 2007; Selb *et al*. 2017) and involved Rep, a repressor protein that functions convergently to phage λ cI repressor protein (Iro *et al*. 2007). During λ lysogeny, the specific binding of cI to its operator sites embedded in promoter sequences induces its own expression but represses the transcription of the lytic operons. A similar repressor protein is also present in the non-integrative myovirus φH1 (Ken and Hackett 1991; Stolt and Zillig 1992) suggesting that it might be implicated in the regulation of virion production rather than in excision control. Proteins similar to the repressor were also found in several integrase-encoding Pleolipoviruses (Chen *et al*. 2014; Liu *et al*. 2015; Atanasova *et al*. 2018a) suggesting that this mechanisms of lysis-lysogeny regulation is widely shared among halophilic viruses.

Integration/excision directionality control in archaeal mobile elements

All the characterized archaeal integrases can catalyze both integration and excision reactions in the absence of any recombination directionality factor (RDF) in sharp contrast to the phage λ directionality regulation (Landy 2015). However, the activity of the halophilic integrase IntSNJ2 is modulated by two proteins Orf2 and Orf3, which increased *in vivo* integration efficiency (Wang *et al*. 2018). Orf1 to 3 are transcribed in an operon with two alternative transcription start sites. Using one or the other transcription site might constitute a control system for lysogeny.

In experimental setups with complete integrase proteins, characterized suicidal integrases catalyzed integration and excision alone (Cossu *et al*. 2017; Badel *et al*. 2020). However, in naturally occurring conditions, suicidal integrases are partitioned after integration. Excision would then require the activity of the split integrase that might be inactive. As a consequence, excision could not proceed after integration in the absence of some external factor (a complete integrase gene). This situation is similar to the directionality control by a RDF except that, for suicide integrases, the RDF is the complete integrase gene. In that sense, the suicidal integrase can be viewed as an 'all in one integration module' that include the integrase gene, the recombination site and the recombination directionality factor.

INTEGRASE EVOLUTION AND SPECIFICITY SWITCH

Tyrosine recombinases evolution

The first extensive tyrosine recombinase alignments revealed that the C-terminal portion carrying the catalytic domain is much more conserved than the N-terminal part responsible for site-specific recognition and protein multimerization (Esposito and Scocca 1997; Guo, Gopaul and van Duyne 1997). This variability reflects the divergence in target site sequence and the capacity of some phage integrases to bind to two distinct DNA segments (Moitoso de Vargas *et al*. 1988). The sequence divergence of tyrosine recombinases clearly illustrates the ancient origin of these proteins. The detailed evolutions mechanisms leading to such a diversity and particularly to the acquisition of different specificities are not fully understood. The naturally occurring change in specificity between related recombinases has been addressed for the integrases of lambdoid phages (Yagil *et al*. 1989). Experimental evidence using mutated and chimeric enzymes from bacteriophages λ and HK022 suggested a multistep process in which integrase specificity first broadens then narrows to permit co-evolution of the target site (Dorgai, Yagil and Weisberg 1995; Yagil, Dorgai and Weisberg 1995). Despite these efforts, the mechanisms underpinning integrase evolution remains somewhat murky. The special case of suicidal integrases presented in the following sections could shed some light on this fundamental process.

Postmortem suicidal integrase excision activity… *in vivo*

After integration, the suicidal integrase gene is split in two inactive int(N) and int(C) pseudogenes potentially encoding the Int(N) N-terminal part and the Int(C) C-terminal part of the integrase respectively. The latter fragment carries the catalytic domain (Figs 2A and 3A). *In vitro* experiments concurred in demonstrating the inability of several truncated integrases to perform recombination reactions. The Int(N) and Int(C) moieties of IntSSV2 did not interact in solution in absence of DNA suggesting that they do not cooperate to assemble as an entire functional enzyme (Zhan, Zhou and Huang 2015). On its own, Int(C) could not form multimers since the N-terminal part of the integrase is responsible for multimerization (IntSSV1 dimerization and IntSSV2 tetramerization) (Zhan *et al*. 2012; Zhan, Zhou and Huang 2015). Contrastingly, *in vitro* recombination could be achieved with the truncated Int(C)SSV1 and Int(C)SSV2 albeit with a significantly reduced efficiency (Zhan *et al*. 2012; Zhan, Zhou and Huang 2015).

Several reports addressed the issue whether the expression of the Int(N) and Int(C) moieties could promote *in vivo* MGE excision in the absence of an intact integrase gene. First, the level of expression of the separate moieties was explored *in vivo*. In *S. solfataricus* P2, only the int(N) moiety is transcribed for the pXQ1 and XQ2 integrated elements (She *et al*. 2001; Jager *et al*. 2014) whereas no expression was detected for the Int(N) moiety nor for the complete integrase gene from integrated plasmid pSSVi in *S. solfataricus* (Ren, She and Huang 2013). In *T. nautili*, the IntpTN3 int(N) fragment lies downstream the integrase promoter and translation could be initiated at the original start codon while

its int(C) moiety could be transcribed from the tRNA^{Leu} gene promoter (Cossu *et al.* 2017). For IntpTN3-related integrases, an in-frame start codon is often present near the beginning of int(C) suggesting that the catalytic part of the integrase could potentially be translated (our observation). So far, no consensus has emerged on the actual expression of the Int(N) and Int(C) moieties. Their level of expression might vary from one suicidal integrase to another resulting in various modes of excision control. In *S. solfataricus* P2 cells carrying the integrated plasmid pSSVi, episomal copies of this plasmid were barely detectable (Ren, She and Huang 2013). While infection of this strain with the related SSV2 virus accumulated free pSSVi plasmids, the increase was due to additional replication of rare episomes rather than to the excision of integrated copies (Ren, She and Huang 2013). However, SSV2 excision was shown to occur in the presence of the episomal MGE coding for the complete integrase (Fusco *et al.* 2015b). Similarly, the heterologous expression of plasmid pTN3 integrase in *T. kodakarensis* promoted excision of the related integrated element TKV4 (Cossu *et al.* 2017). Furthermore, TKV4 excision could be obtained in the same organism by supplying in *trans* a gene encoding inactive IntpTN3 Y428A, therefore suggesting that TKV4 Int(C) is effectively expressed (Cossu *et al.* 2017). This trans-complementation suggests that the Int(C) moiety might play a role in the excision of a mobile element by an exogenous integrase. For both SSV2 and pTN3, fragmented integrase and exogenous integrases were closely related. This excision catalysis by another MGE therefore depends on the widespread occurrence of closely related integrases in the population and in various MGEs that were observed for the pT26-2 family of integrases (Badel *et al.* 2020).

Suicidal integrase maintenance, evolution and specificity switch

As mentioned above, it appears at first glance that integrated MGEs encoding suicidal integrases would remain permanently entrapped in the host chromosome. The inability of their fragmented integrase gene to encode an active enzyme would prevent the excision reaction and further propagation. One would therefore expect these particular MGE populations to decrease progressively and eventually disappear. Paradoxically, it was observed to the contrary that such integrated MGEs pervade entire populations in both *Crenarchaea* and *Euryarchaea* (Pauly *et al.* 2019; Badel *et al.* 2020). Several observations were instrumental in explaining this phenomenon. First, several related or unrelated MGEs can be integrated in the same chromosome at different loci (Badel *et al.* 2020). Second, different integrases present different integration sites and closely related archaeal integrases do not always target the same att site. For example, the classical integrases identified in pT26-2 related plasmids from *Methanococcales* can target tRNA^{SerTGA}, tRNA^{SerGCT} or tRNA^{LeuTAA} (Badel *et al.* 2019). The suicidal integrases identified in pT26-2-related plasmids from *Thermococcales* can target 14 different tRNA genes (Badel *et al.* 2020). In both cases, the most probable evolutionary scenario involves an ancestral integrase with a single DNA substrate specificity followed by target diversification in the descendant lineages. Target switching is however restricted within the two classes of tRNA genes: those encoding tRNAs with a supplementary loop and those that encode tRNAs without such a loop (Badel *et al.* 2019). Furthermore, archaeal suicidal integrases harbor the translation of the att site within their protein sequence (Figs 2A and 3A) deepening the conundrum of specificity change. For them, a change in site specificity is mechanically reflected by a change in protein sequence. One could expect that such a change would compromise protein integrity, but it was shown on the contrary that the att site translation is quite variable in closely related sequences without obvious deleterious effect (Badel *et al.* 2020). Notably, length variations are compensated around the att site avoiding any frameshifts in the C-terminal region.

Accurate DNA and protein comparison of the genes and integrases belonging to the archaeal IntpT26-2 family underlined the differential evolution history of their Int(N), Int(C) and att components (Badel *et al.* 2020). It was argued that the integration of multiple elements sharing extensive sequence conservation could lead to homologous recombination (Redder *et al.* 2009; Gehring *et al.* 2017), generating chimeric integrase genes expressing active integrase (Badel *et al.* 2020). A model was proposed to explain the evolution and specificity switches in suicidal integrases. It is based on the observation of a large chromosomal inversion in a subset of the natural population of *T. kodakarensis* between TKV2 and TKV3, two related MGEs integrated in opposite orientation (Gehring *et al.* 2017). In this model, homologous inversion could generate two chimeric integrase genes by exchanging their int(N) and int(C) moieties and potentially novel attPs (Fig. 13) (Badel *et al.* 2020). Integrated MGEs encoding these chimeric suicidal integrases can be resurrected by superinfection of an incoming MGE with a compatible or more relaxed specificity and generate the observed variability (Cossu *et al.* 2017; Badel *et al.* 2020). This combinatorial mechanism does not only explain the pervasiveness of suicidal enzymes but also identifies the source of their variability.

INTEGRASE-RELATED GENOME EVOLUTION

Mobile genetic element modular evolution

Several reports underlined that bacterial MGE evolution proceeds mainly through module exchange (Botstein 1980; Oberto, Sloan and Weisberg 1994; Hendrix *et al.* 2000). The same rule applies to archaeal plasmids and viruses (Basta *et al.* 2009; Krupovic, Forterre and Bamford 2010; Iranzo *et al.* 2016). For example, the *Pyrococcus yayanosii* PYG1 integrated element shares a module with the MP integrated from *Thermococcus barophilus* element and another module with plasmid pTBMP1 of the same organism (Li *et al.* 2016). This mechanism of module exchange can be explained by homologous recombination involving several MGE integrated in tandem at the same chromosomal locus (Redder *et al.* 2009). Alternatively, homologous recombination between inverted modules of tandem integrated MGEs could lead to integrase-independent excision embarking portion of both integrated elements. MGE integration therefore facilitates this modular evolution. Additionally, integrases are directly involved in this process as several halophilic viruses were identified that encode tyrosine recombinases that seem implicated in viral DNA rearrangements (Rossler *et al.* 2004; Sencilo *et al.* 2013). One of the DNA rearrangements was involved in the generation of protein variants presenting various cell surface adhesion specificities (Klein *et al.* 2012).

Horizontal gene transfer

Horizontal gene transfer (HGT) refers to the transmission of genetic information between individual organisms independently of direct progeny. HGT is recognized as a driving force of archaeal evolution (Wagner *et al.* 2017). Several successive steps are required for effective HGT: (i) DNA is transferred into the cell via transformation, membrane vesicle, viral infection, conjugation, cell fusion or other specialized cellular apparatus (Wagner *et al.* 2017). (ii) The foreign genetic information is incorporated into the host chromosome through homologous recombination

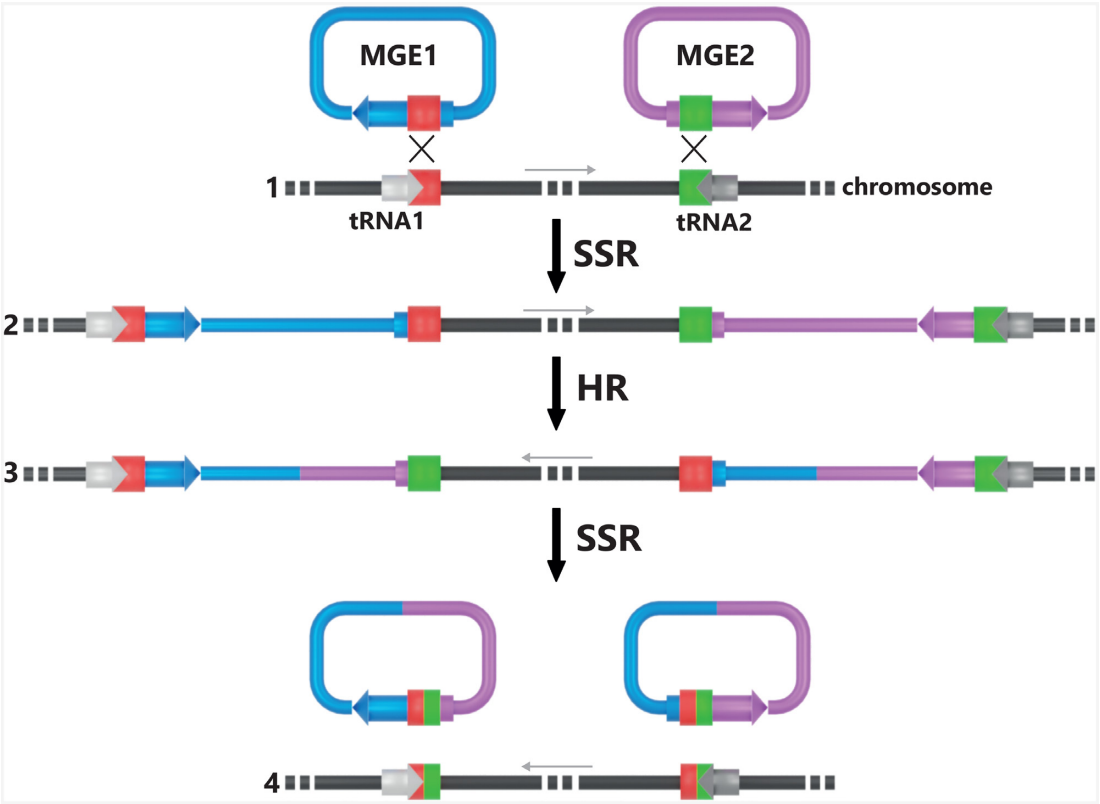

Figure 13. Suicidal integrase recombination. The integration by site-specific recombination (SSR) of multiple related MGE at different chromosomal locations and in inverted orientation **(A)** gives rise to homologous recombination (HR) between conserved MGE sequences **(B, C)**. Such a recombination has been observed in *T. kodakarensis* (Gehring *et al.* 2017). Recombinant integrated MGEs encoding hybrid integrases can then excise if a compatible integrase is provided by a superinfecting MGE **(D)**.

or through site-specific integration if the incoming DNA carries an integration module. (iii) In the case of MGE-catalyzed integration, the DNA should be immobilized in the host chromosome to be effectively characterized as HGT. This can happen through integrase gene mutation or loss. A nonsense mutation present inside the integrase coding-sequence of the integrated element pST4 illustrated this case of HGT in *Sulfolobus* (She, Chen and Chen 2004). Similarly, mutations into the attL and attR sites would lead to MGE sequestration into the host chromosome. Integrated plasmids that lack detectable att sites were recently identified that might correspond to captured elements (Badel *et al.* 2019). Additionally, integration in secondary attachment sites, i.e. att sites with a mismatch, could prevent efficient MGE excision and lead to permanent acquisition of MGE genes by the host chromosome (She, Chen and Chen 2004).

If all functions can be transmitted by MGE-mediated HGT, the most studied was DNA replication. The archaea *Sulfolobus islandicus* and *Haloferax volcanii* possess several active chromosomal origins of replication, some of which were acquired from integrated MGE (Robinson and Bell 2007; Hawkins *et al.* 2013; Samson *et al.* 2013). At the archaeal domain scale, an exhaustive phylogenic analysis of all major replication components showed that chromosomal copies of several components (e.g. MCM, PCNA, PolB) probably arose from MGE integration (Raymann *et al.* 2014; Badel *et al.* 2019). Archaeal MGEs were also proposed to be implicated in the HGT of tRNA gene introns. Sugahara *et al.* (2012) proposed recombination between an intron-free tRNA gene attB and an intron-containing attP as a mechanism of intron acquisition in tRNAs where the MGE attP serves as intron vector between tRNAs and between cells.

Chromosomal inversions

Among MGEs, transposable elements (TEs) are known to be frequently involved in generating inversions in the host chromosome (Eickbush and Furano 2002; Zivanovic *et al.* 2002; Redder and Garrett 2006; Darmon and Leach 2014; Weckselblatt and Rudd 2015; Vandecraen *et al.* 2017). Inversions proceed through homologous recombination between two paralogous integrated elements. Other integrated MGEs are also involved in similar processes. In *T. kodakarensis*, a large-scale inversion was identified that occurred through homologous recombination between the related integrated elements TKV2 and TKV3 (Gehring *et al.* 2017; Badel *et al.* 2019, 2020). Tyrosine recombinases are fundamental in that process because they catalyze the integration of paralogous MGE copies into the chromosome. Another unique mechanism was identified in archaea that led to chromosomal inversions. The tyrosine integrase from *Thermococcus nautili* plasmid pTN3 'catalyzes low sequence specificity recombination reactions with the same outcome as homologous recombination events' between identical DNA segments as short as 104 bp (Cossu *et al.* 2017). This homologous recombination activity resulted in four large-scalet chromosomal inversions over the span of 66 generations in *T. nautili*. (Cossu *et al.* 2017). The broad occurrence of integrated MGEs carrying integrases similar to IntpTN3 is probably one of the major causes of the large chromosomal inversions observed in *Thermococcales*, especially in the *Thermococcus* genus where TEs are absent or rare (Cossu *et al.* 2015, 2017). Archaeal tyrosine recombinases are thus involved in chromosomal inversion either indirectly through the integration of multiple, recombinable, MGE copies or directly through the homologous recombination of original chromosomal segments. As a consequence of both

mechanisms, chromosomes are largely disrupted in their otherwise conserved organization (Cossu *et al.* 2015). The fitness cost or benefice of such inversions is yet unknown.

CONCLUSIONS AND FUTURE PERSPECTIVES

The combination of specific DNA binding and multimerization domains with a module capable to resect and exchange DNA strands gave rise to an enzyme widely adopted in all domains of life. With a remarkable efficiency and without the expense of energy, these recombinases allow the effortless host chromosome integration and excision of a number of MGEs. By resolving chromosomes dimers, tyrosine enzymes also provide the opportunity to correct flaws resulting from circular chromosomal replication in bacteria, archaea and large bacteriophages.

In the last three decades since the identification of the first integrative element in Archaea, the study of tyrosine recombinases revealed common aspects for the three domains of life, such as allowing both MGE integration and excision and their propensity to target tRNA genes in the host genome. The primary sequences of archaeal, bacterial and eukaryal tyrosine recombinases are highly diverse, precluding the use of phylogeny to assess relationships between these enzymes. With a network analysis, we generated a robust classification of archaeal integrases while confirming and extending previous comparisons based on Pfam protein domains. Going further, a global evolutionary analysis of all tyrosine recombinases from archaeal, bacterial and eukaryotes could be undertaken. It would evidence potential transfers between the three domains and would shed some light on the origin of tyrosine recombinases. On other aspects, archaeal integrases differ from bacterial and eukaryal integrases. They do not require essential helper or directionality factors, and hyperthermophilic archaea have developed a particular suicidal integration system where the MGE target site is carried within the integrase gene.

Archaeal integrases have now proved to be important models for understanding tyrosine recombinases. The study of suicidal integrases found exclusively in the archaeal domain provided important clues on the evolution of these enzymes. Additionally, the study of a new tyrosine recombinase led to discovery of an integrase family capable of the dual activity of site-specific and homologous recombination. Both examples warrant the further investigation of archaeal tyrosine recombinases, including the new integrase families identified in our network analysis.

On many other aspects, archaeal integrases still have much to reveal. Archaeal MGE lysogeny was never studied in detail despite the obvious differences with the canonical phage λ lysogeny. Such approaches would lead to a better understanding of the dynamics of MGE integration and excision in natural archaeal communities. The study of archaeal MGE would also help determine whether integration is preferentially implemented in certain environmental or genetic conditions.

Recent advances in the crystal structure resolution of several archaeal tyrosine recombinases successfully demonstrated similarities of the catalytic domain with other known bacterial enzymes. However, the modalities by which archaeal recombinases interact with their DNA substrate remain to be explored and could be approached by solving the complete structure of protein–DNA complexes by co-crystallization or cryo-electron microscopy.

The development of large-scale genome sequencing is bound to improve the knowledge of genome dynamics and might emphasize the already acknowledged importance of MGE integration in this process. Further studies should also investigate the multiple integration of archaeal MGEs and how they shape the genome evolution and diversity of their hosts.

SUPPLEMENTARY DATA

Supplementary data are available at *FEMSRE* online.

DATA AVAILABILITY

The protein sequence data files used in the archaeal network analysis are provided in Fasta format at https://archaea.i2bc.paris-saclay.fr/Archaea_Int.zip.

FUNDING

This work was funded by the Centre National de la Recherche Scientifique and the Agence Nationale de la Recherche (grant ANR-19-CE11-0007). CB was supported by the Ecole Normale Supérieure de Lyon.

Conflict of Interest. None declared.

REFERENCES

Abremski K, Gottesman S. Purification of the bacteriophage lambda Xis gene product required for lambda excisive recombination. *J Biol Chem* 1982;**257**:9658–62.

Abremski K, Wierzbicki A, Frommer B *et al.* Bacteriophage P1 Cre-loxP site-specific recombination. Site-specific DNA topoisomerase activity of the Cre recombination protein. *J Biol Chem* 1986;**261**:391–6.

Abremski KE, Hoess RH. Evidence for a second conserved arginine residue in the integrase family of recombination proteins. *Protein Eng* 1992;**5**:87–91.

Argos P, Landy A, Abremski K *et al.* The integrase family of site-specific recombinases: regional similarities and global diversity. *EMBO J* 1986;**5**:433–40.

Arinkin V, Smyshlyaev G, Barabas O. Jump ahead with a twist: DNA acrobatics drive transposition forward. *Curr Opin Struct Biol* 2019;**59**:168–77.

Atanasova NS, Demina TA, Krishnam Rajan Shanthi SNV *et al.* Extremely halophilic pleomorphic archaeal virus HRPV9 extends the diversity of pleolipoviruses with integrases. *Res Microbiol* 2018a;**169**:500–4.

Atanasova NS, Roine E, Oren A *et al.* Global network of specific virus–host interactions in hypersaline environments. *Environ Microbiol* 2012;**14**:426–40.

Ausubel FM. Radiochemical purification of bacteriophage lambda integrase. *Nature* 1974;**247**:152–4.

Badel C, Da Cunha V, Forterre P *et al.* Pervasive suicidal integrases in deep-sea archaea. *Mol Biol Evol* 2020;**37**:1727–43.

Badel C, Erauso G, Gomez AL *et al.* The global distribution and evolutionary history of the pT26-2 archaeal plasmid family. *Environ Microbiol* 2019;**21**:4685–705.

Barre FX, Aroyo M, Colloms SD *et al.* FtsK functions in the processing of a Holliday junction intermediate during bacterial chromosome segregation. *Genes Dev* 2000;**14**:2976–88.

Basta T, Smyth J, Forterre P *et al.* Novel archaeal plasmid pAH1 and its interactions with the lipothrixvirus AFV1. *Mol Microbiol* 2009;**71**:23–34.

Bastos MC, Murphy E. Transposon Tn554 encodes three products required for transposition. *EMBO J* 1988;**7**:2935–41.

Bayliss CD. Determinants of phase variation rate and the fitness implications of differing rates for bacterial pathogens and commensals. *FEMS Microbiol Rev* 2009;**33**:504–20.

Bigot S, Corre J, Louarn JM *et al*. FtsK activities in Xer recombination, DNA mobilization and cell division involve overlapping and separate domains of the protein. *Mol Microbiol* 2004;**54**:876–86.

Biswas T, Aihara H, Radman-Livaja M *et al*. A structural basis for allosteric control of DNA recombination by lambda integrase. *Nature* 2005;**435**:1059–66.

Blakely G, May G, McCulloch R *et al*. Two related recombinases are required for site-specific recombination at dif and cer in E. coli K12. *Cell* 1993;**75**:351–61.

Bobay LM, Rocha EP, Touchon M. The adaptation of temperate bacteriophages to their host genomes. *Mol Biol Evol* 2013;**30**:737–51.

Bordet J, Ciuca M. Le bactériophage de d'Hérelle, sa production et son interprétation. *Compt Rend Acad* 1920;**83**:1296–8.

Botstein D. A theory of modular evolution for bacteriophages. *Ann N Y Acad Sci* 1980;**354**:484–90.

Brochier-Armanet C, Gribaldo S, Forterre P. A DNA topoisomerase IB in Thaumarchaeota testifies for the presence of this enzyme in the last common ancestor of Archaea and Eucarya. *Biol Direct* 2008;**3**:54.

Brussow H, Desiere F. Comparative phage genomics and the evolution of Siphoviridae: insights from dairy phages. *Mol Microbiol* 2001;**39**:213–22.

Bushman W, Thompson JF, Vargas L *et al*. Control of directionality in lambda site specific recombination. *Science* 1985;**230**:906–11.

Campbell A. Prophage insertion sites. *Res Microbiol* 2003;**154**: 277–82.

Campbell AM. Chromosomal insertion sites for phages and plasmids. *J Bacteriol* 1992;**174**:7495–9.

Campbell AM. Episomes. *Adv Genet* 1963;**11**:101–45.

Castillo F, Benmohamed A, Szatmari G. Xer site specific recombination: double and single recombinase systems. *Front Microbiol* 2017;**8**:453.

Cheng C, Kussie P, Pavletich N *et al*. Conservation of structure and mechanism between eukaryotic topoisomerase I and site-specific recombinases. *Cell* 1998;**92**:841–50.

Chen S, Wang C, Xu JP *et al*. Molecular characterization of pHRDV1, a new virus-like mobile genetic element closely related to pleomorphic viruses in haloarchaea. *Extremophiles* 2014;**18**:195–206.

Chen Y, Narendra U, Iype LE *et al*. Crystal structure of a Flp recombinase–Holliday junction complex: assembly of an active oligomer by helix swapping. *Mol Cell* 2000;**6**:885–97.

Clore AJ, Stedman KM. The SSV1 viral integrase is not essential. *Virology* 2007;**361**:103–11.

Colloms SD, Sykora P, Szatmari G *et al*. Recombination at ColE1 cer requires the Escherichia coli xerC gene product, a member of the lambda integrase family of site-specific recombinases. *J Bacteriol* 1990;**172**:6973–80.

Contursi P, Jensen S, Aucelli T *et al*. Characterization of the Sulfolobus host–SSV2 virus interaction. *Extremophiles* 2006;**10**:615–27.

Cortez D, Quevillon-Cheruel S, Gribaldo S *et al*. Evidence for a Xer/dif system for chromosome resolution in archaea. *PLoS Genet* 2010;**6**:e1001166.

Cossu M, Badel C, Catchpole R *et al*. Flipping chromosomes in deep-sea archaea. *PLoS Genet* 2017;**13**:e1006847.

Cossu M, Da Cunha V, Toffano-Nioche C *et al*. Comparative genomics reveals conserved positioning of essential genomic clusters in highly rearranged Thermococcales chromosomes. *Biochimie* 2015;**118**:313–21.

Craig NL, Nash HA. The mechanism of phage lambda site-specific recombination: site-specific breakage of DNA by Int topoisomerase. *Cell* 1983;**35**:795–803.

Craig NL. The mechanism of conservative site-specific recombination. *Annu Rev Genet* 1988;**22**:77–105.

d'Hérelle F. Sur un microbe invisible antagoniste des bacilles dysentériques. *Compt Rend Acad* 1917;**165**:373–5.

Darmon E, Leach DR. Bacterial genome instability. *Microbiol Mol Biol Rev* 2014;**78**:1–39.

Demarre G, Frumerie C, Gopaul DN *et al*. Identification of key structural determinants of the IntI1 integron integrase that influence attC x attI1 recombination efficiency. *Nucleic Acids Res* 2007;**35**:6475–89.

Dorgai L, Yagil E, Weisberg RA. Identifying determinants of recombination specificity: construction and characterization of mutant bacteriophage integrases. *J Mol Biol* 1995;**252**: 178–88.

Dorman CJ, Bogue MM. The interplay between DNA topology and accessory factors in site-specific recombination in bacteria and their bacteriophages. *Sci Prog* 2016;**99**:420–37.

Echols H. Developmental pathways for the temperate phage: lysis vs lysogeny. *Annu Rev Genet* 1972;**6**:157–90.

Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010;**26**:2460–1.

Eickbush TH, Furano AV. Fruit flies and humans respond differently to retrotransposons. *Curr Opin Genet Dev* 2002;**12**: 669–74.

Eilers BJ, Young MJ, Lawrence CM. The structure of an archaeal viral integrase reveals an evolutionarily conserved catalytic core yet supports a mechanism of DNA cleavage in *trans*. *J Virol* 2012;**86**:8309–13.

Engelstadter J, Harms K, Johnsen PJ. The evolutionary dynamics of integrons in changing environments. *ISME J* 2016;**10**: 1296–307.

Erauso G, Stedman KM, van de Werken HJ *et al*. Two novel conjugative plasmids from a single strain of Sulfolobus. *Microbiology* 2006;**152**:1951–68.

Erdmann S, Tschitschko B, Zhong L *et al*. A plasmid from an Antarctic haloarchaeon uses specialized membrane vesicles to disseminate and infect plasmid-free cells. *Nat Microbiol* 2017;**2**:1446–55.

Escudero JA, Loot C, Nivina A *et al*. The integron: adaptation on demand. *Microbiol Spectr* 2015;**3**: MDNA3-0019-2014.

Esposito D, Scocca JJ. The integrase family of tyrosine recombinases: evolution of a conserved active site domain. *Nucleic Acids Res* 1997;**25**:3605–14.

Farrugia DN, Elbourne LD, Mabbutt BC *et al*. A novel family of integrases associated with prophages and genomic islands integrated within the tRNA-dihydrouridine synthase A (dusA) gene. *Nucleic Acids Res* 2015;**43**:4547–57.

Filee J, Siguier P, Chandler M. Insertion sequence diversity in Archaea. *Microbiol Mol Biol R* 2007;**71**:121–57.

Frols S, Gordon PM, Panlilio MA *et al*. Elucidating the transcription cycle of the UV-inducible hyperthermophilic archaeal virus SSV1 by DNA microarrays. *Virology* 2007;**365**:48–59.

Fusco S, She Q, Bartolucci S *et al*. T(lys), a newly identified Sulfolobus spindle-shaped virus 1 transcript expressed in the lysogenic state, encodes a DNA-binding protein interacting at the promoters of the early genes. *J Virol* 2013;**87**:5926–36.

Fusco S, She Q, Fiorentino G *et al*. Unravelling the role of the F55 regulator in the transition from lysogeny to UV induction of Sulfolobus spindle-shaped virus 1. *J Virol* 2015b;**89**:6453–61.

Gandon S. Why be temperate: lessons from bacteriophage lambda. *Trends Microbiol* 2016;**24**:356–65.
Gaudin M, Krupovic M, Marguet E *et al.* Extracellular membrane vesicles harbouring viral genomes. *Environ Microbiol* 2014;**16**:1167–75.
Gehring AM, Astling DP, Matsumi R *et al.* Genome replication in Thermococcus kodakarensis independent of Cdc6 and an origin of replication. *Front Microbiol* 2017;**8**:2084.
Goodman DA, Stedman KM. Comparative genetic and genomic analysis of the novel fusellovirus Sulfolobus spindle-shaped virus 10. *Virus evolution* 2018;**4**:vey022.
Gorlas A, Koonin EV, Bienvenu N *et al.* TPV1, the first virus isolated from the hyperthermophilic genus Thermococcus. *Environ Microbiol* 2012;**14**:503–16.
Grindley NDF, Whiteson KL, Rice PA. Mechanisms of site-specific recombination. *Annu Rev Biochem* 2006;**75**:567–605.
Grogan D, Palm P, Zillig W. Isolate B12, which harbours a virus-like element, represents a new species of the archaebacterial genus Sulfolobus, Sulfolobus shibatae, sp. nov. *Arch Microbiol* 1990;**154**:594–9.
Guo F, Gopaul DN, van Duyne GD. Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. *Nature* 1997;**389**:40–6.
Happonen LJ, Redder P, Peng X *et al.* Familial relationships in hyperthermo- and acidophilic archaeal viruses. *J Virol* 2010;**84**:4747–54.
Hawkins M, Malla S, Blythe MJ *et al.* Accelerated growth in the absence of DNA replication origins. *Nature* 2013;**503**:544–7.
Hendrix RW, Lawrence JG, Hatfull GF *et al.* The origins and ongoing evolution of viruses. *Trends Microbiol* 2000;**8**:504–8.
Hsu PL, Ross W, Landy A. The lambda phage att site: functional limits and interaction with Int protein. *Nature* 1980;**285**:85–91.
Iranzo J, Koonin EV, Prangishvili D *et al.* Bipartite network analysis of the archaeal virosphere: evolutionary connections between viruses and capsidless mobile elements. *J Virol* 2016;**90**:11043–55.
Iro M, Klein R, Galos B *et al.* The lysogenic region of virus phiCh1: identification of a repressor-operator system and determination of its activity in halophilic Archaea. *Extremophiles* 2007;**11**:383–96.
Jager D, Forstner KU, Sharma CM *et al.* Primary transcriptome map of the hyperthermophilic archaeon Thermococcus kodakarensis. *BMC Genomics* 2014;**15**:684.
Jayaram M, Ma CH, Kachroo AH *et al.* An overview of tyrosine site-specific recombination: from an Flp perspective. *Microbiol Spectr* 2015;**3**:1–28.
Jo CH, Kim J, Han AR *et al.* Crystal structure of Thermoplasma acidophilum XerA recombinase shows large C-shape clamp conformation and cis-cleavage mode for nucleophilic tyrosine. *FEBS Lett* 2016;**590**:848–56.
Jo M, Murayama Y, Tsutsui Y *et al.* *In vitro* site-specific recombination mediated by the tyrosine recombinase XerA of Thermoplasma acidophilum. *Genes Cells* 2017;**22**:646–61.
Ken R, Hackett NR. Halobacterium halobium strains lysogenic for phage phi H contain a protein resembling coliphage repressors. *J Bacteriol* 1991;**173**:955–60.
Klein R, Baranyi U, Rossler N *et al.* Natrialba magadii virus phiCh1: first complete nucleotide sequence and functional organization of a virus infecting a haloalkaliphilic archaeon. *Mol Microbiol* 2002;**45**:851–63.
Klein R, Rossler N, Iro M *et al.* Haloarchaeal myovirus phiCh1 harbours a phase variation system for the production of protein variants with distinct cell surface adhesion specificities. *Mol Microbiol* 2012;**83**:137–50.
Krupovic M, Bamford DH. Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota. *Virology* 2008;**375**:292–300.
Krupovic M, Cvirkaite-Krupovic V, Iranzo J *et al.* Viruses of archaea: structural, functional, environmental and evolutionary genomics. *Virus Res* 2018;**244**:181–93.
Krupovic M, Forterre P, Bamford DH. 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* 2010;**397**:144–60.
Krupovic M, Makarova KS, Wolf YI *et al.* Integrated mobile genetic elements in Thaumarchaeota. *Environ Microbiol* 2019;**21**:2056–78.
Kubo A, Kusukawa A, Komano T. Nucleotide sequence of the rci gene encoding shufflon-specific DNA recombinase in the IncI1 plasmid R64: homology to the site-specific recombinases of integrase family. *Mol Gen Genet* 1988;**213**:30–5.
Landy A. The lambda integrase site-specific recombination pathway. *Microbiol Spectr* 2015;**3**: MDNA3-0051-2014.
La Scola B, Desnues C, Pagnier I *et al.* The virophage as a unique parasite of the giant mimivirus. *Nature* 2008;**455**:100–4.
Letunic I, Bork P. Interactive Tree of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019;**47**: W256–9.
Letzelter C, Duguet M, Serre MC. Mutational analysis of the archaeal tyrosine recombinase SSV1 integrase suggests a mechanism of DNA cleavage in *trans*. *J Biol Chem* 2004;**279**:28936–44.
Levin RL, Stewart FM, Chao L. Resource-limited growth, competition, and predation: a model and experimental studies with bacteria and bacteriophage. *Am Nat* 1977;**111**:3–24.
Lewis JA, Hatfull GF. Control of directionality in integrase-mediated recombination: examination of recombination directionality factors (RDFs) including Xis and Cox proteins. *Nucleic Acids Res* 2001;**29**:2205–16.
Liu D, Huang L. Induction of the Sulfolobus shibatae virus SSV1 DNA replication by mitomycin C. *Chin Sci Bull* 2002;**47**:923–7.
Liu Y, Wang J, Liu Y *et al.* Identification and characterization of SNJ2, the first temperate pleolipovirus integrating into the genome of the SNJ1-lysogenic archaeal strain. *Mol Microbiol* 2015;**98**:1002–20.
Li Z, Li X, Xiao X *et al.* An integrative genomic island affects the adaptations of the piezophilic hyperthermophilic archaeon Pyrococcus yayanosii to High temperature and high hydrostatic pressure. *Front Microbiol* 2016;**7**:1927.
Lu F, Churchward G. Conjugative transposition: Tn916 integrase contains two independent DNA binding domains that recognize different DNA sequences. *EMBO J* 1994;**13**:1541–8.
Luo Y, Pfister P, Leisinger T *et al.* The genome of archaeal prophage PsiM100 encodes the lytic enzyme responsible for autolysis of Methanothermobacter wolfeii. *J Bacteriol* 2001;**183**:5788–92.
Marchler-Bauer A, Bryant SH. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* 2004;**32**:W327–331.
Martin A, Yeats S, Janekovic D *et al.* SAV 1, a temperate u.v.-inducible DNA virus-like particle from the archaebacterium Sulfolobus acidocaldarius isolate B12. *EMBO J* 1984;**3**:2165–8.
McCusker MP, Turner EC, Dorman CJ. DNA sequence heterogeneity in Fim tyrosine-integrase recombinase-binding elements and functional motif asymmetries determine the directionality of the fim genetic switch in Escherichia coli K-12. *Mol Microbiol* 2008;**67**:171–87.

Meinke G, Bohm A, Hauber J *et al.* Cre recombinase and other tyrosine recombinases. *Chem Rev* 2016;**116**:12785–820.
Mendler K, Chen H, Parks DH *et al.* AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids Res* 2019;**47**:4442–8.
Miele V, Penel S, Duret L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* 2011;**12**:116.
Mizuuchi M, Mizuuchi K. Integrative recombination of bacteriophage lambda: extent of the DNA sequence involved in attachment site function. *Proc Natl Acad Sci U S A* 1980;**77**:3220–4.
Moitoso de Vargas L, Pargellis CA, Hasan NM *et al.* Autonomous DNA binding domains of lambda integrase recognize two different sequence families. *Cell* 1988;**54**:923–9.
Muskhelishvili G, Palm P, Zillig W. SSV1-encoded site-specific recombination system in Sulfolobus shibatae. *Mol Gen Genet* 1993;**237**:334–42.
Nash HA. Integrative recombination of bacteriophage lambda DNA *in vitro*. *Proc Natl Acad Sci U S A* 1975;**72**:1072–6.
Nash HA. Purification of bacteriophage lambda Int protein. *Nature* 1974;**247**:543–5.
Nilsson H, Cardoso-Palacios C, Haggard-Ljungquist E *et al.* Phylogenetic structure and evolution of regulatory genes and integrases of P2-like phages. *Bacteriophage* 2011;**1**:207–18.
Nunes-Duby SE, Kwon HJ, Tirumalai RS *et al.* Similarities and differences among 105 members of the Int family of site-specific recombinases. *Nucleic Acids Res* 1998;**26**:391–406.
Nussinov R, Weisberg RA. Bacteriophage lambda int protein may recognize structural features of the attachment sites. *J Biomol Struct Dyn* 1986;**3**:1133–44.
Oberto J, Gaudin M, Cossu M *et al.* Genome sequence of a hyperthermophilic archaeon, Thermococcus nautili 30-1, that produces viral vesicles. *Genome Announc* 2014;**2**:1–2.
Oberto J, Sloan SB, Weisberg RA. A segment of the phage HK022 chromosome is a mosaic of other lambdoid chromosomes. *Nucleic Acids Res* 1994;**22**:354–6.
Oppenheim AB, Kobiler O, Stavans J *et al.* Switches in bacteriophage lambda development. *Annu Rev Genet* 2005;**39**:409–29.
Palm P, Schleper C, Grampp B *et al.* Complete nucleotide sequence of the virus SSV1 of the archaebacterium Sulfolobus shibatae. *Virology* 1991;**185**:242–50.
Papagiannis CV, Sam MD, Abbani MA *et al.* Fis targets assembly of the Xis nucleoprotein filament to promote excisive recombination by phage lambda. *J Mol Biol* 2007;**367**:328–43.
Pargellis CA, Nunes-Duby SE, de Vargas LM *et al.* Suicide recombination substrates yield covalent lambda integrase–DNA complexes and lead to identification of the active site tyrosine. *J Biol Chem* 1988;**263**:7678–85.
Parks DH, Chuvochina M, Chaumeil PA *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol* 2020;**38**:1079–86.
Paul JH. Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *ISME J* 2008;**2**:579–89.
Pauly MD, Bautista MA, Black JA *et al.* Diversified local CRISPR-Cas immunity to viruses of Sulfolobus islandicus. *Philos Trans R Soc Lond B Biol Sci* 2019;**374**:20180093.
Peng X, Holz I, Zillig W *et al.* Evolution of the family of pRN plasmids and their integrase-mediated insertion into the chromosome of the crenarchaeon Sulfolobus solfataricus. *J Mol Biol* 2000;**303**:449–54.
Peng X. Evidence for the horizontal transfer of an integrase gene from a fusellovirus to a pRN-like plasmid within a single strain of Sulfolobus and the implications for plasmid survival. *Microbiology* 2008;**154**:383–91.
Prangishvili D, Stedman K, Zillig W. Viruses of the extremely thermophilic archaeon Sulfolobus. *Trends Microbiol* 2001;**9**:39–43.
Prangishvili D, Vestergaard G, Haring M *et al.* Structural and genomic properties of the hyperthermophilic archaeal virus ATV with an extracellular stage of the reproductive cycle. *J Mol Biol* 2006;**359**:1203–16.
Rajeev L, Malanowska K, Gardner JF. Challenging a paradigm: the role of DNA homology in tyrosine recombinase reactions. *Microbiol Mol Biol Rev* 2009;**73**:300–9.
Raymann K, Forterre P, Brochier-Armanet C *et al.* Global phylogenomic analysis disentangles the complex evolutionary history of DNA replication in archaea. *Genome Biol Evol* 2014;**6**:192–212.
Redder P, Garrett RA. Mutations and rearrangements in the genome of Sulfolobus solfataricus P2. *J Bacteriol* 2006;**188**:4198–206.
Redder P, Peng X, Brugger K *et al.* Four newly isolated fuselloviruses from extreme geothermal environments reveal unusual morphologies and a possible interviral recombination mechanism. *Environ Microbiol* 2009;**11**:2849–62.
Reiter WD, Palm P, Yeats S. Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. *Nucleic Acids Res* 1989;**17**:1907–14.
Ren Y, She Q, Huang L. Transcriptomic analysis of the SSV2 infection of Sulfolobus solfataricus with and without the integrative plasmid pSSVi. *Virology* 2013;**441**:126–34.
Rice G, Tang L, Stedman K *et al.* The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life. *Proc Natl Acad Sci U S A* 2004;**101**:7716–20.
Robinson NP, Bell SD. Extrachromosomal element capture and the evolution of multiple replication origins in archaeal chromosomes. *Proc Natl Acad Sci U S A* 2007;**104**:5806–11.
Roine E, Kukkaro P, Paulin L *et al.* New, closely related haloarchaeal viral elements with different nucleic acid types. *J Virol* 2010;**84**:3682–9.
Rossler N, Klein R, Scholz H *et al.* Inversion within the haloalkaliphilic virus phi Ch1 DNA results in differential expression of structural proteins. *Mol Microbiol* 2004;**52**:413–26.
Rutkai E, Gyorgy A, Dorgai L *et al.* Role of secondary attachment sites in changing the specificity of site-specific recombination. *J Bacteriol* 2006;**188**:3409–11.
Samson RY, Xu Y, Gadelha C *et al.* Specificity and function of archaeal DNA replication initiator proteins. *Cell Rep* 2013;**3**:485–96.
Schleper C, Kubo K, Zillig W. The particle SSV1 from the extremely thermophilic archaeon Sulfolobus is a virus: demonstration of infectivity and of transfection with viral DNA. *Proc Natl Acad Sci U S A* 1992;**89**:7645–9.
Schuch R, Fischetti VA. The secret life of the anthrax agent Bacillus anthracis: bacteriophage-mediated ecological adaptations. *PLoS One* 2009;**4**:e6532.
Selb R, Derntl C, Klein R *et al.* The viral gene ORF79 encodes a repressor regulating induction of the lytic life cycle in the Haloalkaliphilic virus varphiCh1. *J Virol* 2017;**91**:e00206–17.
Sencilo A, Jacobs-Sera D, Russell DA *et al.* Snapshot of haloarchaeal tailed virus genomes. *RNA Biol* 2013;**10**:803–16.

Serre MC, El Arnaout T, Brooks MA *et al.* The carboxy-terminal αN helix of the archaeal XerA tyrosine recombinase is a molecular switch to control site-specific recombination. *PLoS One* 2013;**8**:e63010.

Serre MC, Letzelter C, Garel JR *et al.* Cleavage properties of an archaeal site-specific recombinase, the SSV1 integrase. *J Biol Chem* 2002;**277**:16758–67.

She Q, Brugger K, Chen L. Archaeal integrative genetic elements and their impact on genome evolution. *Res Microbiol* 2002;**153**:325–32.

She Q, Chen B, Chen L. Archaeal integrases and mechanisms of gene capture. *Biochem Soc Trans* 2004;**32**:222–6.

She Q, Peng X, Zillig W *et al.* Gene capture in archaeal chromosomes. *Nature* 2001a;**409**:478.

She Q, Phan H, Garrett RA *et al.* Genetic profile of pNOB8 from Sulfolobus: the first conjugative plasmid from an archaeon. *Extremophiles* 1998;**2**:417–25.

Signer ER, Beckwith JR. Transposition of the Lac region of Escherichia coli III. The mechanism of attachment of bacteriophage phi80 to the bacterial chromosome. *J Mol Biol* 1966;**22**:33–51.

Siguier P, Gourbeyre E, Varani A *et al.* Everyman's guide to bacterial insertion sequences. *Microbiol Spectr* 2015;**3**: MDNA3-0030-2014.

Smith MC, Till R, Brady K *et al.* Synapsis and DNA cleavage in phiC31 integrase-mediated site-specific recombination. *Nucleic Acids Res* 2004;**32**:2607–17.

Soler N, Marguet E, Cortez D *et al.* Two novel families of plasmids from hyperthermophilic archaea encoding new families of replication proteins. *Nucleic Acids Res* 2010;**38**: 5088–104.

Stark WM. The serine recombinases. *Microbiol Spectr* 2014;**2**:1–16.

Stedman KM, She Q, Phan H *et al.* Relationships between fuselloviruses infecting the extremely thermophilic archaeon Sulfolobus: SSV1 and SSV2. *Res Microbiol* 2003;**154**:295–302.

Stern A, Keren L, Wurtzel O *et al.* Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet* 2010;**26**:335–40.

Stolt P, Zillig W. *In vivo* studies on the effects of immunity genes on early lytic transcription in the Halobacterium salinarium phage phi H. *Mol Gen Genet* 1992;**235**:197–204.

Subramanya HS, Arciszewska LK, Baker RA *et al.* Crystal structure of the site-specific recombinase, XerD. *EMBO J* 1997;**16**:5178–87.

Sugahara J, Fujishima K, Nunoura T *et al.* Genomic heterogeneity in a natural archaeal population suggests a model of tRNA gene disruption. *PLoS One* 2012;**7**:e32504.

Sung P, Klein H. Mechanism of homologous recombination: mediators and helicases take on regulatory functions. *Nat Rev Mol Cell Biol* 2006;**7**:739–50.

Sun Y, McCorvie TJ, Yates LA *et al.* Structural basis of homologous recombination. *Cell Mol Life Sci* 2020;**77**:3–18.

Suzuki S, Yoshikawa M, Imamura D *et al.* Compatibility of site-specific recombination units between mobile genetic elements. *iScience* 2020;**23**:100805.

Tagashira K, Fukuda W, Matsubara M *et al.* Genetic studies on the virus-like regions in the genome of hyperthermophilic archaeon, Thermococcus kodakarensis. *Extremophiles* 2013;**17**:153–60.

Tang SL, Nuttall S, Ngui K *et al.* HF2: a double-stranded DNA tailed haloarchaeal virus with a mosaic genome. *Mol Microbiol* 2002;**44**:283–96.

Twort FW. An investigation on the nature of ultramicroscopic viruses. *Lancet* 1915;**11**:1241–3.

Vandecraen J, Chandler M, Aertsen A *et al.* The impact of insertion sequences on bacterial genome plasticity and adaptability. *Crit Rev Microbiol* 2017;**43**:709–30.

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

Wagner A, Whitaker RJ, Krause DJ *et al.* Mechanisms of gene flow in archaea. *Nat Rev Microbiol* 2017;**15**:492–501.

Wang J, Liu Y, Liu Y *et al.* A novel family of tyrosine integrases encoded by the temperate pleolipovirus SNJ2. *Nucleic Acids Res* 2018;**46**:2521–36.

Wang X, Kim Y, Ma Q *et al.* Cryptic prophages help bacteria cope with adverse environments. *Nat Commun* 2010;**1**:147.

Wang Y, Duan Z, Zhu H *et al.* A novel Sulfolobus non-conjugative extrachromosomal genetic element capable of integration into the host genome and spreading in the presence of a fusellovirus. *Virology* 2007;**363**:124–33.

Weckselblatt B, Rudd MK. Human structural variation: mechanisms of chromosome rearrangements. *Trends Genet* 2015;**31**:587–99.

Weil J, Signer ER. Recombination in bacteriophage lambda. II. Site-specific recombination promoted by the integration system. *J Mol Biol* 1968;**34**:273–9.

Wiedenheft B, Stedman K, Roberto F *et al.* Comparative genomic analysis of hyperthermophilic archaeal Fuselloviridae viruses. *J Virol* 2004;**78**:1954–61.

Williams KP. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res* 2002;**30**:866–75.

Wimmer F, Beisel CL. CRISPR-Cas systems and the paradox of self-targeting spacers. *Front Microbiol* 2019;**10**:3078.

Winckler T, Szafranski K, Glockner G. Transfer RNA gene-targeted integration: an adaptation of retrotransposable elements to survive in the compact Dictyostelium discoideum genome. *Cytogenet Genome Res* 2005;**110**:288–98.

Witte A, Baranyi U, Klein R *et al.* Characterization of Natronobacterium magadii phage phi Ch1, a unique archaeal phage containing DNA and RNA. *Mol Microbiol* 1997;**23**:603–16.

Yagil E, Dolev S, Oberto J *et al.* Determinants of site-specific recombination in the lambdoid coliphage HK022. An evolutionary change in specificity. *J Mol Biol* 1989;**207**:695–717.

Yagil E, Dorgai L, Weisberg RA. Identifying determinants of recombination specificity: construction and characterization of chimeric bacteriophage integrases. *J Mol Biol* 1995;**252**: 163–77.

Yang W. Topoisomerases and site-specific recombinases: similarities in structure and mechanism. *Crit Rev Biochem Mol Biol* 2010;**45**:520–34.

Yeats S, McWilliam P, Zillig W. A plasmid in the archaebacterium Sulfolobus acidocaldarius. *EMBO J* 1982;**1**:1035–8.

Zhan ZY, Ouyang SY, Liang WG *et al.* Structural and functional characterization of the C-terminal catalytic domain of SSV1 integrase. *Acta Crystallogr D* 2012;**68**:659–70.

Zhan ZY, Zhou J, Huang L. Site-specific recombination by SSV2 integrase: substrate requirement and domain functions. *J Virol* 2015;**89**:10934–44.

Zissler J. Integration-negative (int) mutants of phage lambda. *Virology* 1967;**31**:189.

Zivanovic Y, Lopez P, Philippe H *et al.* Pyrococcus genome comparison evidences chromosome shuffling-driven evolution. *Nucleic Acids Res* 2002;**30**:1902–10.