Eléments génétiques mobiles et évolution génomique chez les Archées Thermococcales

Catherine Badel

université PARIS-SACLAY

NNT : 2019SACLS168

Thèse de doctorat

Eléments génétiques mobiles et évolution génomique chez les Archées Thermococcales

Mobile genetic elements and genome evolution in the Archaea Thermococcales

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris Sud

École doctorale n°577
Structure et dynamique des systèmes vivants (SDSV)
Spécialité de doctorat: Sciences de la vie et de la santé

Thèse présentée et soutenue à Orsay, le 2 juillet 2019, par

Catherine Badel

Composition du Jury :

Stéphanie Bury-Moné Professeur, Université Paris-Sud (– I2BC UMR9198)	Président
Claire Geslin Maître de Conférences, Université de Bretagne Occidentale (– LM2E UMR6197)	Rapporteur
Céline Loot Chargé de recherche, Institut Pasteur	Rapporteur
Gaël Erauso Professeur, Université Aix Marseille (– MIO UMR7294)	Examinateur
Eduardo Rocha Directeur de recherche, Institut Pasteur	Examinateur
Jacques Oberto Directeur de Recherche, CNRS (– I2BC UMR9198)	Directeur de thèse

Remerciements

Je tiens tout d'abord à remercier Claire Geslin et Céline Loot d'avoir accepté de rapporter mes travaux de thèse et Stéphanie Bury-Moné, Gaël Erauso et Eduardo Rocha d'avoir accepté d'examiner mon travail.

Je remercie sincèrement mon directeur de thèse Jacques Oberto pour sa disponibilité, ses conseils et sa confiance. Jacques a su me guider tout en me donnant de l'indépendance. Merci pour la richesse et l'ouverture de nos discussions scientifiques.

Je remercie également Roxane Lestini et Marc Lavigne pour leur participiation à mes comités de thèse. Vos remarques et conseils m'ont aidé tout au long de la thèse.

Je remercie chaleureusement tous les membres passés et actuels de l'équipe Biologie Cellulaire des Archées à Orsay. Vous avez participé à créer des conditions de travail idéales et je n'aurais pas pu espérer de meilleurs collègues. Nos discussions scientifiques et moins scientifiques autour d'un café, et souvent d'un gâteau, me manqueront. Plus particulièrement, merci Patrick pour tes réflexions scientifiques qui sont toujours passionnantes. Merci Violette pour notre collaboration très enrichissante et tes conseils et suggestions en bioinformatique. Merci Danièle, Tomio et Adeline pour vos précieux conseils en purification de protéines. Merci Ryan de m'avoir accepté dans ton bureau et d'avoir partagé ton immense culture scientifique et expérimentale. Merci Marie-Claire et Tamara pour vos conseils avisés sur la recherche et l'enseignement. Merci Evelyne pour ton enthousiasme et tes encouragements. Merci Adeline et Paul, vous avez été de super compagnons de thèse. Merci à Myriam qui prend soin du laboratoire avec gentillesse et bonne humeur. Merci à Matteo de m'avoir transmis le projet intégrase. Merci également aux stagiaires Nicolas Alexandre et Florian de Ceuyper pour leur contribution. Et un dernier merci à Patrick et Marie-Claire pour leur relecture de mon manuscrit.

Je remercie Herman Van Tilbeurgh, Ines Gallay et Stéphane Plancqueel de la plateforme de cristallisation de l'I2BC, pour leur disponibilité et leur conseils en cristallographie. J'ai beaucoup appris de notre collaboration.

Je remercie le département de biologie de l'ENS de Lyon qui m'a éveillée à la microbiologie et à l'évolution et a fourni mon financemement de thèse.

Je remercie ma famille et ma belle-famille pour leur présence, leur curiosité de mon travail et tous les bons moments passés ensemble. Merci pour les discussions animées qui ont incontestablement participé à ma formation scientifique. Je remercie spécialement mes parents pour leur soutien sans limite. Grâce à vous, je ne connais pas l'impossible. Merci Anne-Céline d'avoir pavé la route devant moi. Merci Charles pour tes jeux d'enfants qui effacent tous les soucis.

Et enfin, je remercie Benjamin pour son soutien au délà des kilomètres. A tes côtés, tout devient simple.

Table des matières

TABLE DES ILLUSTRATIONS **4**

Figures *4*
Tableaux *6*

INTRODUCTION **7**

Chapitre 1. Les archées 7
Les archées et le triptyque du vivant *7*
Ecologie des archées 8
Classification et phylogénie des archées 10
Le chromosome des archées 10
Morceaux choisis de diversité des archées *12*
Les Thermococcales 12
Les Methanococcales 14
Les Methanosarcinales 14
Les Archaeoglobales 14

Chapitre 2. Les éléments génétiques mobiles 15
Diversité des éléments génétiques mobiles dans l'arbre du vivant *15*
Les plasmides 15
Les virus 19
Les éléments transposables 20
Discussion de la classification des éléments génétiques mobiles 23
Fonctions portées par les éléments génétiques mobiles 23
La lysogénie *25*
Le cycle lysogénique typique 26
Coûts et bénéfices évolutifs de la lysogénie 26
La lysogénie chez les archées 27
Diversité des éléments génétiques mobiles chez les archées *27*
Les plasmides d'archées 27
Les virus d'archées 32
Les éléments transposables d'archées 35
Evolution des éléments génétiques mobiles *36*
Mécanismes cellulaires de défense contre les éléments génétiques mobiles *38*
Méthodes d'identification des éléments génétiques mobiles *40*

Chapitre 3. Recombinaisons génétiques 42
Les différents types de recombinaison et leurs enzymes *43*
Recombinaison homologue 43
Transposition par les recombinases à DDE 45
Recombinaison site-spécifique conservative 47
La recombinaison dans le contexte génétique *50*
Différents produits de recombinaison 50
L'intégration et l'excision d'éléments génétiques mobiles 50

Les recombinases à tyrosine en trois enzymes *52*
La recombinase Cre du phage P1, un système simple 52
L'intégrase du phage lambda et le contrôle de la directionnalité de réaction 54
La flipase Flp du plasmide 2µ assemble le site catalytique en trans 56
Les intégrases d'archées *56*
Diversité des intégrases archées 56
Les intégrases des fusellovirus de Sulfolobus 58
L'intégrase du pléolipovirus SNJ2 59

Chapitre 4. Evolution génomique des archées Thermococcales 60
Evolution des génomes et éléments génétiques mobiles *60*
Les chromosomes de Thermococcales subissent de nombreux réarrangements *61*

QUESTIONNEMENTS, OBJECTIFS ET STRATEGIES **63**

Quels sont les mécanismes d'inversions génomiques chez les Thermococcales ? 63
Quels sont les avantages évolutifs de l'activité suicidaire des intégrases d'archées ? 64

RESULTS AND DISCUSSION **65**

Part 1. An unprecedented catalytic activity generates chromosomal inversions in Thermococcus species 65
Article 1. Flipping chromosomes in deep-sea archaea *65*

Part 2. The plasmid family pT26-2 alternatively harbors two distinct integrases 108
Article 2. The global distribution and evolutionary history of the pT26-2 archaeal plasmid family *108*
Characterization of the integrases of Methanococcales *172*

Part 3. The activity and evolution of Thermococcales suicidal integrases 175
Article 3. Pervasive suicidal integrases in archaea *175*
The annotated Int^{TPV1} is a truncated and inactive mutant *218*

Part 4. A new tool for mobile genetic element studies 221
Article 4. WASPS: Web Assisted Symbolic Plasmid Synteny *221*

Part 5. Towards a dissection of IntpTN3 site-specific and homologous recombinational activities 236
No reaction conditions can discriminate the two activities *236*
Integrases similar to Int^{pTN3} catalyze site-specific recombination in one of two distinct sites *239*
Identification of integrases similar to Int^{pTN3} 239
The integrases have distinct site specificities 239
The integrases can be divided in two groups: one for each specificity 239
The two specificities can efficiently be used for integration 240
Searching for the mechanism of specific site recognition 240
Searching for the sequence determinism of the homologous recombination activity *247*
The homologous recombination activity is shared by all the members of the Int^{pTN3} family . 247
Additional sequences of the Int^{pTN3} family 247

CONCLUSIONS AND PERSPECTIVES 250

L'intégrase du plasmide pTN3 catalyse des inversions chromosomiques chez les Thermococcus par une activité de recombinaison homologue 250

Dissection des deux activités catalytiques de l'intégrase du plasmide pTN3 ..... *250*

Conséquences évolutives de l'activité de l'intégrase du plasmide pTN3 ..... *251*

Du côté du plasmide 251

Du côté des hôtes Thermococcales 251

Avantages évolutifs de l'activité suicidaire des intégrases d'archées 254

Un module d'intégration tout en un 255

Coopération salvatrice entre intégrases suicidaires 255

Un suicide coercitif 255

La fragmentation comme mécanisme évolutif 256

Origine évolutive des intégrases de type suicide *257*

Intégrases suicidaires et hyperthermophilie *258*

Reconnaissance de la spécificité par les intégrases suicidaires *259*

Intégrases d'archées, revue de mes résultats et de la littérature 259

Article 5. Archaeal tyrosine integrases *259*

SUPPLEMENTARY MATERIAL AND METHODS 291

IN VITRO RECOMBINATION ACTIVITY ASSAYS 291

Expression vector construction 291

Recombinant protein production and purification 291

Integrase substrate construction and production 292

In vitro integrase enzymatic assay 292

CRYSTALLOGRAPHY 293

REFERENCES 295

Note au lecteur :

L'introduction, les questionnements et les conclusions sont rédigés en français. Les articles et autres résultats sont rédigés en anglais. De plus, les articles sont indépendants du reste du manuscrit, avec leur propre numérotation et bibiographie.

Table des illustrations

Les illustrations des articles ne sont pas inclues.

Figures

Figure 1. Phylogénie des archées 9
Figure 2. Eléments de l'organisation des génomes bactériens. 11
Figure 3. Distribution de la taille (kb) des plasmides de la base de donnée WASPS 16
Figure 4. Les plasmides emploient différents mécanismes de réplication 17
Figure 5. Représentation schématique des cycles lytique et lysogénique 20
Figure 6. Organisation schématique des plusieurs éléments transposables de bactéries 21
Figure 7. Carte du plasmide pTN3 28
Figure 8. Carte du plasmide pT26-2. 29
Figure 9. Comparaison du plasmide pT26-2 avec des éléments intégrés de Thermococcales et Methanococcales. 30
Figure 10. Comparaison des génomes du plasmide pAMT11 et de l'élément intégré TKV1 31
Figure 11. Morphotypes des virus d'archées. 32
Figure 12. Virion et carte du génome de TPV1. 34
Figure 13. Relations entre les plasmides pCIR10 et pIRI48 et les virus PAV1 et TPV1 de Thermococcales 36
Figure 14. Comparaison génétique de différents plasmides conjugatifs de Sulfolobus.. 37
Figure 15. Cycle de vie lytique d'un virus et mécanismes de défense bactériens. 38
Figure 16. Le mécanisme de défense CRISPR-Cas en trois étapes. 39
Figure 17. Les voies de réparation de cassures double brins chez les eucaryotes 44
Figure 18. Synthèse d'ADN et intégration rétrovirale dans le génome hôte. 46
Figure 19. Modèle du mécanisme d'échange de brins catalysé par les recombinases à sérine. 47
Figure 20. Modèle du mécanisme d'échange de brins catalysé par les recombinases à tyrosine 49
Figure 21. Les différents produits de recombinaison 50
Figure 22. Les recombinases comme intégrases et résolvases d'éléments génétiques mobiles 51
Figure 23. La recombinase Cre liée à son site spécifique loxP. 53
Figure 24.Structure de l'intégrase λ et de ses sites spécifiques 54
Figure 25. Représentation schématique de la structure du complexe synaptique de λ lors des réactions d'excision et d'intégration 55
Figure 26. Modèles d'intégration pour les deux types d'intégrases d'archées 57

Figure 27. Représentation « dotplot » de la comparaison deux à deux des génomes de Pyrococcus horikoshii, Pyrococcus abyssi et Pyrococcus furiosus. 62

Figure 28. MAFFT alignment of the integrases identified in pT26-2 like elements of Methanococcales. 173

Figure 29. Size exclusion chromatography profile for the purification of the integrases of *Methanocaldococcus sp.* 406-22 and *Methanocaldococcus fervens* AG86 174

Figure 30. The integrase from the integrated element MspFS406-22_IP1 can catalyze site-specific recombination.. 174

Figure 31. The integrase of the virus TPV1 presented no site-specific recombination activity in vitro 218

Figure 32. MAFFT alignment of Int^{TPV1} with its most closely related integrases and Int^{pT26-2} 219

Figure 33. Size exclusion chromatography profile during the purification of the Int^{TPV1} and Int^{pT26-2}.. 220

Figure 34. Divalent cation influence on Int^{pTN3} integration activity 237

Figure 35. Synteny conservation between the different elements encoding an integrase of the Int^{pTN3} family 241

Figure 36. The Int^{pTN3} family of integrases presents two integration sites in $tRNA^{Leu}CAA$ or $tRNA^{Ser}CGA$ genes 242

Figure 37. PhyML phylogenetic tree of the Int^{pTN3} family of integrases..... 243

Figure 38. MAFFT sequence alignment of the integrases of Int^{pTN3} family 243-89

Figure 39. Plasmid integrations and integrase evolution models.. 245

Figure 40. Int^{pTF1} can catalyze site-specific and homologous recombination in vitro. 245

Figure 41. Integrases of plasmids pTN3 and pTF1 sequence alignment. 246

Figure 42. Int^{pTN3} crystal and its diffraction pattern. 248

Figure 43. The Int^{pTN3} family of integrases presents additional sequences compared to the Int^{pT26-2} family of integrases 248-94

Figure 44. Comparaison par dotplot des chromosomes des deux souches de Thermococcus nautili avec *Thermococcus* 9-3 252

Figure 45. Nombre de répétition d'une taille supérieure à 100 pb par megabase dans les chromosomes de Thermococcales..... 253

Figure 46. Une hypothèse pour l'origine des intégrases de type suicide à partir d'une intégrase classique 257

Figure 47. L'ilot génomique PYG1 et le plasmide intégré PyaCH1_IP16 sont intégrés au niveau de gènes codant pour des ARNt dans le chromosome de Pyrococcus yayanosii CH1..... 258

Tableaux

Tableau 1. Les archées partagent des caractéristiques avec les eucaryotes et les bactéries 8

Tableau 2. Caractéristiques de quatre ordres d'archées 12

Tableau 3. Fonctions portées par les éléments génétiques mobiles 24

Tableau 4. Caractéristiques des éléments génétiques mobiles intégrés utilisables pour leur détection 41

Tableau 5. Caractéristiques et enzymes des différents types de recombinaisons 42

Table 6. Mobile genetic elements presenting an integrase of the Int^{pTN3} family 238

Table 7. Plasmids constructed during the thesis 293

Table 8. Oligonucleotides used during the thesis 294

Introduction

Chapitre 1. Les archées

Les archées et le triptyque du vivant

Les archées ont été définies en 1977 par Carl Woese et George Fox comme un domaine du vivant additionnel auprès des bactéries et des eucaryotes, révélant ainsi une trichotomie du vivant (Woese and Fox, 1977). Cette définition reposait sur l'analyse de la diversité de nombreuses séquences d'ARN ribosomiques 16S et 18S. Auparavant, les espèces archéennes étaient groupées avec les bactéries sur la base de caractères morphologiques communs (Tableau 1). Les archées et les bactéries sont en effet des microorganismes unicellulaires dont le chromosome est présent directement dans le cytoplasme (absence de noyau). D'autres points communs ont par la suite été découverts entre les archées et les bactéries notamment en lien avec leur chromosome et leurs éléments génétiques mobiles. Les archées et la plus part des bactéries présentent un chromosome circulaire dont la composition est organisée et compactée sous l'effet de contraintes similaires (cf. page 10) (Kellner et al., 2018). Les archées et les bactéries mettent également en place des mécanismes de défense similaires contre les éléments génétiques mobiles reflétant la similarité entre leurs éléments génétiques mobiles (mobilome) (cf. Chapitre 2 page 15) (Forterre et al., 2014). Au confluent des similarités de chromosome et mobilome se trouve un autre point commun entre les archées et les bactéries : leur évolution génomique est largement soumise à l'influence des transferts horizontaux de gènes (Touchon and Rocha, 2016). Ces trois points communs concernent particulièrement les aspects abordés par mes travaux de thèse.

Les analyses moléculaires et biochimiques des archées ont mis en évidence une certaine similarité avec les eucaryotes pour différents aspects du fonctionnement cellulaire (traitement des protéines ou présence d'actine par exemple) et pour les systèmes de traitement de l'information (Tableau 1) (Albers et al., 2013; Cavicchioli, 2011; Makarova et al., 2010). Notamment, les archées et les eucaryotes présentent une base commune pour les machineries de réplication de l'ADN, de transcription et de traduction (Ausiannikava and Allers, 2017; Bell and Jackson, 1998). Cette similarité a été exploitée par l'utilisation des archées comme modèle d'étude simplifiée des eucaryotes. Pour autant, les archées ne sont pas de simples chimères de bactéries et d'eucaryotes et présentent des caractères propres. Les archées possèdent par exemple des lipides membranaires uniques (Villanueva et al., 2014) et sont infectées par des virus exceptionnels (Prangishvili et al., 2017). Une autre caractéristique des archées est leur versatilité avec de nombreux types de métabolismes et conditions de vie exploités (Offre et al., 2013; Spang et al., 2017).

Tableau 1. Les archées partagent des caractéristiques avec les eucaryotes et les bactéries

Points communs entre les archées et les eucaryotes	Points communs entre les archées et les bactéries	Spécificités des archées
	‣ Morphologie (taille et forme)	‣ Lipides membranaires
	‣ Chromosome dans le cytoplasme	
	‣ Chromosome circulaire	
	‣ Organisation du chromosome (compaction, opérons, distribution des gènes)	
	‣ Importance évolutive du transfert horizontal de gènes	
‣ Machineries de traitement de l'information (réplication de l'ADN, transcription, traduction)		
‣ Traitement des protéines (sécrétion, modification et dégradation)	‣ Régulateurs de transcription	
‣ Cytosquelette (homologues d'actine et de tubuline)		
‣ Formation des vésicules (système ESCRT)		
	‣ Nombreux plasmides (cf. Chapitre 2 pages 15 et 27)	‣ Virus spécifiques (cf. Chapitre 2 page 32)
	‣ Plasmides et virus codant pour une intégrase à tyrosine (cf. Chapitre 3 pages 48 et 50)	
	‣ Systèmes de défense contre les éléments génétiques mobiles (cf. Chapitre 2 page 38)	

Ecologie des archées

Les archées sont présentes de manière ubiquitaire dans tous les environnements (Offre et al., 2013; Schleper et al., 2005) parmi lesquels : ‣ les sols agriculturaux, forestiers ou gelés du permafrost ‣ les sédiments océaniques, marins ou lacustres plus ou moins salés et plus ou moins chauds ‣ les eaux océaniques, les eaux marines salées à très salées ou les eaux lacustres douces ‣ les systèmes hydrothermaux marins, océaniques et terrestres ‣ les lacs acides géothermaux ‣ les surfaces animales (système digestif, peau) ‣ le cytoplasme d'autres archées. La présence d'archées dans tous ces environnements est rendue possible par la versatilité de leurs modes de vies (aérobiose, anaérobiose, chimioautotrophie, hétérotrophie, acidophilie, halophilie, piézophilie, mésophilie, psycrophilie, thermophilie, hyperthermophilie, etc.). Les organismes connus les plus hyperthermophiles sont d'ailleurs des archées. La palette de métabolismes des archées est aussi très diverse et inclut la réduction du souffre, l'oxydation de l'ammonium, l'oxydation du méthane, ou la méthanogenèse (Offre et al., 2013). Seules des archées catalysent la méthanogenèse. Les archées sont ainsi des acteurs majeurs des cycles biogéochimiques (carbone, azote et souffre).

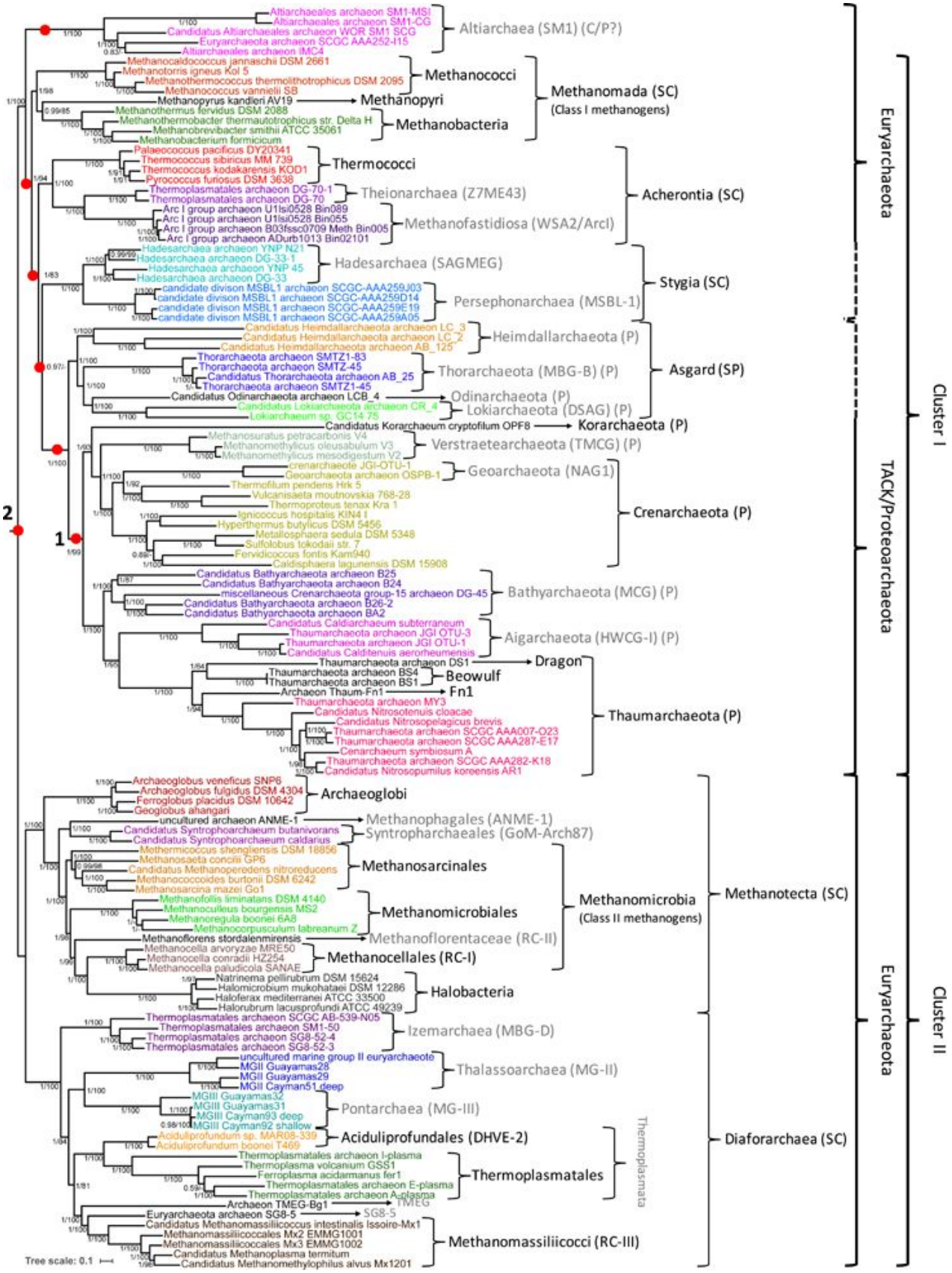


Figure 1. Phylogénie des archées. Phylogénie bayésienne basé sur 41 gènes. La barre d'échelle représente le nombre moyen de substitutions par sites. La valeur des nœuds correspond aux probabilités postérieurs et au à la valeur du bootstrap ultrafast. L'arbre est raciné comme dans (Raymann et al., 2015). Des racines alternatives sont indiquées par les points rouges. La police grise indique les clades pour lesquels aucun isolat n'est disponible. C : classe, P :phylum, SC : superclasse, SP : superphylum. Figure tirée de Adam et al., 2017

Classification et phylogénie des archées

Les archées ont d'abord été divisées en deux phyla : Crenarcheota et Euryarcheota (Woese et al., 1990). Tous les organismes Crenarcheota sont thermophiles à hyperthermophiles et beaucoup sont acidophiles (Offre et al., 2013). Les Euryarcheota sont plus divers et comprennent entre autres des méthanogènes, des halophiles et des réducteurs de souffre. Aujourd'hui encore, la plupart des souches cultivées et caractérisées appartiennent à ces deux phyla. Toutefois, depuis une vingtaine d'années, de nombreuses nouvelles espèces d'archées ont été découvertes et n'appartiennent ni aux Crenarchées ni aux Euryarchées. Certaines espèces peuvent être enrichies ou isolées par exemple parmi les phyla Thaumarcheota et Korarcheota (Elkins et al., 2008; Stieglmeier et al., 2014) mais beaucoup d'autres restent incultivables et ne sont connues que par leur séquences nucléiques détectées lors d'analyses métagénomiques (Spang et al., 2017). La masse de nouvelles espèces putatives d'archées découvertes ces dernières années a chamboulé la phylogénie des archées qu'on pensait pourtant résolue (Figure 1) (Brochier-Armanet et al., 2011).

Les Crenarcheota et Euryarcheota ne sont plus que des phyla parmi d'autres (Figure 1). Pour faciliter la compréhension de la diversité archéenne, des superphyla ont été définis. Le superphylum TACK regroupe les Thaumarcheota, les Aigarcheota non cultivées, les Korarchaeota, les Crenarcheota et les Bathyarcheota non cultivés (Adam et al., 2017; Guy and Ettema, 2011). Le superphylum Asgard regroupe les phyla Heimdallarcheota, Thorarcheota, Odinarcheota et Lokiarcheota, tous non cultivés (Zaremba-Niedzwiedzka et al., 2017). Le superphylum DPANN est parfois proposé comme regroupant des archées au génome très réduit et à évolution rapide comme les Nanoarcheota (Rinke et al., 2013; Spang et al., 2017). L'existence du superphylum DPANN et son positionnement dans la phylogénie sont débattus et pourraient correspondre à un artefact de longues de branches de la reconstruction phylogénétique (Adam et al., 2017). Un autre aspect débattu de la phylogénie des archées est le placement de la racine (Adam et al., 2017; Raymann et al., 2015; Spang et al., 2017). Suivant ce positionnement, la monophylie du superphylum TACK (Cunha et al., 2017) ou des Euryarchées (Figure 1) peut être remise en cause. Ceci conduit notamment à la définition de deux groupes au sein des Euryarchées : le groupe I qui comprend entre autres les Thermococcales, Methanobacteriales, Methanococcales et Methanopyri et le groupe II qui comprend entre autres les Thermoplasmatales, Archaeoglobales, Methanosarcinales, Methanomicrobiales et Methanobacteriales.

Le chromosome des archées

Le chromosome des archées est circulaire et sa taille varie de 0,5 Mb à plus de 5 Mb (Kellner et al., 2018). Certaines archées sont haploïdes comme les Crenarcheota et présentent un cycle cellulaire régulé (Lindås and Bernander, 2013). D'autres archées sont polyploïdes comme les Euryarchées et leur nombre de chromosome varie en fonction du stade de croissance (Ausiannikava and Allers, 2017). *Thermococcus kodakarensis* présente par exemple 7 à 19 chromosomes par cellule (Spaans et al., 2015).

L'organisation du chromosome archéen est similaire à celle du chromosome bactérien présentée en Figure 2. Les chromosomes archéens et bactériens sont denses avec des distances intergéniques courtes et peu de gènes interrompus (Kellner et al., 2018; Koonin and Wolf, 2008). Les gènes sont couramment organisés en opérons qui permettent une coordination d'expression (Figure 2, point 7) (Koonin and Wolf, 2008). Le « GC skew » est différent entre les brins précoces et tardifs (Lopez et al., 1999) et les gènes sont souvent co-orientés avec la direction de réplication (Cossu et al., 2015) (Figure 2, point 4 et 6). Les gènes les plus exprimés sont situés à proximité de l'origine de réplication où ils

peuvent bénéficier d'un plus grand nombre de matrice pour la transcription (Figure 2, point 5) (Andersson et al., 2010). Cette organisation commune résulte vraisemblablement d'une convergence évolutive due à des contraintes génomiques similaires (circularité, absence de noyau, éléments génétiques mobiles notamment) et à une grande taille de population effective (Kellner et al., 2018; Rocha, 2008).

Des différences existent également entre les chromosomes archéens et bactériens. Certains chromosomes archéens présentent plusieurs origines de réplication (Ausiannikava and Allers, 2017). De plus, les origines de réplications sont accessoires en conditions de laboratoires pour certaines espèces et peuvent être supprimées (Figure 2, point 2). La localisation de la terminaison de réplication n'est pas connue pour les archées (Figure 2, point 1) et aucun motif archéen correspondant aux motifs Chi ou KOPS n'a jusqu'à présent été fonctionnellement caractérisé (Cortez et al., 2010; Zivanovic et al., 2009) (Figure 2, point 8). Finalement, en absence de données publiées, nous ne connaissons pour l'instant pas la structuration spatiale du chromosome archéen.

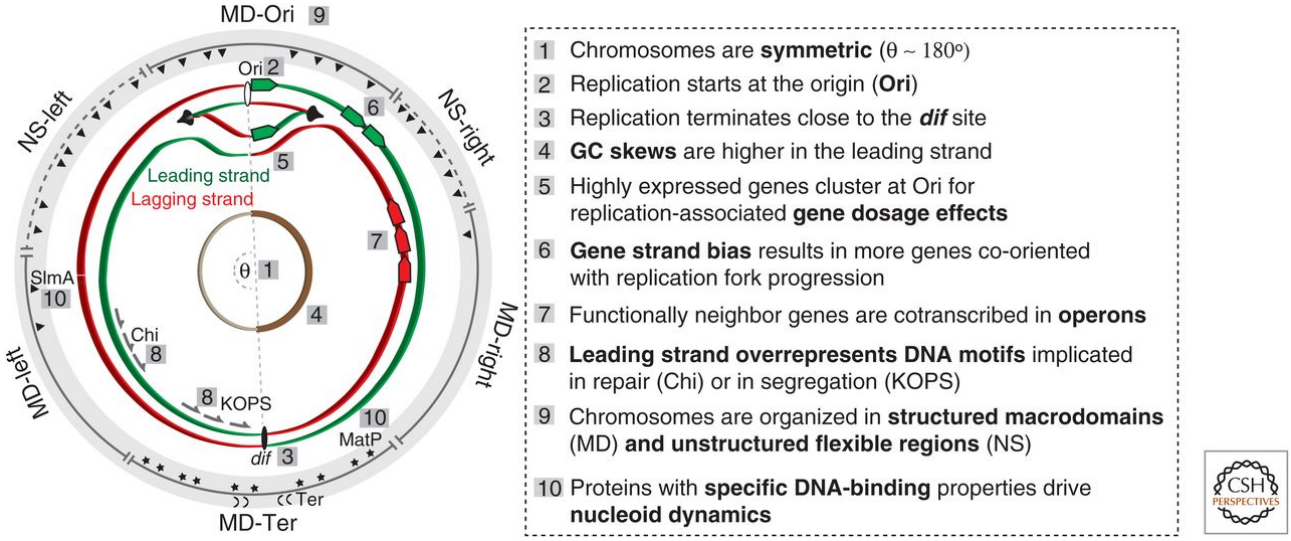


Figure 2. Eléments de l'organisation des génomes bactériens. Les éléments 1 à 8 concernent l'organisation unidimensionnelle du chromosome. Parmi eux, les points 2 et 4 à 6 sont vrais pour les archées. Les éléments 9 et 10 correspondent à la structure spatiale du chromosome qui n'a pas encore été étudiée pour les archées. Figure tirée de (Touchon and Rocha, 2016)

Morceaux choisis de diversité des archées

Nous allons maintenant brièvement détailler la diversité écologique et génomique de quelques ordres rencontrés pendant les travaux de thèse : les Euryarchées de groupe I Thermococcales et Methanococcales et les Euryarchées de groupe II Methanosarcinales et Archaeoglobales. Parmi ceux-ci, les Thermococcales, Methanococcales et Archaeoglobales sont retrouvées dans le même environnement hydrothermal océanique. Leurs caractéristiques sont résumées dans le Tableau 2.

Tableau 2. Caractéristiques de quatre ordres d'archées. D'après Brileya and Reysenbach, 2014; Oren, 2014a, 2014b, 2014c; et Schut et al., 2014.

	Thermococcales	Methanococcales	Methanosarcinales	Archaeoglobales
Classification	Euryarchées de groupe I	Euryarchées de groupe I	Euryarchées de groupe II	Euryarchées de groupe II
Chromosomes séquencés et fermés[1]	39	20	39	8
Taille du génome	Environ 1,8 à 2,1 Mb	Environ 1,7 à 1,9 Mb	2,0 à 5,8 Mb	Environ 2 Mb
Nombre de protéines prédites	Environ 1900 à 2200	Environ 1700	2000 à 4700	Environ 2000
Tolérance à l'oxygène	Anaérobie obligatoire	Anaérobie obligatoire	Anaérobie	Anaérobie obligatoire
Métabolisme	Hétérotrophe Réduction du souffre	Méthanogenèse	Methanogenèse	Variable
Température optimale	Th : environ 85°C Py : environ 95°C Pa : environ 80°C	Mc : 18°C à 55°C Mtc : environ 60°C Mcc : environ 80°C	20°C à 55°C	70 C à 88°C
Habitats majeurs	‣ Environnements hydrothermaux océaniques et marins	‣ Environnements hydrothermaux océaniques et marins ‣ Sédiments marins	‣ Sédiments variés ‣ Environnements thermaux ‣ Systèmes de traitement des déchets ‣ Systèmes digestifs animaux	‣ Environnements hydrothermaux océaniques et marins

[1]Publiés au NCBI au 1 mai 2019
Abréviations : Th : *Thermococcus*, Py : *Pyrococcus*, Pa : *Paleococcus*, Mc : *Methanococcus*, Mtc : *Methanothermococcus*, Mcc : Methanocaldococcaceae

Les Thermococcales

L'ordre des Thermococcales comprend une famille et trois genres : *Thermococcus*, *Pyrococcus* et *Paleococcus* (Schut et al., 2014). Les espèces sont détectées et isolées principalement au niveau des champs hydrothermaux océaniques (cheminées et sédiments) (Canganella et al., 1998; Gorlas et al., 2013, 2014; Lepage et al., 2004). Elles sont aussi présentes dans des environnements hydrothermaux marins et terrestres (Atomi et al., 2004; Fiala and Stetter, 1986; González et al., 1999). Toutes les Thermococcales sont anaérobies et hyperthermophiles mais les espèces du genre *Pyrococcus* présentent des températures de croissance plus élevées, parfois supérieures à 100°C (Callac et al., 2016). Le chromosome a généralement une taille légèrement inférieure à 2 Mb et un peu plus de 2000 gènes codants pour des protéines prédits.

Les Thermococcales font partie des Euryarchées de groupe I (Figure 1 page 9). Leurs plus proches parents cultivables sont les Methanococcales, Methanopyri et Methanobacteria. Récemment, des chromosomes encore plus proches ont été reconstitués à partir de métagénomes. Ils sont classés dans quatre groupes : les Theionarchaea, les Methanofastidiosa, les Hadesarchaea et les Persophonarchaea (Adam et al., 2017). Les Theinoarchaea, Methanofastidiosa et Thermococcales forment la superclasse monophylétique des Acherontia (Figure 1 page 9). Les seules informations disponibles sur ces nouveaux groupes proviennent de leurs génomes non complets. On ne connait notamment pas leurs éléments génétiques mobiles.

Les Thermococcales sont utilisées en laboratoire comme organismes modèles d'étude pour plusieurs raisons. D'abord, leur culture est relativement facile à mettre en place et leur temps de génération est court (environ 1 heure) (Dalmasso et al., 2016; Gorlas et al., 2014; Hileman and Santangelo, 2012). Leur petit génome limite la complexité des fonctions étudiées. De plus, les génomes publiés sont nombreux et permettent des études génomiques comparatives (par exemple : Cossu et al., 2015). Finalement, des outils sont disponibles pour la manipulation génétique de 4 souches : *Thermococcus kodakarensis*, *Thermococcus barophilus, Thermococcus onnurineus* et *Pyrococcus furiosus* (Catchpole et al., 2018; Farkas et al., 2012; Hileman and Santangelo, 2012; Kim et al., 2015; Thiel et al., 2014). L'étude des Thermococcales a ainsi participé à une meilleure compréhension de l'hyperthermophilie et de la biologie des archées (réplication, réparation, transcription, métabolisme) (Atomi et al., 2012; Leigh et al., 2011). Les découvertes d'enzymes thermostables des Thermococcales ont aussi eu des retombées biotechnologiques importantes comme pour l'ADN polymérase Pfu de *Pyrococcus furiosus* (Straub et al., 2018).

Les environnements hydrothermaux océaniques

Les environnements hydrothermaux océaniques sont des écosystèmes productifs et riches basés sur la chimiosynthèse, à une profondeur de 800 à 3500 m où la lumière solaire ne pénètre pas (Prieur, 1997). Ils se trouvent dans des zones tectoniques actives. L'eau de mer pénètre en profondeur dans le plancher océanique où elle est chauffée à proximité de chambres magmatiques et lessive ainsi les minéraux présents (Flores and Reysenbach, 2011). Le fluide hydrothermal acide, réduit et enrichi en métaux lourds rejaillit ensuite et se mélange à l'eau froide de l'océan. Ce mélange entraine la précipitation des minéraux du fluide et la constitution de parois minérales des cheminées hydrothermales. Ces parois constituent l'habitat principal des Thermococcales et sont caractérisées par des gradients physico-chimiques brutaux (température, pH et composition minérale et chimique). Le champ hydrothermal dans son ensemble présente une température localement élevée et une forte pression hydrostatique due à la profondeur.

Les Methanococcales

L'ordre des Methanococcales est composé des deux familles Methanocaldococcaceae et Methanococcaceae qui comprennent chacune deux genres : *Methanocaldococcus* et *Methanotorris* d'un côté, *Methanococcus* et *Methanothermococcus* de l'autre côté (Oren, 2014a, 2014b). Les espèces de Methanocaldococcaceae sont détectées et isolées dans des cheminées hydrothermales marines et sédiments environnants. Ce sont des méthanogènes, anaérobies obligatoires et hyperthermophiles (Tableau 2). La taille du chromosome varie de 1,2 Mb pour *Methanocaldococcus villosus* à 1,85 Mb pour *Methanotorris igneus* avec 1400 à 2000 gènes codants pour des protéines prédits. Les espèces de Methanococcaceae sont détectées et isolées dans des sédiments marins et des environnements marins géothermaux ou hydrothermaux. Ce sont des méthanogènes, mésophiles à thermophiles et anaérobies obligatoires. La taille du chromosome est d'environ 1,7 à 1,9 Mb avec 1700 gènes codants pour des protéines prédits.

Les Methanosarcinales

L'ordre Methanosarcinales comprend une famille et neuf genres : *Methanosarcina*, *Halomethanococcus*, *Methanimicrococcus*, *Methanococcoides*, *Methanohalobium*, *Methanohalophilus*, *Methanolobus*, *Methanomethylovorans* et *Methanosalsum* (Oren, 2014c). Les espèces sont détectées et isolées dans des environnements variées : sédiments d'eau douce, marins ou hypersalins, environnements thermaux, systèmes de traitement des déchets anaérobies ou système digestif des animaux (Tableau 2). Ils sont anaérobies et mésophiles ou thermophiles modérées. Toutes les espèces sont méthanogènes mais utilisent des substrats différents. La taille du chromosome varie entre 2,0 Mb pour *Methanohalophilus mahii* et 5,8 Mb pour *Methanosarcina acetivorans* avec 2000 et 4700 gènes codants pour des protéines prédits, respectivement.

Plus en détail, les espèces du genre *Methanosarcina* sont trouvées dans des sédiments d'eau douce et marins, des systèmes de traitement des eaux usées et des systèmes digestifs animaux (Oren, 2014c). Ils sont anaérobies obligatoires et mésophiles ou légèrement thermophiles. Le chromosome a une taille entre 3 et 5 Mb dans l'ensemble avec environ 3000 à 3500 gènes codants pour des protéines prédits (Lambie et al., 2015). Les espèces du genre *Methanococcoides* sont trouvées dans des sédiments variés (Guan et al., 2014). Ils sont anaérobies obligatoires et mésophiles ou légèrement psychrophiles (Oren, 2014c). Le chromosome a une taille d'environ 2,0 à 2,5 Mb avec 2000 à 2500 gènes codants pour des protéines prédits (d'après la base de données de génomes du NCBI https://www.ncbi.nlm.nih.gov/genome/50813).

Les Archaeoglobales

L'ordre Archaeoglobales comprend une famille et trois genres : *Archaeoglobus*, *Geoglobus* et *Ferroglobus* (Brileya and Reysenbach, 2014). Les espèces sont détectées et isolées principalement dans des systèmes hydrothermaux marins et océaniques (Tableau 2). Ils sont thermophiles ou hyperthermophiles et anaérobies obligatoires. Ils présentent des métabolismes très divers. Le chromosome a une taille d'environ 2 Mb (1,6 Mb pour *Archaeoglobus profundus* à 2,2 Mb pour *Ferroglobus placidus*) avec 2000 gènes codant pour des protéines prédits.

Chapitre 2. Les éléments génétiques mobiles

Diversité des éléments génétiques mobiles dans l'arbre du vivant

Dans leur définition la plus générale, les éléments génétiques mobiles (EGM, ou MGE en anglais) sont des séquences d'acides nucléiques qui « changent leur position de manière répétée au sein d'un génome cellulaire ou entre génomes cellulaires » (Shapiro, 1983). Les éléments génétiques mobiles sont présents sous forme extrachromosomale libre et/ou sous forme intégrée dans le chromosome hôte. Ils correspondent aux plasmides, aux virus et aux éléments transposables et contribuent à la variabilité intraspécifique. L'ensemble des éléments génétiques mobiles présents dans un groupe d'individus donné (population, espèce, clade etc.) est appelé mobilome.

Deux types de mobilités intercellulaire et intracellulaire sont inhérents à la définition d'un élément génétique mobile. La mobilité intercellulaire correspond au transfert de l'élément génétique mobile d'une cellule à une autre. Le transfert met en œuvre des mécanismes communs aux trois domaines du vivant ou spécifiques à certaines espèces (Wagner et al., 2017). Les mécanismes les plus communs et les mieux décrits sont la transformation naturelle, la conjugaison et la transduction (Frost et al., 2005). Ils sont codés par le chromosome ou par l'élément génétique mobile et sont spécifiques ou non. La mobilité intracellulaire correspond au changement de localisation de l'élément génétique mobile à l'intérieur d'un génome cellulaire. Pour cela, les éléments génétiques mobiles utilisent différents mécanismes de recombinaison qui seront détaillés dans le Chapitre 3. Certains éléments génétiques mobiles ne présentent que l'une ou l'autre des mobilités évoquées ci-avant. Par exemple, les virus lytiques ne s'intègrent pas dans le chromosome et ne changent donc pas de localisation intragénomique. A l'inverse, la majorité des éléments transposables ne codent pas pour une fonction de transfert intercellulaire.

Les plasmides

Les plasmides sont des molécules d'ADN extrachromosomales qui se répliquent séparément du chromosome. Ils utilisent toutefois certaines protéines de réplication codées par le chromosome, par exemple l'ADN polymérase. Les plasmides ne codent pas de fonctions essentielles à la survie de la cellule, à l'inverse du chromosome. Les plasmides ont aussi une taille plus petite que le chromosome de l'hôte. La majorité des plasmides font quelques de dizaines de kilobases (Figure 3) même si les plus grands peuvent dépasser 2 Mb (Medema et al., 2010; Salanoubat et al., 2002). Les plasmides sont typiquement circulaires mais il existe aussi des plasmides linéaires, par exemple chez les Streptomyces ou Borrelia (Hinnebusch and Tilly, 1993). Les plasmides sont présents chez les archées, les bactéries et les eucaryotes.

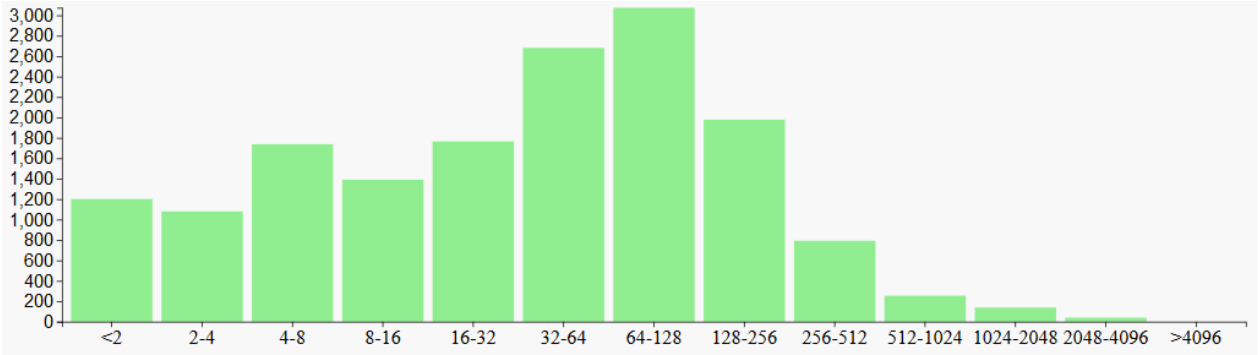


Figure 3. Distribution de la taille (kb) des plasmides de la base de donnée WASPS (https://archaea.i2bc.paris-saclay.fr/wasps/Dashboard.aspx). La base de donnée WASPS contient tous les plasmides naturels séquencés archéens, bactériens et eucaryotes.

La distinction entre plasmide et chromosome telle que définie ci-dessus peut paraître nette : le chromosome est essentiel et le plasmide ne l'est pas. Pourtant, cette différence est difficile à détecter dans certains cas. Certains plasmides confèrent un fort avantage sélectif à leur hôte qui peut être confondu avec une fonction essentielle. Par exemple, un plasmide peut porter un gène de résistance aux antibiotiques. En présence d'un antibiotique, ce gène devient essentiel à la croissance ou à la survie et le plasmide pourrait être confondu avec un chromosome. De même, pour les plasmides qui portent des systèmes toxine-antitoxine, la cellule hôte meurt lorsqu'elle perd le plasmide (Harms et al., 2018). Le plasmide est donc essentiel à la survie et pourrait être confondu avec un chromosome. La non-essentialité du plasmide est surtout à considérer du point du vue des fonctions de ménage (machineries de transcription et traduction notamment) (Bentley and Parkhill, 2004). Les plasmides ne codent pas pour des gènes de ménage ou alors en redondance de la copie du chromosome.

La transmission d'un plasmide s'accomplit de manière horizontale ou verticale. La transmission horizontale correspond à la mobilité intercellulaire des plasmides et permet de recruter de nouveaux hôtes, de la même espèce ou non. La transmission verticale a lieu lors de la division cellulaire où le plasmide est réparti entre les cellules filles (ségrégation). Pour optimiser la ségrégation et assurer son maintien, le plasmide peut employer une gamme de stratégies entre deux extrêmes. D'un côté de la gamme, une réplication abondante produit un grand nombre de copies du plasmide et assure une transmission aléatoire d'au moins une copie par cellule fille. De l'autre côté de la gamme, le plasmide présente un faible nombre de copie qui est plus soutenable pour la cellule hôte. Le plasmide coordonne alors sa réplication avec la division cellulaire et code un système de partition efficace (Pinto et al., 2012).

Historiquement, les plasmides ont été principalement étudiés pour leurs possibles applications en biologie moléculaire (ingénierie génétique). Les plasmides sont petits ce qui les rend facile à manipuler ; ils ont une réplication autonome ; ils sont stables et utilisables pour un très grand spectre d'organismes. Ces caractéristiques en ont fait de très bons candidats pour l'introduction d'ADN modifié dans un organisme et ils ont été utilisés dans de très nombreux systèmes de biologie moléculaire (Atomi et al., 2012; Hileman and Santangelo, 2012)

L'aspect majeur de la définition du plasmide est son autonomie de réplication. Les plasmides circulaires utilisent différents mécanismes de réplication dont le mécanisme thêta (θ), le mécanisme de déplacement de brin et le mécanisme de cercle roulant (rolling-circle RC) (del Solar et al., 1998). Ces trois mécanismes sont détaillés ci-dessous. Pour les plasmides linéaires, les extrémités sont répliquées grâce à différents types de télomères (Chen, 2007; Kobryn, 2007). Dans tous les cas, les protéines plasmidiques majeures de réplication ont été nommées Rep indépendamment de leur activité.

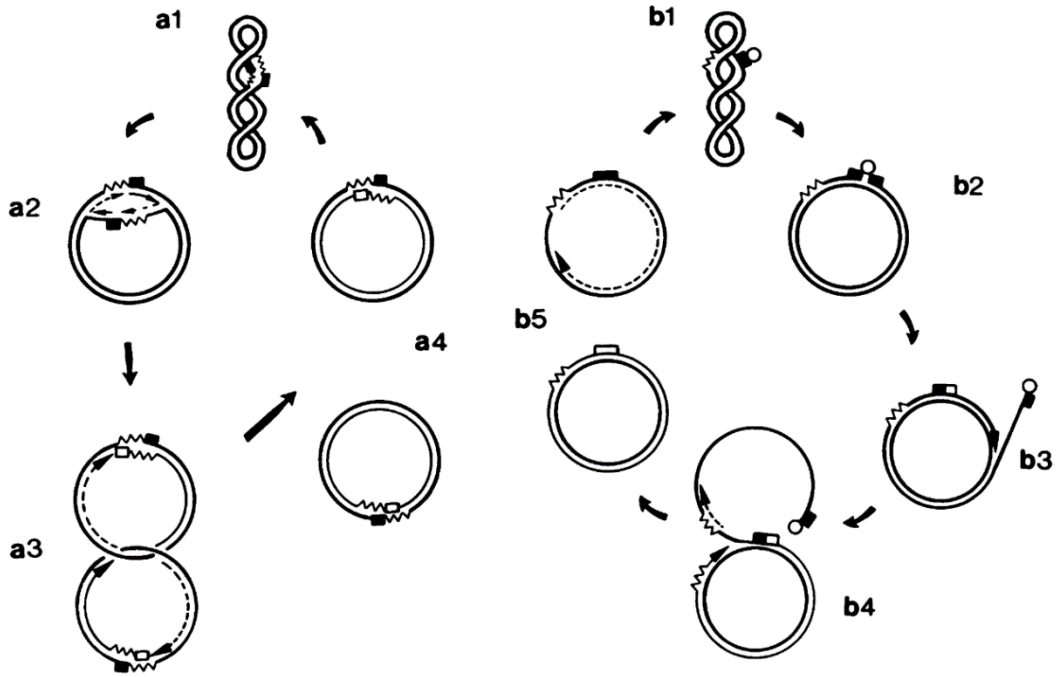


Figure 4. Les plasmides emploient différents mécanismes de réplication. Un modèle simplifié est présenté pour le mécanisme thêta (a1 à a4) et le mécanisme de cercle roulant (b1 à b4). L'origine de réplication du brin précoce (■) est utilisée sous forme double brin pour les deux mécanismes. □ représente l'origine de réplication du brin précoce de la molécule néo-synthétisée. L'origine de réplication du brin tardif (﹏) est utilisée sous forme double brin pour le mécanisme thêta et sous forme simple brin pour le mécanisme de cercle roulant. ○ représente la protéine d'initiation Rep. Les brins matrices sont représentés par des traits épais, les brins précoces par des traits fins et les brins tardifs par des pointillés. Les flèches indiquent la direction de polymérisation. Figure tirée de Viret et al., 1991

Le mécanisme de réplication plasmidique thêta est similaire au mécanisme de réplication bidirectionnelle du chromosome bactérien circulaire (Lilly and Camps, 2015). Pour la mise en place de ce mécanisme, le plasmide présente au minimum une origine de réplication *ori* (Figure 4). Le plasmide peut aussi coder pour une protéine d'initiation Rep, une primase ou une hélicase. Les autres composants du réplisome sont fournis par l'hôte. La réplication est initiée par la dénaturation de l'ADN au niveau de l'origine de réplication. Le réplisome de l'hôte est ensuite recruté et parfois agrémenté de protéines plasmidiques. La synthèse procède par extension 3'OH d'une amorce. Le brin précoce est synthétisé de manière continue à partir d'une amorce initiale. Le brin tardif est synthétisé de manière discontinue à partir de multiples amorces par élongation de fragments d'Okazaki. L'élongation coordonnée des brins précoce et tardif entraine la mise en place d'une boucle de réplication prenant la forme de la lettre grecque thêta (θ). Dans certains cas, la réplication est unidirectionnelle avec une seule fourche active (Lilly and Camps, 2015; Sun et al., 2006). L'initiation de la réplication peut se faire par différents mécanismes. Pour un premier mécanisme, une protéine d'initiation Rep lie

spécifiquement l'origine de réplication *ori* et entraine une courbure de l'ADN suivie par sa dénaturation. Alternativement, l'initiation est médiée par la transcription d'un ARN complémentaire à la séquence *ori*. L'hybridation ADN-ARN entraine la dénaturation de l'ADN et la molécule d'ARN est ensuite traitée pour servir d'amorce au brin précoce. L'amorce du brin précoce peut également être synthétisée par une primase codée par l'hôte ou par le plasmide.

Le mécanisme de déplacement simple brin ressemble au mécanisme thêta mais présente une synthèse continue des deux brins. Cinq éléments portés par le plasmide sont nécessaires : un gène codant pour une hélicase RepA, un gène codant pour une primase RepB et un gène codant pour une protéine d'initiation RepC et deux origines de réplication simple brin (Lilly and Camps, 2015). Dans un premier temps, la protéine d'initiation RepC se lie aux deux origines de réplication et entraine une courbure suivie par la dénaturation de l'ADN. La dénaturation est facilitée par l'hélicase RepA. La primase RepB reconnait ensuite les origines de réplication sous forme simple-brin et synthétise des amorces. La synthèse de chaque brin procède alors à partir de chaque amorce de manière continue.

Le mécanisme de cercle roulant, historiquement nommé mécanisme sigma, met en œuvre trois éléments portés par le plasmide : une origine double brin *dso*, une origine simple brin *sso* et un gène codant pour une protéine d'initiation Rep (Khan, 2005). La protéine Rep présente deux activités : une activité de liaison site-spécifique à l'origine double brun et une activité de coupure simple-brin. La réplication se divise en deux phases distinctes (Figure 4). La première phase correspond à la synthèse du premier brin et se déroule d'après le mécanisme suivant. La protéine Rep se lie à l'origine double brin *dso* et catalyse un clivage simple-brin qui libère une extrémité 3'OH du brin complémentaire. La machinerie de réplication de l'hôte est ensuite recrutée et l'extension est amorcée à partir de l'extrémité 3'OH libre. L'élongation entraine un déplacement du brin complémentaire et continue jusqu'après le site *dso*. Le brin complémentaire est alors libéré sous forme de monomère simple brin circulaire par un clivage catalysé par la protéine Rep. Le brin matrice et le brin néo-synthétisé forment une nouvelle molécule circulaire double-brin et surenroulée grâce à différents réactions de clivage et ligation. Lors de la deuxième phase, l'origine simple brin *sso* du monomère simple-brin circulaire est reconnue par les protéines hôtes qui catalysent la synthèse du brin complémentaire.

Ces mécanismes de réplication et leurs protéines ont été caractérisés en détail pour quelques plasmides et extrapolés pour les protéines homologues. En revanche, l'activité exacte de certaines protéines Rep divergentes est inconnue. De plus, certains plasmides ne présentent pas d'homologues de protéines Rep et nous ne connaissons pas leur mécanisme de réplication.

Les virus

Même si l'identité des virus est clairement établie par la communauté scientifique, la formulation exacte de leur définition est sujette à débat. D'un point de vue centré sur la cellule, le virus est un parasite obligatoire et infectieux qui requiert la machinerie cellulaire de l'hôte pour se reproduire (Ofir and Sorek, 2018). D'un point de vue centré sur le virus, le virus peut être vu comme une particule (ou un organisme) codant pour une capside et qui utilise des organismes cellulaires pour accomplir son cycle de vie (Raoult and Forterre, 2008) ou bien comme une information génétique qui se transmet d'une génération à l'autre sans continuité physique (désintégration et reconstitution) (Rohwer and Barott, 2013; Wolkowicz and Schaechter, 2008). La difficulté à définir les virus provient en partie de la versatilité de leur cycle de vie. Il inclut deux formes obligatoires. Premièrement, le virion ou particule virale est la forme extracellulaire infectieuse. Le virion est composé au minimum d'un acide nucléique enrobé d'une capside protéique. L'acide nucléique peut être de l'ADN ou de l'ARN, sous forme simple brin ou sous forme double brin. Additionnellement, le virion peut contenir des enzymes ou une enveloppe lipidique. Deuxièmement, le virus existe sous une forme intracellulaire qu'on nomme parfois usine virale. Cette forme permet la réplication du virus en parasitant la machinerie de l'hôte. Enfin, le virus présente parfois une troisième forme facultative appelée provirus. Il correspond au génome du virus intégré dans le chromosome de l'hôte et dont l'expression est souvent silencieuse. La classification des virus se base sur trois critères : identité de l'acide nucléique, présence d'une enveloppe lipidique dans le virion et morphologie du virion (King et al., 2011). Les virus infectent les cellules des trois domaines du vivant et sont appelés phages lorsqu'ils infectent une bactérie.

Les virus utilisent un continuum de stratégies de vie dont les extrémités correspondent aux cycles lytique et lysogénique (Figure 5). Le cycle lytique correspond à un cycle infectieux basique en cinq étapes. (1) Lors de la première étape d'attachement, le virion se lie de manière spécifique à un récepteur de la cellule hôte. (2) Le virion délivre son acide nucléique viral dans la cellule hôte. (3) Le virus détourne la machinerie cellulaire pour l'expression des gènes viraux et la réplication du génome viral. (4) Le virion est assemblé. (5) Les virions sont relâchés dans le milieu extérieur entrainant la mort de la cellule hôte, souvent par lyse. En adoptant le cycle lytique, le virus se reproduit de manière horizontale. A l'autre opposé du continuum, le cycle lysogénique correspond à un état de latence entre les étapes (2) et (3) du cycle lytique. Le virus est alors seulement sous la forme d'un acide nucléique et se reproduit de manière verticale avec la cellule hôte. Nous aborderons ce cycle plus en détail dans une prochaine partie (page 25). Entre ces deux extrêmes, il existe une variété de stratégies de vie qui sont parfois appelées infections chroniques ou « état porteur » (carrier state en anglais), où le génome viral est plus ou moins exprimé dans la cellule hôte et des virions sont plus ou moins produits, mais sans aboutir à la mort de la cellule hôte.

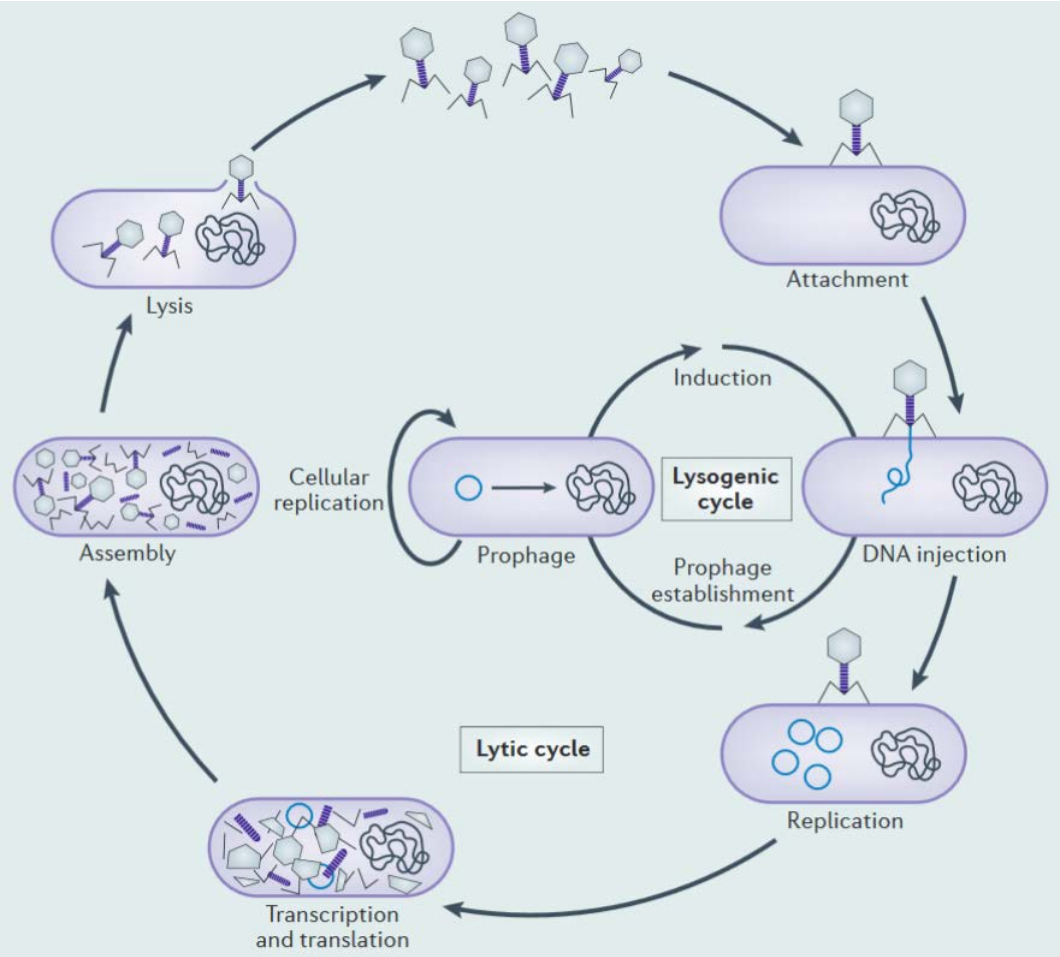


Figure 5. Représentation schématique des cycles lytique et lysogénique. Figure tirée de Salmond and Fineran, 2015

D'un point de vue historique, la recherche sur les virus, et particulièrement les phages, a apporté de grandes avancées, en biologie moléculaire notamment (Salmond and Fineran, 2015). Les virus étaient utilisés comme des systèmes modèles simplifiés du fonctionnement moléculaire du vivant. L'étude des phages a permis de démontrer la nature aléatoire des mutations (Luria and Delbrück, 1943), de déterminer la structure des gènes (Benzer, 1955), de démontrer de l'usage de triplets dans le code génétique (Crick et al., 1961) ou d'identifier des systèmes de régulation de l'expression des gènes (Hain et al., 1992; Reiter et al., 1987). L'étude des virus a aussi permis le développement d'outils pour la biologie moléculaire comme les enzymes de restriction, les ligases ou le système d'expression T7 (Salmond and Fineran, 2015).

Les éléments transposables

Les éléments transposables (TE) sont des séquences d'ADN qui sont capables de changer de position au sein d'un génome (Bourque et al., 2018). Les éléments transposables ne présentent généralement pas de spécificité stricte d'intégration et peuvent s'intégrer à de nombreux loci du génome. Les éléments transposables définis ainsi sont présents dans les trois domaines du vivant. Ils ne sont pourtant pas étudiés comme un ensemble cohérent et une séparation claire est présente avec l'étude des éléments transposables des archées et bactéries d'un côté et de ceux des eucaryotes de l'autre. La classification et la nomenclature sont notamment différentes entre les deux (Bourque et al., 2018; Roberts et al., 2008). Cette séparation est partiellement basée sur une similarité entre les éléments transposables de bactéries et d'archées.

La plupart des éléments transposables présentent une séquence répétée en orientation directe à leurs deux extrémités et qui provient de la duplication du site d'intégration. Les extrémités comprennent parfois également des répétitions inversées reconnues lors de la transposition. Plusieurs critères permettent de classer les éléments transposables dont leur autonomie de transposition. Certains sont autonomes et codent le mécanisme responsable de leur tranposition. D'autres sont non-autonomes et dépendent d'enzymes codées par d'autres éléments transposables Les éléments transposables peuvent également être distingués par l'acide nucléique intermédiaire de transposition. Les rétrotransposons (classe I) utilisent un intermédiaire ARN et les transposons à ADN (classe II) utilisent un intermédiaire ADN. D'autres distinctions existent sur la base de la structure génétique de l'élément, du mécanisme d'insertion, de la stratégie de réplication ou des protéines encodées.

Les rétrotransposons sont très fréquents chez les eucaryotes (International Human Genome Sequencing Consortium, 2001), rares chez les bactéries (Darmon and Leach, 2014) et non détectés chez des archées. Leur mécanisme de réplication fait intervenir la reverse-transcription d'un ARN complémentaire en ADN qui est intégré dans un nouveau site chromosomique lors de la reverse-transcription ou après. Ce mécanisme aboutit à la création d'une nouvelle copie du rétrotransposon à un nouveau locus chromosomique. On parle de transposition « copier-coller ». Les rétrotransposons eucaryotes sont majoritairement répartis entre deux sous-classes : les rétrotransposons à LTR (longue répétition terminale ; Long Terminal Repeat en anglais) apparentés aux rétrovirus et les rétrotransposons non-LTR.

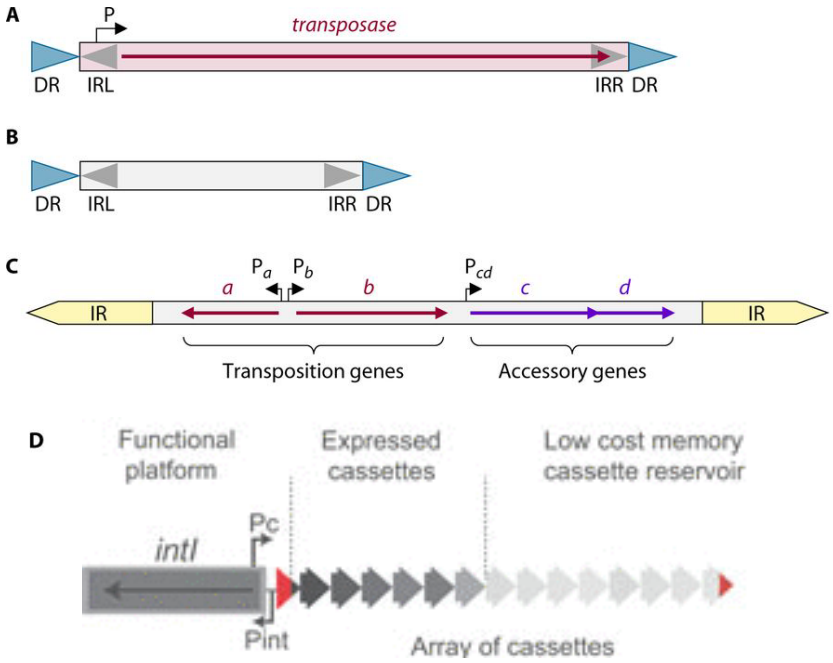


Figure 6. Organisation schématique des plusieurs éléments transposables de bactéries. A. Elément IS. B. Elément MITE. C. Transposon. D. Intégron. Le triangle rouge de gauche correspond au site attI. Les promoteurs P sont indiqués par des flèches noires. Les gènes sont indiqués par des flèches de couleurs. . DR : répétition directe du site chromosomique d'intégration. IRL et IRR : répétition inversée gauche et droite. IR : répétition inversée. Figure tirée de Darmon and Leach, 2014 (A à C) et de Escudero et al., 2015 (D)

Les transposons à ADN sont présents chez les archées, les bactéries et les eucaryotes. Ils présentent deux grands types de transposition. La transposition non-réplicative ne change en principe pas le nombre de copies du transposon. On parle de transposition « couper-coller ». La transposition réplicative crée une nouvelle copie du transposon (transposition « copier-coller »). Pour les

eucaryotes, les transposons à ADN sont répartis en plusieurs catégories suivant leur mécanisme de transposition. Par exemple, les transposons couper-coller utilisent une transposase à motif DDE (Feschotte and Pritham, 2007), les Helitrons utilisent un mécanisme de transposition similaire au cercle roulant (Grabundzija et al., 2016) et les Cryptons utilisent probablement un intermédiaire de transposition circulaire et une recombinase à tyrosine (Poulter and Butler, 2015). Pour les archées et les bactéries, les transposons à ADN sont répartis en 4 groupes principaux. Les trois premiers groupes ont une structure similaire mais varient en taille (Darmon and Leach, 2014). Le premier groupe correspond aux transposons les plus petits : les MITE (élément transposable miniature à répétition inversée, Miniature Inverted Repeat Transposable Element en anglais) (Figure 6.B). Ce sont des séquences riches en AT, contenant une séquence TIR (répétition inversée terminale) à chaque extrémité et bordées par une duplication du site cible (TSD pour Terminal Site Duplication en anglais ou DR pour Direct Repeat en anglais). Ils sont non-autonomes et peuvent être reconnus par une transposase codée par un autre élément. Ils sont présents dans les trois domaines du vivant (Yang et al., 2009). Le deuxième groupe correspond aux éléments IS (séquence d'insertion, Insertion Sequence en anglais) (Figure 6.A). Comme les MITE, les IS contiennent une séquence TIR et DR aux extrémités mais ils codent additionnellement une transposase. Le troisième groupe correspond aux transposons (Tn) (Figure 6.C). Ils sont plus longs que les IS et codent des gènes accessoires. Enfin, le quatrième groupe correspond aux intégrons (Escudero et al., 2015) (Figure 6.D). Ils sont composés de deux parties : (1) une plateforme stable qui inclut un gène *intI* codant une recombinase à tyrosine avec son promoteur Pint, un site de recombinaison attI et un promoteur Pc ; (2) une série variable de cassettes qui contiennent chacune un site de recombinaison attC et un gène sans promoteur qui peut être exprimé à partir du promoteur Pc. Un autre type de transposon à ADN a récemment été découvert chez les archées et les bactéries : les casposons qui utilisent l'enzyme casposase pour leur transposition (Béguin et al., 2016; Hickman and Dyda, 2015a; Krupovic et al., 2014a). La casposase est homologue à l'endonucléase Cas1 des systèmes CRISPR-Cas9 et pourrait être à l'origine de ces systèmes (Krupovic et al., 2017).

Les éléments transposables sont portés à la fois par le chromosome et par d'autres éléments génétiques mobiles (plasmides et virus) (Escudero et al., 2015; Krupovic et al., 2019) qui assurent leur dispersion entre les cellules (Forterre et al., 2014; Peters and Craig, 2001). Les éléments transposables contribuent à plusieurs phénomènes biologiques d'intérêt pour la recherche. Les transposons et les intégrons bactériens sont par exemple des vecteurs de dispersion des gènes de résistance aux antibiotiques (Giedraitienė et al., 2011; Partridge et al., 2018). Les éléments transposables eucaryotes sont impliqués dans le contrôle de l'expression génique (Elbarbary et al., 2016), dans la régulation du développement (Erwin et al., 2014; Gifford et al., 2013) ou dans des syndromes humains (Chuong et al., 2017; Erwin et al., 2014).

Discussion de la classification des éléments génétiques mobiles

La classification historique des éléments génétiques mobiles en 3 catégories présentée ci-avant peut être remise en cause sur la base de deux critères. Premièrement, les trois définitions données ne sont pas mutuellement exclusives. Notamment, les virus se répliquent de manière autonome et rentrent donc dans la définition des plasmides, à l'exception des rétrovirus. Le virus SSV1 avait par exemple dans un premier temps été identifié comme un plasmide (Schleper et al., 1992; Yeats et al., 1982). Similairement, le génome intracellulaire du coliphage N15 est parfois qualifié de plasmide (Ravin, 2011). Deuxièmement, les éléments génétiques mobiles peuvent muter et passer d'une catégorie à l'autre. Par exemple, le plasmide pTN3 est possiblement un ancien virus défectif (Gaudin et al., 2014). De même, les rétrotransposons HERV (rétrovirus humain endogène, Human Endogenous RetroVirus en anglais) sont probablement des anciens rétrovirus intégrés dans des cellules de lignées germinales, inactivés et transmis aux générations suivantes (Bannert and Kurth, 2006). Malgré tout, cette classification permet de faire ressortir une fonction majeure de l'élément considéré : la réplication autonome pour le plasmide, la formation d'un virion pour le virus, le changement de localisation génomique pour le transposon. Les protéines impliquées dans ces fonctions (protéines Rep, protéines de capsides MCP ou transposases) sont utilisées pour délimiter et analyser des familles d'éléments génétiques mobiles (Krupovic et al., 2013; Peng et al., 2000). D'autres protéines peuvent aussi être utilisées comme les protéines de conjugaison VirB4 et VirD4 ou des protéines conservées de fonctions inconnue (Erauso et al., 2006; Gonnet et al., 2011; Soler et al., 2010). On peut néanmoins s'interroger sur la pertinence de l'utilisation de ces protéines pour définir et décrire l'histoire évolutive d'une famille d'éléments génétiques mobiles.

Certains segments du chromosome présentent toutes les caractéristiques d'un élément exogène intégré mais ne peuvent pas être placés dans l'une des catégories présentées ci-avant sur la base de leur séquence. Ce sont probablement des éléments génétiques mobiles intégrés qu'on ne peut pas identifier soit parce qu'ils encodent des protéines fonctionnelles dont on ne connait pas la fonction, soit parce qu'ils ont dégénéré après leur intégration. Ils peuvent dans ce cas avoir perdu leur capacité de mobilité. On les appelle souvent îlot génomique (GI pour Genomic Island en anglais) (Juhas et al., 2009; Langille et al., 2010).

Fonctions portées par les éléments génétiques mobiles

Les éléments génétiques mobiles sont parfois considérés purement comme des éléments génétiques égoïstes (« selfish DNA »). En effet, leur survie dépend avant tout de fonctions qu'ils encodent au détriment de la cellule hôte, que ce soit pour leur stabilité dans leur cellule hôte ou pour leur transmission à de nouvelles cellules (Tableau 3) (Carroll and Wong, 2018; Hülter et al., 2017; Rankin et al., 2011). Les plasmides et les virus assurent leur stabilité dans l'hôte par leur propre machinerie de réplication. Alternativement, ils codent parfois comme les éléments transposables pour une fonction d'intégration dans le chromosome hôte qui permet alors une réplication liée à celle du chromosome. La transmission à de nouvelles cellules peut se faire de manière horizontale ou verticale. La transmission verticale est assurée sous forme intégrée dans le chromosome, par des systèmes de partition (Baxter and Funnell, 2015) ou par des systèmes « d'addiction au plasmide » (Kroll et al., 2010). Un système d'addiction assure l'indispensabilité de l'élément génétique mobile par la mort des cellules filles qui en sont dépourvues. Il correspond par exemple au système toxine-antitoxine . La transmission horizontale est assurée par le virion dans le cas des virus ou par des pili de conjugaison dans le cas des plasmides et éléments transposables (Cabezón et al., 2015; Wozniak and Waldor, 2010). La vision des

éléments génétiques mobiles comme purement égoïstes ne permet cependant pas d'expliquer leur survie dans toutes les situations (Harrison and Brockhurst, 2012).

Tableau 3. Fonctions portées par les éléments génétiques mobiles

Nom de la fonction	Bénéfices (+) et coûts (-) pour le plasmide	Bénéfices (+) et coûts (-) pour l'hôte	Type d'élément génétique mobile concerné
Réplication	**(+)** Stabilité	**(-)** Utilisation de ressources	Plasmide et virus
Transposition	**(+)** Stabilité et transmission verticale	**(-)** Utilisation de ressources	Eléments transposable
Intégration	**(+)** Stabilité et transmission verticale	**(-)** Utilisation de ressources et perturbation du génome	Tous
Système de partition	**(+)** Transmission verticale	**(-)** Utilisation de ressources	Plasmides
Système toxine-antitoxine	**(+)** Indispensabilité	**(-)** Mort de l'hôte	Plasmides et virus
Pili de conjugaison	**(+)** Transmission horizontale	**(-)** Utilisation de ressources	Plasmides et éléments transposables
Libération des virons	**(+)** Transmission horizontale	**(-)** Mort de l'hôte	Virus
Système d'exclusion	**(+)** Compétition inter-EGM	**(+)** Résistance aux surinfections	Plasmides
Résistance à une condition extrême (antibiotique, métal lourd, pression, etc.)	**(-)** Augmente la taille du génome	**(+)** Adaptation	Plasmides et éléments transposables
Utilisation de nouvelles sources d'énergie	**(-)** Augmente la taille du génome	**(+)** Adaptation	Plasmides
Toxine / facteur de virulence	**(-)** Augmente la taille du génome	**(+)** Adaptation	Plasmides et virus
Fonction inconnue	**?**	**?**	Tous

Les éléments génétiques mobiles codent aussi pour des gènes accessoires qui sont bénéfiques pour l'hôte et favorisent leur rétention dans la cellule hôte (Frost et al., 2005; Rankin et al., 2011). Ces gènes apportent des fonctions adaptatives permettant la survie dans certains environnements ou sous certaines conditions. Par exemple, les éléments génétiques mobiles portent des gènes de résistance aux antibiotiques ou de tolérance aux métaux lourds, ou des gènes permettant d'utiliser de nouvelles sources d'énergie (Tableau 3). Certains éléments génétiques archéens sont impliqués dans la résistance aux conditions extrêmes subies par l'hôte (Li et al., 2016). Les éléments génétiques mobiles confèrent parfois une résistance à l'infection par un autre élément par des systèmes d'exclusion (Garcillán-Barcia and de la Cruz, 2008). Les gènes accessoires apportent aussi aux hôtes une meilleure compétitivité dans certains environnements. Par exemple, des gènes codant pour des toxines et des facteurs de virulence ou des facteurs de colonisation peuvent favoriser l'implantation d'un pathogène dans le corps humain (Jamet et al., 2017).

Les stratégies écologiques mises en place par les éléments génétiques mobiles sont variées et peuvent aller du mutualisme au parasitisme en passant par le commensalisme suivant la balance trouvée entre les fonctions égoïstes et les fonctions adaptatives pour l'hôte. Cette balance égoïste versus adaptatif

se reflète souvent dans la balance transmission horizontale / transmission verticale. Le parasitisme est poussé à l'extrême par certains éléments génétiques mobiles qui peuvent être vus comme des parasites d'autres éléments génétiques mobiles dont ils utilisent les mécanismes de réplication (éléments transposables non autonomes) ou de transmission (éléments intégratifs et mobilisables (Guédon et al., 2017) ou plasmides satellites de virus (Arnold et al., 1999)). Par ailleurs, la majorité des gènes des éléments génétiques mobiles archéens codent des protéines de fonctions inconnues. Si aucune fonction n'est détectée pour un élément génétique mobile, on le qualifie de cryptique. La caractérisation des fonctions inconnues améliorerait notre compréhension de la biologie et de l'écologie des éléments génétiques mobiles.

La lysogénie

La lysogénie est une des stratégies de vie mise en place par les virus (Figure 3 page 20). André Lwoff la définit en 1953 comme « le pouvoir héréditaire de produire des bactériophages » (Lwoff, 1953). Cette définition a été posée avant la découverte de l'intégration site-spécifique du bactériophage lambda (Campbell, 1963). Depuis une ambiguïté persiste. Au sens large, la lysogénie correspond à la définition de Lwoff de 1953 et un virus lysogénique peut voir son génome intégré dans le chromosome hôte ou être maintenu de manière libre et stable dans le cytoplasme (Barksdale and Arden, 1974; Stewart and Levin, 1984). Le prophage est définit par Lwoff comme « la forme par laquelle la bactérie lysogénique perpétue le pouvoir de produire des bactériophages » et correspond à la fois à la forme intégrée du virus et à la forme plasmidique libre et stablement maintenue. Le choix entre le cycle lysogénique et le cycle lytique correspond alors à un choix entre transmission verticale et transmission horizontale. Au sens stricte, la lysogénie est restreinte au mode de vie du phage lambda qui intègre son génome dans le chromosome hôte (Echols, 1972). Le prophage correspond alors à la copie intégrée du virus. L'entrée et la sortie de la lysogénie dépendent entièrement de l'enzyme catalysant l'intégration et l'excision du génome viral. Dans les deux cas, le terme lysogénie est appliqué uniquement aux virus. On pourrait argumenter pour son extension aux plasmides qui adoptent également un mode de vie caractérisé par l'intégration de leur génome au chromosome hôte et qui présentent une proximité évolutive avec les virus. Nous décrirons ici la lysogénie en tant que phase intégrative, et particulièrement ses mécanismes de contrôle identifiés pour le phage lambda et sa compréhension évolutive actuelle. L'activité enzymatique d'intégration et d'excision responsable de l'entrée et de la sortie de la lysogénie sera décrite dans le Chapitre 3.

Le cycle lysogénique typique

Le phage lysogénique par excellence est le phage λ. Son cycle lysogénique est caractérisé par trois étapes (Howard-Varona et al., 2017; Oppenheim et al., 2005). (1) La lysogénie est établie par l'intégration du génome viral, sur la base de critères de compatibilité génétique (présence du site d'intégration), de l'état physiologique de la cellule, de signaux environnementaux et de la densité de phage. (2) La lysogénie est stablement maintenue. Cette étape fait intervenir le répresseur phagique CI qui réprime l'expression du prophage. (3) La lysogénie est rompue par l'induction du cycle lytique qui se produit spontanément ou suite à un stress externe. La mise en place du système SOS inactive alors le répresseur CI. Ceci conduit à l'expression du programme génétique phagique et à l'excision du génome. Des régulations complexes interviennent aux différentes étapes et constituent un réseau génétique de la décision entre lysogénie et lyse (Oppenheim et al., 2005). Les paramètres de la décision sont connus en détails pour le phage λ mais peu étudiés pour les autres phages et virus et les variations du cycle lysogénique typique sont peu caractérisées (Ofir and Sorek, 2018).

Coûts et bénéfices évolutifs de la lysogénie

La lysogénie occasionne à la fois des bénéfices et des coûts pour l'hôte et pour le virus. Le maintien évolutif de la lysogénie suggère que les bénéfices sont supérieurs aux coûts dans certaines conditions. Du côté de l'hôte, l'intégration d'un élément exogène dans le chromosome peut perturber son organisation génomique et génétique mais permet en contrepartie l'apparition d'innovations génétiques (nouvelles fonctions, modifications de séquences existantes) et la régulation de l'expression des gènes (Howard-Varona et al., 2017). Ces innovations permanentes ou transitoires peuvent améliorer la fitness de l'hôte. A plus grande échelle, les virus lysogéniques sont des agents de transfert horizontal (Frost et al., 2005). Ils contrôlent également la croissance des populations microbiennes par leur choix entre le cycle lysogénique et le cycle lytique (Howard-Varona et al., 2017).

Pour le phage lysogénique, au moment de l'infection ou lors de la sortie de la lysogénie, le choix entre lyse et lysogénie revient à choisir la stratégie qui donne la plus grande probabilité de survie. La stratégie agressive du cycle lytique peut être avantageuse quand les hôtes potentiels sont nombreux et avec des conditions de croissance soutenant la production de virion (Gandon, 2016; Touchon et al., 2016). Au contraire, le cycle lysogénique est une stratégie prudente qui peut être avantageuse dans des conditions où le virion aurait peu de chances de trouver un hôte convenable. Les fonctions apportées par le virus augmentent alors la probabilité de survie du virus lorsqu'elles augmentent la probabilité de survie de l'hôte (Howard-Varona et al., 2017). L'intégration présente également un coût pour le virus. Le virus fonctionnel prend en effet le risque d'accumuler des mutations délétères dues à une sélection purificatrice relâchée lors de l'intégration. Il prend aussi le risque de rester bloqué dans le chromosome hôte et de disparaître avec le temps.

Les impacts de la lysogénie sont plutôt compris au niveau cellulaire individuel mais beaucoup moins pour les échelons écologiques supérieurs (population, communauté, écosystème). Les conséquences de variations dans le modèle typique de la lysogénie sont aussi peu caractérisées.

La lysogénie chez les archées

Des virus adoptant le mode de vie lysogénique ont été identifiés chez les archées parmi les Fusellovirales infectant les Sulfolobales et Thermococcales, parmi les Pleolipovirales infectant les Halobacteriales (Krupovic et al., 2018) et pour le Myovirus ϕCh1 infectant *Natrialba magadii*. Des plasmides intégratifs sont aussi présents chez les archées Sulfolobales et Thermococcales (Gaudin et al., 2014; Peng, 2008). Il a été proposé que le mode de vie lysogénique est prépondérant dans les environnements hydrothermaux océaniques habités par les Thermococcales (Lossouarn et al., 2015). Récemment, une étude exhaustive a montré que presque la totalité des chromosomes séquencés de Thaumarchées contient un élément intégré (plasmide, virus ou îlot génomique) (Krupovic et al., 2019). L'intégration semble donc être un mode de vie très utilisé par les plasmides et les virus d'archées. Malgré tout, la régulation et les enjeux évolutifs de la lysogénie ont été peu étudiés pour les archées. Notamment, la dynamique d'intégration et d'excision n'a pas été étudiée.

Diversité des éléments génétiques mobiles chez les archées

Des éléments génétiques mobiles ont été identifiés chez les archées avant même leur définition comme domaine du vivant (Simon, 1978; Torsvik and Dundas, 1974). La reconnaissance de la singularité des archées a ensuite favorisé la recherche sur leur mobilome, et particulièrement leurs virus (Snyder et al., 2015). Nous connaissons maintenant de nombreux éléments génétiques mobiles d'archées, principalement pour les Crenarchées et Euryarchées. Ils jouent un rôle majeur pour l'évolution cellulaire et présentent une grande diversité (Wang et al., 2015). La majorité de cette diversité code des fonctions non élucidées représentant un réservoir de découvertes fascinantes et d'innovations biotechnologiques.

Dans cette section, nous présenterons brièvement la diversité des éléments génétiques mobiles identifiés chez les archées. Nous décrirons plus en détail certains éléments d'importance pour ces travaux de thèse, notamment chez les Thermococcales.

Les plasmides d'archées

Près de 200 plasmides naturels d'archées ont été séquencés à ce jour d'après la base de données WASPS. Ces plasmides ont principalement été étudiés chez les Crenarchées Sulfolobales et les Euryarchées Thermococcales et Halobacteriales (Forterre et al., 2014; Wang et al., 2015). Des plasmides sont aussi détectés chez les Euryarchées Methanococcales (Tumbula et al., 1997), Methanobacteriales (Luo et al., 2001), Methanosarcinales (Lambie et al., 2015), Archaeoglobales (López-García et al., 2000) et les Crenarchées Thermoplasmatales (Yamashiro et al., 2006) et Thermoproteales (Anderson et al., 2008). Tous ces hôtes identifiés résident dans des conditions extrêmes. Le spectre d'hôtes des plasmides d'archées est restreint mais des familles de plasmides partageant des protéines homologues ont été identifiées dans des ordres distincts (Krupovic et al., 2013; Soler et al., 2010)

L'ensemble des plasmides d'archées présente une grande diversité de fonctions. Leur réplication est de type thêta ou cercle roulant. Les Sulfolobales présentent des plasmides conjugatifs de la famille pNOB8 sont présents (Erauso et al., 2006; She et al., 1998) et plasmides satellites de virus pSSVx et pSSVi qui utilisent les capsides virales comme véhicule de transmission (Arnold et al., 1999; Wang et al., 2007). Chez les Thermococcales, on trouve des plasmides qui utilisent des vésicules comme véhicule de transmission (Gaudin et al., 2014). Des vésicules sont également utilisées par le plasmide

pR1SE d'Halobactériales pour sa transmission (Erdmann et al., 2017). Elles sont recouvertes de protéines plasmidiques, similairement à un virion. Ce plasmide a ainsi été nommé plasmidion (Forterre et al., 2017) Chez les Halobactériales, on trouve également des mégaplasmides qui pourraient peut-être être qualifiés de minichromosomes puisqu'ils portent des gènes essentiels (Forterre et al., 2014; Wang et al., 2015). Certains d'entre eux codent la formation de vésicules flottantes remplies de gaz (Ng et al., 1991). La majorité des plasmides d'archées sont cependant des plasmides cryptiques pour lesquels les fonctions encodées sont peu caractérisées.

Nous allons maintenant nous intéresser plus en détail à trois plasmides des Thermococcales qui correspondent chacun au membre fondateur d'une famille de plasmide.

Le plasmide pTN3

Le plasmide pTN3 est présent chez *Thermococcus nautili* 30-1(Gaudin et al., 2014). Sa séquence de 18 kilobases comprend 34 ORFs putatives dont 10 avec une fonction putative précise (Figure 7). Notamment, le plasmide code une intégrase qui est probablement responsable de l'intégration du plasmide dans le chromosome hôte au niveau d'un gène codant pour un ARNt (Gaudin et al., 2014). L'élément TKV4 intégré dans le chromosome de *Thermococcus kodakarensis* KOD1 est très similaire à pTN3. Ces deux éléments partagent 12 protéines homologues (Gaudin et al., 2014) et forment probablement une nouvelle famille d'éléments génétiques mobiles (Forterre et al., 2014). Les deux éléments codent une résolvase de jonctions de Holliday typiquement trouvée chez les archées et pour une hélicase de type MCM. Le mécanisme de réplication est donc probablement de type thêta. Additionnellement, pTN3 et TKV4 encodent deux protéines caractéristiques des virus de la lignée PRD1-adenovirus : une protéine majeure de capside (MCP) avec un repliement « double jelly roll » et une ATPase de packaging putative (Krupovič and Bamford, 2008). Puisqu'aucune particule virale n'a été observée dans les cultures de *Thermococcus nautili* 30-1 et de *Thermococcus kodakarensis* KOD1, les deux éléments correspondent probablement à des virus défectifs.

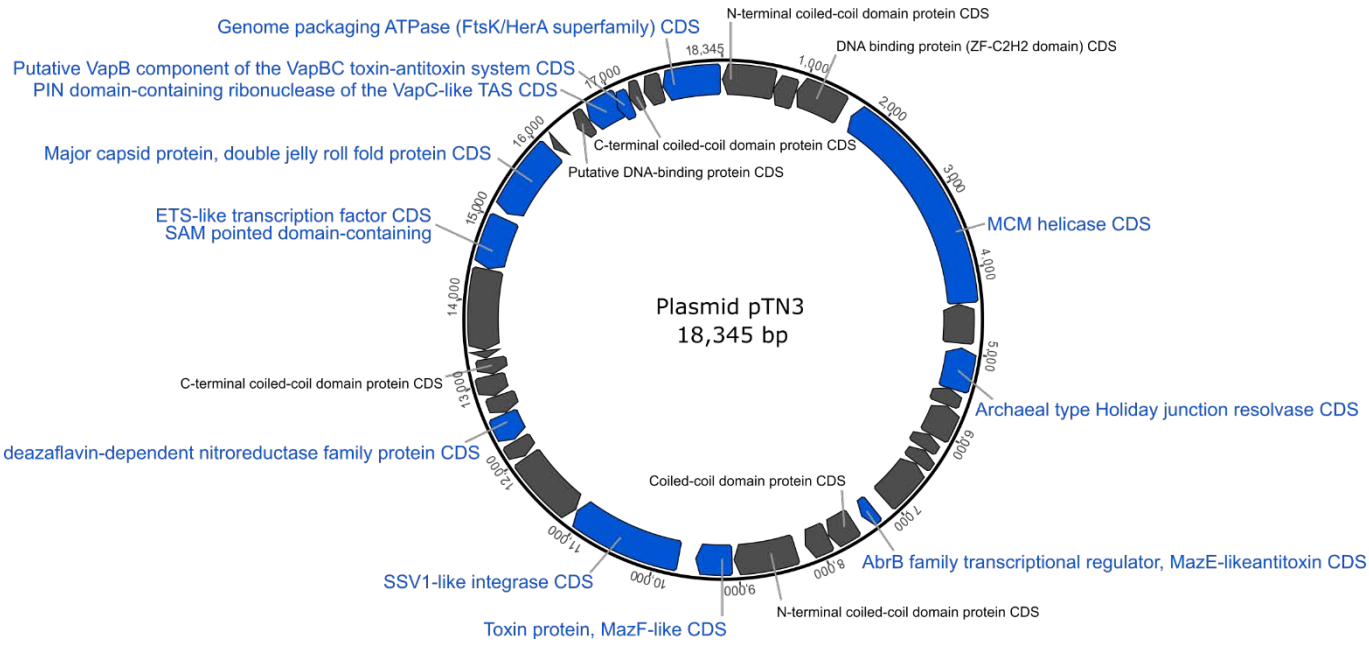


Figure 7. Carte du plasmide pTN3 de *Thermococcus nautili* 30-1 (accession NC_022527.1). Les gènes codant pour une fonction clairement identifiée sont indiqués par une flèche bleue. Les gènes codant pour une fonction inconnue sont indiqués par une flèche grise.

Le plasmide pT26-2

Le plasmide pT26-2 a été isolé à partir de la souche non publiée *Thermococcus sp.* 26-2 (Soler et al., 2010). Sa séquence de 22 kilobases présente 32 ORFs putatives dont 5 avec une fonction putative précise (Figure 8). Notamment, le plasmide code une intégrase similaire à l'intégrase du virus SSV1. Il n'a pas été déterminé si le plasmide présente une copie intégrée dans le chromosome hôte (Soler et al., 2010). En revanche, des intégrases similaires sont présentes dans les éléments intégrés TKV1, TKV2 et TKV3 de *Thermococcus kodakarensis* KOD1 et dans l'élément intégré PHV1 de *Pyrococcus horikoshii* OT3 (Figure 9). Les éléments intégrés TKV2, TKV3 et PHV1 sont d'ailleurs très similaires à pT26-2, ainsi qu'à l'élément intégré TGV1 de *Thermococcus gammatolerans* EJ3 et à 5 éléments intégrés de Methanococcales. Ces 10 éléments partagent un cluster de protéines conservé correspondant à la moitié environ du génome et ils sont très variables sur l'autre partie du génome (Soler et al., 2010). Les gènes conservés codent des protéines présentant des segments hydrophobes sans fonction prédite. Soler et al. ont proposé que ces protéines puissent former une structure de transmission (conjugaison ou virion). La famille de plasmide pourrait en conséquence présenter une dispersion atypique. Cette famille reste pour l'instant cryptique.

Dans sa partie variable, pT26-2 code une hélicase réplicative putative seulement trouvée dans pT26-2 et TKV3 (Soler et al., 2010). Cette hélicase réplicative particulière est remplacée par une hélicase de type MCM dans TKV2 (Forterre et al., 2014). Une origine putative de réplication est détectée à côté de la séquence codant pour l'hélicase. Le plasmide pT26-2 pourrait donc utiliser un mécanisme thêta de réplication.

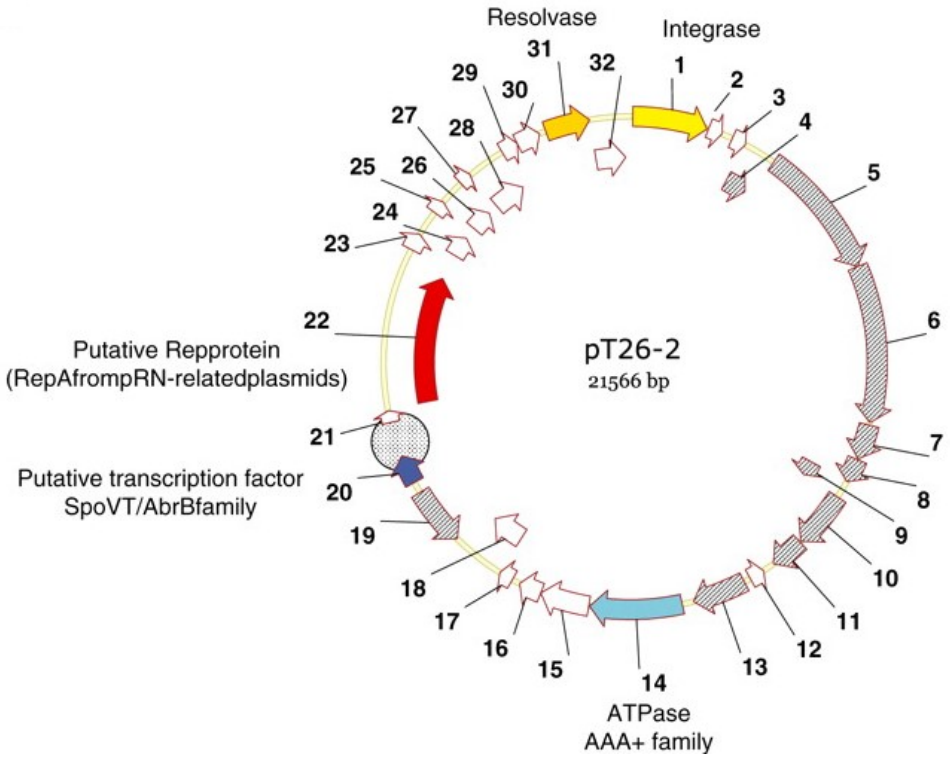


Figure 8. Carte du plasmide pT26-2. Les flèches représentent les ORFs. Chaque ORF est numérotée. Les ORFs colorées présentent une fonction putative. Les ORFs hachurées présentent des segments hydrophobiques putatifs. Le cercle représente l'origine de réplication putative. Figure tirée de Soler et al., 2010

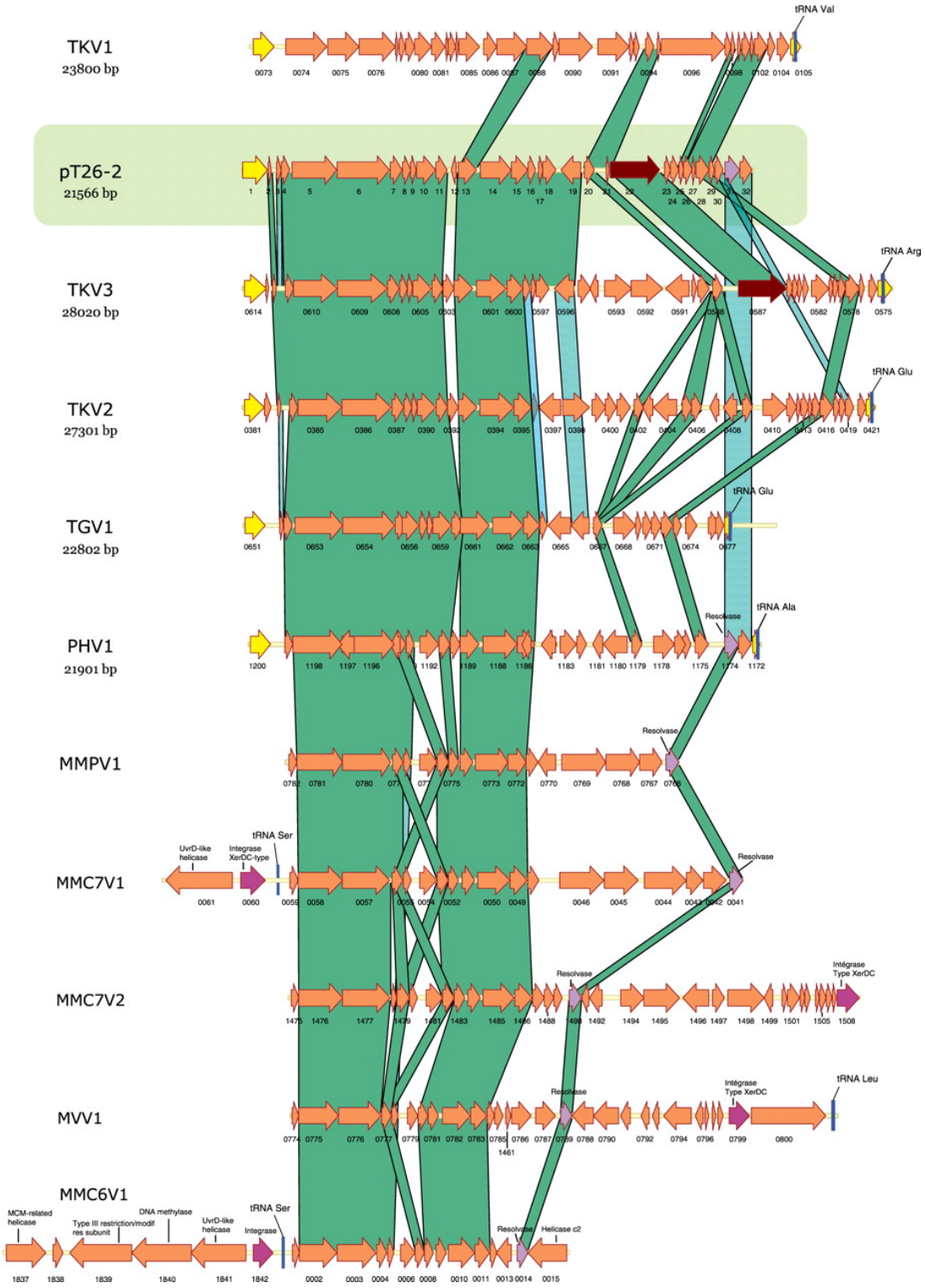


Figure 9. Comparaison du plasmide pT26-2 avec des éléments intégrés de Thermococcales et Methanococcales. Les flèches représentent les ORFs. Les flèches jaunes représentent des intégrases similaire à l'intégrase du virus SSV1 (gènes entiers ou fragments). Les flèches roses représentent des intégrases similaires aux recombinases Xer. Les flèches mauves représentent des résolvases putatives. Les flèches brunes représentent des protéines Rep putatives. Les barres bleues représentent des gènes codant pour les ARNt. Figure tirée de Soler et al., 2010

Le plasmide pAMT11

Le plasmide pAMT11 a été isolé à partir de la souche non publiée *Thermococcus sp.* AMT11 (Gonnet et al., 2011). Sa séquence de 21 kilobases comprend 30 ORFs putatives dont 6 avec une fonction putative précise. pAMT11 est très similaire à l'élément intégré TKV1 de *Thermococcus kodakarensis* KOD1 (Gonnet et al., 2011) mais ne s'intègre pas dans le chromosome de son hôte puisqu'il ne code pas pour une intégrase. Les deux éléments pAMT11 et TKV1 partagent un cluster conservé de 15 protéines homologues dont certaines sont associées à des fonctions virales. (i) Une protéine du cluster est similaire à des laminines globulaires (LamG). Les laminines globulaires sont des glycoprotéines impliquées dans la constitution de matrice extracellulaire et qui pourraient avoir un rôle dans la reconnaissance virus-hôte. (ii) Une protéine du cluster est de fonction inconnue mais avec deux domaines similaires aux laminines globulaires et aux immunoglobulines, respectivement. (iii) Une protéase à sérine de type subtilisine est également présente dans le cluster. Le virus ΨM2 de *Methanothermobacter marburgensis* DSM 2133 code aussi pour une protéase qui facilite son adsorption et la lyse cellulaire. La fonction exacte du cluster reste inconnue. De plus, sa répartition est limitée à pAMT11 et TKV1. La présence facultative de l'intégrase a peut-être freiné l'identification d'autres éléments intégrés. Cette identification faciliterait probablement la détermination de la fonction du cluster conservé.

Parmi les gènes qui ne sont pas partagés entre pAMT11 et TKV1, on retrouve ceux impliqués dans la réplication. pAMT11 code une protéine putative Rep d'initiation de réplication par cercle roulant (Gonnet et al., 2011). Des origines putatives de réplication double brin et simple brin ont aussi été détectées. Au contraire, TKV1 code une hélicase putative de type MCM et utilise probablement le mécanisme thêta de réplication. Pour ces éléments génétiques mobiles, deux types de réplication différents sont couplées à un même module fonctionnel, avec ou sans intégration possible.

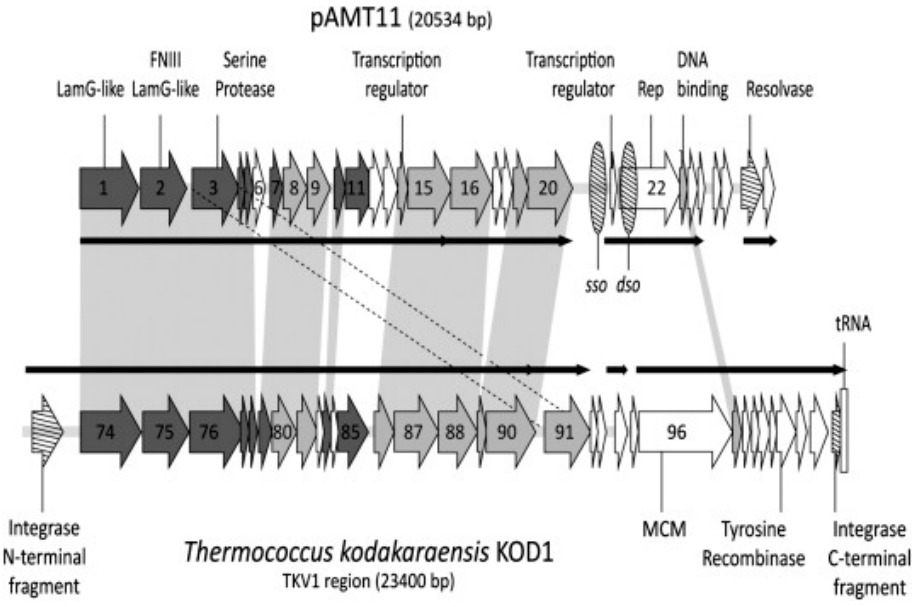


Figure 10. Comparaison des génomes du plasmide pAMT11 (en haut, linéarisé) **et de l'élément intégré** TKV1 (en bas, gènes de *Thermococcus kodakarensis* KOD1 TK0073 to TK0105). Les flèches épaisses représentent des ORFs. Les flèches colorées en gris clair ont une conservation d'identité de 25% à 50M entre les deux éléments. Les flèches colorées en gris foncé ont une conservation de plus de 50%. Les flèches fines représentent des unités de transcription prédites. Figure tirée de Gonnet et al., 2011.

Les virus d'archées

Une centaine de génomes de virus d'archées est publiée au NCBI. Ils proviennent de virus isolés très majoritairement à partir d'hôtes de cinq ordres : les Euryarchées Thermococcales et Halobacteriales et les Crenarchées Sulfolobales, Desulfurococcales et Thermoproteales (Prangishvili et al., 2017). Un virus ΨM1/2 a aussi été identifié chez *Methanothermobacter marburgensis* DSM 2133 (Pfister et al., 1998) et une particule virale A3-VLP a été identifiée chez *Methanococcus voltae* A3 (Wood et al., 1989). Tous ces virus infectent des hôtes extrêmophiles et aucun virus infectant une archée modérée ou d'un autre groupe n'a pour l'instant été isolé (Munson-McGee et al., 2018). Nous savons cependant que des virus existent pour les autres groupes d'archées comme les Thaumarchées, les Nanoarchées ou les Bathyarchées grâce à des données de métagénomique (Munson-McGee et al., 2018; Prangishvili et al., 2017). Ces nouvelles séquences virales détectées par métagénomique augmentent la diversité virale connue mais ne sont pas encore classifiées (Prangishvili et al., 2017).

Les virus archéens isolés ont majoritairement un génome à ADN double brin et certains ont un génome à ADN simple-brin. La taille du génome varie de quelques kilobases pour le virus APBV1 de *Aeropyrum pernix* (Mochizuki et al., 2010) à presque 150 kilobases pour le virus HGTV1 de Halogranum sp. (Senčilo et al., 2013). Les génomes à ADN double-brin sont soit linéaires (Bamford et al., 2005) soit circulaires (Geslin et al., 2003). Aucun virus à génome à ARN n'a été identifié à ce jour mais certains virions comme celui de φCH1 de *Natronobacterium magadii* contiennent de petites molécules d'ARN (Witte et al., 1997).

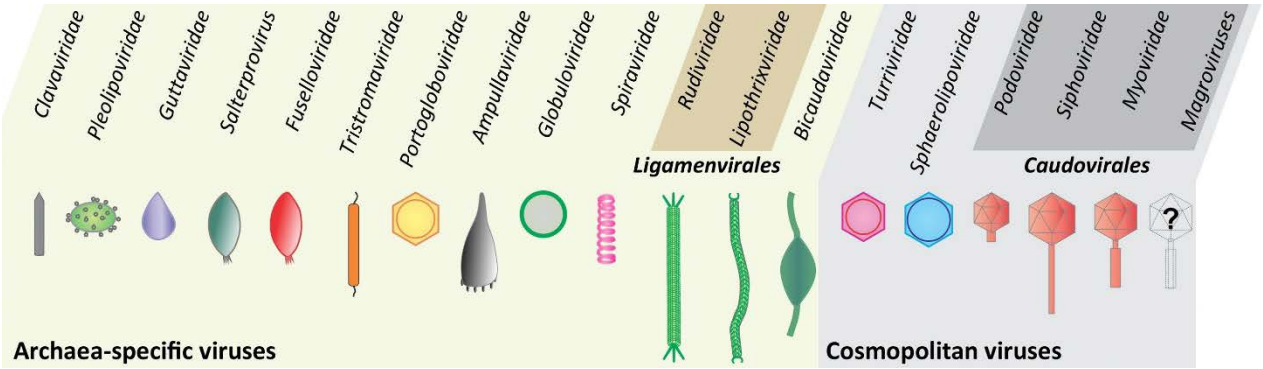


Figure 11. Morphotypes des virus d'archées. Figure tirée de Krupovic et al., 2018

Les virus qui infectent les archées sont typiquement séparés en deux catégories : les virus cosmopolites et les virus spécifiques des archées (Figure 11). Les virus cosmopolites possèdent des caractéristiques structurales et génomiques similaires à des virus de bactéries et d'eucaryotes (Prangishvili et al., 2017). Parmi les virus cosmopolites, on trouve cinq familles officielles (Figure 11). L'ordre Caudovirales regroupe trois familles correspondant aux morphologies de type « tête-queue » et comprend par exemple les bactériophages modèles T4, T7 ou lambda et des virus d'Halobacteriales. Les deux familles Turriviridae et Sphaerolipoviridae possèdent une protéine majeure de capside avec un repliement similaire (Krupovic et al., 2018). Le plasmide pTN3 et l'élément intégré TKV4 de Thermococcales sont apparentés aux virus de la famille Turriviradae (Krupovič and Bamford, 2008). Les virus spécifiques des archées possèdent des caractéristiques structurales et génomiques seulement retrouvées chez les virus d'archées. Ils arborent une grande diversité morphologique : forme de fuseau (ou citron) pour les Fuselloviridae et les Bicaudaviridae, forme de bouteille de champagne pour les Ampullaviridae, forme de goutte pour les Guttaviridae, forme torsadée pour les Spiraviridae, forme filamenteuse pour

les Ligamenvirales et les Tristomaviridae ou encore forme sphérique pour les Globuloviridae et les Portoglobaviridae (Prangishvili et al., 2017). La très grande majorité des virus spécifiques d'archées infectent des Crenarchées à l'exception notamment des Pleolipoviridae qui infectent des Halobacteriales. Les virus spécifiques d'archées constituent un très grand réservoir de découvertes exceptionnelles qu'on commence seulement à explorer.

La diversité des morphologies de virus d'archées se reflète dans la diversité des modes d'interaction mis en place avec l'hôte (Prangishvili et al., 2017). Seulement quelques virus ont été étudiés en détail. Les données expérimentales sur les mécanismes d'attachement et d'entrée du virus, de réplication et de morphogenèse du virion sont rares mais spectaculaires (Bize et al., 2009; Quemin et al., 2013). Les mécanismes sont souvent extrapolés à partir des séquences des virus mais cette méthodologie est limitée par leur très grande diversité et faible redondance.

La famille spécifique des archées Pleolipoviridae est singulière : le virion est similaire à une vésicule membranaire sans capside particulière même si des protéines structurales sont localisées au niveau de la membrane parfois sous forme de de pic (Pietilä et al., 2012). Ils correspondent par exemple aux virus SNJ1 et SNJ2 isolés à partir de *Natrinema sp.* J7-1 (Liu et al., 2015; Zhang et al., 2012). Les génomes des pléolipovirus sont divers : génome d'ADN double brin linéaire ou circulaire ou génome d'ADN simple brin. Certains pléolipovirus comme SNJ2 de *Natrinema sp.* ou HRPV9 de *Halorubrum sp*. codent une intégrase (Atanasova et al., 2018; Liu et al., 2015) et s'intègrent dans le chromosome (Wang et al., 2018).

Les Fuselloviridae

La famille virale Fuselloviridae est spécifique des archées. Le virus modèle de cette famille est le virus SSV1 (Sulfolobus Spindle Shaped Virus 1) isolé à partir de *Sulfolobus shibatae* B12 (Martin et al., 1984). Il a été particulièrement étudié pour ses aspects génétiques (Fröls et al., 2007; Fusco et al., 2015; Reiter et al., 1987). La découverte d'une dizaine d'autres fusellovirus semblables chez les Sulfolobus a permis d'étudier leurs aspects évolutifs et biogéographiques (Goodman and Stedman, 2018; Held and Whitaker, 2009; Pauly Matthew D. et al., 2019; Wiedenheft et al., 2004). De manière générale, les Fusellovirus sont présents chez les Sulfolobales (SSV1 à SSV10 et ASV) (Goodman and Stedman, 2018), chez les Desulfurococcales (virus APSV1), chez les Thermococcales (virus TPV1 et particule virale PAV1) et chez les Methanococcales (particule virale A3-VLP) (Iranzo et al., 2016; Krupovic et al., 2014b). La classification dans la famille des Fuselloviridae se fait sur la base de similarités de séquences des protéines majeures de capside (MCP). Les MCP de Fuselloviridae présentent notamment deux domaines hydrophobiques. L'ensemble des Fusellovirus ne présente pas d'autre protéine commune, même si ils codent tous une ATPase AAA^+ (Krupovic et al., 2014b) qui pourrait servir à la réplication (Iranzo et al., 2016; Krupovic et al., 2018). La majorité des Fusellovirus code également une intégrase spécifique des archées (cf. Chapitre 3 page 56).

Le virus TPV1

Le virus TPV1 a été isolé à partir de *Thermococcus prieuri* (Gorlas et al., 2012). Le virion long de 140 nm a une forme de fuseau avec une petite queue terminée par des fibres (Figure 12.A). Il est tolérant à la température (thermo- et psychro-tolérant). Les virions sont libérés de manière continue et spontanée sans lyse détectée de la cellule hôte (Gorlas et al., 2012). Comme pour la plupart des virus d'archées, les mécanismes des différentes étapes du cycle viral ne sont pas connus.

Le génome d'ADN circulaire double brin de TPV1 est présent dans le virion et comme épisome dans le cytoplasme de la cellule hôte. Le génome a une taille de 22 kilobases et comprend 28 ORFs putatives dont 9 avec une fonction putative précise (Figure 12.A). Notamment, le virus code une intégrase similaire à l'intégrase du virus SSV1. Il n'a pas été déterminé si le plasmide présente une copie intégrée dans le chromosome hôte (Gorlas et al., 2012). Aucun élément intégré similaire au virus n'est présent dans les génomes publiés d'archées bien que TPV1 soit capable d'infecter plusieurs souches de Thermococcales. Le virus code également une hélicase de type MCM séparée en deux morceaux (ORF-1 et -2) et qui pourrait être impliquée dans la réplication du génome. Une origine putative de réplication contenant des répétitions riches en AT a également été détectée (Gorlas et al., 2012). L'ORF15 code probablement pour la protéine majeure de capside d'après sa similarité avec les autres MCP de Fusellovirus et la présence de deux domaines hydrophobes (Krupovic et al., 2014b). Deux protéines de structures similaires au LamG sont également codées par le virus TPV1 et par la particule virale PAV1 de *Pyrococcus abyssi* GE23 (Gorlas et al., 2012).

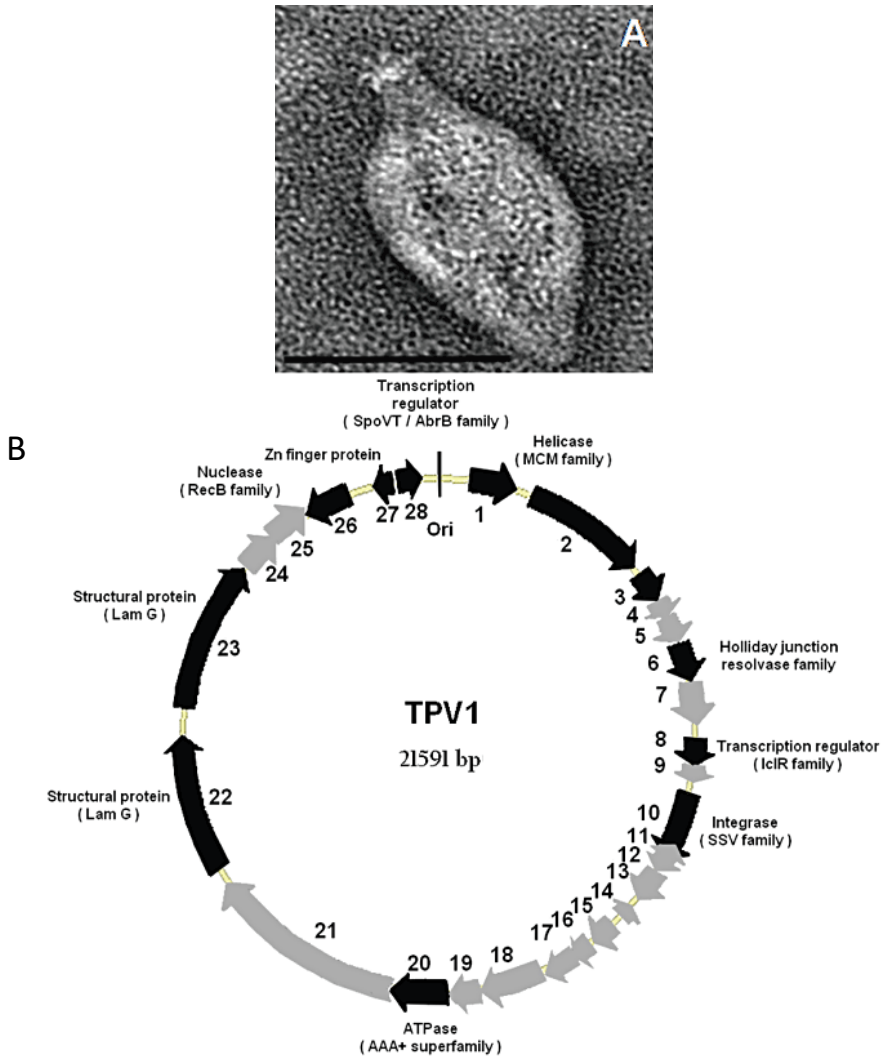


Figure 12. Virion (A) et carte du génome (B) de TPV1. A. Micrographie électronique de TPV1. La barre d'échelle représente 120 nm. B. Les ORFs prédites sont représentés par des flèches. Les flèches noires correspondent aux ORFs avec une fonction attribuée et aux ORFs conservées dans d'autres génomes. Les flèches grises correspondent aux ORFans sans fonction assignée. La localisation prédite de l'origine de réplication est indiquée. Figure tirée de Gorlas et al., 2012

Les éléments transposables d'archées

La majorité des éléments transposables présents chez les archées sont des éléments IS (Filée et al., 2007). Ces éléments encodent une transposase de type DDE, de type sérine ou de type relaxase qui n'ont pas été caractérisées chez les archées (Filée et al., 2007). Ils peuvent être classifiés en différents groupes spécifiques aux archées ou similaires aux IS bactériens et aucun transposon similaire à ceux des eucaryotes n'a été détecté (Filée et al., 2007). Ils sont majoritairement présents dans quatre ordres : Halobacteriales, Sulfolobales, Methanosarcinales et Thermoplasmatales (Filée et al., 2007). Ils y représentent généralement quelques pourcents de la taille du chromosome, ou parfois jusqu'à 10%. Ils sont également présents sur des plasmides où ils peuvent constituer jusqu'à 20% du génome (Filée et al., 2007). La distribution des éléments IS est très variable entre souches proches. Par exemple, *Saccharolobus solfataricus* P2 contient 200 à 350 éléments IS suivant les estimations (Filée et al., 2007; She et al., 2001a) alors que *Sulfolobus tokodaii* strain 7 en contient seulement 34 (Brügger et al., 2002). Similairement, *Pyrococcus furiosus* DSM 3638 contient 35 éléments IS alors que la souche dérivée *Pyrococcus furiosus* COM1 en contient 45 (Bridger et al., 2012). Les éléments additionnels de COM1 sont responsables de de plusieurs changements génomiques par rapport à la souche type DSM 3638 (cf. Chapitre 4). Les deux souches de *Pyrococcus furiosus* apparaissent comme des exceptions parmi les Thermococcales qui portent pour la plupart seulement quelques éléments IS (Fukui et al., 2005; Lecompte et al., 2001). D'autres éléments transposables sont aussi présents chez les archées comme les casposons (Krupovic et al., 2017) ou des éléments de type MITE (Filée et al., 2007). Ces derniers emploient une transposase à DDE encodée par un élément IS et ont été détectées principalement pour les Sulfolobales et Methanosarcinales.

Evolution des éléments génétiques mobiles

Les éléments génétiques mobiles présentent une évolution modulaire (ou mosaïque) (Botstein, 2006; Hendrix et al., 1999; Oberto et al., 1994) caractérisée par la perte ou l'échange entre éléments génétiques mobiles de clusters de gènes appelés modules. Chaque module d'un élément génétique mobile a une histoire évolutive distincte (Cury et al., 2017). Les modules présentent parfois une cohérence fonctionnelle (module de conjugaison, module d'intégration, module de structure du virion, module de réplication, etc.). Par exemple, le génome du virus PAV1 de *Pyrococcus abyssi* (Geslin et al., 2007) présente deux grands modules (Figure 13). Le premier module de type virus est partagé avec le virus TPV1 (Gorlas et al., 2012) et code des protéines de structure virale. Le deuxième module de type plasmide est partagé avec les plasmides pIRI48 et pCIR10 de Thermococcales (Krupovic et al., 2013) et code des régulateurs transcriptionnels à domaine wHTH (hélice-tour-hélice ailé) et à domaine superhélice, une ATPase de type ABC et des protéines de fonctions inconnue. La fonction de ce module reste pour l'instant cryptique. Deux hypothèses sont envisageables pour expliquer l'évolution de ces modules (Krupovic et al., 2013). D'un côté, le génome du virus PAV1 pourrait résulter de l'arrangement unique des deux modules plasmidique et viral. De l'autre côté, les plasmides pIRI48 et pCIR10 pourraient être des virus dégénérés ayant perdu le module structural. Les deux hypothèses mettent en évidence une connexion évolutive entre différents types d'éléments génétiques mobiles. Ces connexions sont fréquemment observées chez les archées (Iranzo et al., 2016)

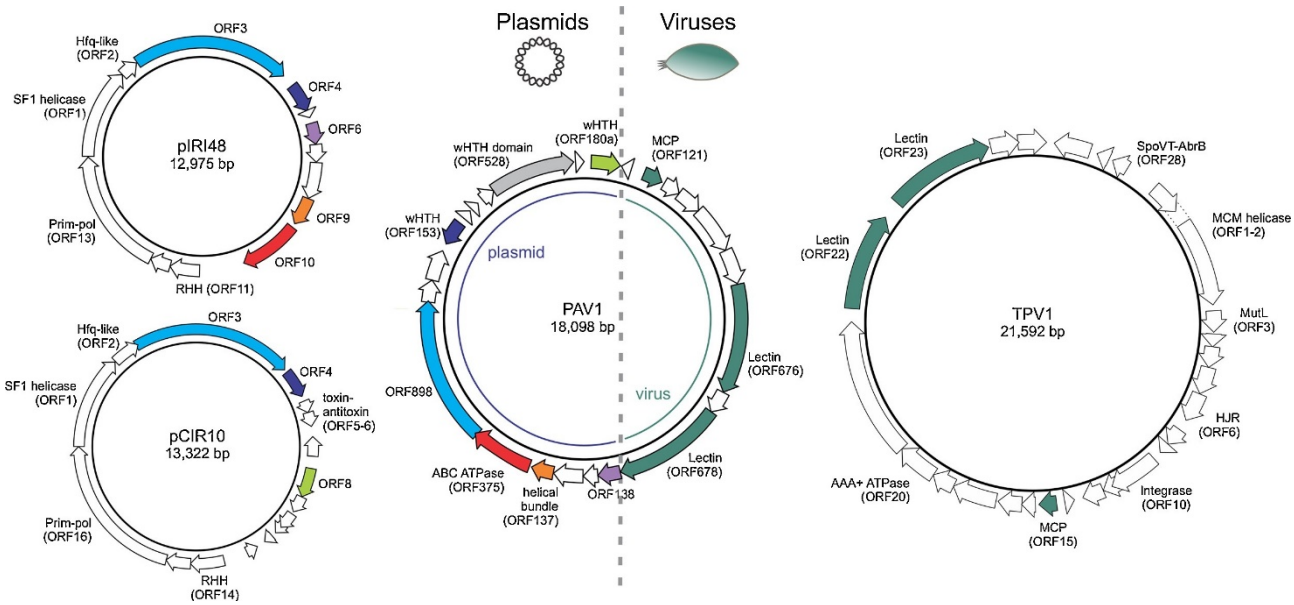


Figure 13. Relations entre les plasmides pCIR10 et pIRI48 et les virus PAV1 et TPV1 de Thermococcales. Une carte de chaque génome est présentée. Les flèches correspondent à des gènes. Les gènes homologues sont indiqués par une même couleur. Abréviations. MCP : protéine majeure de capside. SF1 : superfamille 1. wHTH : hélice-tout-hélice ailé. RHH : ruban-hélice-hélice. HJR: résolvase de jonction de Holliday. Prim-pol : primase-polymérase. Figure tirée de Krupovic et al., 2018

Les éléments génétiques mobiles présentent souvent une organisation génétique conservée, en dépit de leur évolution modulaire. Par exemple, les plasmides conjugatifs de Sulfolobus de la famille pNOB8 présentent tous une succession de 4 grands modules (Figure 14). Le premier et le troisième module sont conservés. Les deux autres modules sont variables mais à une position intercalaire conservée (Erauso et al., 2006). Les plasmides de la famille pT26-2 présentent quant à eux deux grandes régions (Figure 8.B page 29) : une région conservée correspondant au module core et une région variable (Soler et al., 2010). Pour les familles pNOB8 et pT26-2, les régions conservées contiennent les ORFs les plus grandes (Erauso et al., 2006). Cette tendance d'organisation des génomes de virus et plasmides avec une région conservée à grandes ORFs et une région non conservée à plus petites ORFs est courante (Hendrix, 2002; Selb et al., 2017). Les régions non conservées pourraient correspondre à des incubateurs où de de nouvelles séquences codantes sont formées (Hendrix, 2002).

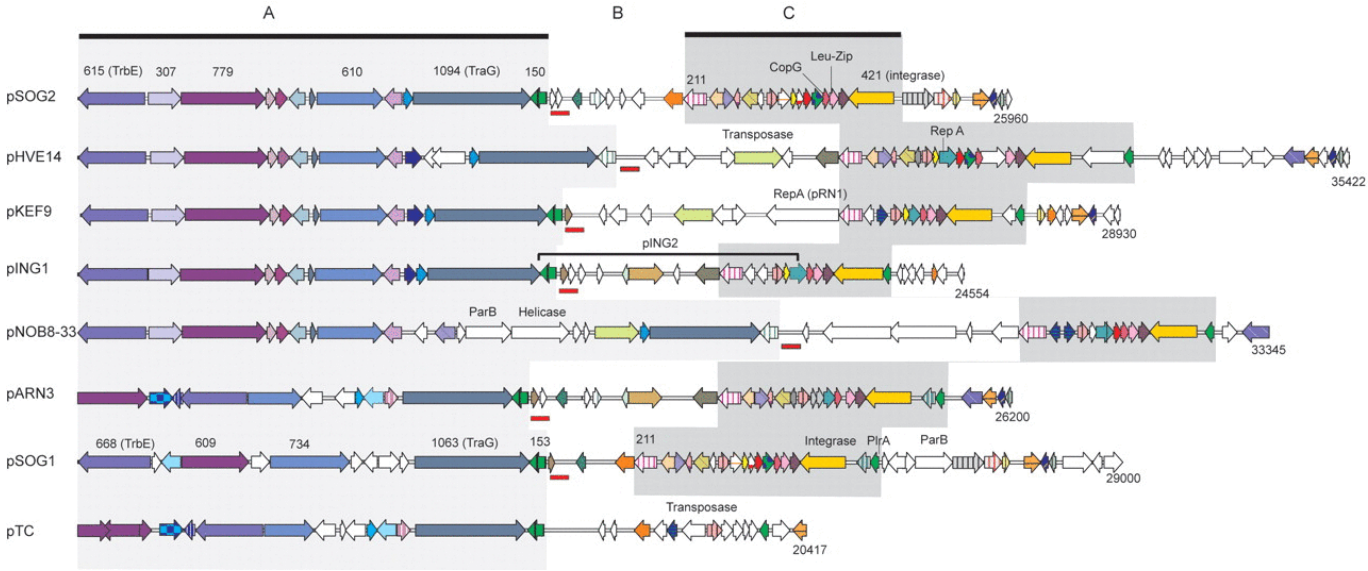


Figure 14. Comparaison génétique de différents plasmides conjugatifs de Sulfolobus. Les flèches représentent les ORFs. Les ORFs homologues entre les différents plasmides sont identifiées par une couleur et un motif particulier. Les ORFs sans homologues parmi les autres plasmides sont représentées en blanc. La région A encode les fonctions de conjugaison. La région B contient l'origine de réplication putative (barre rouge). La région C encode des fonctions de réplication. Figure tirée de Erauso et al., 2006.

Mécanismes cellulaires de défense contre les éléments génétiques mobiles

Pour éviter la diminution de fitness induite par certains éléments génétiques mobiles, les hôtes cellulaires mettent en place une variété de systèmes de défense. Ils sont similaires chez les archées et les bactéries, à l'image de la similarité de leurs éléments génétiques mobiles (Koonin et al., 2017). En revanche, les systèmes de défense sont plus nombreux chez les archées que chez les bactéries. Ils sont également enrichis chez les archées et bactéries thermophiles, par rapport aux mésophiles et psychrophiles (Koonin et al., 2017). Singulièrement, les hôtes tendent à recruter leurs mécanismes de défense à partir de fonctions encodées par des éléments génétiques mobiles (Koonin et al., 2017). Les éléments génétiques mobiles ripostent aux attaques des hôtes entrainant une course aux armements. La course entre l'élément parasitique et son hôte continue perpétuellement sans qu'aucun vainqueur ne se détache (hypothèse de la reine rouge) (Brockhurst et al., 2014; Valen, 1974; Weitz et al., 2005). Cette co-évolution antagoniste est considérée comme un mécanisme majeur de l'évolution.

Trois degrés de défense peuvent être mis en place (Koonin et al., 2017). Le premier degré correspond à la résistance. L'élément infectieux est arrêté avant de pouvoir pénétrer dans la cellule. Le deuxième degré correspond à l'immunité. L'élément est détruit grâce à la distinction entre le soi et le non-soi. L'immunité peut être innée (plutôt non spécifique) ou adaptative (très spécifique). Le troisième degré correspond à un avortement de l'infection par dormance ou mort cellulaire programmée. La cellule évite ainsi la propagation de l'élément génétique mobile au reste de la population. Chaque degré est mis en place à une étape différente de l'infection par un virus (Figure 15) ou par un plasmide (Dy et al., 2014; Labrie et al., 2010). Les mécanismes de résistance agissent lors des premières étapes de pénétration de l'ADN exogène et l'immunité et la mort cellulaire sont mises en jeu lors des étapes suivantes. Nous allons maintenant détailler certains systèmes de défense.

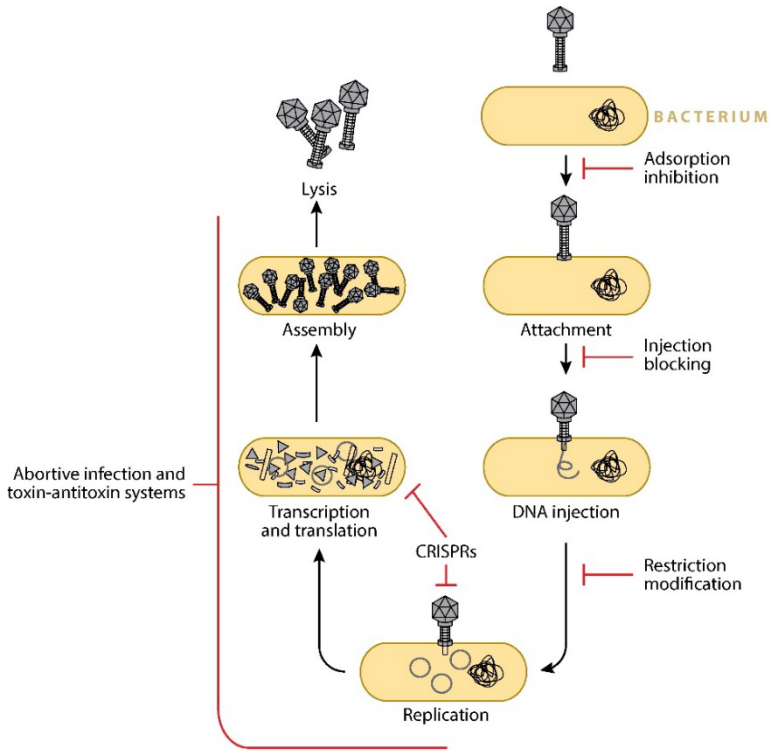


Figure 15. Cycle de vie lytique d'un virus et mécanismes de défense bactériens. Les différentes stratégies de défense ciblent des étapes différentes du cycle de vie du virus. Figure tirée de Dy et al., 2014

Le mécanisme d'immunité innée le mieux caractérisé est celui des modifications de restrictions présent chez les archées et les bactéries. Une modification est apportée à l'ADN hôte (soi) et l'ADN ne portant pas cette modification (non-soi) est dégradé par des endonucléases de restriction. La modification correspond à une méthylation de l'ADN en général (Dy et al., 2014) ou à sa phosphorothioation (Koonin et al., 2017). Il existe des cas où le marquage du non-soi entraine sa dégradation (Koonin et al., 2017). Un autre mécanisme d'immunité inné proposé utilise les protéines Argonaute (pAgo) chez les archées et les bactéries (Dy et al., 2014; Koonin et al., 2017). La reconnaissance du non-soi serait effectuée par une molécule d'ADN ou ARN guide et la protéine Argonaute cliverait ensuite la molécule reconnue. L'activité d'endonucléase dépendante d'un guide a été montrée pour plusieurs protéines Argonaute dont une de *Pyrococcus furiosus* (Swarts et al., 2015). Le mécanisme détaillé d'action de ce système reste encore à déterminer.

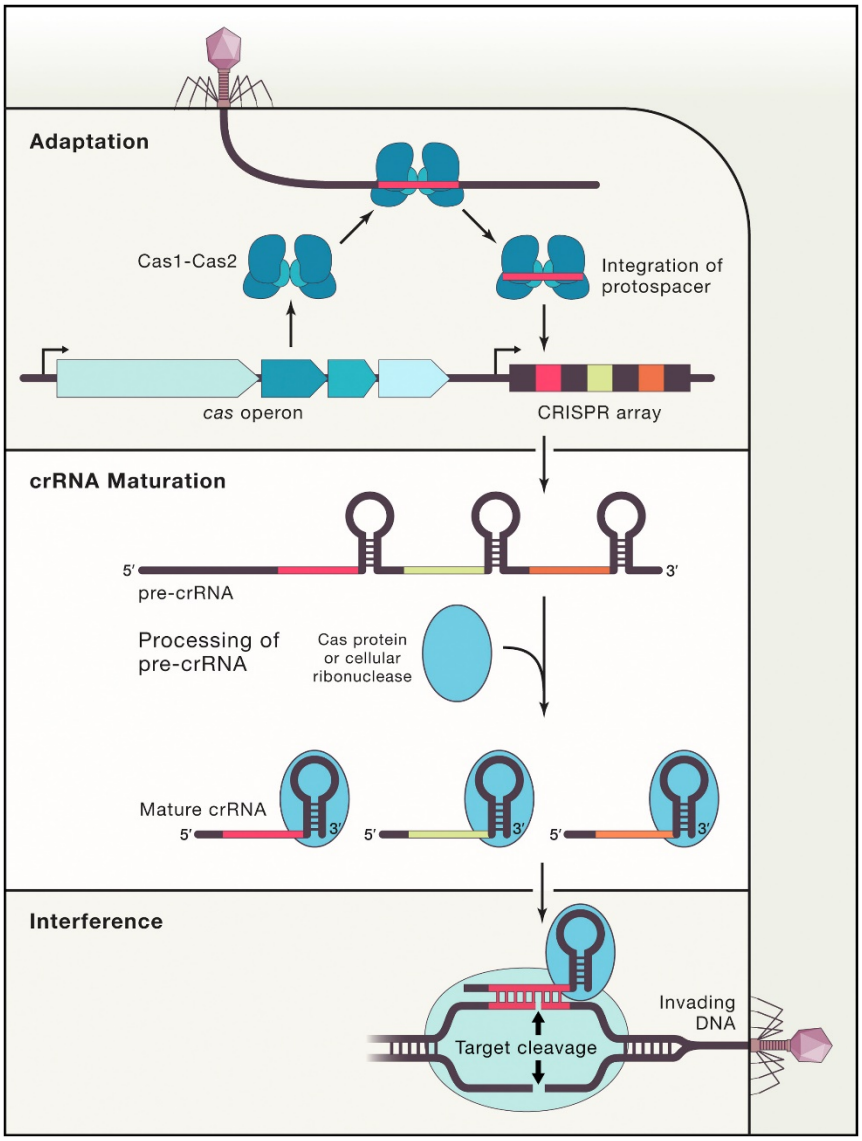


Figure 16. Le mécanisme de défense CRISPR-Cas en trois étapes. Figure tirée de (Hille et al., 2018)

Le système d'immunité acquise CRISPR-Cas a été récemment découvert chez les archées et les bactéries et est fortement étudié pour ses applications dans la modification de l'ADN (Adli, 2018). Il est très présent chez les archées (Mojica et al., 2000), et encore plus chez les hyperthermophiles (Koonin et al., 2017). Grâce au système CRISPR-Cas, le génome garde une mémoire nucléique des

infections passées au niveau de loci « CRISPR array » (Hille et al., 2018). Le « CRISPR array » est composé d'une alternance de répétitions et de « spacers » correspondant aux séquences mémorisées (Marraffini and Sontheimer, 2010). La cellule peut ensuite reconnaître et cliver l'acide nucléique mémorisé lors d'une nouvelle infection. Les protéines mises en jeu dans ce mécanisme sont appelées Cas. Trois étapes successives participent à l'immunité (Figure 16). Premièrement, lors de l'étape d'adaptation, une portion de l'ADN exogène est sélectionnée et intégrée au « CRISPR array » sous forme d'un nouveau « spacer » adjoint d'une répétition (Marraffini, 2015). Deuxièmement, lors de l'étape de maturation, le « CRISPR array » est transcrit et traité en petites séquences d'ARNcr contenant chacune un spacer. Troisièmement, lors de l'étape d'interférence, l'ARNcr guide la reconnaissance et la dégradation d'un nouvel acide nucléique exogène. Du point de vue de la recherche sur les éléments génétiques mobiles, les « CRISPR arrays » permettent d'identifier des éléments génétiques mobiles qui ont probablement été en contact avec un génome donné. Le contact est seulement probable et non certain car les « CRISPR arrays » peuvent être acquis par transfert horizontal (Horvath and Barrangou, 2010).

La mort cellulaire programmée d'une cellule infectée par un élément mobile est une sorte de suicide altruiste. Elle peut être contrôlée par un module toxine-antitoxine (Harms et al., 2018). Le module est composé d'une toxine, souvent une enzyme qui inhibe la croissance cellulaire, et d'une antitoxine correspondant à une protéine ou un ARN qui contrôle la toxine directement ou à travers la régulation de son expression. Le module toxine-antitoxine est exprimé avant l'infection par le phage qui perturbe par la suite l'équilibre toxine antitoxine. Ceci entraine l'activation de la toxine et la mort de la cellule.

Méthodes d'identification des éléments génétiques mobiles

Les méthodes d'identification des éléments génétiques mobiles sont différentes suivant que leur génome soit séparé ou intégré dans le chromosome hôte. Dans le cas des éléments génétiques mobiles séparés du chromosome hôte (plasmides, virus), l'acide nucléique peut être identifié dans des extraits totaux d'ADN (Geslin et al., 2003) ou extrait spécifiquement (Bimboim and Doly, 1979). Additionnellement, les virus sous forme de virion peuvent être isolés et caractérisés à partir de leur cellule hôte. La relation hôte-virus peut alors être étudiée. Toutes ces méthodes reposent sur la culture en laboratoire de l'hôte du plasmide ou du virus. Cependant une très grande proportion des microorganismes, dont certains groupes entiers d'archées, sont actuellement difficilement cultivables ou incultivables (Offre et al., 2013; Spang et al., 2017). Il est donc très difficile d'isoler leurs virus et des plasmides. Une nouvelle méthode qui permet d'obtenir des séquences virales sans cultiver l'hôte a été développée : la métagénomique virale (Delwart, 2007; Rosario and Breitbart, 2011). Les virions sont purifiés à partir d'échantillons environnementaux puis séquencés. Cette méthode permet seulement d'identifier les séquences virales présentes dans un environnement sans information sur le virion ou le cycle de vie du virus. Un hôte putatif peut être identifié s'il présente des spacers CRIPSR correspondant à la séquence du virus.

La détection des éléments génétiques mobiles intégrés dans le chromosome implique le séquençage du génome de l'hôte et leur identification comme élément exogène intégré dans le chromosome. La séquence du génome de l'hôte peut être obtenue à partir d'une culture pure ou par reconstruction d'un MAG (Metagenome-Assembled Genome) à partir de données métagénomiques. Le Tableau 4 résume les différentes méthodes utilisées individuellement ou en coordination pour détecter des éléments mobiles intégrés (Cortez et al., 2009; Krupovic et al., 2019). Ces méthodes sont mises en

œuvre par des logiciels pour faciliter la détection des éléments intégrés (Arndt et al.; Bertelli et al., 2017; Langille et al., 2010). En pratique, ces logiciels sont souvent peu efficaces chez les archées pour lesquels ils n'ont pas été optimisés.

Tableau 4. Caractéristiques des éléments génétiques mobiles intégrés utilisables pour leur détection. Adapté de Langille et al., 2010

Caractéristique	Méthode de détection	Limitations
Distribution sporadique due à la mobilité	Identification de longs sauts dans la synténie entre deux souches proches	Détection d'intégrations postérieures à la divergence entre les deux souches étudiées
Distribution sporadique due à la mobilité	Chute dans la couverture d'un génome séquencé par des reads métagénomiques (Anderson et al., 2014)	
Distribution sporadique due à la mobilité	Segments de chromosome sans gènes core (Cossu et al., 2015)	Il faut suffisamment de génomes pour définir précisément les gènes core
Evolution modulaire	Recherche de segments synténiques entre des éléments génétiques mobiles connus et des génomes	La détection basée sur les éléments déjà connus ne permet d'identifier d'éléments nouveaux
Surreprésentation de certains groupes de gènes dont : ‣ gènes de mobilité (recombinase, transposase …) ‣ gènes signatures (Protéine Rep, système de conjugaison, protéines core …) ‣ ORFans	Détection par analyse de similarité ou grâce aux annotations	La détection basée sur les éléments déjà connus ne permet d'identifier d'éléments nouveaux
Composition personnelle de la séquence (utilisation des codons, pourcentage en GC, utilisation des tétranucléotides …)	Détection de biais de composition dans le génome	La composition du génome cellulaire n'est pas uniforme. Le biais de composition disparait après un temps long d'intégration.
Répétition directes et/ou répétition inversées aux extrémités	Détection de répétition	Tous les éléments intégrés ne sont pas bordés par des répétitions et les répétitions de petites tailles peuvent être difficiles à identifier
Intégration préférentielle dans des gènes codant pour des ARNt	Analyse détaillée des environs des gènes codant pour les ARNt	Tous les éléments mobiles ne sont pas intégrés au niveau de gènes codant pour des ARNt

Chapitre 3. Recombinaisons génétiques

La recombinaison correspond à un échange coordonné de brins entre deux doubles hélices d'ADN. Elle aboutit la création de molécules chimériques d'ADN qui présentent de nouvelles combinaisons d'allèles ou de gènes. Il existe trois types de recombinaisons (Tableau 5). (1) La recombinaison homologue correspond à l'échange de brin ente deux séquences identiques et catalysé par un complexe protéique. (2) La transposition correspond à l'échange de brin entre une séquence spécifique et une autre séquence non ou peu spécifique. (3) La recombinaison site-spécifique correspond à l'échange de brin entre deux séquences spécifiques et identiques. La transposition et la recombinaison site-spécifique présentent des similarités, contrastant avec la recombinaison homologue. Elles font intervenir des recombinases qui catalysent le clivage et la jonction des brins d'ADN au moyen de différents motifs catalytiques (tyrosine, sérine, aspartate-aspartate-glutamate DDE).

Nous allons maintenant décrire les mécanismes enzymatiques mis en jeux par les différentes recombinases. Nous présenterons ensuite la fonction de la recombinaison qui correspond à l'intégration et l'excision d'éléments génétiques mobiles. Nous terminerons par une description de trois recombinases à tyrosine typiques et de celles caractérisées chez les archées.

Tableau 5. Caractéristiques et enzymes des différents types de recombinaisons.

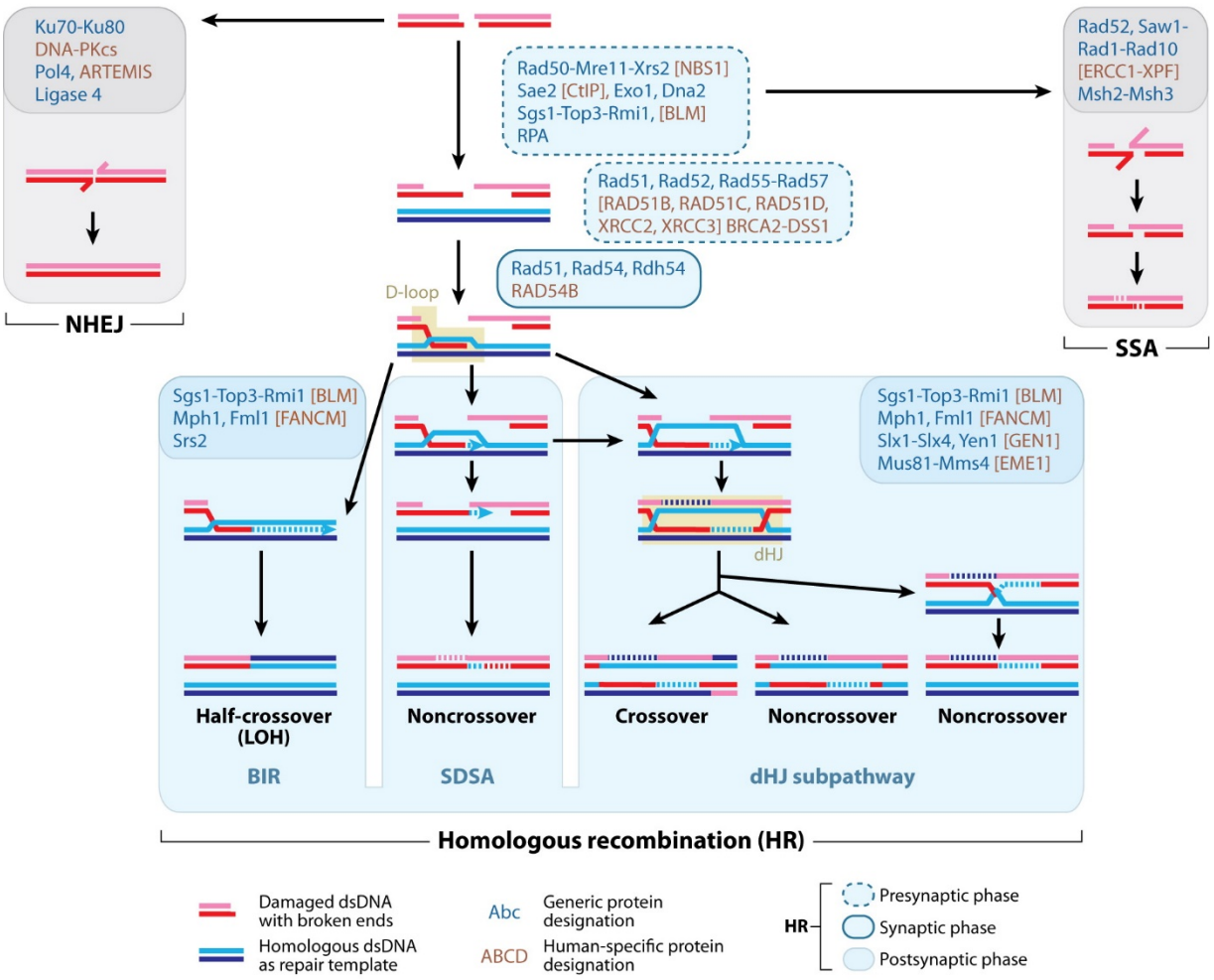
	Recombinaison homologue	Transposition	Recombinaison site-spécifique
Site de recombinaison	Deux séquences homologues (identiques) de taille suffisante (centaines de paires de base) Défini par des interactions ADN-ADN	Une séquence spécifique et une séquence non spécifique Défini par des interactions protéine-ADN	Deux séquences définies, similaires et courtes (dizaines de paires de base) Défini par des interactions protéine-ADN
Enzymes	Complexe protéique incluant la recombinase Rad51-RecA-RadA	Recombinase à DDE	Recombinase à tyrosine ou à sérine
Activité de la recombinase	Formation d'un filament nucléoprotéique et invasion de brins	Clivage et jonction de brins d'ADN	Clivage et jonction de brins d'ADN
Particularités	Requiert de l'ATP Requiert une synthèse d'ADN	Requiert des cations métalliques divalents Requiert une synthèse d'ADN	Possibilité de protéines accessoires RDF pour contrôler la directionnalité
Fonctions biologiques	‣ Réparation de l'ADN ‣ Déblocage des fourches de réplication ‣ Création de nouvelles combinaisons d'allèles	‣ Intégration et excision d'éléments génétiques mobiles ‣ Variation de phase	‣ Intégration et excision d'éléments génétiques mobiles ‣ Résolution de dimères de chromosome ou de plasmide ‣ Variation de phase

Les différents types de recombinaison et leurs enzymes

Recombinaison homologue

La recombinaison homologue a été particulièrement décryptée chez les eucaryotes (San Filippo et al., 2008). Elle est mise en place à la suite d'une cassure double brin d'ADN spontanée ou délibérée (Figure 17) et peut-être subdivisée en trois grandes phases qui font toutes intervenir la recombinase Rad51 (Krejci et al., 2012). Premièrement, lors de la phase présynaptique, les extrémités 5' de la cassure double brin sont dégradées et Rad51 est chargée de manière ATP-dépendante sur le brin simple d'ADN dégagé (Sung and Klein, 2006). Ceci aboutit à la formation d'un filament nucléoprotéique présynaptique. La reconnaissance de l'ADN simple-brin par Rad51 est facilitée par différent facteurs de recombinaison (Heyer et al., 2010). Deuxièmement, lors de la phase synaptique, Rad51 catalyse la recherche d'homologie et l'invasion du brin enrobé dans le duplex homologue d'ADN. Ceci aboutit à la formation d'un filament nucléoprotéique synaptique contenant le brin invasif et le brin matrice complémentaire dans une structure génétique appelée boucle-D (D-loop en anglais). Troisièmement, lors de la phase post-synaptique, la dissociation de Rad51 du filament synaptique libère l'extrémité 3' envahissante. Elle est alors utilisée comme amorce pour la synthèse d'ADN. Trois voies sont ensuite possibles pour éliminer la boucle-D. La première voie dHJ implique la création d'un intermédiaire avec deux jonctions de Holliday. Sa résolution aboutit à la formation soit de molécules identiques au point de départ en absence de cross-over, soit de molécules chimères dans le cas d'un cross-over. Cette voie est principalement mise en œuvre lors de la méiose. La deuxième voie SDSA implique le déplacement du brin invasif hors de la boucle-D et son appariement à son brin complémentaire. Elle est principalement mise en œuvre lors de la mitose. La troisième voie BIR implique la transformation de la boucle-D en fourche de réplication. Elle peut être utilisée dans certains cas pour initier la réplication. Deux autres mécanismes existent pour réparer les cassures doubles brins : le mécanisme SSA et le mécanisme NHEJ. Ils entrainent une réparation non fidèle contrairement à la recombinaison homologue.

La recombinaison homologue a aussi été étudiée chez les archées (White, 2011). Le mécanisme et les protéines mise en jeu sont très conservés avec la recombinaison homologue eucaryote. Notamment, la recombinase Rad51 a pour homologue archéen la recombinase RadA.



Heyer W-D, et al. 2010.
Annu. Rev. Genet. 44:113–39

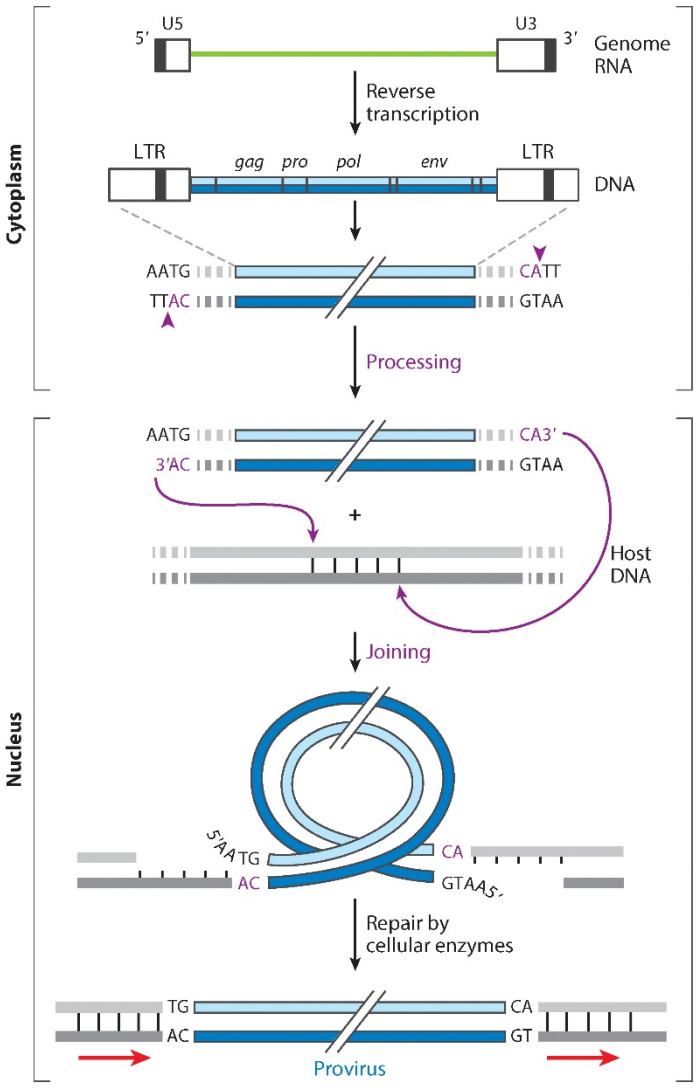
Figure 17. Les voies de réparation de cassures double brins chez les eucaryotes. Les protéines impliquées sont indiquées à chaque étape suivant la nomenclature en vigueur pour la levure *Saccharomyces cerevisiae* (bleu) ou pour l'homme (brun si différent). Les pointillés indiquent les séquences d'ADN néosyntétisées. Abréviations. BIR : réplication induite par une cassure. dHJ : double jonction de Holliday. NHEJ : jonction d'extrémités non homologues. LOH : perte d'hétérozygotie. SDSA : appariement de brin dépendant de la synthèse. SSA : appariement simple-brin. Figure tirée de (Heyer et al., 2010)

Transposition par les recombinases à DDE

Les recombinases à triade catalytique aspartate-aspartate-glutamate (DDE) ou aspartate-aspartate-aspartate (DDD) catalysent la transposition de nombreux éléments transposables (cf. Chapitre 2, page 20) et l'intégration de rétrovirus (Hickman et al., 2010). La triade catalytique est localisée dans une poche où elle coordonne deux cations divalents (Mg^{2+} ou Mn^{2+}) qui activent une molécule d'eau ou un groupement 3'OH pour une attaque nucléophile d'une liaison phosphodiester (Hickman and Dyda, 2015b). Cette activation intervient pour deux activités catalytiques distinctes (Curcio and Derbyshire, 2003). Premièrement, des molécules d'eau activées peuvent hydrolyser des liaisons phosphodiester aux extrémités de l'élément transposable (ou rétrovirus), libérant des groupements 3'OH. Cette activité permet dans certains cas d'exciser l'élément transposable. Deuxièmement, deux groupements 3'OH libérés peuvent attaquer une séquence du chromosome hôte de manière non ou peu spécifique, entrainant la jonction de l'élément transposable avec le chromosome. Cette trans-estérification se produit en absence de similarité de séquence (homologie) entre les deux sites (Curcio and Derbyshire, 2003). Aucun intermédiaire nucléoprotéique covalent n'est mis en œuvre.

Les recombinases à DDE utilisent ces deux activités sur différents substrats et de manière associée ou non à une synthèse d'ADN. Ceci aboutit à une très grande variété de mécanismes de transposition (Curcio and Derbyshire, 2003; Montaño and Rice, 2011) qui est reflétée dans la diversité des séquences des recombinases à DDE (Nesmelova and Hackett, 2010). Nous allons détailler le mécanisme utilisé pour l'intégration de rétrovirus (Figure 18). Ce mécanisme est plus simple que ceux de transposition car il ne nécessite pas l'excision de l'élément génétique mobile. La molécule d'ADN rétrovirale, résultant de la transcription inverse de l'ARN viral, est hydrolysée de manière site-spécifique à l'extrémité des séquences LTR (Skala, 2014). Les groupements 3'OH générés attaquent deux liaisons phosphodiester à une distance de 4 à 6 paires de bases sur les deux brins du chromosome hôte (Andrake and Skalka, 2015). L'ADN rétroviral est alors intégré au chromosome mais présente encore deux extrémités 5'OH libres. La séquence qui était localisée ente les deux liaisons phosphodiester attaquées (site cible) est alors présente sous forme simple brin de chaque côté du rétrovirus. Cette situation est réparée par les enzymes hôtes et aboutit à la formation d'une répétition directe du site cible d'intégration de chaque côté du rétrovirus.

D'autres transposons utilisent des recombinases pour leur transposition (Hickman and Dyda, 2015b). Les transposases HUH catalysent des cassures et jonction pour des intermédiaires simple-brin de transposons (Chandler et al., 2013). Les transposases à sérine et à tyrosine sont utilisées par des éléments transposables présentant un intermédiaire circulaire. Ces transposases sont homologues aux recombinases à tyrosine et sérine détaillées ci-après mais peuvent présenter une spécificité de site relâchée caractérisée par une faible similarité ente les deux sites de recombinaison (Hickman and Dyda, 2015b; Poulter and Butler, 2015; Rajeev et al., 2009).



Andrake MD, Skalka AM. 2015.
Annu. Rev. Virol. 2:241–64

Figure 18. Synthèse d'ADN et intégration rétrovirale dans le génome hôte. La reverse transcription de l'ARN rétroviral en ADN double brin est contrôlée par le génome viral. Une recombinase DDE catalyse les deux étapes suivantes. D'abord, les extrémités de la molécule d'ADN sont hydrolysées de manière site-spécifique (« processing ») puis les extrémités 3'OH libérées attaquent deux liaisons phosphodiester du chromosome hôte (« joining »). La dernière étape de réparation est catalysée par des enzymes de l'hôte et aboutit à la formation d'une répétition directe de la séquence d'intégration hôte. Figure tirée de (Andrake and Skalka, 2015)

Recombinaison site-spécifique conservative

La recombinaison site-spécifique conservative correspond à l'échange réciproque de brins entre deux sites d'ADN définis (Grindley et al., 2006). Le nombre de nucléotides et l'énergie sont conservés pendant la réaction. La recombinaison site-spécifique met seulement en jeu les deux sites d'ADN et un tétramère de recombinase. Dans un premier temps, les recombinases se lient aux sites et forment un complexe synaptique regroupant tous les partenaires. Ensuite, elles catalysent le clivage de certaines liaisons phosphodiester spécifiques puis un échange de brins et la jonction des extrémités libres dans un nouvel arrangement. Ce mécanisme ne fait intervenir aucune hydrolyse. L'énergie des liaisons phosphodiester est conservée par une trans-estérification avec une chaîne latérale de la recombinase (Grindley et al., 2006). Un intermédiaire covalent ADN-recombinase est donc formé. Ces réactions présentent théoriquement un équilibre avec autant de substrat que de produit recombinant mais de nombreux systèmes existent qui déplacent l'équilibre vers le produit souhaité (Stark, 2015). Deux types de recombinases site-spécifiques mettent en jeu un résidu catalytique différent : les recombinases à tyrosine et les recombinases à sérine.

Les recombinases à sérine

Les recombinases à sérine représentent une famille protéique diverse dont tous les membres partagent un domaine catalytique d'environ 150 acides aminés (Stark, 2015). Ce domaine est majoritairement localisé à l'extrémité N-terminale de la protéine et comprend le résidu sérine catalytique ainsi que plusieurs motifs conservés (Grindley et al., 2006). Les recombinases à sérine présentent en outre différents domaines de taille variable pour la liaison spécifique à l'ADN. La recombinase à sérine « modèle » la plus étudiée est la résolvase du transposon γδ de *E. coli* (Stark, 2015). Les autres recombinases à sérine caractérisées sont la résolvase du transposon Tn3, l'invertase Hin de la Salmonelle, l'invertase Gin du phage Mu, et les intégrases des phages ϕC31, Bxb1 and ϕRv1 (Duyne, 2008). Les recombinases à sérine sont principalement trouvées chez les bactéries. Elles ont aussi été identifiées dans des transposons d'Archées (Filée et al., 2007).

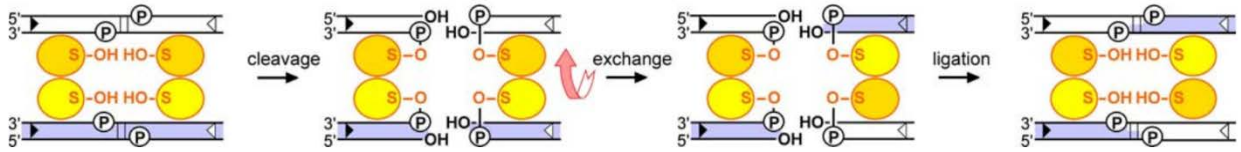


Figure 19. Modèle du mécanisme d'échange de brins catalysé par les recombinases à sérine. Les ronds jaunes représentent des monomères de recombinase. Les résidus sérine catalytique sont notés S-OH. Les liaisons phosphodiester cessibles sont représentées par un P encerclé. Figure tirée de Stark, 2015

La réaction catalysée par les recombinases à sérine se décompose en quatre étapes (Figure 19). La première étape correspond à la formation du complexe synaptique avec les quatre recombinases au centre et l'ADN à l'extérieur (Grindley et al., 2006; Stark, 2015). Au sein du complexe synaptique, les quatre recombinases sont activées. Lors de la deuxième étape, elles catalysent quatre clivages coordonnés qui aboutissent à une cassure double brin en quinconce. Le clivage a lieu par l'attaque nucléophile de liaisons phosphodiester par un résidu sérine. Le résidu sérine est alors lié avec l'extrémité 5' du brin d'ADN et les extrémités 3'OH sont libres. Les liaisons cessibles sont séparées par deux nucléotides sur chaque brin. Lors de la troisième étape, deux recombinases du complexe synaptique effectuent une rotation de 180°C qui conduit à un échange de brins. Enfin, la quatrième étape correspond à une religation. Les extrémités 3'OH attaquent la liaison protéine-ADN entrainant

la reformation de deux brins d'ADN continus. Dans de nombreux cas, la directionnalité de la réaction est contrôlée par une structure élaborée du complexe synaptique permise par des sites accessoires d'ADN et des domaines de la recombinase.

Les recombinases à tyrosine

Les recombinases à tyrosine représentent une famille très diverse qui partage un repliement et des résidus conservés dont la tyrosine catalytique dans le domaine catalytique (Esposito and Scocca, 1997; Grindley et al., 2006; Nunes-Düby et al., 1998). Le domaine catalytique ne présente en revanche aucune similarité globale de séquence. Il est situé à l'extrémité C-terminale de la protéine. Les recombinases à tyrosine contiennent également des domaines variés de liaison à l'ADN en position N-terminale. Elles sont présentent de manière ubiquitaire chez les archées, les bactéries et les eucaryotes mais ont été étudiées principalement chez les bactéries et aussi chez les eucaryotes.

Un modèle mécanistique séquentiel des recombinases à tyrosine a été proposé sur la base de l'étude de la recombinase Cre du phage P1 (Guo et al., 1997) (Figure 20). Un premier brin est d'abord traité puis le deuxième. Au début du modèle, le complexe synaptique est formé avec au minimum un tétramère de recombinases et les deux sites spécifiques (Grindley et al., 2006). Des sites accessoires et des protéines régulatrices de la directionnalité de réaction (RDF) peuvent également intervenir. Deux tyrosines nucléophiles portées par deux recombinases actives catalysent ensuite deux clivages à travers la formation d'un intermédiaire covalent 3' phosphotyrosine. Ces clivages ont lieu aux extrémités 5' d'une séquence « spacer » longue de 6 à 8 paires de base. Les extrémités 5'OH ainsi libérées attaquent la séquence d'ADN opposée entrainant la formation d'un intermédiaire de Holliday. Le complexe synaptique subi alors une isomérisation qui permet l'activation des deux autres recombinases. Une deuxième série de clivage et religation est ensuite catalysée de la même manière sur les deux brins restés intacts et donne le produit final de recombinaison. Le mécanisme des recombinases à tyrosine est similaire à celui des topoisomérases de type I (Cheng et al., 1998).

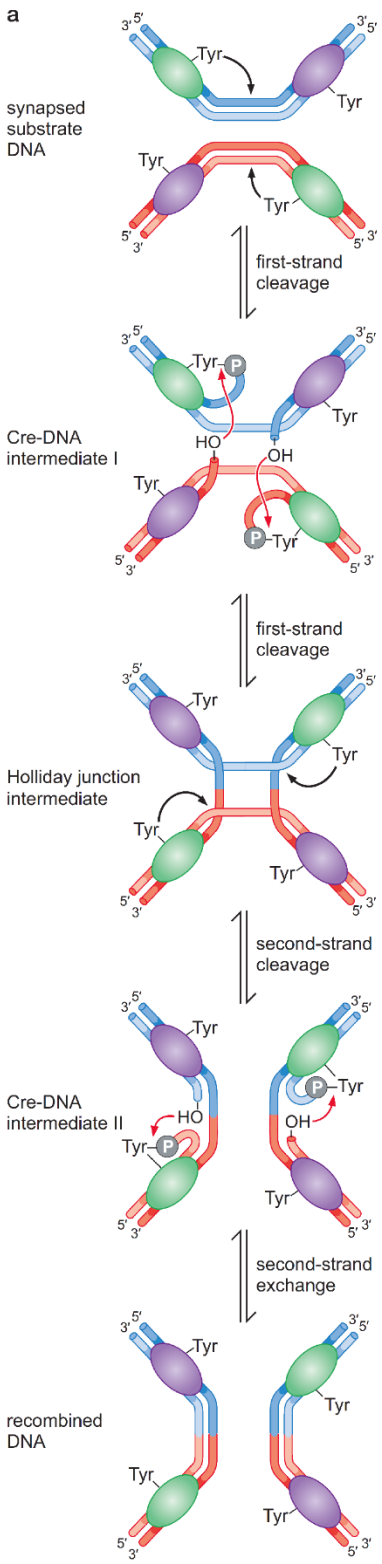


Figure 20. Modèle du mécanisme d'échange de brins catalysé par les recombinases à tyrosine. Les recombinases sont représentées par un ovale vert pour la conformation active et par un ovale violet pour la conformation inactive. Figure tirée de (Watson, 2014)

La recombinaison dans le contexte génétique

Différents produits de recombinaison

Les recombinaisons homologue et site-spécifique ont des effets différents suivant la localisation relative des deux séquences recombinantes et la topologie des molécules d'ADN (Figure 24). De manière générale, la recombinaison entre deux séquences conduit à la formation de deux séquences chimères (Figure 21.A). Lorsque les deux séquences sont portées par deux molécules distinctes dont au moins l'une des deux est de topologie fermée (circulaire), la recombinaison conduit à l'intégration des deux molécules en une seule (Figure 21.B). La molécule intégrée peut être linéaire ou circulaire suivant la topologie de la deuxième molécule recombinante. Les séquences recombinantes sont alors portées en orientation directe sur la même molécule. La recombinaison entre ces séquences conduit à une excision. Enfin, lorsque les deux séquences sont portées par une même molécule en orientation inverse, la recombinaison conduit à une inversion (Figure 21.C).

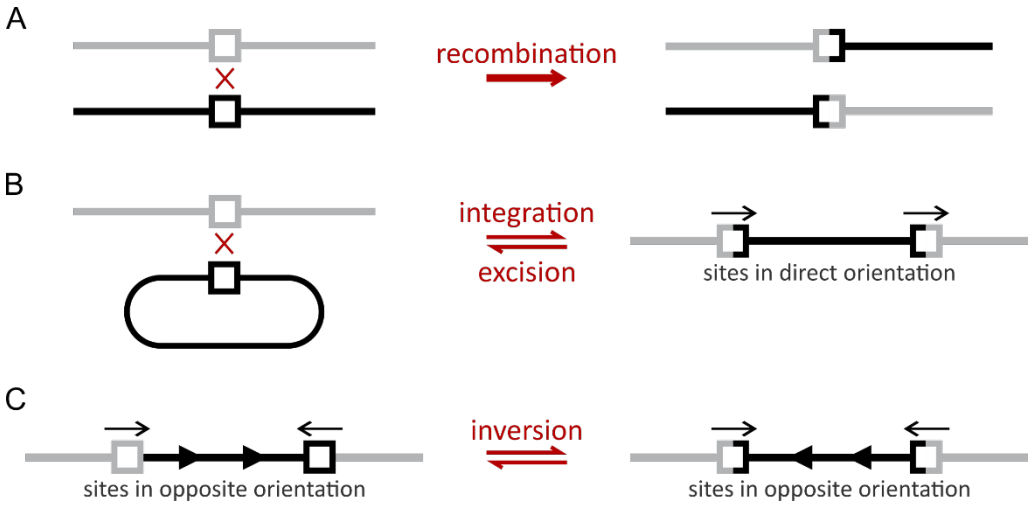


Figure 21. Les différents produits de recombinaison. Les carrés représentent les séquences recombinantes. A. La recombinaison intermoléculaire entre deux séquences recombinantes portées par des molécules linéaires aboutit à deux molécules linéaires chimères. B. La recombinaison intermoléculaire entre deux séquences dont l'une est portée par une molécule circulaire aboutit à une intégration avec les séquences recombinantes en orientation directe. Dans le sens contraire, la réaction intramoléculaire aboutit à une excision C. La recombinaison intramoléculaire entre deux séquences portées par une molécule d'ADN en orientation inverse aboutit à une inversion.

L'intégration et l'excision d'éléments génétiques mobiles

Ces différents arrangements de la recombinaison sont utilisés pour des différentes fonctions (Tableau 5, page 42). Par exemple, l'inversion permet des variations de phase en déplaçant les séquences promotrices d'un gène (Henderson et al., 1999). L'excision permet de résoudre des dimères de chromosomes (Castillo et al., 2017). Nous nous intéresserons ici aux fonctions d'intégration et d'excision d'éléments génétiques mobiles du chromosome de l'hôte. Ces fonctions constituent notamment l'entrée et la sortie de la phase lysogénique d'un virus (cf. Chapitre 2, page 25). Elles sont catalysées par des recombinases site-spécifiques à tyrosine et sérine ou par des recombinases à DDE (Figure 22).

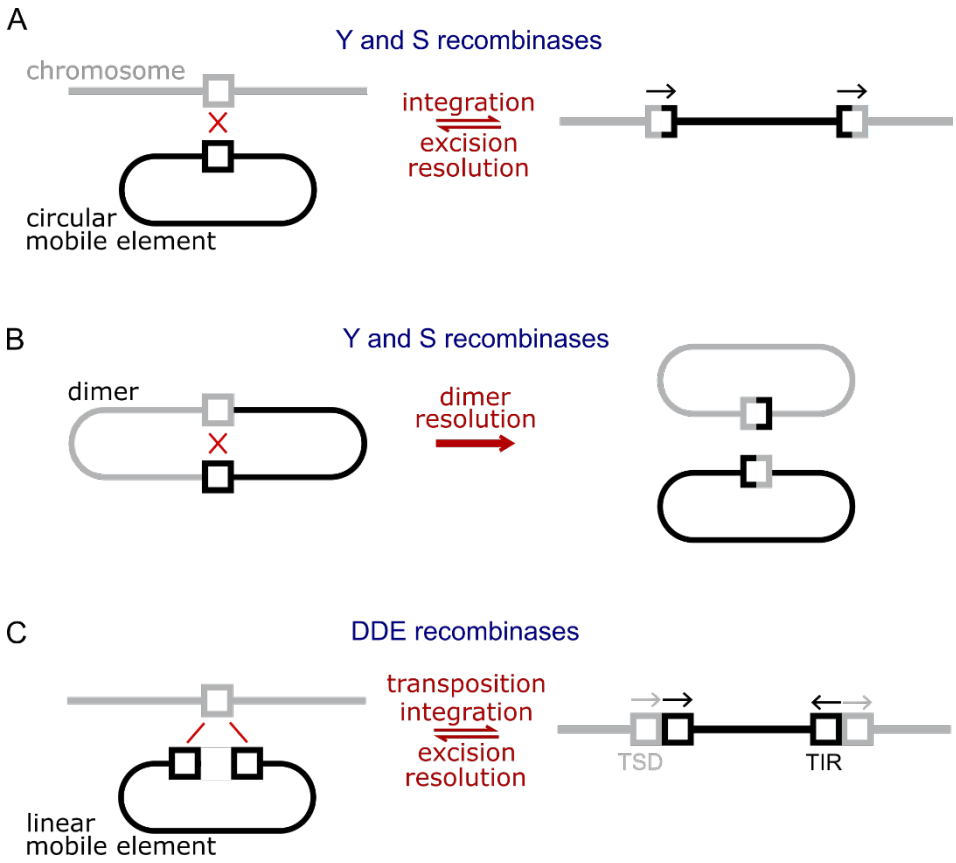


Figure 22. Les recombinases comme intégrases et résolvases d'éléments génétiques mobiles. Les carrés représentent les séquences recombinantes. Le chromosome gris peut être circulaire mais paraît linéaire à l'échelle de l'élément génétique mobile. A. Les recombinases à tyrosine et sérine catalysent l'intégration et l'excision (ou résolution) d'éléments génétiques mobiles présentant un intermédiaire circulaire (plasmide, virus ou élément transposable). B. Les recombinases à tyrosine catalysent la résolution de dimères d'éléments génétiques mobiles circulaires. C. Les recombinases à DDE catalysent l'intégration (ou transposition) et l'excision (ou résolution) d'élément mobiles génétiques présentant un intermédiaire linéaire (principalement des rétrovirus, des rétrotransposons et des transposons). L'élément génétique mobile est encadré par des répétitions inversées terminales (TIR, carrés noirs). L'intégration aboutit à la duplication du site cible gris (TSD) en orientation directe.

Les recombinases à tyrosine et sérine catalysent l'intégration et l'excision de virus à ADN double-brin circulaire et de plasmides chez les archées, les bactéries et dans une moindre mesure chez les eucaryotes (Figure 25.A). Par exemple, l'intégration et l'excision du phage λ d'*E. coli* sont catalysées par sa recombinase à tyrosine (Landy, 2015). Les recombinases à tyrosine et sérine catalysent aussi l'intégration d'intermédiaires circulaires de transposition pour certains transposons (Filée et al., 2007; Poulter and Butler, 2015). Par exemple, l'intégrase d'intégrons catalyse l'intégration de nouvelles cassettes circulaires dans un intégron (Escudero et al., 2016). Finalement, les recombinases à tyrosine et à sérine sont impliquées dans une activité d'excision (ou résolution) d'un dimère ADN circulaire en deux monomères (Figure 25.B). Par exemple, la recombinase Cre catalyse la résolution de dimères du génome circulaire du bactériophage P1 par recombinaison spécifique au site lox (Duyne, 2015). Ce mécanisme est similaire à celui de résolution des dimères de chromosomes circulaires bactériens et archéens catalysé par les recombinase à tyrosine Xer par recombinaison site-spécifique entre deux séquences *dif* (Barre and Midonet, 2015). Certains éléments génétiques mobiles bactériens utilisent les recombinases hôtes XerC et XerD pour catalyser leur intégration dans la chromosome hôte au niveau du site dif par une recombinaison à spécificité relâchée (Barre and Midonet, 2015). Cette

exploitation du système de recombinaison site-spécifique de l'hôte n'a jamais été observé chez les archées.

Les recombinases à DDE catalysent l'intégration de phages bactériens et de virus eucaryotes à génome linéaire (Figure 25.C). Par exemple, chez l'homme, l'intégration du virus VIH est catalysée par une intégrase à DDE virale (Craigie and Bushman, 2012). L'intégration du phage Mu est similairement catalysée par sa recombinase à DDE MuA (Chaconas and Harshey, 2002). MuA catalyse aussi la transposition du phage Mu à l'intérieur du chromosome hôte. Cette transposition conserve la copie originale du phage Mu si elle est couplée à la réplication du chromosome. La formation de virion est généralement contrôlée par la copie intégrée du génome et la recombinase à DDE ne catalyse pas l'excision du virus. Aucun virus d'archées codant pour une intégrase à DDE n'a été identifié jusqu'à présent. Les recombinases à DDE catalysent également la transposition d'éléments transposables chez les archées, les bactéries et les eucaryotes (Curcio and Derbyshire, 2003; Hickman and Dyda, 2015b). Certains mécanismes de transposition ne conservent pas la copie originale du transposon qui est alors excisée.

Les recombinases à tyrosine en trois enzymes

Le mécanisme des recombinases à tyrosine a principalement été élucidé grâce à trois recombinases modèles : la recombinase Cre du phage P1 (Duyne, 2015), l'intégrase du phage λ (Landy, 2015) et la flipase Flp du plasmide 2µ de la levure (Jayaram et al., 2015).

La recombinase Cre du phage P1, un système simple

La recombinase Cre du phage P1 constitue un système simple de recombinaison site-spécifique (Van Duyne, 2005). Il est seulement constitué de la recombinase Cre et de son site spécifique *loxP*. La recombinase Cre est composée de deux domaines NTD et CTD formant une pince en forme de C qui attrape l'ADN par les deux côtés (Figure 23.A) (Van Duyne, 2001). La recombinase est monomérique en solution et forme des dimères en présence d'ADN grâce à la liaison coopérative au site *loxP*. La recombinase présente 7 résidus importants pour son activité catalytique qui sont conservés chez les autres recombinases à tyrosine (Duyne, 2015). La tyrosine catalytique est parfaitement conservée et nécessaire à l'activité. Une lysine est également nécessaire pour stabiliser l'extrémité 5'OH libérée. Deux résidus arginine et un résidu histidine stabilisent les intermédiaires de réactions. Enfin, un tryptophane et un acide glutamique sont conservés pour des raisons structurales.

Le site loxP a une longueur de 34 paires de bases et est constitué de deux séquences RBE (élément de liaison de recombinase) séparées par une séquence centrale de 6 paires de bases (Figure 23.B) (Duyne, 2015). Les séquences RBE correspondent à des répétitions inversées presque parfaites. Chaque séquence RBE est liée par une recombinase. La séquence centrale ne présente aucune symétrie. Son orientation permet donc de définir la position gauche ou droite des séquences RBE et l'orientation directe ou opposée des deux sites de recombinaison dans le cas d'une recombinaison intramoléculaire. Dans le complexe synaptique, l'alignement des deux sites *loxP* se fait de manière antiparallèle avec l'extrémité gauche du premier site faisant face à l'extrémité droite du deuxième site (Duyne, 2015). Le clivage a lieu aux extrémités de la séquence centrale.

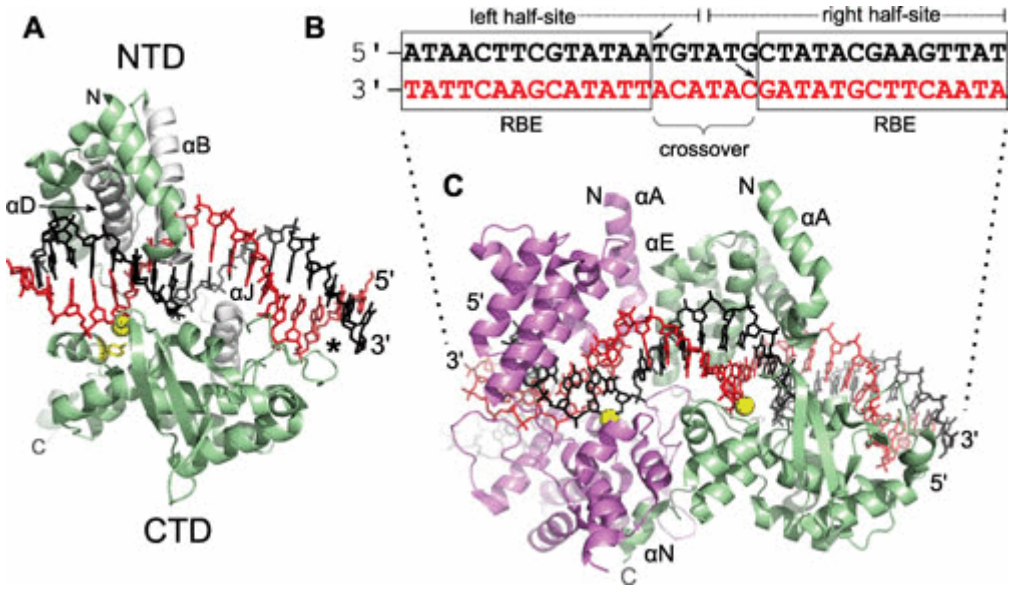


Figure 23. La recombinase Cre liée à son site spécifique *loxP*. A. Monomère de Cre lié à un demi-site *loxP*. NTD : domaine N-terminale. CTD : domaine C-terminal. Les boules jaunes indiquent les phosphates cessibles. B. Séquence du site *loxP*. RBE : élément de liaison à la recombinase. Les flèches indiquent les sites de clivage. C. Dimère de Cre lié à un site *loxP*. Un monomère est coloré en vert, l'autre en rose. Image tirée de Duyne, 2015

Dans le contexte du phage P1, la recombinase Cre catalyse une recombinaison intramoléculaire entre deux sites *loxP* en orientation directe qui permet la résolution d'un dimère du génome du phage (Duyne, 2015). Elle est aussi capable de réaliser la réaction inverse d'intégration en absence de cofacteur additionnel. L'équilibre de la réaction est en revanche fortement biaisé vers l'excision.

Aspects conservés des recombinases à tyrosine

Le système de recombinaison Cre ainsi décrit correspond à « un système de base » conservé pour les autres recombinases à tyrosines. Ainsi, toutes les tyrosines recombinases présentent les 7 résidus importants et la même structure globale que Cre, notamment au niveau du site catalytique (Esposito and Scocca, 1997; Grainge and Jayaram, 1999; Nunes-Düby et al., 1998). Les sites spécifiques de recombinaison présentent également au minimum la même ossature que le site loxP : deux répétitions inversées auxquelles se lient la recombinase et séparées par une séquence de 6 à 8 paires de bases orientée et à l'extrémité de laquelle ont lieu les clivages simple-brins. Des domaines protéiques et/ou séquences nucléiques additionnelles viennent dans certains cas compléter la recombinase ou le site spécifique, respectivement.

L'intégrase du phage lambda et le contrôle de la directionnalité de réaction

Le système de recombinaison du phage λ est similaire à celui du phage P1 mais présente une complexité supplémentaire avec un mécanisme de contrôle de la directionnalité de recombinaison (Van Duyne, 2005). L'intégrase λ présente une partie C-terminale (CTD) dont la structure à deux domaines est similaire à celle de la recombinase Cre (Figure 24). Cette partie CTD permet la liaison de l'intégrase au site spécifique « core » long de 21 paires de bases. Comme pour le site *loxP,* le site « core » est composé de deux répétitions inversées séparées par 7 paires de bases qui subissent l'échange de brin (Azaro and Landy, 2002). L'intégrase λ présente un domaine N-terminal (NTD) supplémentaire relié à la partie CTD par un linker (Figure 24). Ce domaine se lie à des séquences spécifiques situées à proximité du site-spécifique « core » (Azaro and Landy, 2002). Le système de recombinaison du phage λ inclut également des protéines RDF de liaison et courbure de l'ADN : les protéines IHF et Fis encodées par le chromosome hôte et la protéine virale Xis. Elles possèdent des sites de liaison spécifique à l'ADN à proximité du site « core ».

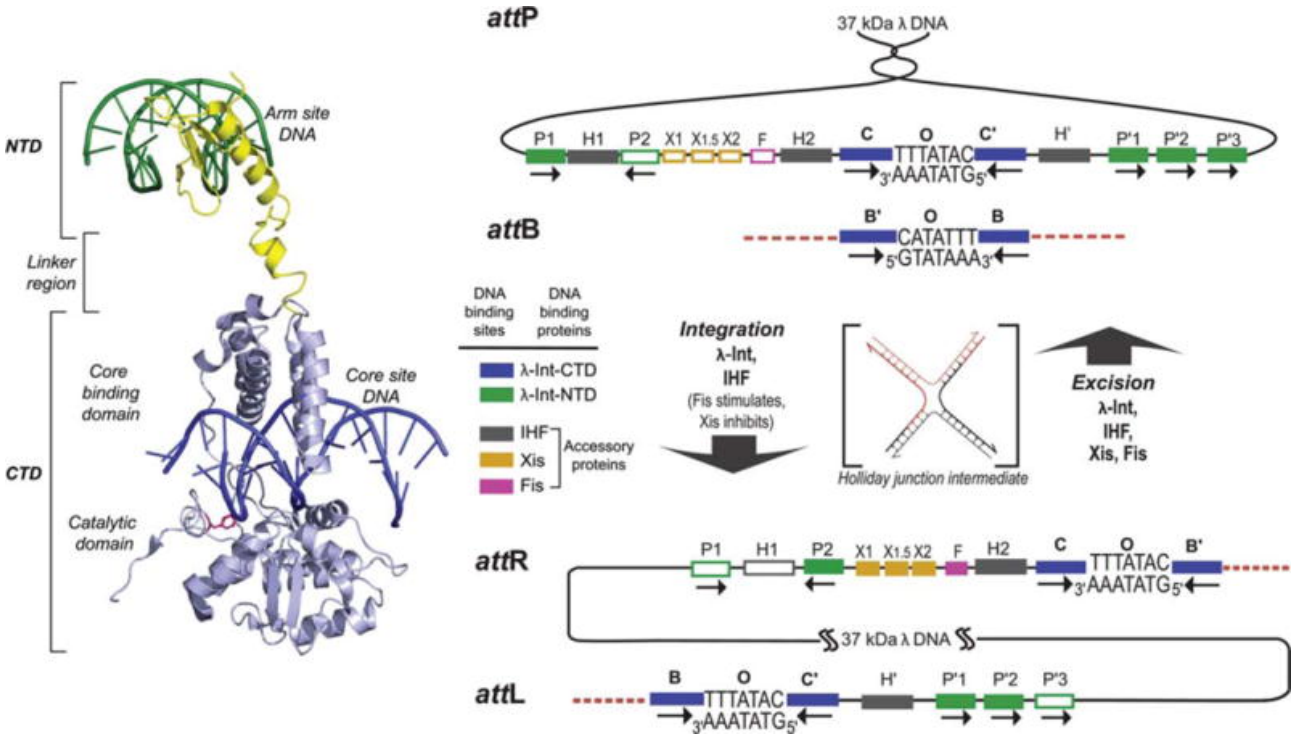


Figure 24.Structure de l'intégrase λ et de ses sites spécifiques. Le panneau de gauche présente la structure d'un monomère d'intégrase λ lié à l'ADN. Le panneau de droite présente les réarrangements générés par les réactions de recombinaisons catalysées par l'intégrase λ. La recombinaison intégrative entre un site attP et un site attB se produit en présence de l'intégrase et de la protéine structurale hôte IHF. Le produit d'intégration est un génome phage intégré dans le chromosome hôte et flanqué par les séquences attL et attR. Le génome phagique libre est reconstitué par la recombinaison entre les sites attL et attR en présence de l'intégrase, de la protéine structurale hôte IHF et des protéines structurales phagiques Xis et Fis. Les rectangles coloriés correspondent aux sites de liaison spécifique des protéines utilisés pour la réaction décrite, les rectangles blancs ne sont pas utilisés. Figure tirée de Landy, 2015

La directionnalité de la réaction de recombinaison est contrôlée par la disposition des différentes séquences de liaison à l'ADN (Figure 24). Le site spécifique d'intégration « vide » du chromosome est appelé attB (attachement à la bactérie) et est composé seulement du site « core ». Le site spécifique d'intégration du phage est appelé attP (attachement du phage) et est composé du site « core » entouré par deux bras correspondants aux sites de liaison au domaine NTD de l'intégrase et aux trois protéines

IHF, Fis et Xis. Cette disposition permet l'intégration du génome phage dans le chromosome hôte en présence de l'intégrase et de la protéine IHF. Le génome phage est alors flanqué par les séquences attL à gauche et attR à droite. Chacune de ces séquences est composée par un site « core » et un bras de sites de liaison. Cette disposition permet l'excision du génome phage en présence de l'intégrase et des protéines IHF, Fis et Xis.

Pour toutes ces dispositions, l'intégrase λ ne peut catalyser la recombinaison que si elle est liée à la fois aux deux sites « core » et aux sites des bras. L'intégrase seule ne peut pas faire ces liaisons pour des raisons structurales (Landy, 2015). La liaison des protéines IHF, Fis et Xis à leurs sites spécifiques des bras entraine une courbure de l'ADN qui permet alors à l'intégrase de se lier aux sites « core » et bras (Figure 25). Suivant la disposition des sites de liaison, des protéines différentes réalisent la courbure. Notamment, Xis est utilisée seulement pour l'excision. La régulation de la quantité produite de Xis et de l'intégrase permet en conséquence de réguler l'excision (Casjens and Hendrix, 2015).

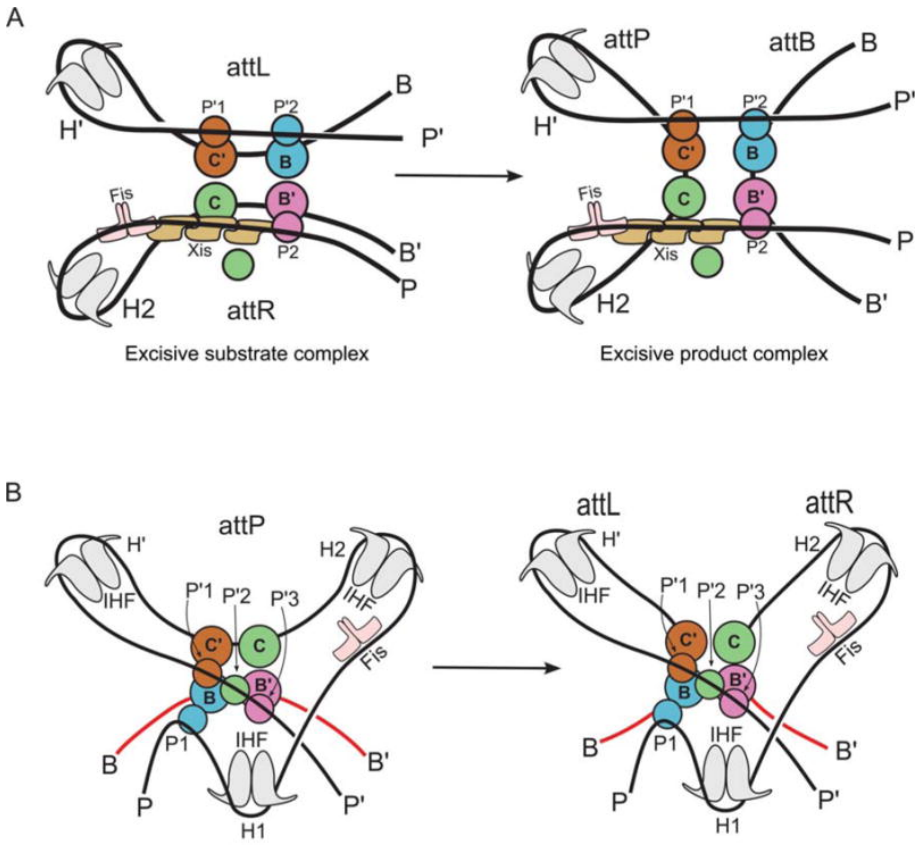


Figure 25. Représentation schématique de la structure du complexe synaptique de λ lors des réactions d'excision (A) et d'intégration (B). Chaque monomère d'intégrase est représenté par deux cercles colorés de la même couleur. Le grand cercle correspond au domaine C-terminal, le petit cercle correspond au domaine N-terminal de liaison aux bras. Les sites nucléiques de liaison des protéines sont indiqués par une majuscule, éventuellement suivie d'un numéro. Figure tirée de (Landy, 2015)

Dans le cas du phage λ, la séquence attB est située entre les gènes gal et bio (Azaro and Landy, 2002). D'autres intégrases présentent également un site attB intergénique (Bobay et al., 2013; Krupovic et al., 2019). La majorité des intégrases présente en revanche un site attB intragénique, en particulier dans les gènes codant pour les ARNt (Reiter et al., 1989; Williams, 2002). Plusieurs hypothèses ont été proposées pour expliquer le ciblage prépondérant des ARNt (Williams, 2002) : (1) Les ARNt présentent des répétition inversées séparées par une séquence orientée qui sont fonctionnellement nécessaires pour les recombinases à tyrosine. (2) Les ARNt sont conservés dans le temps limitant le risque de disparition du site d'intégration. (3) Les ARNt sont conservés entre espèces proches et permettent donc d'intégrer une large gamme d'hôtes. (4) Les gènes codant pour les ARNt sont une grande famille multigénique qui offre de nombreux sites d'intégrations différents avec un nombre limité de nucléotides différents (Winckler et al., 2005).

La flipase Flp du plasmide 2µ assemble le site catalytique en trans

La recombinase eucaryote Flp du plasmide 2µ de *Saccharomyces cerevisiae* présente une particularité structurale. Pour les deux recombinases citées précédemment, un même monomère fourni les résidus nécessaires à l'activation du phosphate cessible et la tyrosine responsable de l'attaque nucléophile. Pour la recombinase Flp, la tyrosine catalytique est fournie en trans par un autre monomère (Chen and Rice, 2003). L'assemblage du site actif se fait donc à l'interface entre deux monomères. D'autres recombinases semblent assembler leur site actif en trans comme l'intégrase du virus SSV1 de *Sulfolbus shibatae* (Eilers et al., 2012).

Les intégrases d'archées

Diversité des intégrases archées

Chez les archées, nous avons identifié de nombreuses recombinases à tyrosine encodées par des plasmides, des virus et les versions intégrées au chromosome de ces éléments génétiques mobiles (Atanasova et al., 2018; Goodman and Stedman, 2018; Krupovic et al., 2019; Wang et al., 2018). Elles présentent les 7 résidus conservés des recombinases à tyrosines bactériennes et eucaryotes ou des résidus aux propriétés équivalentes (She et al., 2004)

Deux types d'intégrases d'archées ont été définis par She et al. en 2002 (She et al., 2002). La plus grande partie des intégrases sont de type II (Figure 26). Elles présentent un système d'intégration classique avec un gène codant pour l'intégrase porté par l'élément génétique mobile et les sites de recombinaison attP et attB portés par l'élément génétique mobile et le chromosome, respectivement. Parmi les intégrases de type II, on trouve notamment les intégrases de plasmides conjugatifs de type pNOB8 et des intégrases phylogénétiquement proches des recombinases XerA. L'autre système d'intégration (type I) a été jusqu'à présent exclusivement détectée pour des archées hyperthermophiles. Le site de recombinaison attP est localisé dans la séquence codant pour l'intégrase (Figure 26). Après intégration, le gène de l'intégrase est donc séparé en deux parties *int(N)* et *int(C)*. *int(N)* code la partie N-terminale représentant environ un quart de l'intégrase entière et est sous le contrôle du promoteur de l'intégrase. *int(C)* code la partie C-terminale représentant environ les trois quarts de l'intégrase entière dont la partie catalytique. *int(C)* ne présente a priori aucun promoteur pour son expression. Il a été remarqué à plusieurs reprises que les éléments mobiles ainsi intégrés ne présentent pas d'excision spontanée (She et al., 2001a; Wang et al., 2007). Cela laisse penser que la

fragmentation inactive l'intégrase. Pour les intégrases de type I, l'intégration correspond donc à un suicide fonctionnel et l'élément intégré se retrouve piégé dans le chromosome. Nous proposons donc de renommer le type I comme intégrases suicidiaires. Des intégrases suicidaires intactes ont été détectées chez les fusellovirus de *Sulfolobus* (Goodman and Stedman, 2018), le virus TPV1 de *Thermococcus* (Gorlas et al., 2012) et les plasmides pT26-2 et pTN3 de *Thermococcus* (Gaudin et al., 2014; Soler et al., 2010). Des intégrases fragmentées ont été détectées dans des chromosomes des Sulfolobales et de Thermococcales (She et al., 2001b; Soler et al., 2010). Les intégrases suicidaires ont été peu étudiées malgré leur découverte il y a une trentaine d'année. Leur activité singulière reste pour l'instant mystérieuse.

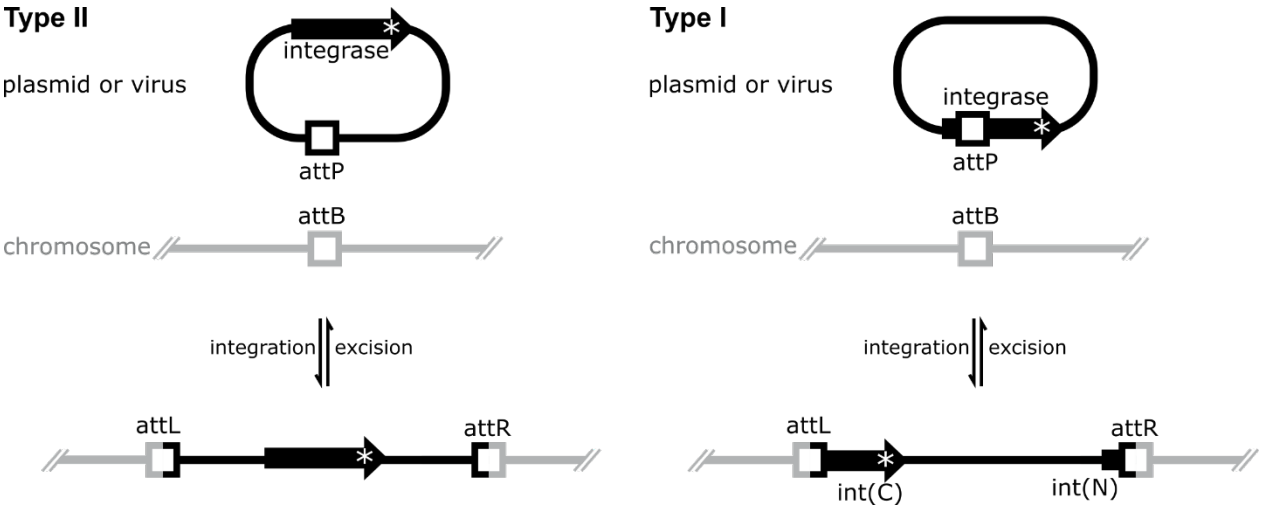


Figure 26. Modèles d'intégration pour les deux types d'intégrases d'archées. Les carrées correspondent aux sites spécifiques de recombinaison. La flèche correspond au gène codant pour l'intégrase et l'astérisque au codon codant pour la tyrosine catalytique. Pour les intégrases de type I, le site attP est situé à l'intérieur de la séquence codant pour l'intégrase. En conséquence, le gène codant pour l'intégrase est séparé en deux parties *int(N)* et *int(C)* après intégration.

Seulement quelques intégrases à tyrosine ont été caractérisées parmi les plusieurs centaines dont les séquences sont connues chez les archées : l'intégrase du fusellovirus SSV1 (Eilers et al., 2012; Letzelter et al., 2004; Serre et al., 2002; Zhan et al., 2012), l'intégrase du fusellovirus SSV2 (Zhan et al., 2015) et l'intégrase du pléolipovirus SNJ2 (Wang et al., 2018). L'activité des recombinases XerA de *Pyrococcus abyssi* et de *Thermoplasma acidophilum* a aussi été charactérisée (Cortez et al., 2010; Jo et al., 2016, 2017; Serre et al., 2013). Nous allons maintenant présenter ces résultats.

Les intégrases des fusellovirus de Sulfolobus

Structure de l'intégrase de SSV1

Le virus SSV1 de *Sulfolobus shibatae* est présent pendant l'infection à la fois sous forme plasmidique (Martin et al., 1984) et sous forme d'un provirus intégré au chromosome hôte (Schleper et al., 1992). La production de virion peut être induite par un traitement UV ou à la mitomycine C sans entrainer l'excision du provirus. La libération des virions se produit alors sans lyse. (Liu and Huang, 2002; Schleper et al., 1992). Le provirus est intégré au niveau d'un gène codant pour un $ARNt^{Arg}$ (Reiter et al., 1989). Le système de recombinaison est composé d'une intégrase Int^{SSV1}de type I (suicidaire) et de sites attB et attP identiques, correspondant à la moitié 3' de l'ARNt, tiges anti-codon inclues, et longs de 44 paires de bases. L'activité de recombinaison *in vitro* de l'intégrase Int^{SSV1} (Muskhelishvili et al., 1993) n'a pas pu être démontrée de manière reproductible (Serre et al., 2002; Zhan et al., 2012). En revanche, Int^{SSV1} présente une activité de coupure simple brin *in vitro* (Serre et al., 2002) qui a permis de localiser le site de coupure aux extrémités de la boucle anti-codon de l'ARNt. Cette localisation correspond aux extrémités d'un spacer entre deux inversions répétées et est cohérente avec ce qui est connu pour les recombinases à tyrosines de bactéries.

Le rôle des différents domaines de l'intégrase Int^{SSV1}a été déterminé. La moitié N-terminale est nécessaire à la multimérisation (Zhan et al., 2012). La moitié C-terminale contient les 7 résidus conservés des tyrosines recombinases qui sont nécessaires à l'activité catalytique de coupure (Letzelter et al., 2004). La structure de la moitié C-terminale de Int^{SSV1} (160/335 résidus) a été résolue (Eilers et al., 2012; Zhan et al., 2012). Le repliement du domaine catalytique est similaire à celui observé pour les autres recombinases à tyrosines. La position de la tyrosine catalytique et des tests de complémentation d'activité indiquent un assemblage du site actif en trans (Eilers et al., 2012; Letzelter et al., 2004). La structure de l'intégrase entière et notamment de la partie N-ter restent inconnues.

Activité de l'intégrase de SSV2

Le fusellovirus SSV2 de *Sulfolobus islandicus* REY15 (Stedman et al., 2003) est caractérisé par une production spontanée de virion sans lyse de la cellule infectée (Contursi et al., 2006). Le virus est présent sous forme épisomale et comme provirus intégré dans le chromosome de l'hôte au niveau d'un $ARNt^{Gly}CCC$ (Contursi et al., 2006). Similairement à SVV1, le système de recombinaison est composé d'une intégrase Int^{SSV2} de type I (suicidaire) et de sites attB et attP identiques, correspondant à la moitié 3' de l'ARNt, tiges anti-codon inclues, et longs de 49 paires de bases. L'intégrase Int^{SSV2} catalyse des recombinaisons sites-spécifiques *in vitro* (Zhan et al., 2015). La partie Int(N) de Int^{SSV2} permet sa tétramerisation, les 80 premiers acides aminés de Int(C) permettent la liaison spécifique à l'ADN et le reste de la séquence Int(C) catalyse la recombinaison (Zhan et al., 2015). La partie Int(C) est capable de catalyser l'intégration et l'excision du virus seule *in vitro* mais avec une efficacité plus faire que l'intégrase entière. Cette activité remet en question la notion de suicide des intégrases lors de leur intégration, en contradiction avec l'absence d'excision observée *in vivo*.

Une coopération entre l'intégrase du virus SSV2 et le plasmide pSSVi a été observée en laboratoire (Wang et al., 2007). Le plasmide pSSVi est intégré dans le chromosome de la souche *Saccharolobus solfataricus* P2 au niveau d'un $ARNt^{Arg}$ (Wang et al., 2007). L'intégration a probablement été catalysée par l'intégrase de type I (suicidaire) présente en deux parties dans le plasmide intégré. pSSVi est effectivement bloqué dans le chromosome à cause du suicide de son intégrase qui ne peut plus catalyser l'excision. En laboratoire, l'infection de *Saccharolobus solfataricus* P2 par SSV2 a entrainé

l'excision du plasmide pSSVi (Wang et al., 2007). Il semblerait donc que l'introduction d'une intégrase relativement proche permette la résurrection d'un élément mobile intégré par une intégrase suicide.

L'intégrase du pléolipovirus SNJ2

Le pléolipovirus SNJ2 infecte l'haloarchée *Natrinema sp*. J7-1 (Liu et al., 2015). SNJ2 produit peu de virions, sauf en cas de co-infection de l'hôte avec le virus SNJ1. SNJ2 est aussi capable de s'intégrer dans le chromosome hôte au niveau d'un gène codant pour un $ARNt^{Met}$. Le système de recombinaison est composé d'une intégrase de type II (classique) et de sites attB et attP identiques, sans inversions répétées et longs de 14 paires de bases (Liu et al., 2015). Il a été montré que l'intégrase est capable de catalyser l'intégration et l'excision du virus *in vivo* (Wang et al., 2018). De plus, le gène de l'intégrase forme un opéron avec deux autres gènes dont les produits augmentent l'efficacité de recombinaison. Ces analyses ont été réalisées sur des séquences de recombinaison plus larges que les 14 paires de bases de attB et attP. On ne sait donc pas si ces sites courts et sans inversions répétées sont suffisants pour la recombinaison. Finalement, des provirus ont été détectés dans de nombreux chromosomes d'haloarchées avec une intégrase similaire (Liu et al., 2015).

Chapitre 4. Evolution génomique des archées Thermococcales

Evolution des génomes et éléments génétiques mobiles

Les mécanismes de modification des génomes peuvent être groupés en deux grandes catégories. Premièrement, les mutations ponctuelles correspondent à des insertions, délétions ou modifications d'une ou de quelques bases. Elles permettent de créer et d'optimiser une séquence nucléique pour une fonction particulière de manière très précise mais lente. Deuxièmement, les réarrangements correspondent à des insertions, délétions ou inversions de grandes séquences préexistantes (à l'échelle de milliers de paires de bases). Les réarrangements permettent de rapidement gagner de nouvelles fonctions, perdre des fonctions inutiles ou générer de nouvelles organisations génomiques. Une balance existe entre l'organisation rigoureuse du chromosome résultant de contraintes variées (cf. Chapitre 1 page 10) et l'apparition de réarrangements qui désorganisent le génome mais peuvent apporter des innovations évolutives (Touchon and Rocha, 2016)

Les éléments génétiques mobiles sont reconnus comme des forces évolutives majeures par deux aspects (Koonin Eugene V., 2016). Premièrement, la course aux armements entre l'hôte et ses éléments génétiques mobiles constitue une force évolutive puissante. Deuxièmement, les éléments génétiques mobiles modifient directement le génome de l'hôte en créant des réarrangements génomiques. Les éléments génétiques mobiles fournissent notamment à leur hôte des séquences nucléiques fonctionnelles personnelles ou provenant d'un autre individu. C'est ce qu'on appelle le transfert horizontal de gènes entre un élément génétique mobile et un chromosome ou ente deux chromosomes médié par un élément génétique mobile (Frost et al., 2005). Cela permet la dispersion de fonctions utiles dont les séquences sont intégrées dans le chromosome hôte grâce à une recombinase à site-spécifique encodée par l'élément génétique mobile ou par recombinaison homologue. Les éléments génétiques mobiles, principalement les éléments transposables, entrainent aussi des inversions et des délétions dans le chromosome hôte (Bourque et al., 2018). L'intégration multiple d'un même élément génétique mobile dans le chromosome fournit des régions d'homologie qui peuvent être utilisées comme substrat pour la recombinaison homologue. Suivant l'orientation des répétitions, la recombinaison aboutit à une inversion ou délétion en cas de crossing-over (cf. Chapitre 3 page 43). Certains mécanismes de transposition peuvent aussi directement générer des inversions et délétions (Curcio and Derbyshire, 2003). Finalement, les éléments génétiques mobiles entrainent des interruptions de séquences lors de leur intégration qui peuvent désactiver ou déréguler un gène (Dubin et al., 2018). Les modifications engendrées directement par les éléments génétiques mobiles sont parfois défavorables ou létales et s'éteignent en conséquence. Ces modifications sont au contraire parfois favorables dans l'environnement immédiat (modification adaptative) et subsistent.

Les chromosomes de Thermococcales subissent de nombreux réarrangements

Dès l'obtention des trois premières séquences de génome de Thermococcales (Lecompte et al., 2001; Zivanovic et al., 2002), la présence d'un fort niveau de réarrangements était évidente. Cette tendance a ensuite été largement confirmée par le séquençage de souches supplémentaires (Bridger et al., 2012; Cossu et al., 2015; Fukui et al., 2005; White et al., 2008; Zivanovic et al., 2009). Le nombre de réarrangements est tel qu'on ne peut dans certains cas pas détecter de segments synténiques entre deux chromosomes présentant par ailleurs une forte identité de séquence (Cossu et al., 2015; Zivanovic et al., 2009). Ces réarrangements correspondent pour partie à des intégrations et excisions d'éléments génétiques mobiles mais principalement à des inversions chromosomiques (Cossu et al., 2015; White et al., 2008). Par exemple, les chromosomes de *Pyrococcus horikoshii* et *Pyrococcus abyssi* présentent de longs segments synténiques avec des orientations dans les deux diagonales sur la réprésentation dotplot (Figure 27.A) Chaque rupture d'orientation correspond à une inversion. Un grand nombre d'inversions mélange tellement le chromosome qu'il n'existe plus de segments synténiques avec les souches proches.

Pour *Pyrococcus furiosus*, les réarrangements ont pu être liés à la présence de nombreuses séquences IS (Bridger et al., 2012; Zivanovic et al., 2002). Notamment, la souche de laboratoire *Pyrococcus furiosus* COM1 présente plusieurs inversions et excisions par rapport à la souche type et parentale *Pyrococcus furiosus* DSM 3638 (Bridger et al., 2012). Ces réarrangements sont clairement liés à des séquences IS qui se trouvent à leurs extrémités. Ce processus est toujours en cours avec une nouvelle transposition détectée dans COM1 et qui a entrainé une inversion. En revanche, les autres espèces de Thermococcales ne présentent qu'une très faible quantité de séquences IS qui ne peuvent donc pas expliquer les nombreux réarrangements observés (Fukui et al., 2005; Zivanovic et al., 2002, 2009). Il a été proposé que ces réarrangements pourraient être dus à l'environnement mutagène des cheminées hydrothermales (White et al., 2008; Zivanovic et al., 2009) mais aucune preuve expérimentale n'appuie cette hypothèse. Le mécanisme des nombreux réarrangements des Thermococcales reste non élucidé.

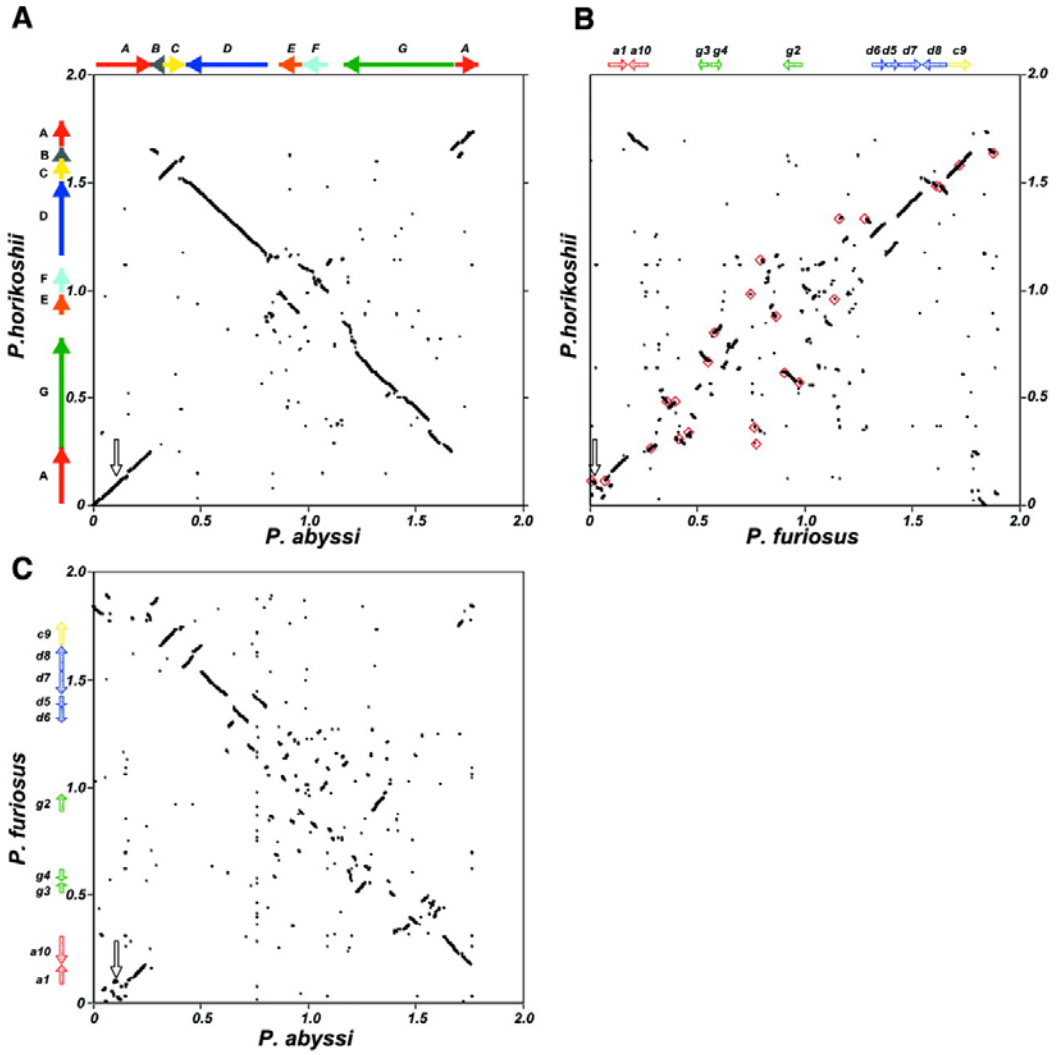


Figure 27. Représentation « dotplot » de la comparaison deux à deux des génomes de *Pyrococcus horikoshii*, *Pyrococcus abyssi* et *Pyrococcus furiosus*. Chaque axe représente le génome de l'espèce indiquée. Chaque point représente un segment de 100 nucléotides significativement similaire entre les deux génomes. La flèche blanche verticale correspond à la position de l'origine de réplication *oriC*. La numérotation en mégabases correspond à la séquence de référence. A. Les flèches correspondent à des segments synténiques entre les deux génomes. B et C. Les flèches correspondent aux segments conservés entre les trois génomes. Les carrées rouges indiquent la localisation des séquences IS de *Pyrococcus furiosus*. Figure tirée de (Zivanovic et al., 2002)

Questionnements, objectifs et stratégies

Mes travaux de thèse s'inscrivent dans une thématique majeure de l'équipe « Biologie cellulaire des archées » qui s'intéresse à l'influence des éléments génétiques mobiles sur l'évolution génomique des archées hyperthermophiles. Les éléments génétiques mobiles et leurs hôtes interagissent en particulier par le biais des intégrases à tyrosine plasmidiques et virales. Mes travaux de thèse ont principalement porté sur ces enzymes à travers deux questions détaillées ci-dessous (Parties 1 à 3 et 5). De plus, face à certaines lacunes dans les outils bioinformatiques disponibles pour l'étude des éléments génétiques mobiles, j'ai participé au développement d'une nouvelle banque de données comprenant tous les plasmides naturels des trois domaines du vivant (Partie 4).

Quels sont les mécanismes d'inversions génomiques chez les Thermococcales ?

Les chromosomes de Thermococcales subissent de nombreuses inversions génomiques qui ont probablement des conséquences évolutives majeures puisqu'elles détruisent l'organisation ordonnée du chromosome (cf. Chapitre 4). Ces inversions ont été évidentes dès la comparaison des trois premiers génomes Thermococcales (Zivanovic et al., 2009). Depuis une dizaine d'années, le séquençage de nombreuses souches supplémentaires a confirmé cette tendance mais n'avait jusqu'à présent pas permis d'en élucider le mécanisme. Cette élucidation a motivé mon travail de thèse. A mon arrivée au laboratoire, une nouvelle séquence chromosomique de Thermococcales avait été obtenue. Elle correspond à la souche *Thermococcus nautili* 66G dont l'ADN avait été ré-extrait plusieurs années après un premier séquençage. Remarquablement, la comparaison des deux séquences du chromosome de *Thermococcus nautili* a révélé plusieurs réarrangements. La présence dans *Thermococcus nautili* de trois plasmides dont le plasmide pTN3 codant une recombinase à tyrosine (l'intégrase Int^{pTN3}) a conduit à proposer un modèle pour expliquer les inversions chromosomiques reposant sur l'activité de cette intégrase. Mon travail a alors consisté à décortiquer les réarrangements observés dans le chromosome de *Thermococcus nautili* et à tester si l'intégrase Int^{pTN3} pouvait les catalyser (Partie 1).

J'ai démontré que Int^{pTN3} possède une activité catalytique double : une première activité de recombinaison site-spécifique ordinaire pour une intégrase et une deuxième activité de recombinaison entre séquences identiques qui est responsable des inversions chromosomiques (Partie 1). Cette deuxième activité était totalement inédite et nous avons cherché à identifier son assise catalytique par deux approches expérimentales. Par une première approche biochimique, j'ai comparé les activités de l'intégrase Int^{pTN3} avec deux autres intégrases des plasmides de Thermococcales pT26-2 (Partie 3) et pTF1 (Partie 5). Par une deuxième approche structurale, en collaboration avec l'équipe d'Herman Van Tilbeurgh (I2BC), j'ai cherché à résoudre la structure de l'intégrase Int^{pTN3} dans le but de déterminer ses particularités éventuelles (Partie 5). Ces deux approches expérimentales ouvrent la voie à une future analyse structure-fonction précise de l'intégrase Int^{pTN3}.

Quels sont les avantages évolutifs de l'activité suicidaire des intégrases d'archées ?

Les trois intégrases de Thermococcales que j'ai caractérisées biochimiquement ont une activité de recombinaison suicidaire. La catalyse de l'intégration de l'élément génétique mobile aboutit à la désactivation de l'intégrase qui ne peut plus catalyser l'excision. L'existence d'une telle fonction suicidaire est déroutante d'un point de vue évolutif. Comment l'existence et l'évolution d'une enzyme est-elle possible si elle détruit sa séquence codante dès qu'elle est active ? Quel est l'intérêt pour l'élément génétique mobile de coder pour sa séquestration dans le chromosome ? L'activité a priori fatale de l'intégrase doit bien présenter des avantages qui expliquent sa persistance évolutive. Ces questionnements avaient très peu été abordés par la littérature avant le début de ma thèse. J'ai donc décidé de mettre à profit la grande collection de génomes de Thermococcales disponible au laboratoire ainsi que mon arsenal d'intégrases caractérisées biochimiquement pour étudier l'histoire évolutive des intégrases de type suicide et de leurs éléments génétiques mobiles. J'ai commencé par choisir un élément génétique mobile portant une intégrase suicidaire entière. Trois candidats étaient connus en début de thèse. Le plasmide pTN3 porte une intégrase suicidaire mais son activité catalytique double en fait un mauvais candidat pour une étude évolutive de l'activité suicidaire (Partie 1). Le virus TPV1 porte une intégrase suicidaire qui s'est révélée être un variant inactif *in vitro* (Partie 3). J'ai donc choisi le dernier candidat : le plasmide pT26-2.

La famille de plasmides de type pT26-2 est caractérisée par 7 gènes en commun constituant le génome core et est présente chez les Thermococcales et Methanococcales (Soler et al., 2010). Chez les Thermococcales, le génome core est associé à une intégrase de type suicidaire alors qu'il est associé à une intégrase classique chez les Methanococcales (Part 2). Pour mieux comprendre cette distribution de deux stratégies d'intégration, j'ai entrepris la caractérisation biochimique d'une intégrase de Methanococcales (Partie 2) qui pourra être comparée à l'intégrase du plasmide pT26-2 (Partie 3). J'ai aussi analysé l'histoire évolutive des intégrases de Thermococcales par génomique comparative (Partie 3). Cette dernière analyse a notamment révélé des avantages évolutifs des intégrases suicidaires.

Results and Discussion

Part 1. An unprecedented catalytic activity generates chromosomal inversions in *Thermococcus* species

Article 1. Flipping chromosomes in deep-sea archaea

The Thermococcales chromosome frequently undergoes inversions. Their mechanisms remained elusive until recently when we detected four inversions in the strain *Thermococcus nautili* that harbors plasmid pTN3 coding for a tyrosine recombinase Int^{pTN3}. In the following article, we demonstrated that the four inversions occurred between four couples of identical inverted repeats longer than 100 bp. The *bona fide* site-specific integrase Int^{pTN3} can catalyze *in vitro* recombination between these sequences. It is therefore most probably responsible for the chromosomal inversions observed in *Thermococcus nautili*. Additionally, integrases similar to Int^{pTN3} are present in several *Thermococcus* species. We therefore propose a model where these integrases are responsible for inversions in *Thermococcus* species by catalyzing recombination between long-enough identical inverted repeats. No similar integrase was detected in *Pyrococcus* or *Paleococcus* species and another mechanism might explain their chromosomal inversions.

Int^{pTN3} catalyzes chromosomal inversions through an unprecedented recombinational activity. It can catalyze *in vitro* recombination between any identical sequences longer than 100 bp in a manner dependent of its catalytic tyrosine. This activity presents the same recombination site requirement as homologous recombination but with the same enzymatic catalysis as conservative site-specific recombination. We will refet to it as “homologous recombination activity” in this manuscript. Modeling of Int^{pTN3} structure evidenced additional loops compared to the related known structure of Int^{SSV1}, in agreement with their primary sequences. Those additional loops might be involved in Int^{pTN3} unprecedented recombinational activity.

Int^{pTN3} is a suicide integrase whose activity is probably impaired after integration. The integrated element TKV4 of *Thermococcus kodakarensis* is similar to plasmid pTN3 and codes for a fragmented integrase similar to Int^{pTN3}. We evidenced that TKV4 does not excise spontaneously, confirming the absence of activity for integrated suicide integrase fragments. However, the introduction of an inactive Int^{pTN3} catalytic mutant in *Thermococcus kodakarensis* resulted in the complementation of the fragment $Int(C)^{TKV4}$ leading to TKV4 excision. This result suggested that integrated integrase fragments and foreign integrases can cooperate in the synaptic complex for the excision of integrated elements.

PLOS | GENETICS

RESEARCH ARTICLE

Flipping chromosomes in deep-sea archaea

Matteo Cossu[1], Catherine Badel[1], Ryan Catchpole[1], Danièle Gadelle[1], Evelyne Marguet[1], Valérie Barbe[2], Patrick Forterre[1], Jacques Oberto[1]*

1 Institute for Integrative Biology of the Cell (I2BC), Microbiology Department, CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, Gif-sur-Yvette, France, **2** Genoscope, Laboratoire de Biologie Moléculaire pour l'Etude des Génomes C.E.A., Institut de Génomique - 2 rue Gaston Crémieux, EVRY, France

* jacques.oberto@i2bc.paris-saclay.fr

OPEN ACCESS

Citation: Cossu M, Badel C, Catchpole R, Gadelle D, Marguet E, Barbe V, et al. (2017) Flipping chromosomes in deep-sea archaea. PLoS Genet 13(6): e1006847. https://doi.org/10.1371/journal.pgen.1006847

Editor: Lotte Søgaard-Andersen, Max Planck Institute for Terrestrial Microbiology, GERMANY

Received: December 16, 2016

Accepted: June 1, 2017

Published: June 19, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. The NCBI database accession for Thermococcus. sp. 5-4 genomic sequence is CP021848.

Funding: This work was funded by the European Research Council under the European Union's Seventh Framework Program (FP/2007-2013)/ Project EVOMOBIL - ERC Grant Agreement no. 340440 (MC, PF). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

One of the major mechanisms driving the evolution of all organisms is genomic rearrangement. In hyperthermophilic Archaea of the order *Thermococcales*, large chromosomal inversions occur so frequently that even closely related genomes are difficult to align. Clearly not resulting from the native homologous recombination machinery, the causative agent of these inversions has remained elusive. We present a model in which genomic inversions are catalyzed by the integrase enzyme encoded by a family of mobile genetic elements. We characterized the integrase from *Thermococcus nautili* plasmid pTN3 and showed that besides canonical site-specific reactions, it catalyzes low sequence specificity recombination reactions with the same outcome as homologous recombination events on DNA segments as short as 104bp both *in vitro* and *in vivo*, in contrast to other known tyrosine recombinases. Through serial culturing, we showed that the integrase-mediated divergence of *T. nautili* strains occurs at an astonishing rate, with at least four large-scale genomic inversions appearing within 60 generations. Our results and the ubiquitous distribution of pTN3-like integrated elements suggest that a major mechanism of evolution of an entire order of Archaea results from the activity of a selfish mobile genetic element.

Author summary

Mobile elements (MEs) such as viruses, plasmids and transposons infect most living organisms and often encode recombinases promoting their insertion into cellular genomes. These insertions alter the genome of their host according to two main mechanisms. First, MEs provide new functions to the cell by integrating their own genetic information into the DNA of the host, at one or more locations. Secondly, cellular homologous recombination will act upon multiple integrated copies and produce a variety of large-scale chromosomal rearrangements. If such modifications are advantageous, they will spread into the population by natural selection. Typically, enzymes involved in cellular homologous recombination and the integration of MEs are distinct. We describe here a novel plasmid-encoded archaeal integrase which in addition to site-specific recombination can catalyze low sequence specificity recombination reactions akin to homologous recombination.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Large-scale genomic rearrangements allow organisms to evolve much more rapidly than through random mutation alone. Rearrangements can result in the movement of genes within genomes, changes in coding strand use, loss of nonessential functions and the incorporation of foreign DNA. As a result, the organization, content and processing of genetic information can be deeply altered. In all three domains of life, chromosomal reorganization is mainly promoted by recombination between homologous sequences, for example between redundant ribosomal operons [1,2] or integrated copies of mobile elements (ME) such as prophages [3,4], transposons [5,6] and insertion sequences (IS) [7]. Such recombination can result in the DNA inversions readily observed in closely related genomes [8,9]. In addition to homologous recombination, chromosomes can undergo rearrangement through retrotransposon-associated non-homologous recombination [10]. Other elements like integrons confer rapid adaptation to bacteria in changing environments by shuffling cassette arrays encoding a variety of functions, a process involving a site-specific recombinase and two types of attachment sites [11]. Further genomic rearrangement/reorganization can occur through the acquisition of new genetic material, predominantly by lateral gene transfer. Such gene transfer occurs in all organisms through infection by mobile elements such as viruses or plasmids, or through the uptake of free or encapsulated DNA from the environment [12,13]. Genomes can acquire novel genes in a fashion ranging from transient to permanent depending on the type of element and the physiological conditions of the host. When ME succeed in stably inserting their genome, the inserted DNA is then replicated as part of the host chromosome. The transactions between ME DNA and host genome are catalyzed by recombinases typically encoded by the elements themselves. These recombinases rank in different classes based on their enzymatic activity and the specificity of their DNA targets. The smallest ME are insertion sequences (IS) composed of a short DNA segment encoding only the enzymes involved in their transposition which can occur at many different genomic locations [14]. The related transposons are larger DNA segments which can be transposed by two flanking IS and frequently carry additional genes such as antibiotic resistance determinants [15]. The most frequent IS recombinases are DDE transposases which do not form covalent transposase-DNA intermediates during transposition [16]. Other and typically larger ME such as plasmids and viruses encode recombinases promoting DNA transactions with a stronger DNA sequence specificity. Such site-specific recombination is not only used for mobile element integration and excision in bacteria but also in the spread of antibiotic resistance by transposable elements, the control of plasmid copy number, regulation of gene expression and the resolution of concatenated chromosomes [17]. Site-specific recombinases can be categorized into the serine recombinases and tyrosine recombinases (Y-recombinases); which, in contrast to DDE transposases, form covalent enzyme-DNA intermediates during recombination, albeit with markedly different mechanisms of action. Before religation of the two recombining DNA strands, serine recombinases generate breaks in all strands while Y-recombinases produce two sequential single-strand breaks [17]. As a rule, site-specific integration/excision reactions promoted by Y-recombinases occur via a synaptic complex composed of two DNA duplexes carrying the specific sites bound by four recombinase protomers [17]. The two-recombinase pairs are activated sequentially, allowing one strand from each duplex to be exchanged at a time via two consecutive and symmetrical Holliday junctions. A notable exception is *Vibrio cholerae* phage CTX. Not only does this phage integrate into its host genome in single stranded form where two sites fold into a hairpin structure, mimicking a recombination target for the cellular XerCD chromosome resolvase; but also only requires XerC for integration [18].

One of the best-studied Y-recombinases is the integrase of phage λ. The primary function of this enzyme is the integration of phage DNA into the chromosome of its bacterial host (and its excision). This function is achieved by promoting site-specific recombination between the phage attachment site *attP* and its chromosomal counterpart *attB* [19]. Under particular circumstances, the integrase of the lambdoid phage HK022 is capable of generating inversions between *attP* and a secondary attachment site in the HK022 left operon [3]. Similarly, the primary function of the yeast FLP protein is the control of the 2μ plasmid copy number [20] by DNA inversion between two divergent 34bp FRT sites located on the plasmid [21]. FLP recombinase activity has also been successfully used for integration and excision of synthetic DNA in mammalian genomes [22]. The recombination activities of both λ integrase and FLP recombinase are summarized as shown in S1 Fig. Historically, this reciprocal and conservative recombination between two stringently defined double-stranded DNA sequences in each chromosome was denominated the Campbell model [23].

The sequences of a considerable number of Y-recombinases have been compared to reveal the position of conserved residues and infer the location of the catalytic active site [24]. They share in their C-terminal moiety a rather well conserved region of ~120 amino acids containing up to six nearly invariant amino acids R..K..HxxR..[W/H]..Y forming the active site [25,26]. A small number of Y-recombinases have been characterized biochemically in Archaea, for example the XerA recombinase of the hyperthermophilic euryarchaeon *Pyrococcus abyssi* which exhibits a perfect active site consensus [27]. Sequence alignments have revealed that other archaeal active sites diverge slightly from the bacterial consensus R..HxxR..Y [28]. The integrases of viruses SSV1 isolated from the hyperthermophilic crenarchaeon *Sulfolobus shibatae* [29] and SSV2 from *Sulfolobus islandicus* [30] share the consensus R..KxxR..Y while the plasmidic integrase of *Sulfolobus* sp. NOB8H2 displays R..YxxR..Y [28].

Mobile elements therefore contribute to genome evolution through both site-specific and homologous recombination, which usually operate by distinct mechanisms and enzymatic activities. Homologous recombination is also known to occur frequently between multiple IS copies resulting in large scale archaeal genomic rearrangements, as observed in both *Crenarchaeota* e.g. *Sulfolobus islandicus* [31] and *Euryarchaota* e.g. *Pyrococcus abyssi* [32]. The distribution of archaeal ISs is patchy not only at the phylum level but also at genus level [9]. Interestingly, genome shuffling occurs in *Thermococcus* [33] even if ISs are seldom found in this genus suggesting that alternative recombination mechanisms are capable of producing large-scale genomic rearrangements.

If site-specific recombination only requires specific nucleotide sequences targeted by a dedicated recombinase, homologous recombination on the other hand is a much more complex process. In all organisms, homologous recombination constitutes one of several pathways to repair double-strand breaks. In addition to DNA synthesis, it requires dedicated recombinases and their accessory factors which act on stretches of near-sequence-identical DNA. In eukaryotic and bacterial cells, the enzymes and pathways involved in homologous recombination have been extensively studied (see [34,35] for reviews), whereas archaeal homologous recombination is still an active field of investigation. It is known that the initial resectioning step after double-strand break involves the Rad50–Mre11–HerA–NurA complex to generate 3' single-strand substrates [36,37]. The RecA paralog RadA and its accessory functions associate with this ssDNA to constitute the presynaptic filament, which will scan and pair with homologous sequences [38]. In the archaeon *Thermococcus kodakarensis*, homologous recombination has been detected experimentally between stretches of identical DNA sequences equal to or greater than 500bp [39].

To our knowledge, a direct overlap between site-specific and homologous recombination processes has not been described so far. In the present work, we report the discovery and

characterization of a new integrase from the hyperthermophilic archaeon *Thermococcus nautili* [40,41] capable of catalyzing both site-specific recombination and low sequence specificity recombination reactions mimicking homologous recombination. The wide distribution of this particular Y-recombinase among the *Thermococcus* genus provides a valid rationale for the observed genomic rearrangements in these Archaea.

Results

Dotplot comparisons identify synteny breakpoints in *Thermococcus* chromosomes

We compared the chromosomes of the 13 completely sequenced *Thermococcus* species available to date by dotplot analysis and observed high levels of genome scrambling as shown in Fig 1A. Strikingly, comparison of *T*. *onnurineus* and *T*. sp. 4557 chromosomes by this approach revealed only two large inversions of 139/143Kb and 102/74Kb respectively (Fig 1B & 1C). This relatively small number of inversions facilitated the investigation of the synteny breakpoints bordering both inversions. Using the SyntTax web tool [42], a composite representation was obtained as shown in Fig 1C. Gene order is conserved immediately upstream and downstream of each inversion border and was used to identify the synteny breakpoints. For each inversion, the breakpoints are located within tRNA gene pairs, transcribed in opposite orientations. Interestingly, *T*. *nautili* plasmid pTN3 integrates in the $tRNA^{Leu}$ gene BD01_0018 [41,43] (S2 Fig) and this gene displays over 97% sequence identity with $tRNA^{Leu}$ (GQS_t10759), which borders a large chromosomal inversion between *T*. *onnurineus* and *T*. sp. 4557 (Fig 1B). The concordance between the chromosomal attachment site of the pTN3 integrase (Int^{pTN3}) and the recombination targets bordering each inversion (in opposite orientations) led us to define a working model to explain the formation of genomic inversions observed in the *Thermococcus* genus. We hypothesize that the frequent genomic inversions observed in the evolution of the *Thermococcales* order are a result of enzymatic activity of the integrase encoded by horizontally mobile elements, such as pTN3.

Int^{pTN3} is a *bona fide* tyrosine recombinase

The integrase of pTN3 shares significant sequence similarity with canonical Y-recombinases and its predicted active site can be defined as R..K..AxxR..Y which only slightly diverges from the consensus (S3A Fig). In addition, Int^{pTN3} displays a high degree of conservation with two biochemically characterized hyperthermophilic Y-recombinases, the archaeal Int^{SSV1} [44] and Int^{SSV2} [30] (S3B Fig). Thus, it seemed worthwhile to compare the enzymatic activities of Int^{pTN3} to those of other enzymes of the same family such as phage λ integrase and *Saccharomyces cerevisiae* 2μ plasmid FLP protein and to validate them against the canonical Y-recombinase model.

Int^{pTN3} is an active site-specific tyrosine recombinase

In order to characterize the activities of Int^{pTN3}, it was necessary to over-produce and purify the enzyme (S4 Fig) and to construct DNA substrates carrying appropriate attachment sites (as determined by sequential deletions (S5 Fig). An integrase variant (Int^{pTN3}Y428A) in which the catalytic tyrosine is substituted with an alanine was constructed, purified and tested (S6 Fig). We used these proteins and DNA components in a series of *in vitro* and *in vivo* experiments, detailed below, to ascertain the properties of Int^{pTN3}.

Int^{pTN3} catalyzes *attP*-*attB* integration. In order to measure the activity of purified Int^{pTN3}, we initially developed a simple test in which integrase-catalyzed integration of one

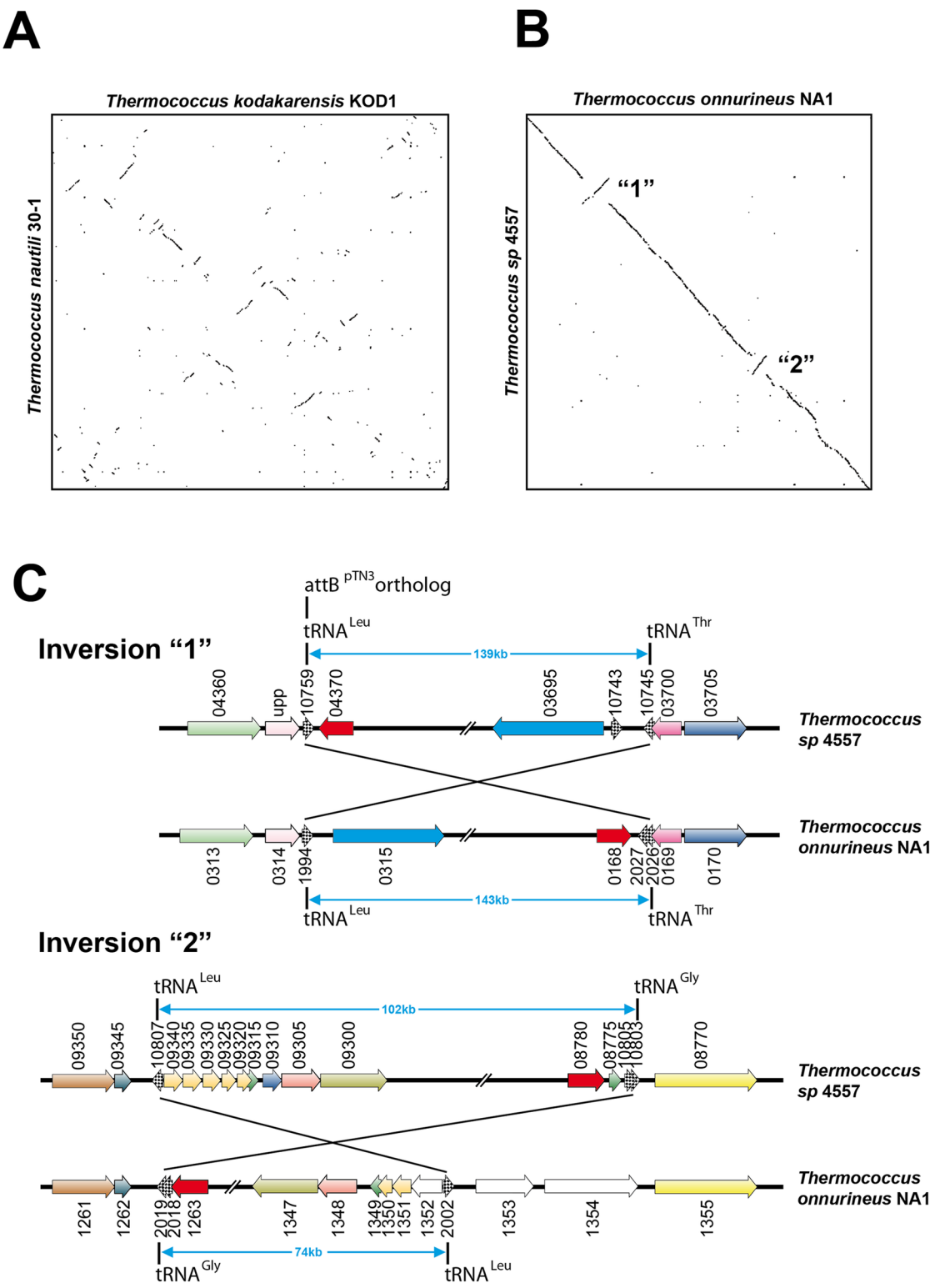


Fig 1. Genomic dotplots and synteny analysis. Genomic dotplots (**A**) between *T. kodakarensis* and *T. nautili* and (**B**) between *T. onnurineus* and *T. sp.* 4557. All genomes are centered on their putative predicted origin of replication [33]. **C.** The two synteny breaks in the genomic alignment between *T. onnurineus* and *T.* sp. 4557 (Panel B) were further analyzed. Gene order conservation and recombination endpoints of the two major inversions were identified using composite images generated by the SyntTax web tool. Inversion "1" occurred between $tRNA^{Leu}$ (GQS_t10759) and $tRNA^{Thr}$ (GQS_t10745)

genes; *T.* sp. 4557 GQS_t10759 gene is orthologous to the *T. nautili* $tRNA^{Leu}$ gene (BD01_0018) which corresponds to the chromosomal attachment site of plasmid pTN3. Inversion "2" (Panel B) occurred between $tRNA^{Leu}$ (GQS_t10807) and $tRNA^{Gly}$ (GQS_t10803) genes.

https://doi.org/10.1371/journal.pgen.1006847.g001

plasmid-encoded *attB* site in an identical site on a second plasmid results in formation of a plasmid-plasmid dimer (S1A Fig), which can be detected by gel electrophoresis. In accordance with our identification of $tRNA^{Leu}$ as a potential *attB* site, we generated a supercoiled DNA template carrying a quasi-full-length *T. nautili* $tRNA^{Leu}$ gene, Leu2-88 (see below). We observed the formation of dimeric DNA molecules only with DNA templates carrying *attB* $tRNA^{Leu}$, and only in the presence of Int^{pTN3} (Fig 2). Thus, the Int^{pTN3} is able to catalyze the site-specific recombination of one *att* site with another.

Int^{pTN3} catalyzes *attL*-*attR* excision. The capacity of Int^{pTN3} to catalyze the inverse reaction i.e. the excision of a DNA segment located between *attL* and *attR* sites was tested using the template pMC479, which carries a Leu2-88 site and a minimal Leu2-44 site in the same orientation, separated by a 762bp segment. In the presence of Int^{pTN3}, the restriction digestion pattern revealed the presence of two bands of 2358 and 849bp, consistent with the excision of a circular DNA species between two *attB* sites (Fig 3). The recombination reaction also generated an additional band of 4056bp, explainable by the integration of the 849bp circular product into the initial pMC478 template. This demonstrates that Int^{pTN3} is able to efficiently catalyze both DNA integration and excision reactions.

Int^{pTN3} can re-activate related integrated mobile elements. The species *T. kodakarensis* carries in its genome the stably integrated element TKV4 [45], which is closely related to pTN3 of *T. nautili*. As shown for pTN3 (S2 Fig), the integration of TKV4 into the *T. kodakarensis* genome has disrupted the gene encoding Int^{TKV4}, rendering TKV4 incapable of spontaneous chromosomal excision. Considering that Int^{pTN3} and Int^{TKV4} display extensive sequence similarity (S3 Fig) and promote integration in orthologous $tRNA^{Leu}$ genes [45], we investigated the capacity of Int^{pTN3} to excise TKV4 *in vitro*. Excision and circularization of a DNA molecule is detectable by PCR amplification using suitably oriented primers (Fig 4A). Treatment of *T. kodakarensis* genomic DNA with purified Int^{pTN3} resulted in products consistent with TKV4 circularization (Fig 4B), demonstrating that Int^{pTN3} could excise, and hence re-activate this dormant mobile element. In light of this *in vitro* activity, we endeavored to test this TKV4 resurrection reaction *in vivo*. This experiment involved the construction of specialized *T. kodakarensis* expression vectors pRC524 and pRC526 expressing wild type Int^{pTN3} and mutant Int^{pTN3}Y428A respectively (Fig 4C) (see Material and methods). Surprisingly, both Int^{pTN3} and the active site mutant Int^{pTN3}Y428A were able to revive TKV4 *in vivo* (Fig 4D). Not only does this result demonstrate the ability of pTN3 to excise, and therefore re-activate integrated mobile elements, it also strongly suggests that the activity of mutated Int^{pTN3}Y428A could be complemented by the truncated Int^{TKV4} encoded by the integrated element, whereas both variants are inactive on their own. A similar phenomenon of complementation has been reported between a DNA-binding impaired mutant and a catalytic tyrosine residue mutant of Int^{SSV1} [44].

Int^{pTN3} catalyzes DNA inversion between *att* sites. The ability of Int^{pTN3} to catalyze the inversion of DNA sequences is key in our model of large-scale integrase-mediated chromosomal rearrangements in the *Thermococcus* genus. To test the Int^{pTN3} invertase activity, we constructed a plasmid (pMC478) with two attachment sites in inverted orientation: the full-length $tRNA^{Leu}$ gene and the minimal Leu2-44. The restriction digestion pattern showed the presence of two new bands corresponding to the inversion of the DNA segment between the

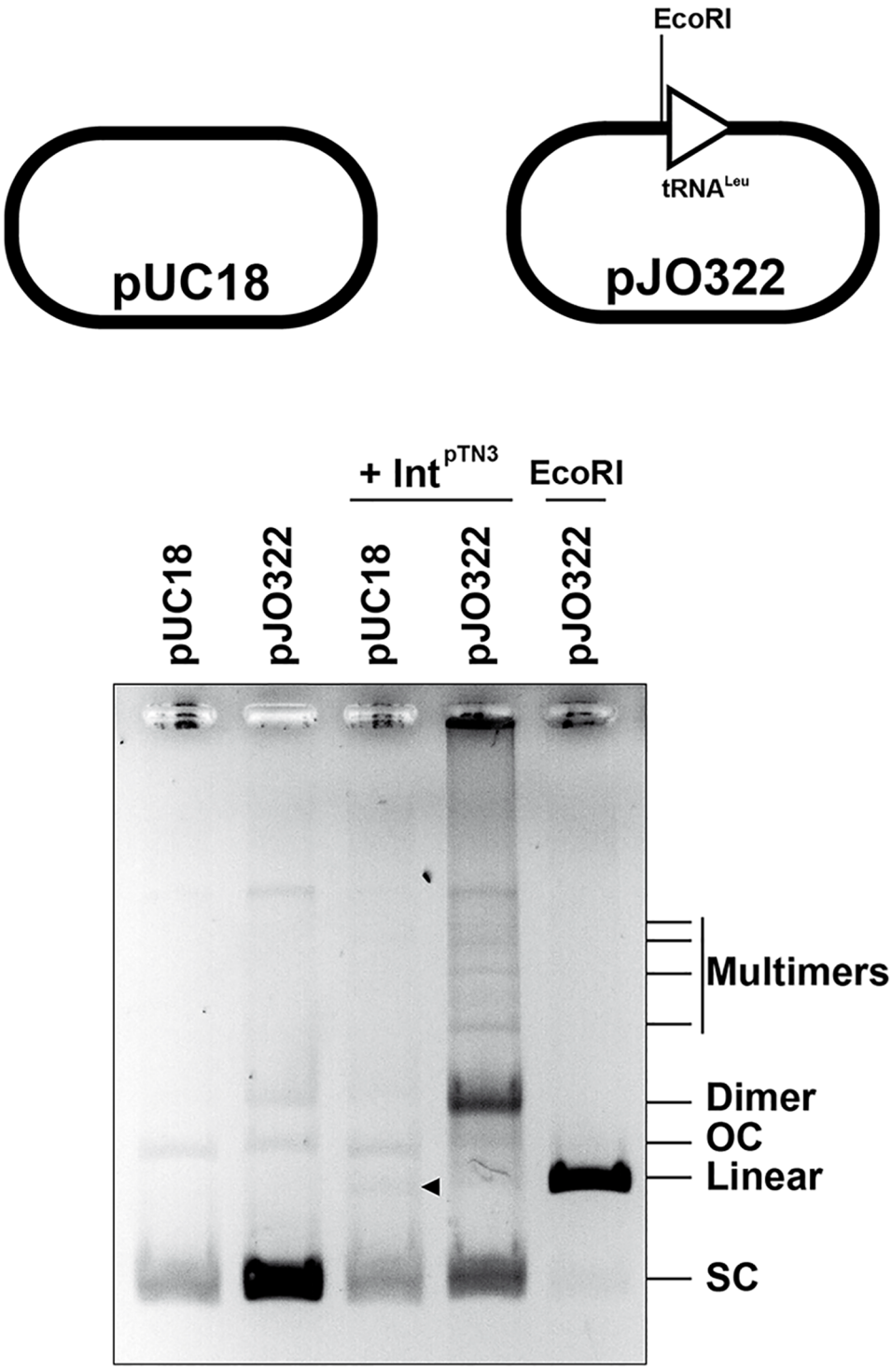


Fig 2. Dimer formation. Supercoiled (SC) plasmids pUC18 and pJO322 carrying the Leu2-88 fragment (S5A Fig) were incubated with Int^{pTN3} in a standard reaction (see Materials and methods) and compared with linearized pJO322 by agarose gel electrophoresis. The integrase has no effect on pUC18 with the exception of the production of a faint linear species (indicated by an arrow). The integrase increases considerably the formation of plasmid pJO322 dimers and to a lower extent that of multimers. No increase in the formation of open circular (OC) form was observed.

https://doi.org/10.1371/journal.pgen.1006847.g002

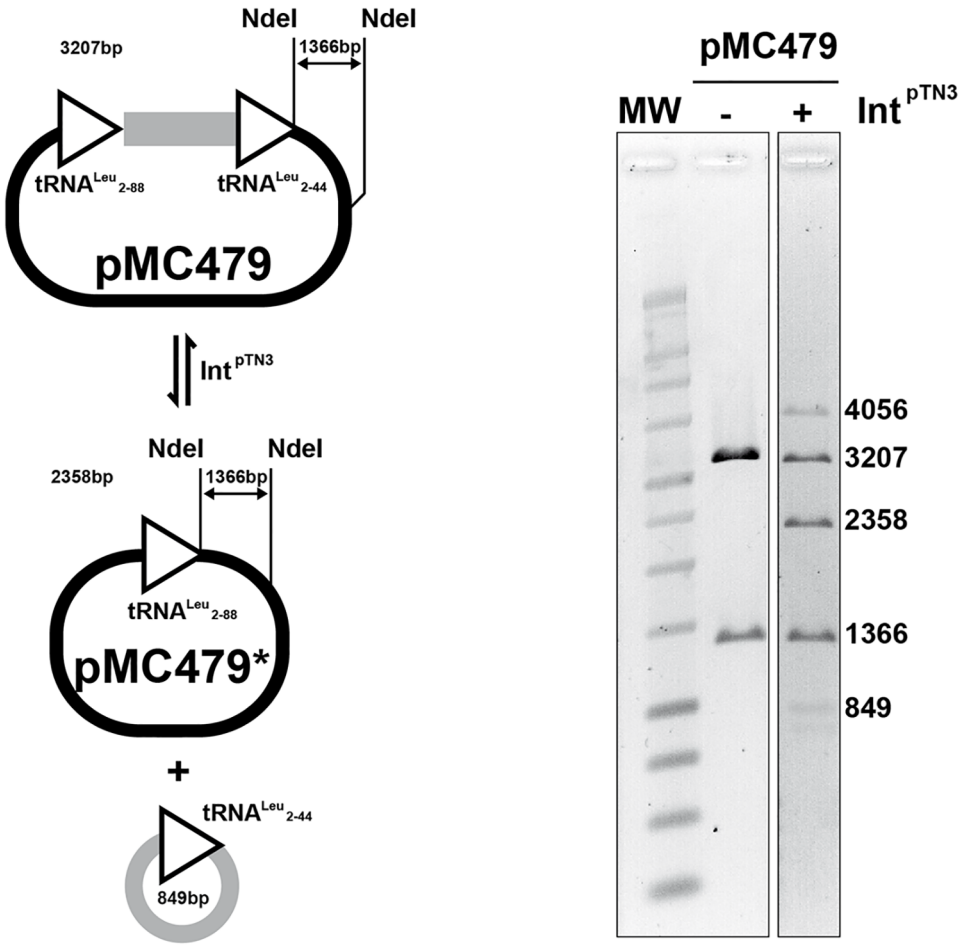


Fig 3. Int^{ptn3} excision and integration. Plasmid pMC479 carries two copies of tRNA^{Leu} cloned in direct orientation and separated by a 762bp spacer fragment (see Material and methods). The direct repeats consist of the minimal tRNA^{Leu} 2–44 and the longer tRNA^{Leu} 2–88, both proficient in dimerization reactions. Plasmid pMC479 was incubated with Int^{pTN3} in a standard reaction (see Materials and methods). The *Nde*I restriction enzyme generates two fragments of 3207 and 1366bp respectively in pMC479. Upon incubation with Int^{pTN3}, *Nde*I digestion generates additional fragments of 2358bp corresponding to recombined pMC479* and 849bp corresponding to the circularized spacer and recombined *att* site. Both constitute the products of the excision reaction. A larger 4056bp fragment is generated as well and corresponds to the recombination product generated by integration of the 3207 and 849bp species. The relative intensity of the bands is compatible with an expected equilibrium reaction.

https://doi.org/10.1371/journal.pgen.1006847.g003

attB sites only when DNA was treated with the integrase (Fig 5). This result indicates that, like the *S. cerevisiae* FLP recombinase, Int^{pTN3} is capable of efficiently performing all three canonical reactions characteristic of site-specific Y-recombinases: integration, excision and inversion. No recombination products could be observed in inversion reactions performed with the inactivated integrase variant Int^{pTN3}Y428A (S6 Fig).

Synteny analysis of the inversion endpoints observed between *T*. sp. 4557 and *T. onnurineus* indicates that recombination may have occurred between different tRNA genes, namely between tRNA^{Leu} (GQS_t10759) and tRNA^{Thr} (GQS_t10745) as well as between tRNA^{Leu} (GQS_t10807) and tRNA^{Gly} (GQS_t10803). Interestingly, inversion templates combining tRNA^{Leu} and tRNA^{Thr} failed to produce recombination products (Fig 5).

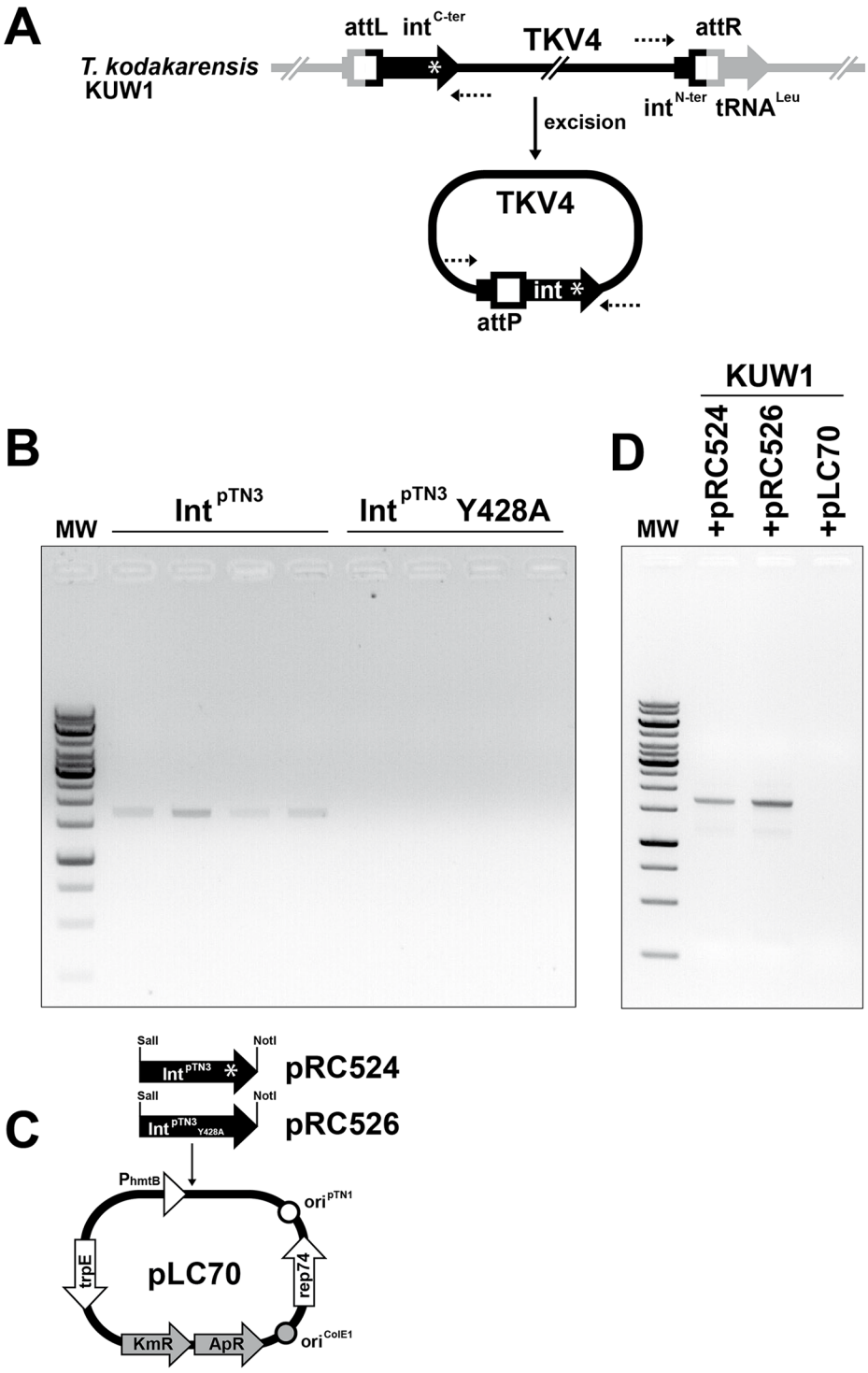


Fig 4. TKV4 excision *in vitro* and *in vivo*. A PCR amplification assay was designed to assert artificial Int^{pTN3}-mediated TKV4 circularization (Panel A). The assay was first performed *in vitro* on four samples of purified *T. kodakarensis* genomic DNA incubated with wild type Int^{pTN3} or inactive Int^{pTN3} Y428A mutated enzyme in a standard reaction analyzed by agarose gel electrophoresis (see Materials and methods). Only reactions using wild-type enzyme generated a 1710bp band of the expected excision size (Panel B). The same TKV4 excision reaction was tested *in vivo* by transforming *T. kodakarensis* KUW1 with shuttle plasmids pRC524 (expressing wild type integrase) and pRC526 (expressing mutated Int^{pTN3}Y428A) or with the vector alone (Panel C). Total DNA was extracted from the transformants and amplified as described above. In this *in vivo* experiment, both enzymes were TKV4 excision-proficient (Panel D).

https://doi.org/10.1371/journal.pgen.1006847.g004

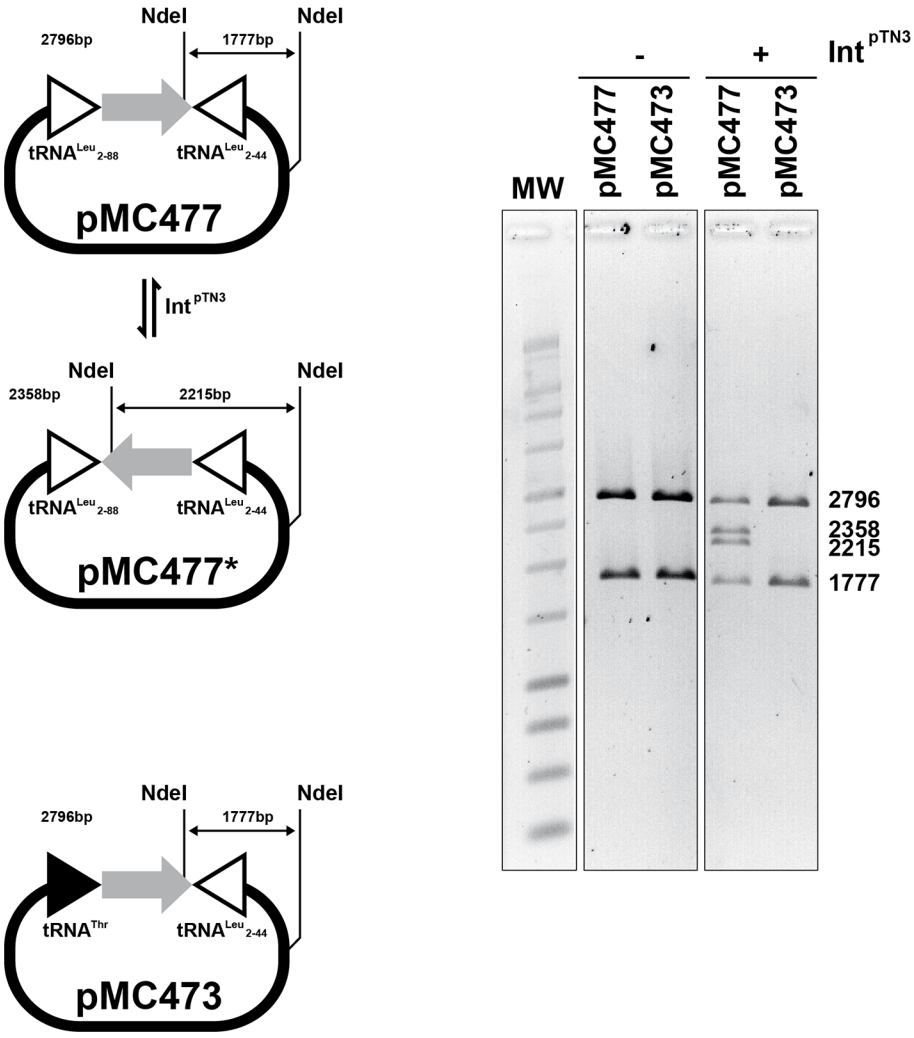


Fig 5. Int^{ptn3} inversion. Plasmid pMC477 carries two copies of tRNA^{Leu} cloned in inverted orientation and separated by a 892bp spacer fragment (see Material and methods). The inverted repeats consist of the minimal tRNA^{Leu} 2–44 and the longer tRNA^{Leu} 2–88, both proficient in dimerization reactions. Plasmid pMC473 carries tRNA^{Leu} 2–44 and tRNA^{Thr} GQS_t10745, in inverted orientation as well. Both plasmids were incubated with Int^{pTN3} in a standard reaction (see Materials and methods). The *Nde*I restriction enzyme generates in each case two fragments of 2796 and 1777bp. Upon incubation with Int^{pTN3}, *Nde*I digestion of pMC477 generates additional fragments of 2358 and 2215bp corresponding to the recombinant pMC477*. As for the integration/excision reactions, the relative intensity of the bands is compatible with an expected equilibrium reaction. We could not detect any inversion between tRNA^{Leu} and tRNA^{Thr} in plasmid pMC473.

https://doi.org/10.1371/journal.pgen.1006847.g005

Thermococcus nautili undergoes rapid genomic rearrangement under laboratory conditions

The large-scale genomic inversions observed between *T*. sp. 4557 and *T*. *onnurineus* display minor gene order rearrangements near the recombination endpoints indicating that these events are not recent and might have undergone remodeling (Fig 1C). In order to identify more recent rearrangements, we investigated whether large-scale genomic inversions could occur spontaneously under laboratory conditions. *T*. *nautili* carrying its natural plasmids was sub-cultured in two independent experiments for 60 and 66 generations (therefore termed *T*. *nautili* 60G and 66G) in rich liquid medium with intermittent storage at 4˚C and the

metagenomes of the resulting populations were completely re-sequenced. We observed in both *T. nautili* 60G and 66G sub-cultures a high proportion of a novel rearranged genome exhibiting four new large-scale chromosomal inversions when compared to the original published *T. nautili* genome (GenBank accession NZ_CP007264) [41] (Fig 6A). By mapping the frequency of the Illumina reads around the four inversion sites, we measured the incidence of the rearranged genome in the *T. nautili* 66G population, which was found in most cases to exceed that of the original genome (S3 Table). Both *T. nautili* 60G and 66G rearranged chromosomes were remarkably similar when compared by dotplot analysis (S7 Fig). Additionally, plasmid pTN3 was largely underrepresented in the *T. nautili* 66G sub-culture (S3 Table), whereas the smaller pTN1 and pTN2 were conserved. The chromosomally-integrated pTN3 copy carrying the disrupted integrase gene was also retained. The chain of nested inversion events leading to these new recombined genomes could be reconstructed (Fig 6C) and allowed us to analyze and precisely map the recombination endpoints. Each of the four genomic inversions occurred between paralogous gene pairs: between $tRNA^{Gly}$ genes BD01_1557 and BD01_1976, between methyl accepting chemotaxis genes BD01_1166 and BD01_1584, between transposase genes BD01_1317 and BD01_1763 and finally between UDP-glucose-6 dehydrogenase genes BD01_1333 and BD01_1481. For each pair of paralogous genes, the inversion events always occurred between two inverted segments of DNA sharing extensive sequence identity (S8 Fig). However, we could not detect significant similarity between inverted DNA segments corresponding to different pairs of paralogous genes using BLAST (e-value $\geq$ 0.075). Furthermore, none of these sequences could be aligned with the original pTN3 attachment site, $tRNA^{Leu}$ (e-value $\geq$ 10). In a control experiment, in contrast to *T. nautili*, the genome of a closely related organism, the plasmid-less *Thermococcus sp*. 5–4 (GenBank accession CP021848) remained stable when sub-cultured for 36 or 66 generations in two separate experiments (Fig 6B and S7 Fig).

Int^{pTN3} also catalyzes DNA inversion between non-*att* sites on the archaeal chromosome

The remarkable differences in the outcome of *T. nautili* and *T. sp* 5–4 sub-culturing experiments and the observation that $tRNA^{Gly}$ genes could recombine in these conditions suggested a causal link between Int^{pTN3} and genome shuffling. To ascertain if the new recombinations in *T. nautili* 60G and 66G could have been indeed generated by Int^{pTN3}, we decided to test whether this integrase was able to catalyze *in vitro* inversions using the sequences detected at the borders of these recombination events. New inversion templates pCB548 and pCB552 were thus constructed respectively carrying sequences encompassing $tRNA^{Gly}$ genes BD01_1557 and BD01_1976 or sequence fragments from chemotaxis genes BD01_1166 and BD01_1584 (S8 Fig). To limit the number and size of generated fragments, an *in vitro* inversion assay was conducted on linear fragments originating from these plasmids and compared to a linear fragment carrying inverted *attP* sites derived from pCB524. Inversions could be detected with all three templates albeit with significantly longer incubation times or higher Int^{pTN3} concentrations for pCB548 and pCB552-derived templates as compared to pCB524 (Fig 7). To confirm this recombination event, one of the products of the pCB548 template inversion reaction was further characterized by DNA sequencing and corresponded to a *bona fide* cross-over between BD01_1557 and BD01_1976 (S9 Fig). We conclude that Int^{pTN3} is able to catalyze low sequence specificity recombination reactions between sites that differ in sequence from its cognate *att* site, with the same outcome as homologous recombination events. It is to be noted that Int^{pTN3} catalyzes these two types of reactions with a different efficiency. Site-specific recombination reactions reach the equilibrium within 30 minutes whereas

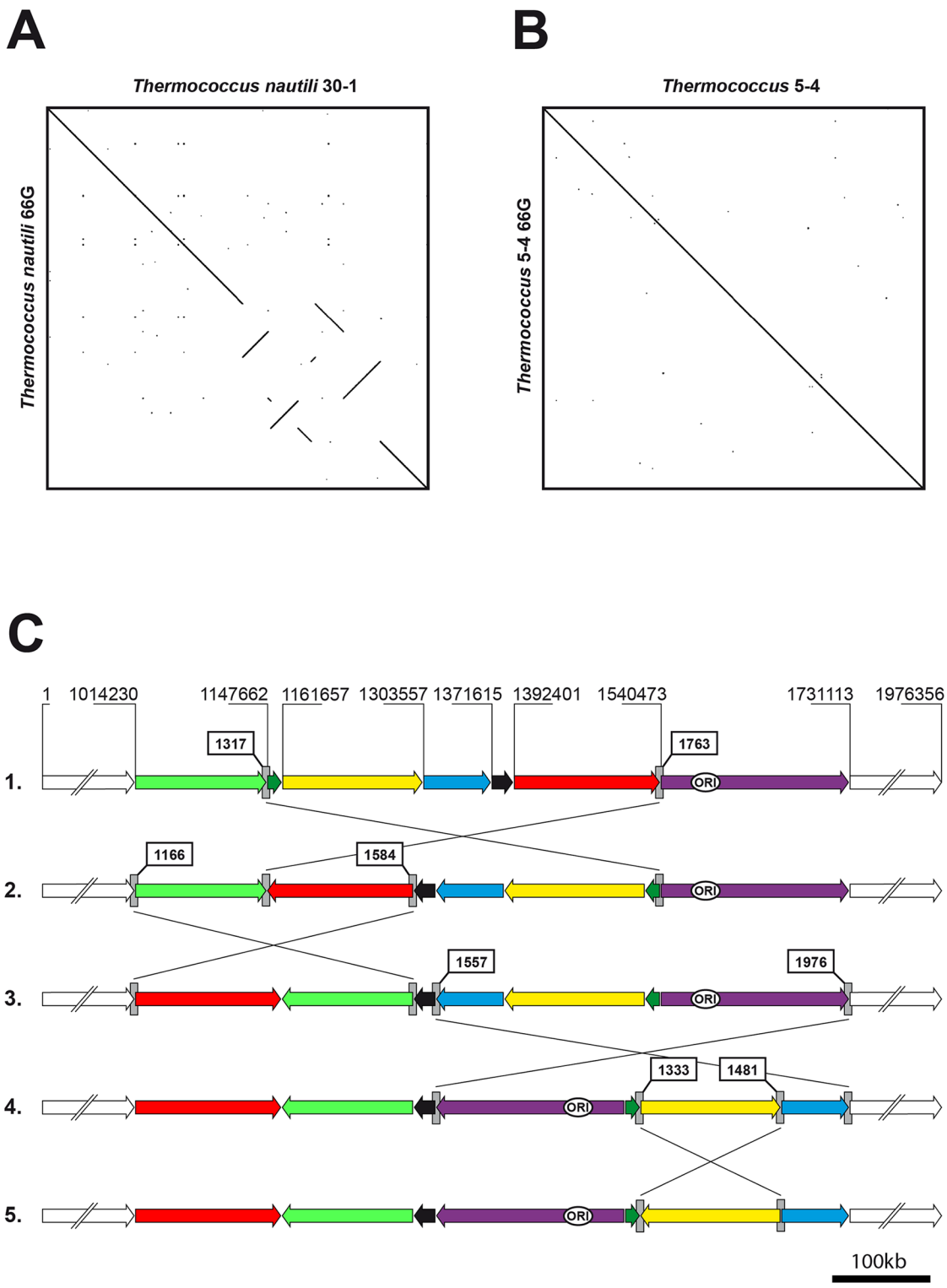


Fig 6. Laboratory inversions events. A. Dotplot analysis of the original isolate of *T. nautili* (GenBank accession NZ_CP007264) and the same organism after 66 generations (S2 Dataset). **B.** Dotplot analysis of the original isolate of *T.* 5–4 (GenBank accession CP021848) and the same organism after 66 generations (S4 Dataset). **C.** One of the possible sequential inversion scenarios leading to *T. nautili* 66G (Panel A), drawn to scale. The arrows direction reflects the chromosomal segment orientation in the original *T. nautili* strain. Genomic coordinates are indicated and the identifiers of the genes bordering each inversion are boxed.

https://doi.org/10.1371/journal.pgen.1006847.g006

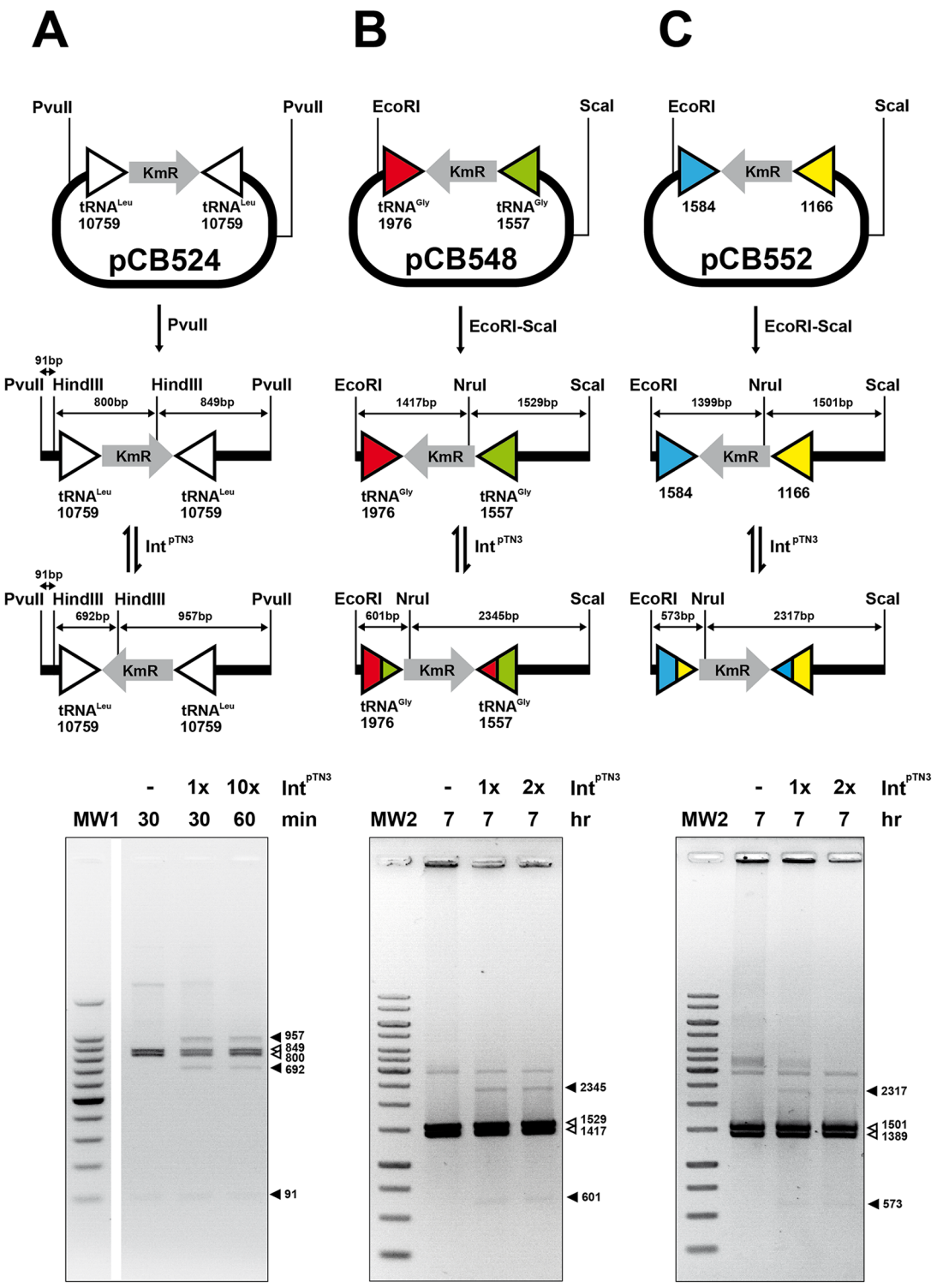


Fig 7. Int^{pTN3}-promoted low sequence specificity reactions on archaeal sequences. Int^{pTN3} catalyzes inversion on linear DNA substrates between archaeal gene pairs separated by a Kanamycin resistance determinant. White arrowheads refer to original fragments and black arrowheads indicate inversions products. **A.** Inversion between two identical copies of tRNA^{leu} gene GQS_t10759 from *T. sp.* 4557. **B.** Inversion between tRNA^{Gly} genes BD01_1557 and BD01_1976 from *T. nautili*. **C.** Inversion between chemotaxis genes BD01_1166 and BD01_1584 from *T. nautili*. Int^{pTN3} concentration multipliers refer to the standard assay described in Materials and Methods. The detailed DNA sequences involved in these reactions are illustrated in S8 Fig.

https://doi.org/10.1371/journal.pgen.1006847.g007

several hours and higher enzyme concentrations are required to detect all low sequence specificity recombinations.

Int^{pTN3} catalyzes low sequence specificity recombination reactions mimicking homologous recombination between any DNA sequence pairs

The absence of inter-pair DNA similarity observed in *T. nautili* 60G and 66G chromosomal inversions prompted us to test whether Int^{pTN3} could catalyze recombination between homologous non-archaeal sequences. The simplest experiment consisted of the incubation of cloning vector pBR322 DNA with the integrase in the same conditions as described above. This recombination reaction promoted by Int^{pTN3} yielded a ladder of plasmid multimers produced by sequential integration, which could be readily observed by eletrophoretic migration whereas no homologous integration reaction was detected with the mutated Int^{pTN3}Y428A (Fig 8A). Surprisingly, Int^{pTN3} generated also a double-strand cut at the pBR322 ColE1 origin of replication for which we have no explanation at this stage (S10 Fig). This cleavage does not constitute an intermediate step in the recombination reaction since none of Int^{pTN3} linear substrates shown in Fig 7 carries the ColE1 origin. In addition to the homologous integration reaction, we investigated the capacity of Int^{pTN3} to promote inversions between homologous sequences of bacterial origin. Short DNA segments of decreasing length (250, 175 and 100bp, see S11 Fig) originating from the *E. coli lacZ* gene were cloned in opposite orientations respective to the *lacZα* gene of pUC18 to generate plasmids pCB574, pCB571 and pCB558, respectively. These templates were linearized, incubated with Int^{pTN3} and tested by subsequent restriction analysis. In each case, Int^{pTN3} generated additional bands consistent with homologous inversion reactions displaying efficiencies proportional to the extent of DNA identity (Fig 8B).

Discussion

The major mechanism producing chromosomal rearrangements is recombinational exchange between homologous sequences [46]. These rearrangements often consist of DNA inversions between IS elements [9,46,47]. The observation that, in the *Thermococcus* genus, large chromosomal inversions occur even in the absence of IS elements prompted us to investigate the molecular mechanism behind these rearrangements. The presence of tRNA genes at recombination endpoints in genomes as diverse as plant chloroplasts [48,49] and *Thermococcales* [9], combined with the fact that integrases often target tRNA genes [50], lead us to propose a precise molecular model involving Int^{pTN3} to explain large-scale genomic rearrangements. Using a combination of comparative genomics, *in vitro* analyses, and serial culturing experiments, we uncovered a mechanism and enzymatic activity responsible for the shuffling-driven chromosomal evolution in *Thermococcales*. By means of deep comparative genomic analyses, we were able to correlate genome scrambling with the presence of a mobile element. This mobile element has been identified as plasmid pTN3, naturally present in *T. nautili* both as an episome and integrated in the genome [41,43]. Plasmid pTN3 encodes the Int^{pTN3} integrase of the Y-recombinase superfamily capable of promoting its site-specific plasmid integration at a tRNA^{Leu} gene of its host. Due to perfect DNA conservation between *attB* and *attP* attachment sites (S2B Fig), an intact and presumably expressed tRNA^{Leu} is reconstituted upon pTN3 chromosomal integration. We successfully reproduced, with high efficiency in a purified *in vitro* system, the canonical DNA reactions of integration and excision expected from a *bona fide* integrase. Site-specific mutation of the active site tyrosine to alanine abolished these activities. A positive excision reaction was also obtained *in vivo* by expressing wild-type Int^{pTN3} and the catalytic tyrosine mutant Int^{pTN3}Y428A in *T. kodakarensis* KOD1 cells. The genome of this

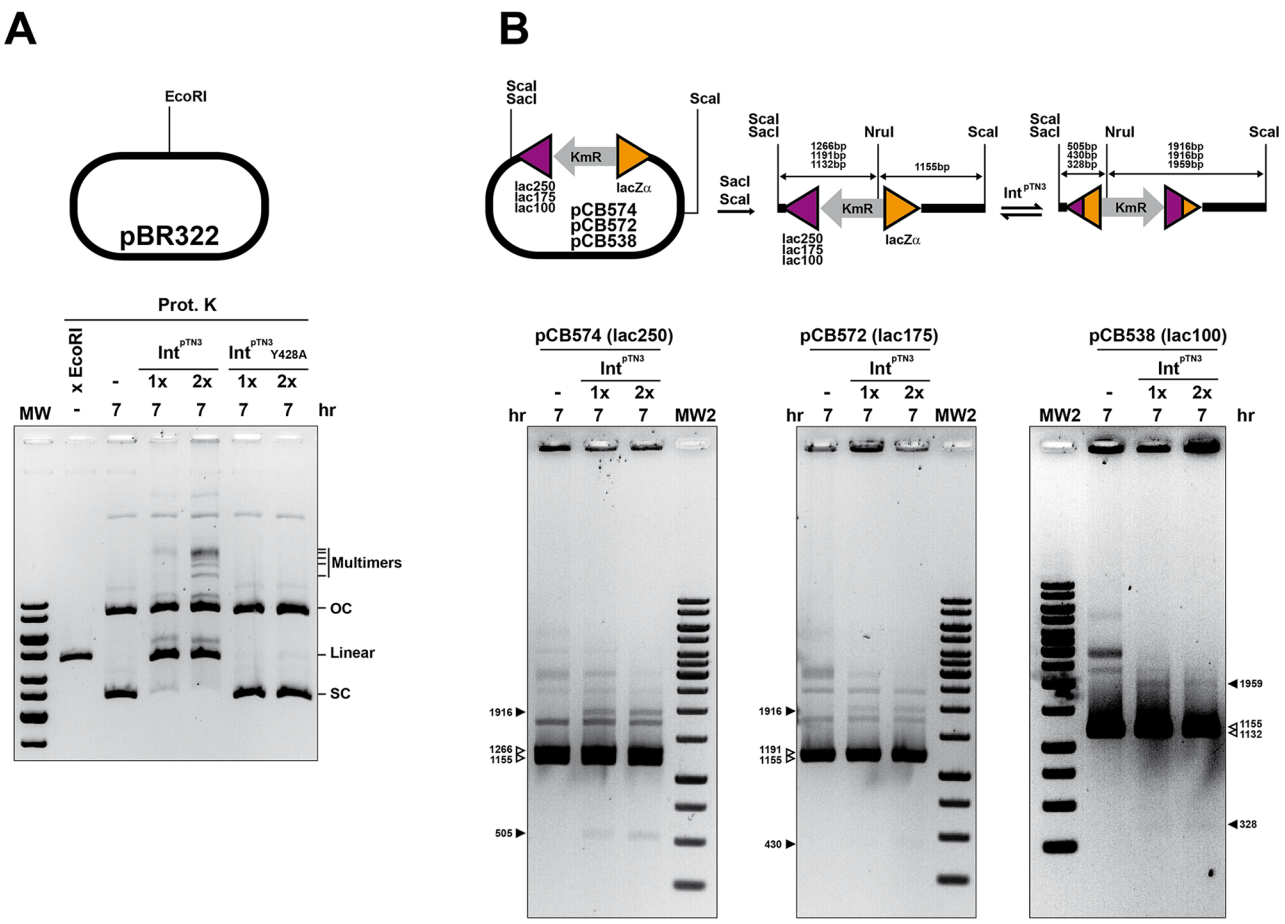


Fig 8. Int^{pTN3}-promoted low sequence specificity reactions on exogenous sequences. A. Low sequence specificity reactions mimicking homologous DNA integration are visualized by the accumulation of multimers of increasing size only when the reaction occurs in the presence of wild-type Int^{pTN3}. A linear pBR322 species generated by Int^{pTN3}-generated double-strand cleavage is visible and migrates close to a control plasmid digested by the EcoRI endocnuclease. OC and SC refer to the open circle and supercoiled DNA forms, respectively. **B.** Int^{pTN3} catalyzes inversion on linear DNA substrates between two inverted *E. coli lacZ* gene segments of varying sizes separated by a Kanamycin resistance determinant. The sequence identity between the inverted segment amounts to 250, 175 and 100bp respectively in plasmids pCB574, pCB572 and pCB538 (see Materials and methods). White arrowheads refer to original fragments and black arrowheads indicate inversions products. Int^{pTN3} concentration multipliers refer to the standard assay described in Materials and Methods.

https://doi.org/10.1371/journal.pgen.1006847.g008

strain carries the integrated episome TKV4 [45] which is remarkably similar to pTN3 (Fig 9). Surprisingly, both wild-type and mutant forms of the integrase excised TKV4 in circular form. This suggests that a truncated C-terminal Int^{TKV4}, presumably impaired in DNA-binding but carrying the catalytic tyrosine, can complement Int^{pTN3}Y428A. A plausible explanation invokes the participation of integrase dimers in the recombination reaction. In this case, only the heterodimeric form would possess an active catalytic site where Tyr428 is provided by the first monomer while the second monomer contributes the remaining conserved residues. This cleavage *in trans* was initially reported for the FLP recombinase [51,52]. Similarly, the complementation of activity between a DNA-binding impaired mutant and a catalytic tyrosine residue mutant has been described for another archaeal integrase, Int^{SSV1} [44].

The peculiar location of tRNA^{Leu} GQS_t10759 at the exact border of a large DNA inversion observed between the genomes of *T. onnurineus* and *T.* sp. 4557 suggested that this inversion could have occurred by the recombinase activity of Int^{pTN3}. In our purified system, we could

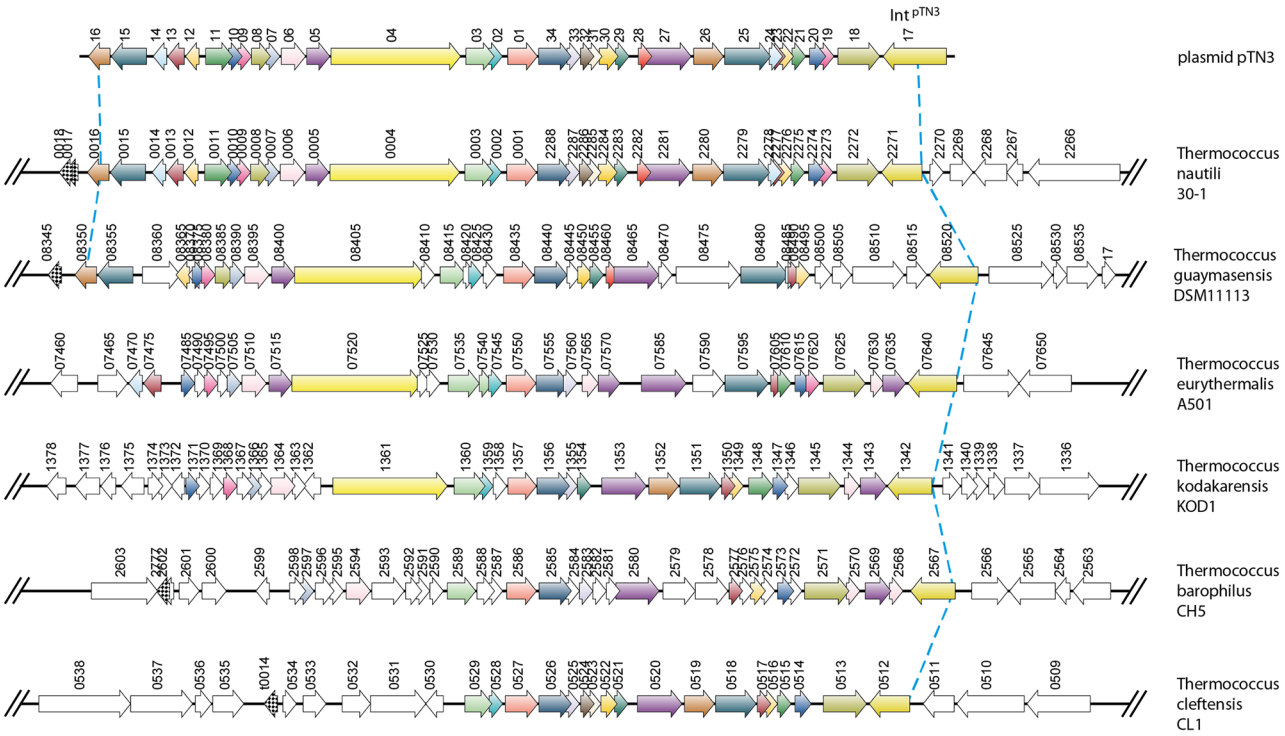


Fig 9. pTN3-like integrated elements in *Thermococcales*. The presence of pTN3-like integrated elements was investigated in all completely sequenced *Thermococcales* genomes by synteny analysis using the SyntTax web server [42]. In addition to *T. nautili*, the genomes of *T. guaymasensis* DSM11113, *T. eurythermalis* A501, *T. kodakarensis* KOD1, *T. barophilus* CH5, and *T. cleftensis* CL1 carry an extensive genomic region corresponding to plasmid pTN3 shown on top. Each arrow corresponds to an individual gene numbered according to GenBank annotations. The consistent gene color code illustrates orthology across organisms while white color indicates its absence. As indicated by a blue dotted line, conservation of synteny is clearly visible on the right border and limited by the gene encoding pTN3 C-ter integrase and its remnants. Truncated N-terminal-encoding integrase genes constitute pseudogenes lacking a stop codon and are therefore not annotated. Genetic divergence appears stronger on the left border.

https://doi.org/10.1371/journal.pgen.1006847.g009

obtain highly efficient DNA inversions between two inverted copies of GQS_t10759. Paradoxically, we were unable to promote inversion between $tRNA^{Leu}$ GQS_t10759 and $tRNA^{Thr}$ GQS_t10745 contrary to what the genomic comparisons between *T. onnurineus* and *T.* sp. 4557 suggested. An experiment of prolonged *T. nautili* cultivation was instrumental in elucidating the large-scale inversion mechanism in *Thermococcus*. The strain carrying its natural plasmids was cultivated during 60 or 66 generations; total DNA was extracted from this population and sequenced in a manner similar to a metagenome. We observed the high incidence in the resulting populations of a particular recombined genome with four large chromosomal inversions and a very low copy number of plasmid pTN3 encoding active Int^{pTN3} (< 2/chromosome) (S3 Table). This plasmid loss could have contributed to the higher fitness and spread of a particular clone in the population. The four large-scale inversions occurred between four pairs of naturally occurring paralogous genes sharing at least 104bp of sequence identity in inverted orientation (S8 Fig). No significant sequence conservation could be detected between the four pairs. We did not observe chromosomal rearrangements after prolonged incubation of *Thermococcus* sp. 5–4, which does not carry plasmids. The potential causal link between pTN3 and a number of unrelated sequence pairs involved in large scale genomic shuffling in *T. nautili* was difficult to conciliate with the classical site-specific recombination properties we described for Int^{pTN3}. Remarkably, by *in vitro* assays with this integrase, we succeeded in producing inversions between several pairs of inverted paralogous genes detected in our *T. nautili*

sub-culturing experiments. These results suggested that the recombination properties of Int^{pTN3} could be extended to virtually any homologous pair of DNA sequences. Using exogenous pBR322 plasmid DNA or genes segments from bacterial origin, we demonstrated *in vitro* that Int^{pTN3} actively promotes low sequence specificity reactions mimicking homologous integration and inversion of any sequence pair as short as 100bp. The catalytic site mutation in variant Int^{pTN3}Y428A abolishes this particular recombination reaction as well. Interestingly, cellular homologous recombination in Archaea operates according to a different pathway with dedicated enzymes [36,37] and in *Thermococcus kodakarensis* has only been reported between DNA segments of 500bp or more [39].

These reactions unveiled a specific Int^{pTN3}-generated double-strand cut at the ColE1 origin of replication carried by pBR322 and its derivatives (S10 Fig). At this moment, we do not have a precise rationale to explain this observation other than a potential distant secondary structure similarity between the small RNAI and RNAII encoded by the ColE1 origin and the $tRNA^{Leu}$ encoded by Int^{pTN3} *attB* substrate. Biological interactions between tRNAs and ColE1 RNAs have been reported [53]. Clearly, this double-strand cleavage does not participate in any recombination reaction since we demonstrated all *in vitro* Int^{pTN3} inversions on linear DNA segments devoid of ColE1 origin.

The positive *in vitro* Int^{pTN3}-promoted low sequence specificity recombination results explain the failure of this enzyme to promote inversion between $tRNA^{Leu}$ GQS_t10759 and $tRNA^{Thr}$ GQS_t10745. These sites were initially thought to constitute inversion endpoints between the genomes of *T. onnurineus* and *T.* sp. 4557 but do not share sufficient sequence similarity to be efficiently recombined *in vitro*. The particular positioning of these sequences in opposite orientations could have occurred through previous overlapping inversions between a different set of paralogs or by less frequent native homologous recombination. We observed a similar situation in the sequence of the *T. nautili* 60G and 66G populations. In several cases, homologous segments were in direct orientation in the original genome but became opposed due to a previous overlapping inversion therefore indicating that *T. nautili* 60G and 66G inversions occurred sequentially.

In order to investigate whether pTN3 could account for large-scale rearrangements in the *Thermococcus* genus, we examined by synteny analysis the distribution of pTN3-like integrated element among completely sequenced *Thermococcales*. Out of 17 sequenced *Thermococcus*, and in addition to the previously reported *T. kodakarensis* TKV4 element [45], five isolates were found to harbor a pTN3-related element (Fig 9). The natural competence for DNA uptake of some *Thermococcales* such as *T. kodakarensis* [39] and the capacity of pTN3 to be transferred between cells using membrane vesicles [43] could explain the ubiquitous presence of this mobile element.

Protein sequence and structural comparisons between Int^{pTN3} and other hyperthermophilic archaeal integrases such as that of crenarchaeal virus SSV1 indicate that these proteins are clearly related. However, Int^{pTN3} possesses several additional interspersed domains relative to SSV1 (S2 and S12 Figs). We surmise that these additional domains contribute to the low sequence specificity recombination reactions akin to homologous recombination events that we have observed.

By summing up all direct and indirect evidence reported here, it is very likely that the integrase encoded by pTN3-like plasmids can account for the genomic shuffling observed in the *Thermococcus* genus. Plasmids of the pTN3 class are genetically closely related to viruses as they encode a capsid protein and a DNA packaging ATPase [43] but pTN3 virions have not be observed to date. It is not clear at this stage whether plasmids or viruses equipped with an Int^{pTN3}-like integrase have a better fitness either due to provirus maintenance or by virion spreading. An integrase mimicking homologous recombination could promote viral

integration into the host genome only if both viral and cellular chromosomes share significant DNA similarity. This enzyme however, could facilitate integration of a virus into the genome of a closely related provirus.

The question arises whether an enzyme promoting genome shuffling using very short repeated segments as substrates, would be beneficial for a cellular organism. On one hand, 'wrongly' recombined genomes would result in suboptimal gene expression programs and cells carrying scrambled genomes would display a reduced fitness and clearly be counter-selected in the population. Interestingly, the presence of a pTN3-specific spacer in a *T. nautili* CRISPR locus strongly suggests that the presence of this plasmid is deleterious [41]. On the other hand, it is also possible to envision situations where high-level genome shuffling by inversion could be advantageous. Alternate gene expression patterns could increase, for instance, adaptation to rapid environmental changes. In addition, for organisms such as *Thermococcales* where highly-expressed essential housekeeping genes maintain invariable positions [33], genome scrambling could be beneficial by relocating "less desirable" integrated elements to chromosomal areas of reduced gene expression, therefore minimizing their impact on cellular physiology.

Materials and methods

Bacterial, archaeal strains, plasmids and media

Escherichia coli strain XL1-Blue was used for cloning, plasmid amplification and site-directed mutagenesis. Overexpression of recombinant wild-type or mutant Int^{pTN3} was carried out in strain BL21 (DE3) (Novagen). All *E. coli* strains were grown in Luria-Bertani medium supplemented with 100μg/mL ampicillin or/and 50μg/mL kanamycin when necessary. *T. kodakarensis* KUW1 (*ΔpyrF ΔtrpE*) was grown anaerobically in ASW-YT medium [54] at 85˚C. Long term *Thermococcus* sub-culturing experiments were carried out in the same conditions by sequential 50x dilutions of stationary phase cultures into fresh media. The number of generations was assessed statistically at each dilution step using a Thoma cell counting chamber under 400x magnification. The plasmids used or constructed in this work are listed in S1 Table. Transformation with pRC524 and pRC526 plasmids (see below) was performed following standard protocols [55]. Plasmid-containing KUW1 strains were grown in ASW-CH medium [54] supplemented with uracil (10 μg/mL). *T. nautili* sp. 30–1 (CP007264) was grown anaerobically at 85˚C in Zillig's broth [56].

Bioinformatics and sequencing

Genomic sequences were compared and aligned by dotplot analysis using Gepard [57]. Conservation of gene order was assessed by synteny analysis using Absynte [58] and SyntTax [42]. The original genome of *Thermococcus* 5–4 JCM31817 (GenBank accession CP021848) and the genomes of sub-cultured *T. nautili* 60G and 66G and *T. sp.* 5–4 36G and 66G were sequenced by Genoscope (Centre National de Séquençage, France), using Illumina MiSeq. Reads were assembled with Newbler (release 2.9) and gap closure was performed by PCR, Sanger sequencing and Oxford Nanopore MinION. The primary genomic sequences of rearranged *T. nautili* 60G, 66G and *T.* 5–4 36G, 66G are available in S1, S2, S3 and S4 Datasets, respectively. These genomic sequences are compared by dotplot analysis in S7 Fig.

Metagenome analysis

Genomic regions corresponding to ~2000bp upstream and downstream of inversion breakpoints were extracted from both the ancestral *T. nautili* sequence, and the sub-cultured *T.*

nautili 66G sequence. Illumina sequencing reads were mapped to the ancestral sequence, and the pool of unmapped reads were mapped to the 66G sequence (Geneious 6.1.8). Two positions close to the break-point which differ in base composition between ancestral and 66G sequences were chosen to classify reads as resulting from original or inverted genome sequences. Bases were enumerated at these positions, and the percentage of reads corresponding to original sequences or inversions were calculated. The prevalence of pTN3 in the population was determined by comparing read depth across the entire *T*. *nautili* 66G genome (excluding the integrated pTN3 region) to that of pTN3 (S3 Table).

Recombinant protein expression and purification

The gene encoding the integrase of the plasmid pTN3 of *T*. *nautili* 30–1, (gene ID: 17125032) was codon-optimized for expression in *E*. *coli* and synthesized by GenScript. The synthetic gene contained a Strep-Tag at the 5' end and was cloned into pET26b+ expression vector (Novagen) to yield pJO344. Plasmid pJO496 carrying the mutated Int^{pTN3}Y428A was obtained by site directed mutagenesis of pJO344 with primers Int_A and Int_B (S2 Table) using the Agilent QuikChange Lightning Site-Directed Mutagenesis Kit. Wild-type Int^{pTN3} and mutated Int^{pTN3}Y428A were purified from *E*. *coli* BL21 (DE3) strain (Novagen) harboring respectively pJO344 or pJO496 by affinity chromatography and gel filtration (S4 Fig). All integrase enzymatic assays were conducted with strep-tagged protein derivatives.

Integrase plasmid substrates

Plasmids used for the integrase dimerization assays were constructed as follows. *Eco*RI and *Bam*HI restriction sites were added respectively at the 5' and 3' end of the various oligonucleotides shown in S5 Fig. Each oligonucleotide (Sigma-Aldrich) was annealed to its complementary sequence and the resulting double-stranded segments were cloned between the corresponding restrictions sites of pUC18. To generate plasmid pMC451, the Leu2-88 fragment was cloned in pBR322 instead of pUC18. Plasmids pMC477 and pMC479 used respectively for *att* integration/excision and inversion assays were constructed using pMC451 as backbone. The insertion fragment was amplified with primers Leu43scaI_fw and Leu43scaI_rev using pMC449 plasmid DNA as template. It contains $tRNA^{Leu}$ gene (2-44bp) and *lacZa* gene for blue-white screening. This amplified region was cloned in pMC451 in both possible orientation using *Sca*I and *Nru*I blunt sites. Plasmid pCB538 was obtained by amplifying with primers LacZ100-Sac1-For and KanR-Xba1-Rev (S2 Table) a 1364bp fragment from pUC4K and subsequent cloning between the XbaI-SacI sites of pUC18. The other plasmids: pCB548, pCB552, pCB572 and pCB574 used for non-att inversion assays were generated by Gibson Assembly [59]. Briefly, for pCB548, the genomic region corresponding to -80 to +245 of BD01_1557 (*T*. *nautili*) was amplified by PCR (Phusion Polymerase, ThermoScientific) using primers 1557_fwd and 1557_rev (S2 Table); the region from –80 to +245 of BD01_1976 was amplified using primers 1976_fwd and 1976_rev. The KmR gene was amplified from plasmid pUC4K using primers KanR_fwd and KanR_rev. Fragments were assembled into EcoRI + SalI digested pUC18 using the NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs) following the manufacturer's protocols. Similarly, for pCB552, part of the genes BD01_1166 and BD01_1584 (S8 Fig) were amplified by PCR and assembled into EcoRI + SalI digested pUC18 with the KmR gene sequence. To construct pCB538, a fragment containing KmR and the beginning of the *lacZ* gene (lac100) was PCR-amplified from pUC4K with the primers LacZ100-Sac1-For and KanR-Xba1-Rev containing the restriction sites for SacI and XbaI, respectively, at the 5' end. The adequately digested fragment was then ligated into a SacI-XbaI digested pUC18. For plasmids pCB572 and pCB574, part of the *lacZ* gene was

amplified from pUC18 and the KmR gene sequence was amplified from plasmid pUC4K. The two fragments were then assembled into the EcoRI digested pUC18. Purified plasmids pCB548, pCB552, were digested using ScaI and EcoRI and plasmids pCB572 and pCB574 were digested using ScaI. The fragments containing the non *att*-sites were then gel purified using the kit NucleoSpin Gel and PCR Clean-up (Macherey Nagel). All plasmid constructs were confirmed by DNA sequencing (Beckman Coulter Genomics).

In vitro/in vivo integrase enzymatic assay

Standard *in vitro* integrase assays were performed as follows: 165ng (8.25ng/μL, 3.1pmol) purified Int^{pTN3} and 0.5μg (25ng/μL, 10pmol) supercoiled plasmid substrates were incubated 30 min at 65˚C in a reaction buffer containing 300mM KCl, 27 mM Tris HCl pH8, 0.17mM DTT and 1mM $MgSO_4$. Depending on the size of the plasmid substrate, the DNA/integrase molar ratio varied from 30 to 60. For substrates with non-*att* sites, the integrase concentration was increased up to 50pmol. To assay dimer formation, the reaction products were separated by gel electrophoresis and visualized with ethidium bromide. For the excision and inversion assays, reaction products were purified with the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) and digested with appropriate restriction enzymes (Thermo Scientific) prior to eletrophoretic separation. *In vitro* circularization of TKV4 was performed in a standard integrase assay with genomic DNA of *T*. *kodakarensis* isolated as described previously [60]. The reaction products were purified using NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel). Recircularized products were scored by amplifying a reconstituted full-length TKV4 integrase gene. PCR was performed using Phusion Polymerase (ThermoScientific) and primers TKV4_FW and TKV4_REV (S2 Table) in conditions recommended by the supplier. *In vivo* circularization of TKV4 was obtained using total DNA from *T*. *kodakarensis* KUW1 transformed with plasmid pRC524 or pRC526. These plasmids express constitutively wild type integrase and mutated Int^{pTN3}Y428A from the P_{hmtB} promoter present in parental pLC70. DNA extraction and PCR reactions was performed as per the *in vitro* assay described above. To generate plasmids pRC524 and pRC526, the Int^{pTN3} integrase gene was amplified by PCR with primers int_fwd and int_rev (S2 Table), using total *T*. *nautili* genomic DNA as a template. The amplification product was cloned into pJET1.2 using the CloneJET PCR Cloning Kit (Thermo Fischer Scientific). The Y428A mutation was introduced into the integrase gene using the QuickChange II Site Directed Mutagenesis Kit (Agilent Technologies) with primer intY428A_fwd and its reverse complement. Both the wild-type and Y428A alleles were digested from pJET1.2 using *Sal*I and *Not*I and cloned into the corresponding sites of pLC70. All *in vitro* and *in vivo* recombination junctions and plasmid constructs were confirmed by DNA sequencing (Beckman Coulter Genomics).

Supporting information

S1 Table. Plasmids used in this work.
(DOCX)

S2 Table. Oligonucleotides used in this work.
(DOCX)

S3 Table. Metagenomic reads mapping *(T. nautili* 66G).
(DOCX)

S1 Fig. Classical site-specific recombination model. A. The intermolecular site-specific integration between cognate *attP* and *attB* sites generates a co-integrate with recombined *attL* and *attR* sites in direct orientation. The reverse reaction of excision regenerates the original

components. **B.** In the intramolecular site-specific inversion reaction, the *att* sites are in opposite orientation. This reaction is reversible as well.
(PDF)

S2 Fig. pTN3 integration. A. The comparison between the replicative and the chromosomal integrated forms of plasmid pTN3 enabled us to reconstitute the integration event. A stretch of 41bp is shared by both *attP* and *attB* sites. The nucleotides corresponding to the leucine anticodon are underlined. Upon integration, the integrase gene is disrupted and a full length $tRNA^{Leu}$ gene is reconstituted although separated from its original promoter. An excision event would regenerate the original recombination partners. **B.** DNA sequence alignment between the integrase gene of pTN3 (black) and the $tRNA^{Leu}$ gene (red). The start and stop codons of the integrase open reading frame are boxed in blue. The integration sites attP and attB as defined by Krupovic & Bamford [45] are boxed in their respective color.
(PDF)

S3 Fig. Tyrosine recombinases sequence comparison. A. Alignment of Int^{pTN3} with tyrosine recombinases from the three domains of life. The protein sequence of Int^{pTN3} (WP_022547007.1) is aligned using Praline [Reference 4 in S1 Text] with the reconstituted integrase from *T. kodakarensis* TKV4 and other previously characterized tyrosine recombinases from the three domains of life. These recombinases consist of the integrases from Sulfolobus Spindle Viruses SSV1 (P20214.1) and SSV2 (NP_944456.1), phage λ integrase (ALA45781.1), phage HP1 integrase (NP_043466.1), XerD resolvase from *Escherichia coli* (NP_417370.1) and FLP recombinase from *Saccharomyces cerevisiae* 2μ plasmid (P03870.1). The region corresponding to the catalytic signatures (BoxI, K_β, BoxII) of crystallized tyrosine recombinases are boxed in light gray. The predicted residues composing Int^{pTN3} catalytic site are shown (R..K..AxxR..Y) and the catalytic tyrosine residue is indicated by a black arrow. The color code refers to the extent of residue conservation at each position as show in the color scale. **B.** Alignment of Int^{pTN3} with Int^{TKV4} and the hyperthermophilic tyrosine recombinases Int^{SSV1} and Int^{SSV2}. Global protein sequence similarities were computed with the Needleman-Wunsch algorithm (Needle EMBOSS, http://www.ebi.ac.uk/Tools/psa/emboss_needle/): Int^{pTN3}-Int^{TKV4}: 93.6%; Int^{pTN3}-Int^{SSV1}: 33.0% and Int^{pTN3}-Int^{SSV2}: 31.2%.
(PDF)

S4 Fig. Int^{ptn3} overexpression and purification. A. Protein expression was induced with 1mM IPTG in 1L of LB medium; cells harvested by centrifugation, and lysed by sonication. The soluble fraction of the sonicate was heated at 65˚C for 10 minutes, and denatured proteins removed by centrifugation and by passing through a 0.45 μm filter. Strep-tagged proteins were purified by affinity fractionation using a Strep-Tactin column (IBA Lifesciences) as recommended by the supplier. **B.** Strep-Tactin fractions 4 and 5 were pooled and submitted to gel filtration (Superdex 200 16/600, GE Healthcare). **C.** Gel filtration fractions 21 to 31 were pooled and the purified protein was concentrated with an Amicon 3kDa cutoff concentrator (Millipore), aliquoted and stored at -80˚C.
(PDF)

S5 Fig. AttB nested deletions. The Integrase dimerization test was used to determine the minimal site required for Int^{pTN3} $tRNA^{Leu} \times tRNA^{Leu}$ recombination on nested deletions carried by plasmid templates. **A.** DNA sequence of the nested deletions. DNA segments corresponding to theses sequences were annealed and cloned directionally in pUC18. **B.** The resulting supercoiled plasmids were incubated with purified Int^{pTN3} in a standard reaction and scored for dimer formation by agarose gel electrophoresis where only relevant reactions are shown. The dimerization-proficient sequences in Panel A are marked as positive. It is noteworthy that the

Leu41 site, a site corresponding to the 41bp of sequence identity shared by both *attP* and *attB* is not a sufficient substrate for this reaction. Therefore, the minimal site for efficient dimerization is Leu2-44 with a size of 43bp. The asterisks indicate the extent of sequence identity between chromosomal *att*B and pTN3 *att*P. The leucine CAA anticodon is underlined.
(PDF)

S6 Fig. Mutated IntY428A assay. Increasing amounts of wild type Int^{pTN3} and mutated Int^{pTN3}Y428A enzymes were incubated with plasmid pMC477 as substrate to analyze the inversion properties. The experimental conditions are those of the standard integrase assay (see Material and methods) except that increasing amounts of enzyme were used: 0.5, 1, 1.5, 2.5 and 5μg, respectively. No inversion is detectable with Int^{pTN3}Y428A.
(PDF)

S7 Fig. Subcultures genome comparisons. Dotplot alignment of the prominent genomes obtained after *T*. *nautili* 60G and 66G subculturing (left) and *T*. 5–4 36G and 66G (right).
(PDF)

S8 Fig. Detailed mapping of the Int^{pTN3}-promoted *in vivo* inversions between four pairs of *T*. *nautili* paralogs. The sequences corresponding to the four genomic crossovers observed in *T*. *nautili* 60G and 66G were identified each time in pairs of paralogous genes shown aligned here. The sequences blocked in grey throughout the figure refer to perfectly conserved DNA segments in each paralogous pair where recombination occurred. Short sequences boxed in red refer to open reading frames start and stop codons when applicable (see also Fig 7 for throughout consistent color-coding). **Panel A** shows the alignment between segments overlapping $tRNA^{Gly}$ genes BD01_1557 and BD01_1976. The precise regions corresponding to both $tRNA^{Gly}$ genes are boxed in black. DNA segments cloned in pCB548 indicated by green blocks refer to BD01_1557-related sequences while red blocks correspond to BD01_1976-related sequences. The BD01-1976 nucleotide highlighted in black corrects a sequencing error in the original *T*. *nautili* genome sequence. A 176bp segment (grayed) is perfectly conserved between BD01_1557 and BD01_1976. Gly anticodons are boxed in yellow color. **Panel B** displays the alignment between methyl accepting chemotaxis genes BD01_1166 and BD01_1584. DNA segments cloned in pCB552 indicated by yellow blocks refer to BD01_1166-related sequences while blue blocks correspond to BD01_1594-related sequences A 176bp segment (grayed) is perfectly conserved between BD01_1166 and BD01_1584. **Panel C** displays the alignment between UDP-glucose-6 dehydrogenase genes BD01_1333 and BD01_1481. The two separate regions of extended sequence identity (I and II) are found between these genes respectively 284 and 620bp long (greyed). The presence of gene conversion in the interval between these two regions suggests that both were presumably involved in distinct crossover events. **Panel D** shows the alignment between transposase genes BD01_1317 and BD01_1763. The shortest recombination segment (104bp, grayed) is shared between these two paralogous genes.
(PDF)

S9 Fig. Detailed characterization of Int^{pTN3}-promoted *in vitro* inversion event by DNA sequencing. Specific sequences surrounding tRNA gene BD01_1976 are blocked in red while specific sequences surrounding tRNA gene BD01_1557 are blocked in green. Relevant anticodon sequences are boxed in yellow color. Two nucleotide mismatches between these tRNA genes are blocked in black. The tripartite composition of these DNA segments is further highlighted by blocking in grey color the stretch of identical sequenced shared by the DNA fragments carrying BD01_1976 and BD01_1557. **Panel A** depicts the sequence of steps involved in generating a suitable recombinant fragment for DNA sequencing. Plasmid

pCB548 carries DNA segments containing *T*. *nautili* $tRNA^{Gly}$-encoding genes BD01_1976 and BD01_1557 in inverted orientation and separated by a Kanamycin resistance determinant originating from pUC4K. The exact sequence of the cloned DNA segments encompassing BD01_1976 & BD01_1557 is displayed in S8A Fig. The inversion reaction was performed as shown in Fig 7B: an EcoRI-ScaI fragment originating from pCB548 was incubated with Int^{pTN3} after which the 601bp EcoRI-NruI fragment generated by Int^{pTN3} recombination was gel-purified, PCR-amplified with the forward primer 5'-ccgtttaatcgtcgcgcggaagc-3' targeting the upstream sequence of the $tRNA^{Gly}$ gene BD01_1976 and the reverse primer 5'-cccgttgaatatggctcataacaccc-3' targeting the beginning of the KanR cassette. The resulting fragment was submitted to Sanger DNA sequencing using the forward primer mentioned above. **Panels B** and **C** display also the alignment between the 5' half of both tRNA genes and the minimal Leu2-44 segment involved in Int^{pTN3} site-specific recombination. **Panel D** shows the result of the DNA sequencing reaction. The crossover point in the recombination reaction occurred precisely downstream of the two nucleotide mismatches mentioned above, in the sequence blocked in grey corresponding to the 3' half of the tRNA genes and strictly conserved sequences immediately following. The sequences boxed in black in Panels B,C and D correspond to the exact extents of $tRNAs^{Gly}$.
(PDF)

S10 Fig. Integrase-promoted double-strand cut at ori ColE1. Circular plasmid pCB548 (4675bp) treated with Int^{pTN3} and digested with XhoI-NdeI endonucleases generates bands of 2966 and 1709bp due to integrase-promoted low sequence specificity recombination (white arrowheads). The original larger 3896bp XhoI-NdeI fragment undergoes an additional double-stranded cut at the plasmid ColE1 origin of replication to generate fragments of ~2400 and ~1500bp (black arrowheads). Int^{pTN3} concentration multipliers refer to the standard assay described in Materials and Methods.
(PDF)

S11 Fig. *LacZ* gene segments used for low sequence specificity reactions mimicking homologous recombination. DNA sequence of the *lacZ* gene segments cloned in plasmids pCB538 (lac100), pCB572 (lac175) and pCB574 (lac250) (Fig 8B).
(PDF)

S12 Fig. Integrase structure comparisons. The catalytic domain of Int^{pTN3} (**B**) was modeled using Phyre2 [Reference 5 in S1 Text] and compared using PyMol (The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.) with the tridimensional structure of the integrase of *Sulfolobus solfataricus* virus SSV1 (PDB 3VCF) (**A**) determined by Zhan et al [Reference 6 in S1 Text]. The Int^{pTN3} catalytic tyrosine residue is highlighted.
(PDF)

S1 Dataset. *Thermococcus nautili* 60G nucleotide sequence. Predominant *T*. *nautili* chromosome sequence obtained after sub-culturing for 60 generations.
(FASTA)

S2 Dataset. *Thermococcus nautili* 66G nucleotide sequence. Predominant *T*. *nautili* chromosome sequence obtained after sub-culturing for 66 generations.
(FASTA)

S3 Dataset. *Thermococcus* 5–4 36G nucleotide sequence. Predominant *T*. 5–4 chromosome sequence obtained after sub-culturing for 36 generations.
(FASTA)

S4 Dataset. *Thermococcus* 5–4 66G nucleotide sequence. Predominant *T*. 5–4 chromosome sequence obtained after sub-culturing for 66 generations.
(FASTA)

S1 Text. Supporting information references.
(DOCX)

Author Contributions

Conceptualization: JO.

Data curation: VB.

Funding acquisition: PF.

Investigation: MC CB RC DG.

Resources: EM.

Supervision: JO.

Visualization: JO.

Writing – original draft: JO.

Writing – review & editing: MC CB RC DG EM VB PF JO.

References

1. Lim K, Furuta Y, Kobayashi I (2012) Large Variations in Bacterial Ribosomal RNA Genes. Molecular Biology and Evolution 29: 2937–2948. https://doi.org/10.1093/molbev/mss101 PMID: 22446745
2. Anderson P, Roth J (1981) Spontaneous tandem genetic duplications in Salmonella typhimurium arise by unequal recombination between rRNA (rrn) cistrons. Proceedings of the National Academy of Sciences of the United States of America 78: 3113–3117. PMID: 6789329
3. Dorgai L, Oberto J, Weisberg RA (1993) Xis and Fis proteins prevent site-specific DNA inversion in lysogens of phage HK022. J Bacteriol 175: 693–700. PMID: 8423145
4. Iguchi A, Iyoda S, Terajima J, Watanabe H, Osawa R (2006) Spontaneous recombination between homologous prophage regions causes large-scale inversions within the Escherichia coli O157: H7 chromosome. Gene 372: 199–207. https://doi.org/10.1016/j.gene.2006.01.005 PMID: 16516407
5. Busseau I, Pelisson A, Bucheton A (1989) I-Elements of Drosophila-Melanogaster Generate Specific Chromosomal Rearrangements during Transposition. Molecular & General Genetics 218: 222–228.
6. Cui LZ, Neoh HM, Iwamoto A, Hiramatsu K (2012) Coordinated phenotype switching with large-scale chromosome flip-flop inversion observed in bacteria. Proceedings of the National Academy of Sciences of the United States of America 109: E1647–E1656. https://doi.org/10.1073/pnas.1204307109 PMID: 22645353
7. Daveran-Mingot ML, Campo N, Ritzenthaler P, Le Bourgeois P (1998) A natural large chromosomal inversion in Lactococcus lactis is mediated by homologous recombination between two insertion sequences. Journal of Bacteriology 180: 4834–4842. PMID: 9733685
8. Schindler D, Echols H (1981) Retroregulation of the int gene of bacteriophage lambda: control of translation completion. Proceedings of the National Academy of Sciences of the United States of America 78: 4475–4479. PMID: 6457302
9. Zivanovic Y, Lopez P, Philippe H, Forterre P (2002) Pyrococcus genome comparison evidences chromosome shuffling-driven evolution. Nucleic Acids Res 30: 1902–1910. PMID: 11972326
10. Carbone L, Harris RA, Gnerre S, Veeramah KR, Lorente-Galdos B, et al. (2014) Gibbon genome and the fast karyotype evolution of small apes. Nature 513: 195–201. https://doi.org/10.1038/nature13679 PMID: 25209798
11. Escudero JA, Loot C, Nivina A, Mazel D (2015) The Integron: Adaptation On Demand. Microbiology Spectrum 3: MDNA3-0019-2014

12. Skippington E, Ragan MA (2011) Lateral genetic transfer and the construction of genetic exchange communities. Fems Microbiology Reviews 35: 707–735. https://doi.org/10.1111/j.1574-6976.2010.00261.x PMID: 21223321

13. Dupressoir A, Vernochet C, Bawa O, Harper F, Pierron G, et al. (2009) Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. Proceedings of the National Academy of Sciences of the United States of America 106: 12127–12132. https://doi.org/10.1073/pnas.0902925106 PMID: 19564597

14. Siguier P, Gourbeyre E, Varani A, Bao TH, Chandler M (2015) Everyman's Guide to Bacterial Insertion Sequences. Microbiology Spectrum 3: MDNA3-0030-2014.

15. So M, Heffron F, McCarthy BJ (1979) The E. coli gene encoding heat stable toxin is a bacterial transposon flanked by inverted repeats of IS1. Nature 277: 453–456. PMID: 368646

16. Hickman AB, Chandler M, Dyda F (2010) Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. Critical Reviews in Biochemistry and Molecular Biology 45: 50–69. https://doi.org/10.3109/10409230903505596 PMID: 20067338

17. Grindley NDF, Whiteson KL, Rice PA (2006) Mechanisms of site-specific recombination. Annual Review of Biochemistry 75: 567–605. https://doi.org/10.1146/annurev.biochem.73.011303.073908 PMID: 16756503

18. Val ME, Bouvier M, Campos J, Sherratt D, Cornet F, et al. (2005) The single-stranded genome of phage CTX is the form used for integration into the genome of Vibrio cholerae. Molecular Cell 19: 559–566. https://doi.org/10.1016/j.molcel.2005.07.002 PMID: 16109379

19. Landy A (2015) The lambda Integrase Site-specific Recombination Pathway. Microbiology spectrum 3: MDNA3-0051-2014.

20. Yen Ting L, Sau S, Ma C-H, Kachroo AH, Rowley PA, et al. (2014) The partitioning and copy number control systems of the selfish yeast plasmid: an optimized molecular design for stable persistence in host cells. Microbiology spectrum 2: PLAS-0003-2013.

21. Mcleod M, Craft S, Broach JR (1986) Identification of the Crossover Site during Flp-Mediated Recombination in the Saccharomyces-Cerevisiae Plasmid 2-Mu-M Circle. Molecular and Cellular Biology 6: 3357–3367. PMID: 3540590

22. Dymecki SM (1996) Flp recombinase promotes site-specific DNA recombination in embryonic stem cells and transgenic mice. Proceedings of the National Academy of Sciences of the United States of America 93: 6191–6196. PMID: 8650242

23. Campbell AM (1963) Episomes. Advances in Genetics 11: 101–145.

24. Esposito D, Scocca JJ (1997) The integrase family of tyrosine recombinases: evolution of a conserved active site domain. Nucleic Acids Research 25: 3605–3614. PMID: 9278480

25. Yang W, Mizuuchi K (1997) Site-specific recombination in plane view. Structure 5: 1401–1406. PMID: 9384556

26. Grainge I, Jayaram M (1999) The integrase family of recombinases: organization and function of the active site. Molecular Microbiology 33: 449–456. PMID: 10577069

27. Cortez D, Quevillon-Cheruel S, Gribaldo S, Desnoues N, Sezonov G, et al. (2010) Evidence for a Xer/dif system for chromosome resolution in archaea. PLoS Genet 6: e1001166. https://doi.org/10.1371/journal.pgen.1001166 PMID: 20975945

28. She Q, Chen B, Chen L (2004) Archaeal integrases and mechanisms of gene capture. Biochemical Society Transactions 32: 222–226. https://doi.org/10.1042/ PMID: 15046576

29. Serre MC, Letzelter C, Garel JR, Duguet M (2002) Cleavage properties of an archaeal site-specific recombinase, the SSV1 integrase. Journal of Biological Chemistry 277: 16758–16767. https://doi.org/10.1074/jbc.M200707200 PMID: 11875075

30. Zhan ZY, Zhou J, Huang L (2015) Site-Specific Recombination by SSV2 Integrase: Substrate Requirement and Domain Functions. Journal of Virology 89: 10934–10944. https://doi.org/10.1128/JVI.01637-15 PMID: 26292330

31. Jaubert C, Danioux C, Oberto J, Cortez D, Bize A, et al. (2013) Genomics and genetics of Sulfolobus islandicus LAL14/1, a model hyperthermophilic archaeon. Open Biol 3: 130010. https://doi.org/10.1098/rsob.130010 PMID: 23594878

32. Bridger SL, Lancaster WA, Poole FL 2nd, Schut GJ, Adams MW (2012) Genome sequencing of a genetically tractable Pyrococcus furiosus strain reveals a highly dynamic genome. J Bacteriol 194: 4097–4106. https://doi.org/10.1128/JB.00439-12 PMID: 22636780

33. Cossu M, Da Cunha V, Toffano-Nioche C, Forterre P, Oberto J (2015) Comparative genomics reveals conserved positioning of essential genomic clusters in highly rearranged Thermococcales chromosomes. Biochimie 118: 313–321. https://doi.org/10.1016/j.biochi.2015.07.008 PMID: 26166067

34. Filippo JS, Sung P, Klein H (2008) Mechanism of eukaryotic homologous recombination. Annual Review of Biochemistry 77: 229–257. https://doi.org/10.1146/annurev.biochem.77.061306.125255 PMID: 18275380

35. Michel B, Leach D (2012) Homologous Recombination-Enzymes and Pathways. EcoSal Plus 5.

36. White MF (2011) Homologous recombination in the archaea: the means justify the ends. Biochemical Society Transactions 39: 15–19. https://doi.org/10.1042/BST0390015 PMID: 21265740

37. Constantinesco F, Forterre P, Elie C (2002) NurA, a novel 5 '-3 ' nuclease gene linked to rad50 and mre11 homologs of thermophilic Archaea. Embo Reports 3: 537–542. https://doi.org/10.1093/embo-reports/kvf112 PMID: 12052775

38. Graham WJt, Rolfsmeier ML, Haseltine CA (2013) An archaeal RadA paralog influences presynaptic filament formation. DNA Repair (Amst) 12: 403–413.

39. Sato T, Fukui T, Atomi H, Imanaka T (2005) Improved and versatile transformation system allowing multiple genetic manipulations of the hyperthermophilic archaeon Thermococcus kodakaraensis. Applied and Environmental Microbiology 71: 3889–3899. https://doi.org/10.1128/AEM.71.7.3889-3899.2005 PMID: 16000802

40. Gorlas A, Croce O, Oberto J, Gauliard E, Forterre P, et al. (2014) Thermococcus nautili sp. nov., a hyperthermophilic archaeon isolated from a hydrothermal deep sea vent (East Pacific Ridge). Int J Syst Evol Microbiol 64: 1802–1810. https://doi.org/10.1099/ijs.0.060376-0 PMID: 24556637

41. Oberto J, Gaudin M, Cossu M, Gorlas A, Slesarev A, et al. (2014) Genome Sequence of a Hyperthermophilic Archaeon, Thermococcus nautili 30–1, That Produces Viral Vesicles. Genome Announc 2: e00243–00214. https://doi.org/10.1128/genomeA.00243-14 PMID: 24675865

42. Oberto J (2013) SyntTax: a web server linking synteny to prokaryotic taxonomy. BMC Bioinformatics 14: 4–13. https://doi.org/10.1186/1471-2105-14-4 PMID: 23323735

43. Gaudin M, Krupovic M, Marguet E, Gauliard E, Cvirkaite-Krupovic V, et al. (2013) Extracellular membrane vesicles harbouring viral genomes. Environ Microbiol 16: 1167–1175. https://doi.org/10.1111/1462-2920.12235 PMID: 24034793

44. Letzelter C, Duguet M, Serre MC (2004) Mutational analysis of the archaeal tyrosine recombinase SSV1 integrase suggests a mechanism of DNA cleavage in trans. Journal of Biological Chemistry 279: 28936–28944. https://doi.org/10.1074/jbc.M403971200 PMID: 15123675

45. Krupovic M, Bamford DH (2008) Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota. Virology 375: 292–300. https://doi.org/10.1016/j.virol.2008.01.043 PMID: 18308362

46. Raeside C, Gaffe J, Deatherage DE, Tenaillon O, Briska AM, et al. (2014) Large Chromosomal Rearrangements during a Long-Term Evolution Experiment with Escherichia coli. Mbio 5: e01377–01314. https://doi.org/10.1128/mBio.01377-14 PMID: 25205090

47. Eisen JA, Heidelberg JF, White O, Salzberg SL (2000) Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biology 1: research0011.0011–0011.0019.

48. Haberle RC, Fourcade HM, Boore JL, Jansen RK (2008) Extensive rearrangements in the chloroplast genome of Trachelium caeruleum are associated with repeats and tRNA genes. Journal of Molecular Evolution 66: 350–361. https://doi.org/10.1007/s00239-008-9086-4 PMID: 18330485

49. Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, et al. (1989) The Complete Sequence of the Rice (Oryza-Sativa) Chloroplast Genome—Intermolecular Recombination between Distinct Transfer-Rna Genes Accounts for a Major Plastid DNA Inversion during the Evolution of the Cereals. Molecular & General Genetics 217: 185–194.

50. Williams KP (2002) Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. Nucleic Acids Research 30: 866–875. PMID: 11842097

51. Chen JW, Lee J, Jayaram M (1992) DNA Cleavage in Trans by the Active-Site Tyrosine during Flp Recombination—Switching Protein Partners before Exchanging Strands. Cell 69: 647–658. PMID: 1586945

52. Lee J, Jayaram M, Grainge I (1999) Wild-type Flp recombinase cleaves DNA in trans. Embo Journal 18: 784–791. https://doi.org/10.1093/emboj/18.3.784 PMID: 9927438

53. Wang ZJ, Le GW, Shi YH, Wegrzyn G, Wrobel B (2002) A model for regulation of colE1-like plasmid replication by uncharged tRNAs in amino acid-starved Escherichia coli cells. Plasmid 47: 69–78. https://doi.org/10.1006/plas.2001.1562 PMID: 11982328

54. Santangelo TJ, Cubonova L, Reeve JN (2008) Shuttle vector expression in Thermococcus kodakaraensis: contributions of cis elements to protein synthesis in a hyperthermophilic archaeon. Appl Environ Microbiol 74: 3099–3104. https://doi.org/10.1128/AEM.00305-08 PMID: 18378640

55. Marguet E, Gaudin M, Gauliard E, Fourquaux I, le Blond du Plouy S, et al. (2013) Membrane vesicles, nanopods and/or nanotubes produced by hyperthermophilic archaea of the genus Thermococcus. Biochem Soc Trans 41: 436–442. https://doi.org/10.1042/BST20120293 PMID: 23356325

56. Lepage E, Marguet E, Geslin C, Matte-Tailliez O, Zillig W, et al. (2004) Molecular diversity of new Thermococcales isolates from a single area of hydrothermal deep-sea vents as revealed by randomly amplified polymorphic DNA fingerprinting and 16S rRNA gene sequence analysis. Appl Environ Microbiol 70: 1277–1286. https://doi.org/10.1128/AEM.70.3.1277-1286.2004 PMID: 15006744

57. Krumsiek J, Arnold R, Rattei T (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23: 1026–1028. https://doi.org/10.1093/bioinformatics/btm039 PMID: 17309896

58. Despalins A, Marsit S, Oberto J (2011) Absynte: a web tool to analyze the evolution of orthologous archaeal and bacterial gene clusters. Bioinformatics 27: 2905–2906. https://doi.org/10.1093/bioinformatics/btr473 PMID: 21840875

59. Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA, et al. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. Nature Methods 6: 343–U341. https://doi.org/10.1038/nmeth.1318 PMID: 19363495

60. Sato T, Fukui T, Atomi H, Imanaka T (2003) Targeted gene disruption by homologous recombination in the hyperthermophilic archaeon Thermococcus kodakaraensis KOD1. Journal of Bacteriology 185: 210–220. https://doi.org/10.1128/JB.185.1.210-220.2003 PMID: 12486058

S1 Table. Plasmids used in this work

Name	Backbone	Insertion	Selection	Reference
pUC18	-	-	ApR	[1] in S1 Text
pBR322	-	-	ApR, TcR	[2] in S1 Text
pUC4K	-	-	ApR, KmR	[3] in S1 Text
pET-26b+	pBR322	-	KmR	Novagen®, USA
pLC70	pCR2.1-TOPO pTN1	-	ApR, KmR, trpE, HMG-CoA red.	[54]
pJO344	pET-26b+	integrase gene from plasmid pTN3	KmR	This work
pJO496	pJO344	Int^{pTN3} Y428A allele	KmR	This work
pJO322	pUC18	$tRNA^{Leu}$ gene (2-88bp) from *T. nautili*	ApR	This work
pMC451	pBR322	$tRNA^{Leu}$ gene (2-88bp) from *T. nautili*	ApR, TcR	This work
pMC449	pUC18	$tRNA^{Leu}$ gene (2-44bp)	ApR	This work
pMC477	pMC451	$tRNA^{Leu}$ gene (2-44bp)+lacZα from pMC449	ApR	This work
pMC479	pMC451	$tRNA^{Leu}$ gene (2-44bp)+lacZα from pMC449	ApR	This work
pRC524	pLC70	integrase gene from plasmid pTN3	ApR, KmR, trpE, HMG-CoA red.	This work
pRC526	pLC70	Y428A mutant of integrase gene from plasmid pTN3	ApR, KmR, trpE, HMG-CoA red.	This work
pCB538	pUC18	LacZ100 inverted fragment	ApR, KmR	This work
pCB548	pUC18	genomic region of $tRNA^{Gly}$ (gene ID: BD01_1557) + genomic region of $tRNA^{Gly}$ (gene ID: BD01_1976)	ApR, KmR	This work
pCB552	pUC18	genomic fragment of BD01_1166 + genomic fragment of BD01_1584	ApR, KmR	This work
pCB572	pUC18	LacZ175 inverted fragment	ApR, KmR	This work
pCB574	pUC18	LacZ250 inverted fragment	ApR, KmR	This work

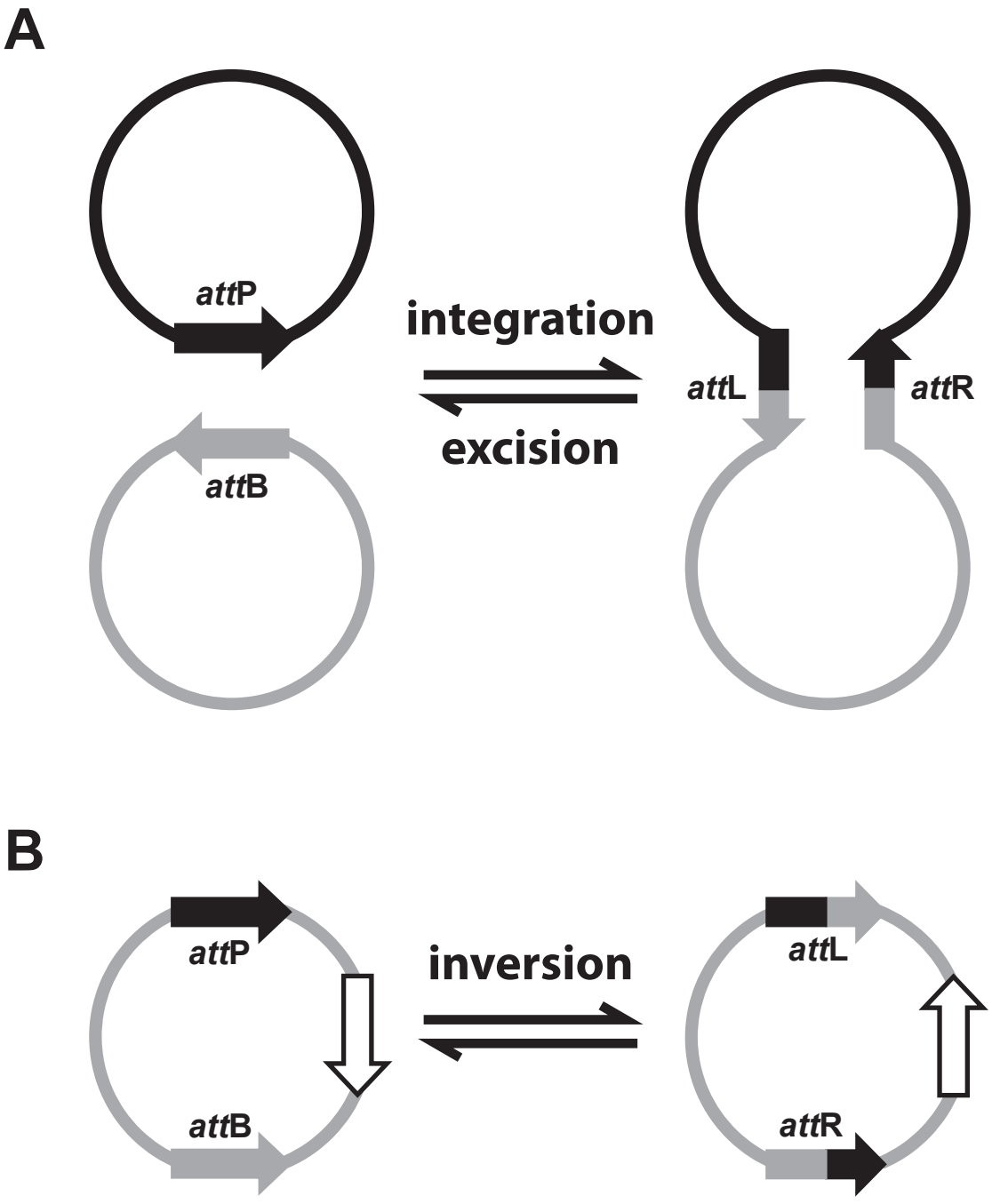
S2 Table. Oligonucleotides used in this work.

Name	Sequence 5'-3'	Usage	Description
LacZ100-Sac1-For	ctgacgtcctattacgccagctggcgaaagg	pCB538	PCR amplification lacZ100-KmR
KanR-XBA1-Rev	gcgtctagaagccacgttgtgtctcaaaatctctg	pCB538	PCR amplification lacZ100-KmR
1557_fwd	agtgccaagcttgcatgcctgcaggtcgacGGTAGCTCAGCCTGGGAGAG	pCB548	Gibson cloning
1557_rev	acacaacgtggctAGTAAGTGAGGAGTGAAGCTCCAC	pCB548	Gibson cloning
KanR_fwd	ctcctcacttactAGCCACGTTGTGTCTCAAAATC	pCB548	Gibson cloning
KanR_rev	atctgctctacttCGCTGAGGTCTGCCTCGT	pCB548	Gibson cloning
1976_fwd	gcagacctcagcgAAGTAGAGCAGATTTTGCTCataaatcg	pCB548	Gibson cloning
1976_rev	ggaaacagctatgaccatgattacgaattcAATCGTCCGTTTAATCGTCgc	pCB548	Gibson cloning
1166_fwd-GIBSON	agtgccaagcttgcatgcctgcaggtcgacGAGCACCGAGAAGGGCGT	pCB552	Gibson cloning
1166_rev-GIBSON	acacaacgtggctGGTCAGAAGAAAAGGAAAATACGAG	pCB552	Gibson cloning
KanR_fwd-GIBSON	ttttcttctgaccAGCCACGTTGTGTCTCAAAATC	pCB552	Gibson cloning
KanR_rev-GIBSON	tttgaggggtgacgctgaggtctgcctcgt	pCB552	Gibson cloning
1584_fwd-GIBSON	gcagacctcagcgTCACCCCCTCAAAGTGGAAAGG	pCB552	Gibson cloning
1584_rev-GIBSON	ggaaacagctatgaccatgattacgaattcGAGCACCCAGCGCGGTGT	pCB552	Gibson cloning
GRep-KanR-F	atccccgggtaccgagctcgAAGCCACGTTGTGTCTCAAAATC	pCB572/574	Gibson cloning
GRep-KanR-R	acggccagtgCGCTGAGGTCTGCCTCGT	pCB572/574	Gibson cloning
GRep-Repeat-F	gacctcagcgCACTGGCCGTCGTTTTAC	pCB572/574	Gibson cloning
GRep-Lac175-ScaI_F	cagctatgaccatgattacgAGTACTAAATACCGCATCAGG	pCB572	Gibson cloning
GRep-Lac250-ScaI_R	cagctatgaccatgattacgAGTACTTATGCGGCATCAG	pCB574	Gibson cloning
Int_A	cacgttccaacgagcattggcgaccgccaacgtttttcggg	Mutagenesis pJO496	The Y428A mutation is underlined
Int_B	cccgaaaaacgttggcggtcgcaatgctcgttggaacgtg	Mutagenesis pJO496	The Y428A mutation is underlined
int_fwd	GATCGTCGACagcgatatatttatatagggatatagtaatagataatatcacaggtggtataga**ATGGTAAAATCGGGTGGTGTGTACG**	pRC526/548	Bold nucleotides refer to the coding sequence of the integrase gene, underline indicates *Sal*I and *Not*I sites, and lower case indicates P*hmtB* promoter sequence added for expression in *T. kodakarensis*
int_rev	GATCGCGGCCGC**TAAAGCTCCAGAATCCCCAGC**	pRC526/548	idem
intY428A_fwd	CGTGGGTGGCAGGAACGCCCGCTGGAACGTCAAAAACG	pRC526/548 mutagenesis	The Y428A mutation is underlined
TKV4_FW	catgtgtcgttcctggtcgg	*In vitro/in vivo* excision assay	
TKV4_REV	gggaggtaagacgggtaaggc	*In vitro/in vivo* excision assay	
Leu43*sca*I_fw	caagtactctatgcggcatcagagcag	pMC477/479	Underline indicates *Sca*I site
Leu43*sca*I_rev	caaagtactctggaaagcgggcagtgag	pMC477/479	Underline indicates *Sca*I site

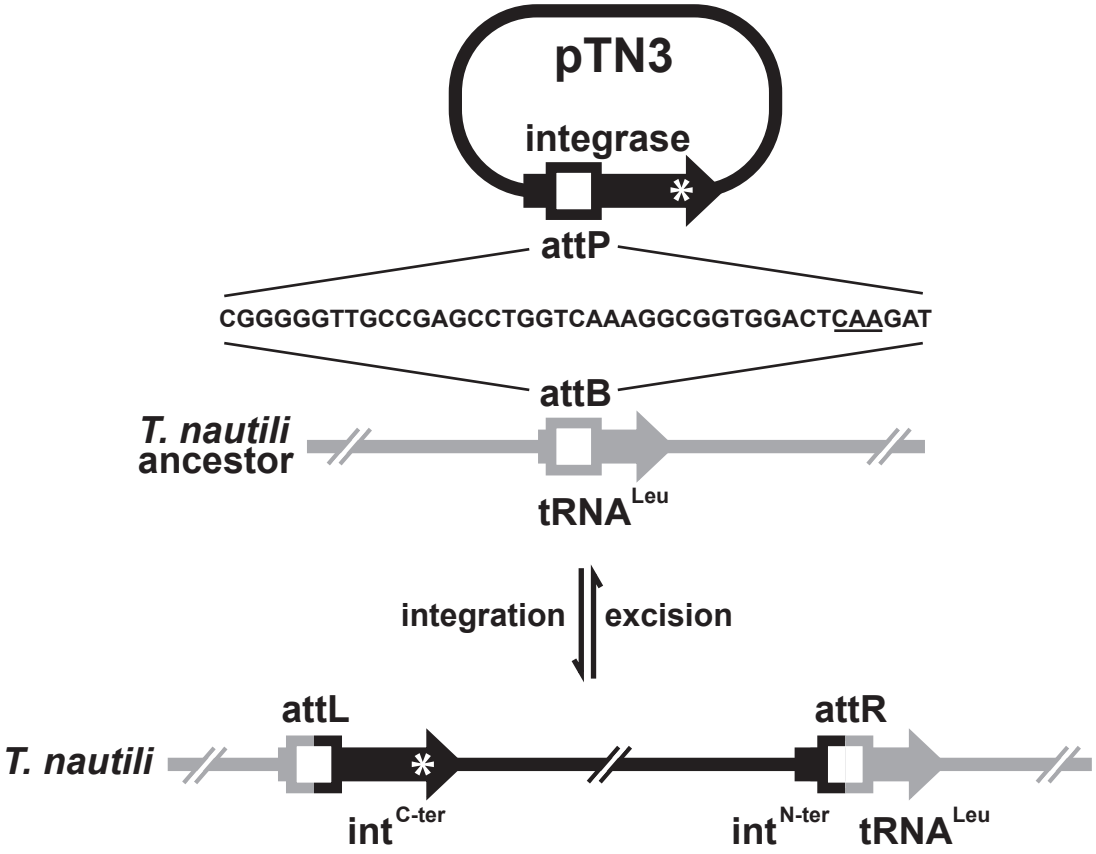
S3 Table. Metagenomic reads mapping *(T. nautili* 66G)

Gene inversion	1166	1333	1317	1557
original reads (position 1)	46%	47%	58%	55%
inversion reads (position 1)	54%	53%	41%	45%
original reads (position 2)	45%	47%	50%	39%
inversion reads (Position 2)	54%	53%	49%	60%

	Average	SD
Reads mapped to *T. nautili* 66G chromosome excluding integrated pTN3.	224.9	27.5
Reads mapped to pTN3	310.2	35.6
Average number of pTN3 per chromosome	1.38	


A
attP
attB
integration
excision
attL
attR
B
attP
attB
inversion
attL
attR

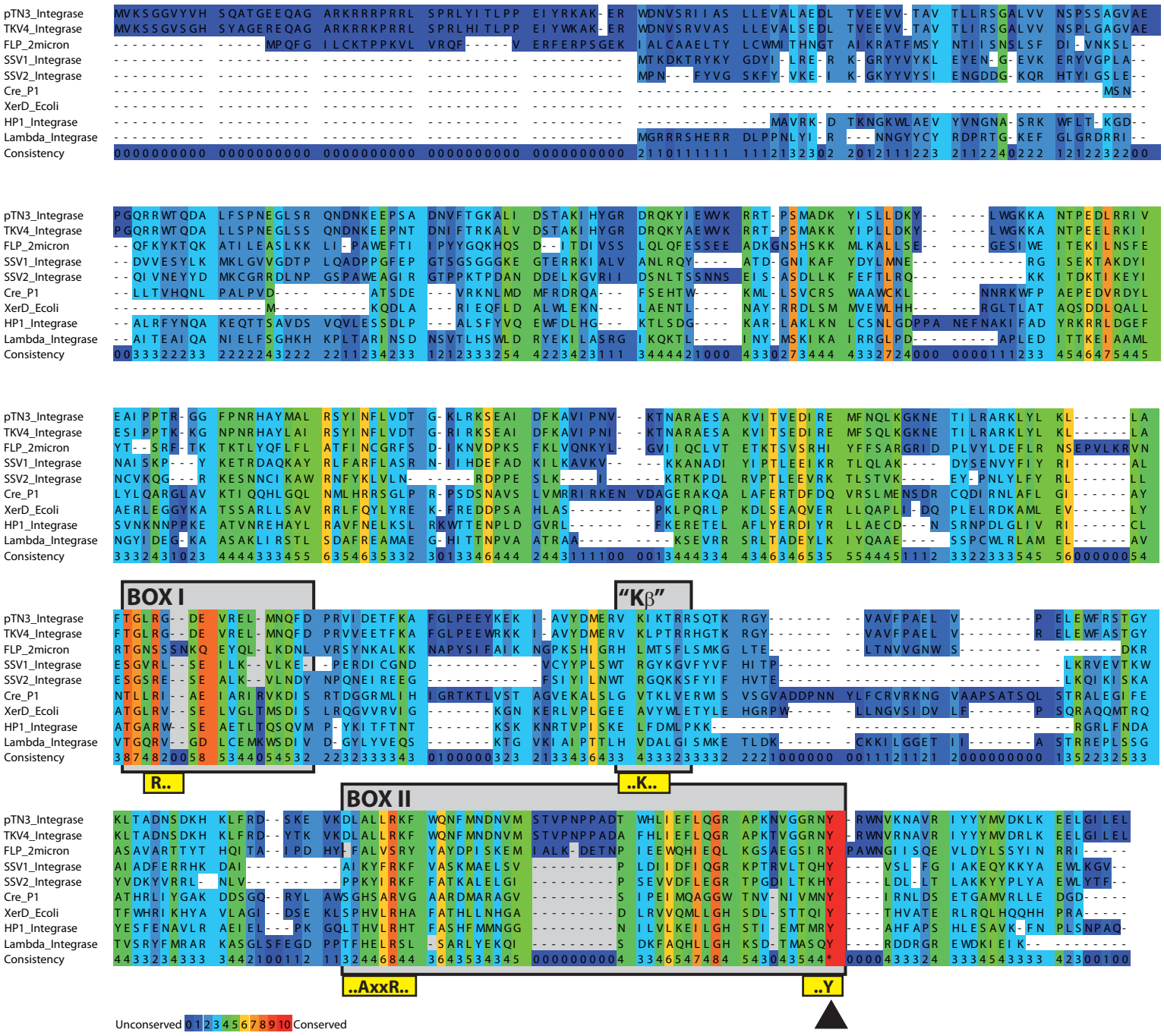
A



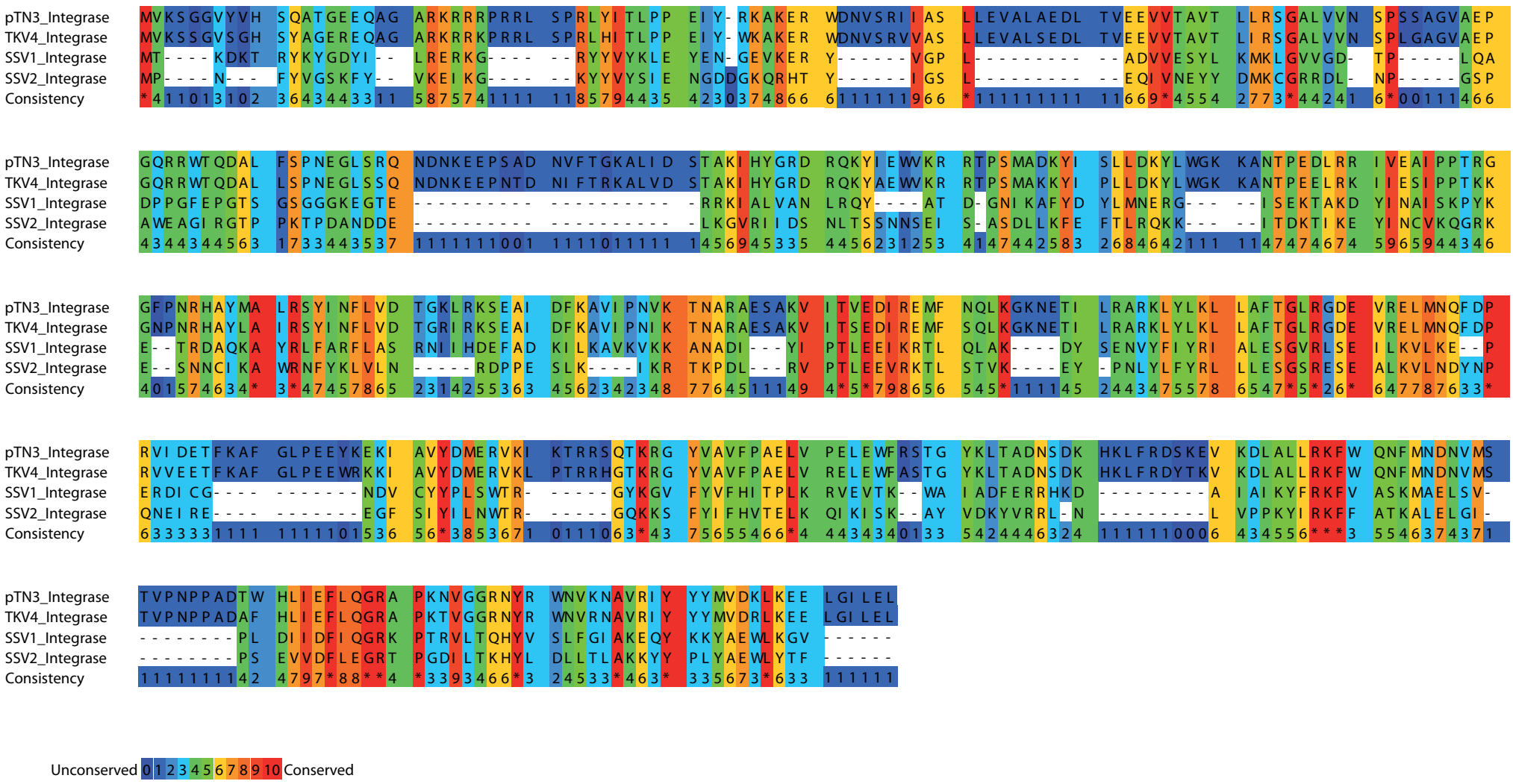
B

```
intpTN3                                                                                                ATGGTAAAATCGGGT
intpTN3   GGTGTGTACGTACACTCCCAAGCGACCGGAGAGGAGCAGGCCGGAGCGCGGAAGCGGAGGCGTCCGAGGCGCCTTTCCCCGCGTCTGT
intpTN3   ACATTACGCTACCGCCAGAAATCTATCGGAAGGCCAAGGAGCGCTGGGATAACGTGAGCCGAATCATCGCAAGCCTGCTTGAGGTGGC
intpTN3   TTTGGCTGAGGATTTAACGGTCGAGGAGGTCGTGACGGCCGTAACGCTCCTTAGGAGTGGCGCTTTGGTGGTGAATTCGCCTTCGAGC
                                attB
tRNALeu   GCGGGGGTTGCCGAGCCTGGTCAAAGGCGGTGGACTCAAGATCCACTCCCGCAGGGGTTCCGGGGTTCAAATCCCCGCCCCCGCACCA
          ****************************************** * **         *      **       *  *    *  ** **
intpTN3   GCGGGGGTTGCCGAGCCTGGTCAAAGGCGGTGGACTCAAGATGCTCTTTTTTCCCCGAATGAAGGCCTCTCCCGTCAGAACGACAACA
                                attP
intpTN3   AAGAAGAGCCGAGCGCCGATAACGTTTTTACAGGAAAGGCTTTGATAGACTCAACGGCCAAAATCCACTATGGTCGTGATAGACAGAA
intpTN3   ATACATCGAATGGGTGAAACGGCGCACGCCAAGCATGGCCGACAAATACATTTCTCTGCTTGACAAGTACCTCTGGGGAAAGAAAGCC
intpTN3   AATACTCCAGAGGACCTCCGGCGCATTGTAGAAGCTATCCCTCCCACCAGGGGAGGCTTTCCCAATAGGCATGCCTACATGGCGTTGA
intpTN3   GGAGCTACATTAACTTCCTTGTGGATACCGGAAAGCTGAGGAAGAGTGAGGCCATTGACTTCAAGGCCGTGATTCCGAACGTTAAGAC
intpTN3   CAACGCTCGCGCTGAATCCGCGAAGGTCATAACGGTTGAGGACATTCGTGAGATGTTCAACCAGCTCAAGGGGAAGAACGAGACGATT
intpTN3   CTCAGAGCGCGCAAGCTTTACCTCAAGCTTCTCGCCTTTACAGGTCTCAGGGGAGACGAGGTCCGCGAGCTGATGAACCAGTTCGACC
intpTN3   CGAGGGTTATTGACGAGACATTCAAGGCCTTTGGCCTTCCTGAGGAATACAAGGAGAAGATAGCGGTCTATGATATGGAGCGGGTGAA
intpTN3   GATTAAGACGAGGAGGAGTCAGACGAAGCGTGGCTATGTCGCGGTCTTTCCCGCTGAGCTCGTTCCCGAGCTGGAGTGGTTCAGGAGC
intpTN3   ACTGGGTACAAACTCACTGCGGACAACTCTGATAAGCATAAGCTGTTCAGGGATTCCAAGGAGGTTAAGGACCTGGCCTTGCTGAGAA
intpTN3   AGTTCTGGCAGAACTTCATGAACGACAATGTGATGAGCACGGTTCCAAACCCTCCTGCTGATACCTGGCACCTCATTGAGTTCCTCCA
intpTN3   GGGACGCGCTCCCAAAAACGTGGGTGGCAGGAACTACCGCTGGAACGTCAAAAACGCCGTGAGAATCTATTATTACATGGTGGACAAA
intpTN3   TTGAAAGAGGAGCTGGGGATTCTGGAGCTTTAG
```

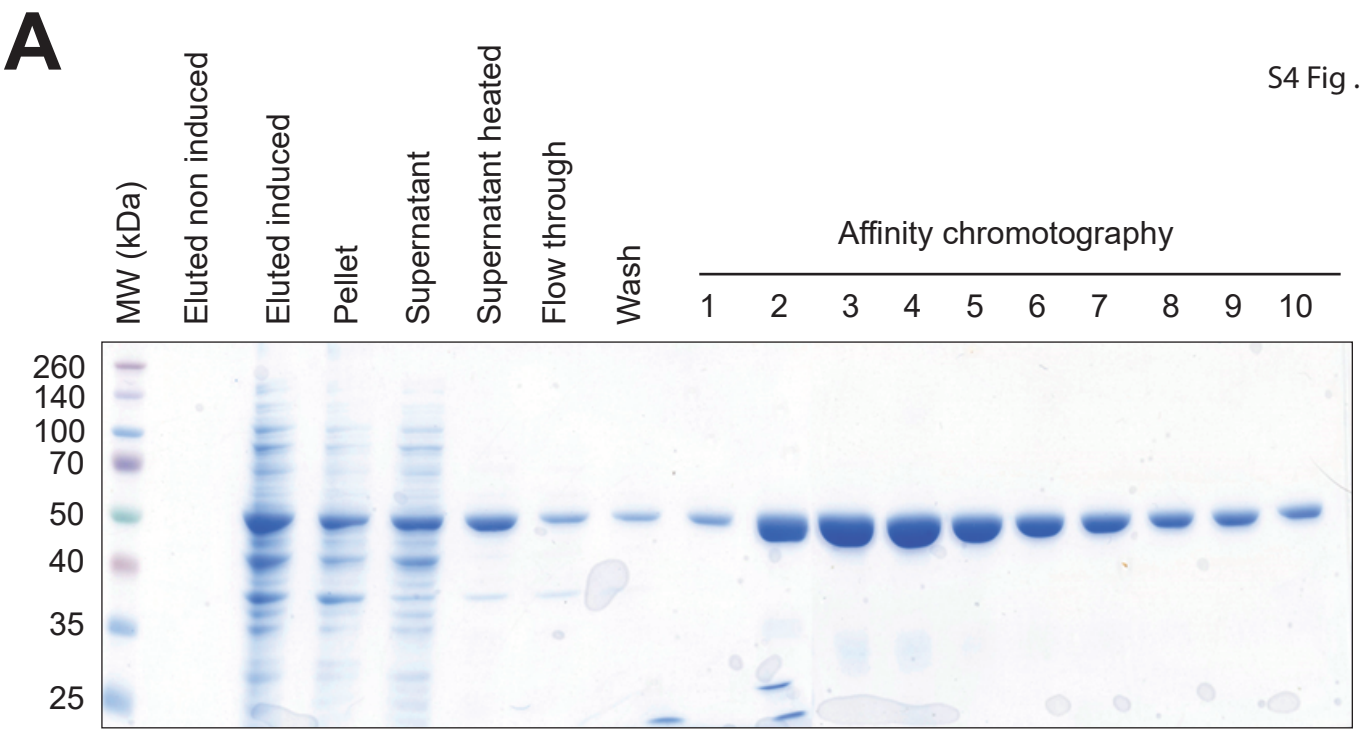
A

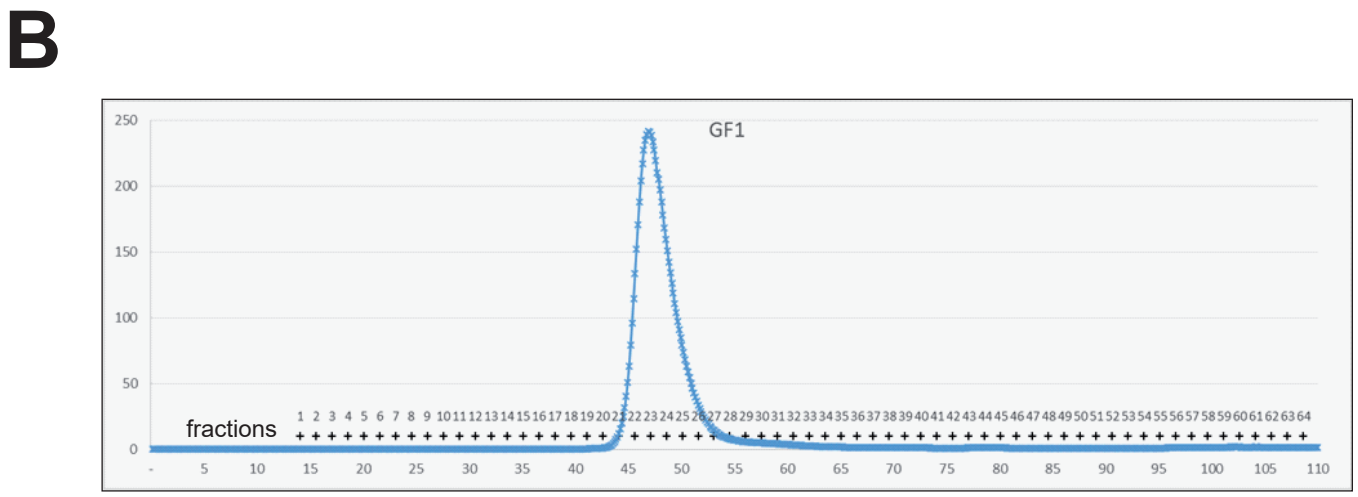

pTN3_Integrase
TKV4_Integrase
FLP_2micron
SSV1_Integrase
SSV2_Integrase
Cre_P1
XerD_Ecoli
HP1_Integrase
Lambda_Integrase
Consistency
BOX I
"Kβ"
R..
..K..
BOX II
..AxxR..
..Y
Unconserved 0 1 2 3 4 5 6 7 8 9 10 Conserved

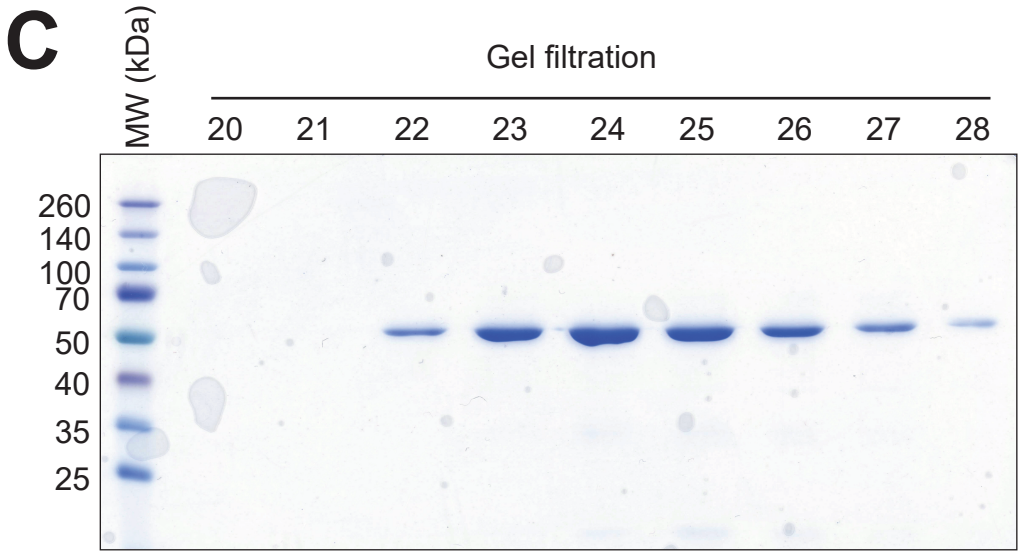
B


pTN3_Integrase
TKV4_Integrase
SSV1_Integrase
SSV2_Integrase
FLP_2micron
Cre_P1
XerD_Ecoli
HP1_Integrase
Lambda_Integrase
Consistency
Unconserved 0 1 2 3 4 5 6 7 8 9 10 Conserved

S4 Fig .


A
MW (kDa)
Eluted non induced
Eluted induced
Pellet
Supernatant
Supernatant heated
Flow through
Wash
Affinity chromotography
1 2 3 4 5 6 7 8 9 10
260
140
100
70
50
40
35
25

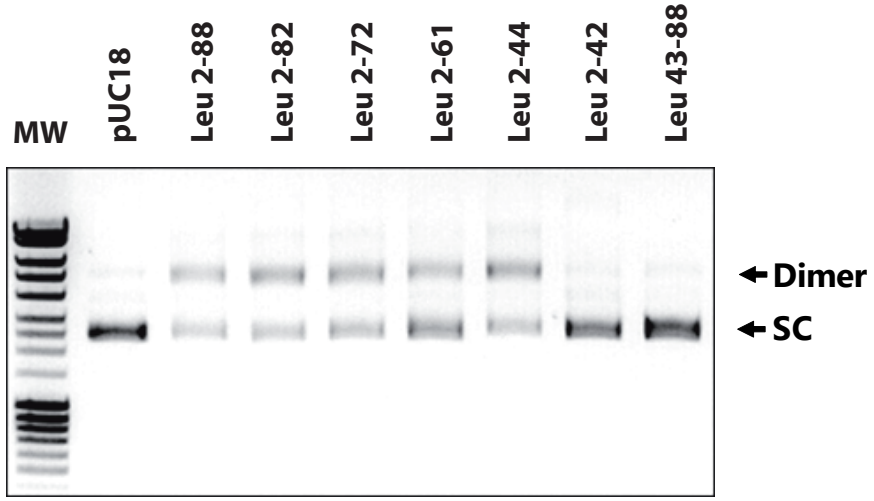

B
GF1
fractions


C
MW (kDa)
Gel filtration
20 21 22 23 24 25 26 27 28
260
140
100
70
50
40
35
25

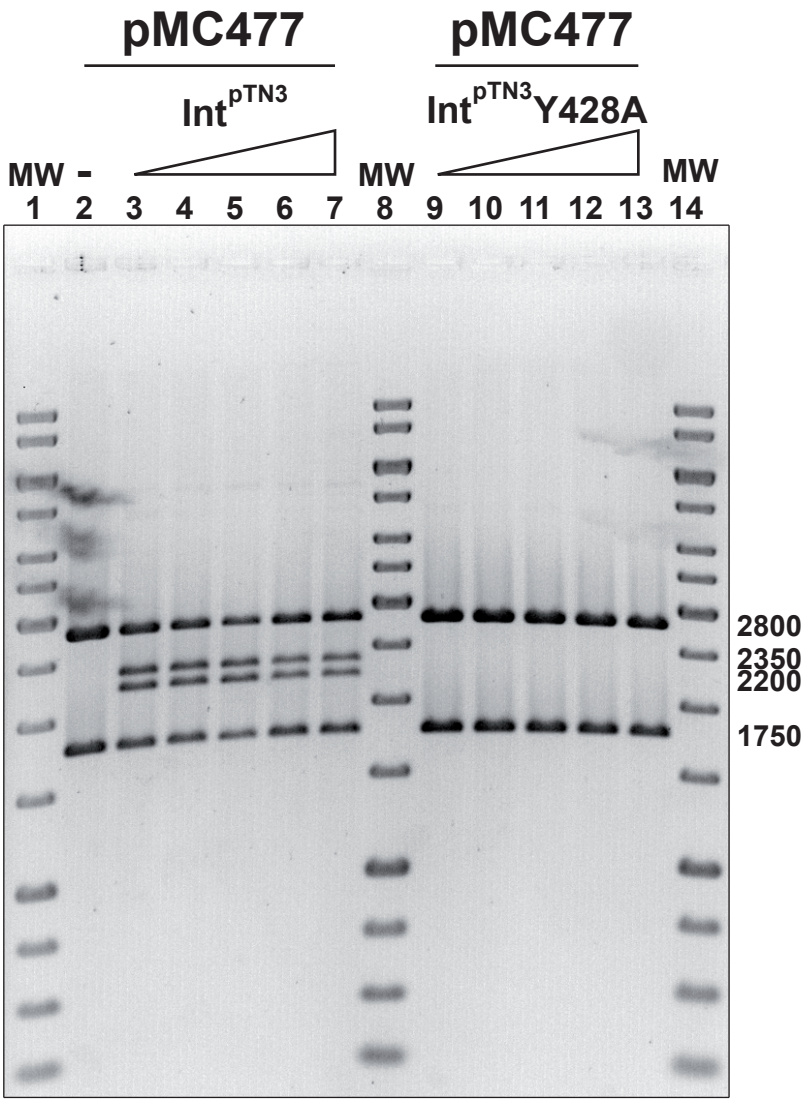
A

attB deletion	Dimerisation	Sequence	Plasmid
Leu WT	ND	GCGGGGGTTGCCGAGCCTGGTCAAAGGCGGTGGACTCAAGATCCACTCCCGCAGGGGTTCCGGGGTTCAAATCCCCGCCCCCGCACCA	–
Leu 43-88	–	CCACTCCCGCAGGGGTTCCGGGGTTCAAATCCCCGCCCCCGCACCA	pJO425
Leu 2-88	+	CGGGGGTTGCCGAGCCTGGTCAAAGGCGGTGGACTCAAGATCCACTCCCGCAGGGGTTCCGGGGTTCAAATCCCCGCCCCCGCACCA	pJO322
Leu 2-82	+	CGGGGGTTGCCGAGCCTGGTCAAAGGCGGTGGACTCAAGATCCACTCCCGCAGGGGTTCCGGGGTTCAAATCCCCGCCCCC	pMC435
Leu 2-72	+	CGGGGGTTGCCGAGCCTGGTCAAAGGCGGTGGACTCAAGATCCACTCCCGCAGGGGTTCCGGGGTTCAAAT	pMC433
Leu 2-61	+	CGGGGGTTGCCGAGCCTGGTCAAAGGCGGTGGACTCAAGATCCACTCCCGCAGGGGTTCC	pMC431
Leu 2-47	+	CGGGGGTTGCCGAGCCTGGTCAAAGGCGGTGGACTCAAGATCCACT	pMC429
Leu 2-45	+	CGGGGGTTGCCGAGCCTGGTCAAAGGCGGTGGACTCAAGATCCA	pMC443
Leu 2-44	+	CGGGGGTTGCCGAGCCTGGTCAAAGGCGGTGGACTCAAGATCC	pMC449
Leu 2-43	–	CGGGGGTTGCCGAGCCTGGTCAAAGGCGGTGGACTCAAGATC	pMC441
Leu 2-42-*attP*	–	CGGGGGTTGCCGAGCCTGGTCAAAGGCGGTGGACTCAAGAT	pJO421
Leu 5-44	–	GGGTTGCCGAGCCTGGTCAAAGGCGGTGGACTCAAGATCC	pJO459
Leu 8-44	–	TTGCCGAGCCTGGTCAAAGGCGGTGGACTCAAGATCC	pJO461
Leu 12-44	–	CGAGCCTGGTCAAAGGCGGTGGACTCAAGATCC	pJO463
Leu 17-44	–	CTGGTCAAAGGCGGTGGACTCAAGATCC	pJO465
		**	
attP		CGGGGGTTGCCGAGCCTGGTCAAAGGCGGTGGACTCAAGAT	

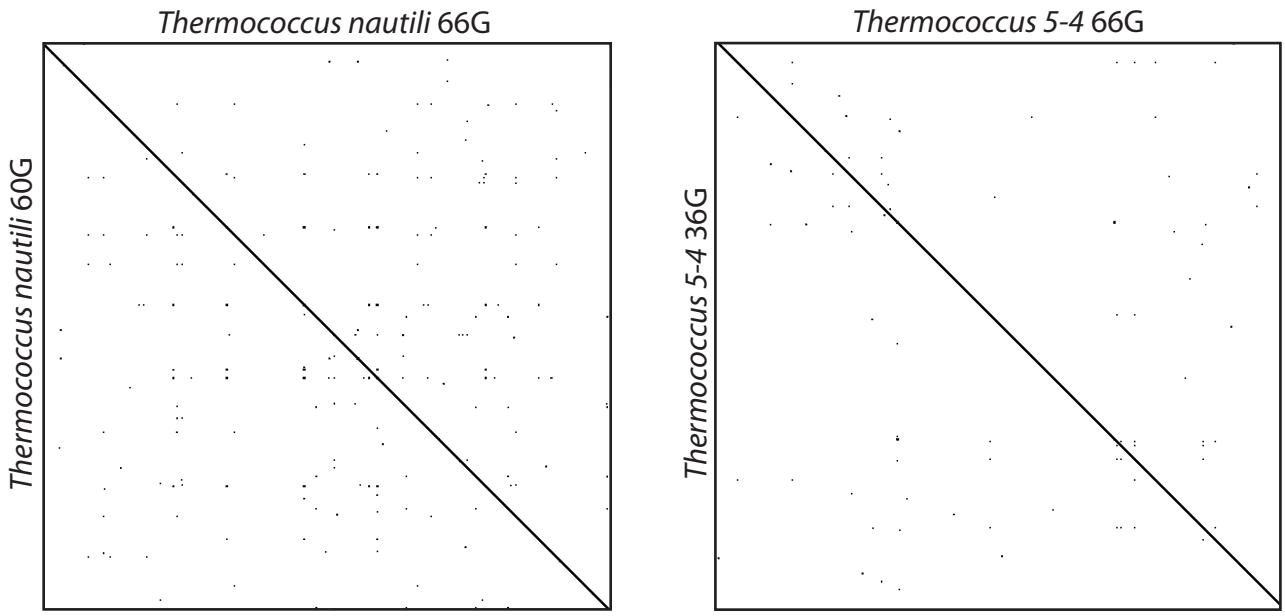
B



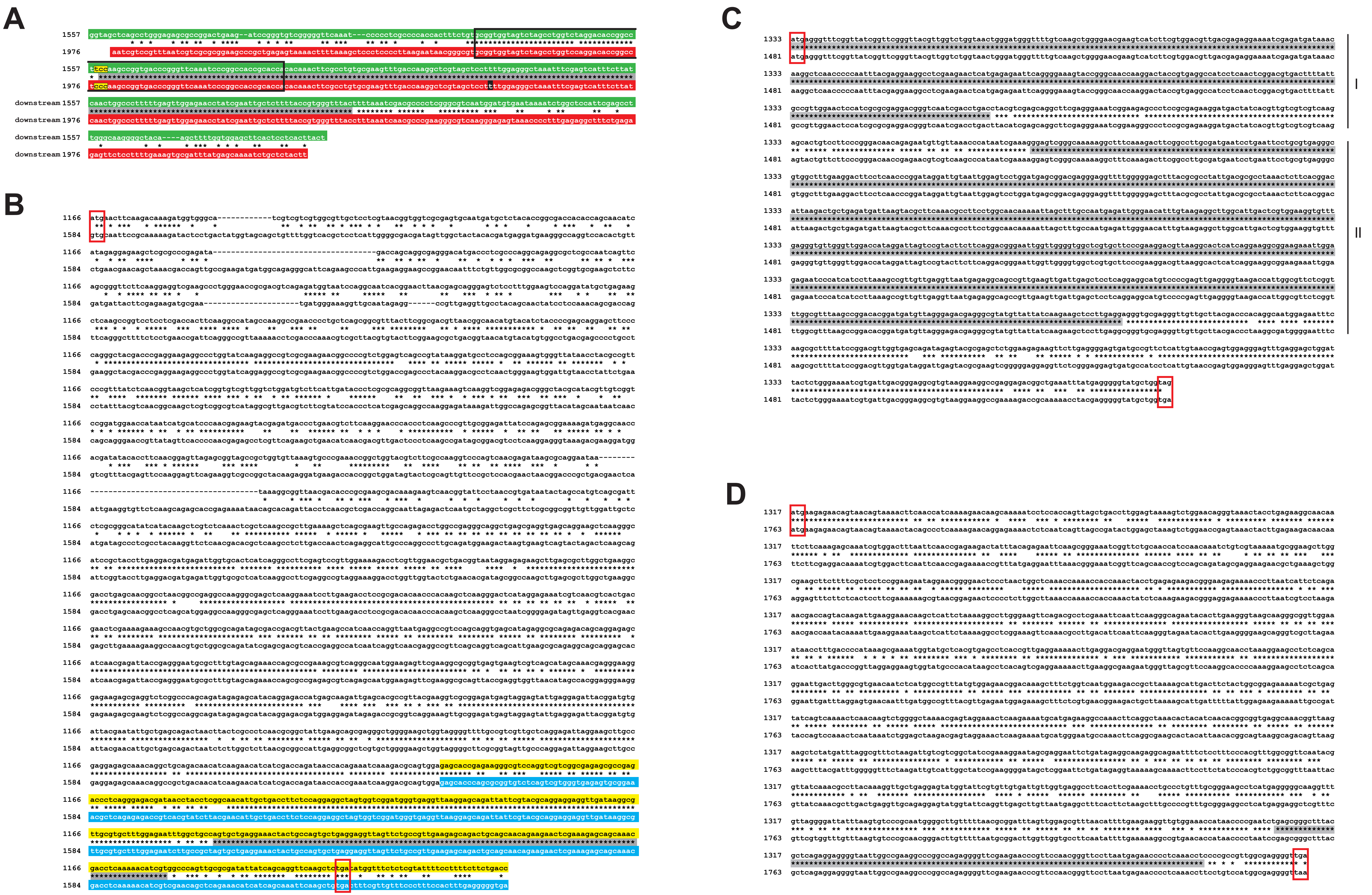
S6 Fig.

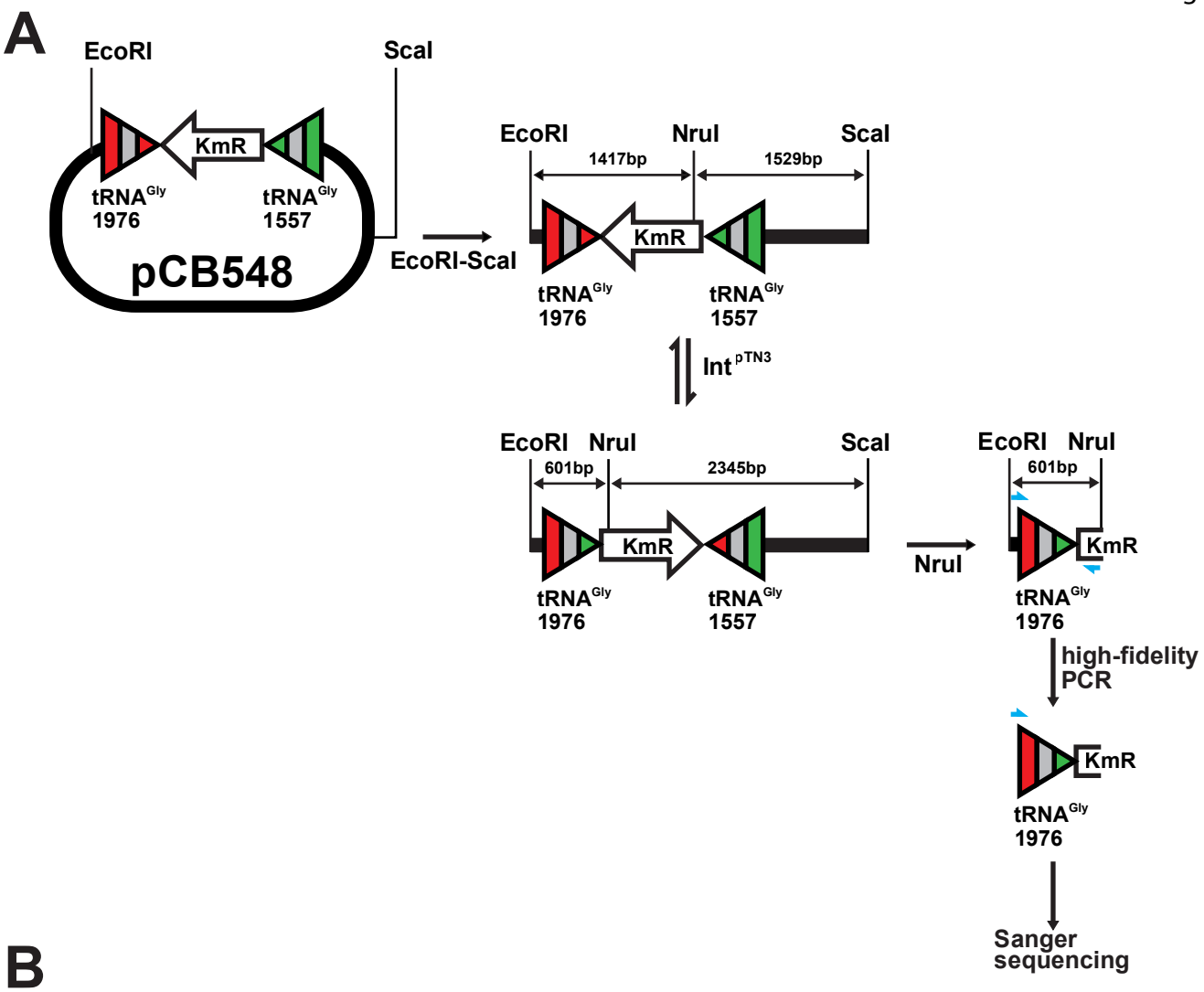

pMC477
pMC477
Int^pTN3
Int^pTN3Y428A
MW - MW MW
1 2 3 4 5 6 7 8 9 10 11 12 13 14
2800
2350
2200
1750

S7 Fig.


Thermococcus nautili 66G
Thermococcus nautili 60G
Thermococcus 5-4 66G
Thermococcus 5-4 36G

S8 Fig.





B

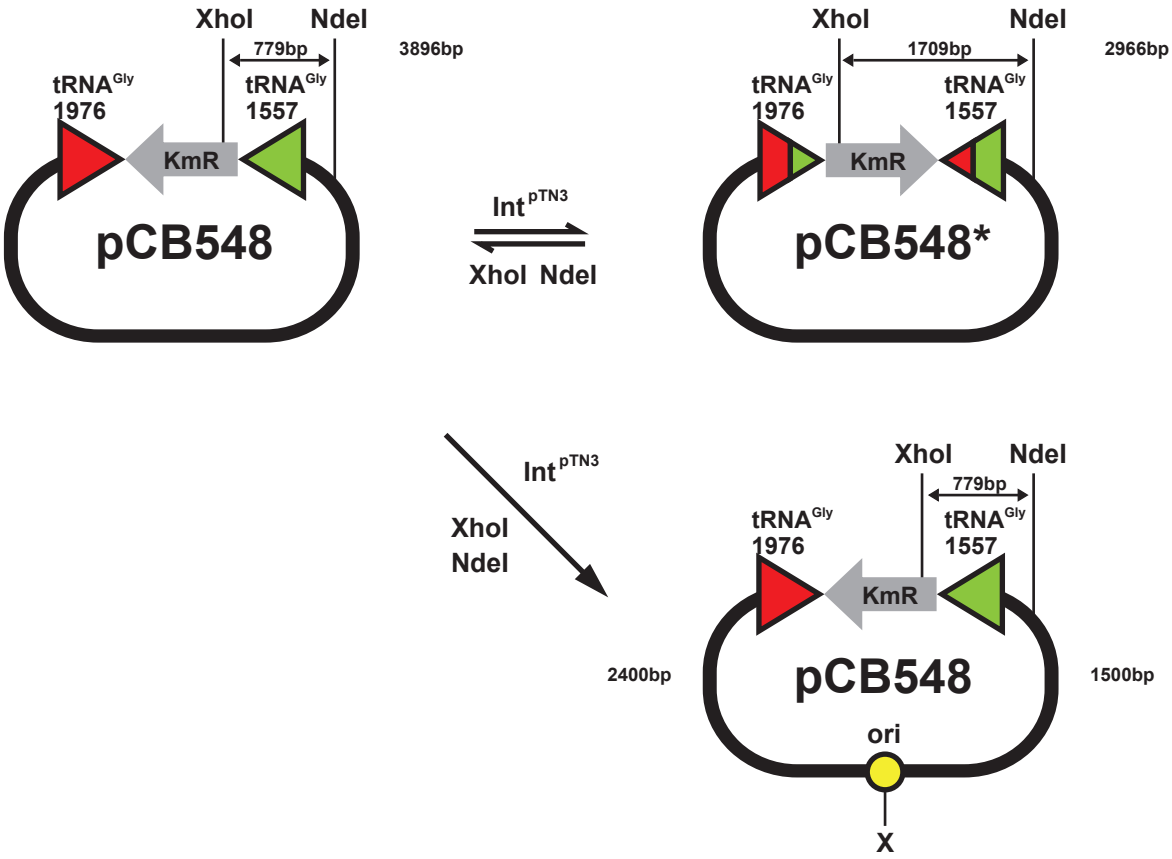
```
          >LEU 2-44
          CGGGGGTTGCCGAGCCTGGTCAAAGGCGGTGGACTCAAGATCC
>1976     *** *** * * ********* * * *   ** ***   * **
AACGGGCGTGCGGTGGTAGTCTAGCCTGGTCCAGGACACCGGCCTCCCAAGCCGGTGACCCGGGTTCAAATCCCGGCCACCGCACCA
CACAAACTTCGCCTGTGCGAAGTTTGACCAAGGCTCGTAGCTCCTGTTGGAGGGCTAAATTTCGAGTCATTTCTTATCAACTGGCCC
TTTTTGAGTTGGAGAACCTATCGAATTGCTCTTTTACCGTGGGTTTACCTTTAAATCAACGCCCGAAGGGCGTCAAGGGAGAGTAAA
CCCCTTTGAGAGGCTTTCTGAGAGAGTTCTCCTTTTGAAAGTGCGATTTATGAGCAAAATCTGCTCTACTT
```

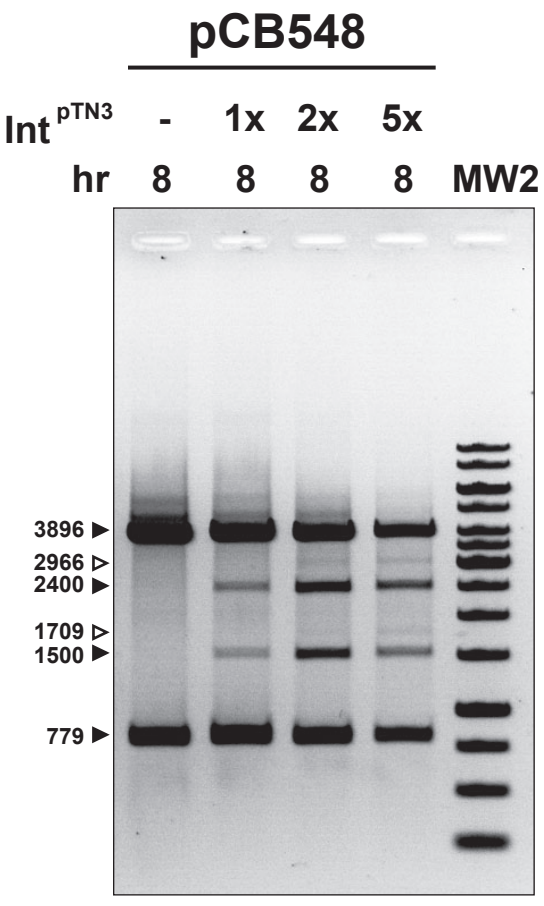
C

```
          >LEU 2-44
          CGGGGGTTGCCGAGCCTGGTCAAAGGCGGTGGACTCAAGATCC
>1557     *** *** * * ********* * * *   ** **     * **
ACTTTCTGTGCGGTGGTAGTCTAGCCTGGTCTAGGACACCGGCCTTCCAAGCCGGTGACCCGGGTTCAAATCCCGGCCACCGCACCA
CACAAACTTCGCCTGTGCGAAGTTTGACCAAGGCTCGTAGCTCCTTTTGGAGGGCTAAATTTCGAGTCATTTCTTATCAACTGGCCC
TTTTTGAGTTGGAGAACCTATCGAATTGCTCTTTTACCGTGGGTTTACTTTTAAATCGACGCCCCTCGGGCGTCAATGGATGTGAAT
AAAATCTGGCTCCATTCGAGCCTTGGGCAAGGGGCTACAAGCTTTTGGTGGAGCTTCACTCCTCACTTACT
```

D

```
>1976/1557 SEQUENCING
AACGGGCGTGCGGTGGTAGTCTAGCCTGGTCCAGGACACCGGCCTCCCAAGCCGGTGACCCGGGTTCAAATCCCGGCCACCGCACCA
CACAAACTTCGCCTGTGCGAAGTTTGACCAAGGCTCGTAGCTCCTTTTGGAGGGCTAAATTTCGAGTCATTTCTTATCAACTGGCCC
TTTTTGAGTTGGAGAACCTATCGAATTGCTCTTTTACCGTGGGTTTACTTTTAAATCGACGCCCCTCGGGCGTCAATGGATGTGAAT
AAAATCTGGCTCCATTCGAGCCTTGGGCAAGGGGCTACAAGCTTTTGGTGGAGCTTCACTCCTCACTTACTAGCCACGTTGTGTCTC
AAAATCTCTGATGTTACATTGCACAAGATAAAAATATATCATCATGAACAATAAAACTGTCTGCTTACATAA KmR →
```


XhoI
NdeI
779bp
3896bp
tRNA Gly 1976
tRNA Gly 1557
KmR
pCB548
Int pTN3
XhoI NdeI
1709bp
2966bp
pCB548*
Int pTN3
XhoI
NdeI
2400bp
1500bp
ori
X

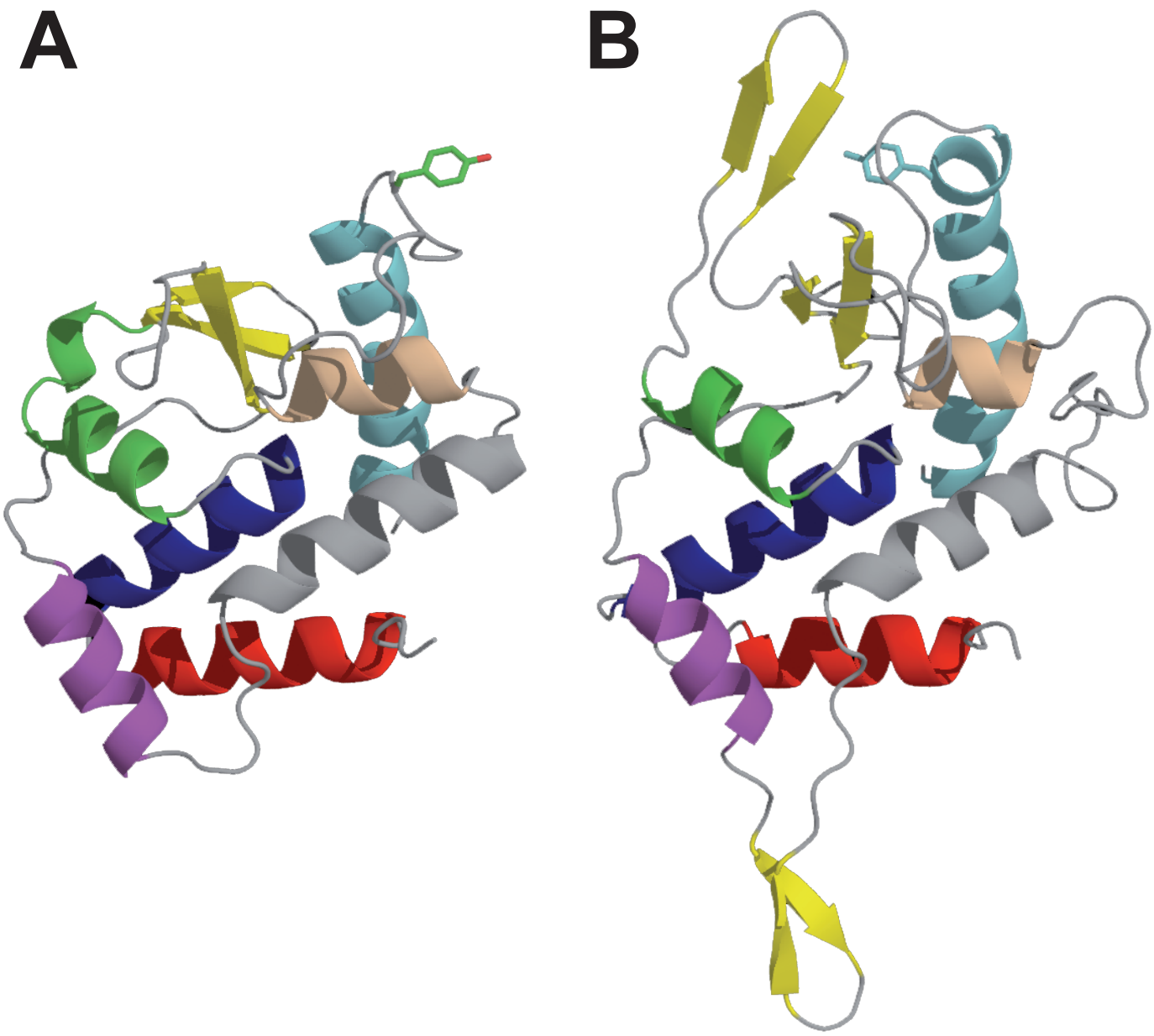

pCB548
Int pTN3 - 1x 2x 5x
hr 8 8 8 8 MW2
3896
2966
2400
1709
1500
779

>lac100
CACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACC
CAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAGCTGGCGTAATAG

>lac175
CACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACC
CAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAGCTGGCGTAATAG
CGAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGCGCAGCCTGAATG
GCGAATGGCGCCTGATGCGGTATTT

>lac250
CACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACC
CAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAGCTGGCGTAATAG
CGAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGCGCAGCCTGAATG
GCGAATGGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATT
TCACACCGCATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATA

S12 Fig.



Part 2. The plasmid family pT26-2 alternatively harbors two distinct integrases

Article 2. The global distribution and evolutionary history of the pT26-2 archaeal plasmid family

In order to uncover the evolutionary advantages of suicidal integrases, we decided to analyze the evolutionary history of a family of mobile genetic elements encoding one: the plasmid family pT26-2. The family is defined by a common core genome of 7 genes of unknown function and correspond to both free plasmids and integrated elements from Methanococcales and Thermococcales. The plasmid family evolved independently in Thermococcales and Methanococcales in the absence of inter-order transfer, despite the common oceanic hydrothermal environment. In both orders, more than 50% of the isolates harbor either a pT26-2-like element or a CRISPR spacer corresponding to a known member of the family, meaning that they likely already encountered a member of the family at least once. The pT26-2 family is therefore prevalent both for Methanococcales and Thermococcales.

The members of the pT26-2 family present an integration module, albeit distinct for Methanococcales and Thermococcales. The Methanococcales integration module is composed of a classical integrase (type II *sensu* She (She et al., 2004)) and a specific recombination site. The site corresponds to the 5' half of genes coding for $tRNA^{Ser}$ or $tRNA^{Leu}$. The two tRNAs harbor a supplementary loop. The Thermococcales integration module is composed of a suicidal integrase, either intact or fragmented, whose sequence includes the specific recombination site. This site corresponds to the 5' half of genes coding for a wide variety of tRNAs without any supplementary loop. Integration modules differ in both the integrase type and the integration target between Methanococcales and Thermococcales. They represent two phylogenetically distinct integrase families.

The analysis of these two distinct integrase families brought insights into their evolutionary patterns. Notably, both families present integration sites in tRNA genes. However, they are restricted to a tRNA subgroup for each family: tRNA with a supplementary loop for Methanococcales and tRNA without any supplementary loop for the Thermococcales. It appears that not all specificity switches are possible.

Plasmids from the pT26-2 family implemented two different integration strategies depending on their host: a classical strategy in Methanococcales and a suicidal strategy for Thermococcales. The two strategies could have been acquired independently after the divergence of the two subfamilies or one strategy could have been acquired in replacement of the other. In any case, the same plasmidic core backbone can sustain the presence of a classical or a suicidal integrase. Therefore, suicidal integrases do not present an advantage specific to certain mobile element functions. Additionally, in this case, the two integrases types are present in the same oceanic hydrothermal environment. Suicidal integrases are not exclusive to an environment. This concomitant presence of the two integrase types for the same plasmidic backbone and in the same environment further highlights the enigmatic existence of the suicidal integrases.

Article

The global distribution and evolutionary history of the pT26-2 archaeal plasmid family

Badel C. [1], Erauso G. [2,3], Gomez A.[4], Catchpole R. [1], Gonnet M. [2], Oberto J. [1], Forterre P[1,4]*., Da Cunha V[1,4]*.

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

[2] Université de Bretagne Occidentale (UBO, UEB), Institut Universitaire Européen de la Mer (IUEM) – UMR 6197, Laboratoire de Microbiologie des Environnements Extrêmes (LM2E), Place Nicolas Copernic, F-29280 Plouzané, France

[3] Aix-Marseille Université, CNRS/INSU, Université de Toulon, IRD, Mediterranean Institute of Oceanography (MIO) UM 110, Marseille, France

[4] Institut Pasteur, Unité de Biologie Moléculaire du Gène chez les Extrêmophiles (BMGE), Département de Microbiologie, Paris, France

* corresponding authors

patrick.forterre@pasteur.fr

Violette.DACUNHA@i2bc.paris-saclay.fr

Abstract

Although plasmids play an important role in biological evolution, the number of plasmid families well characterized in terms of geographical distribution and evolution remains limited, especially in Archaea. Here, we describe the distribution, biogeography and host-plasmid co-evolution patterns of 26 integrated and 3 extrachromosomal plasmids related to the pT26-2 plasmid from the hyperthermophilic archaeon *Thermococcus* sp.26-2. The pT26-2 family plasmids are widespread in Thermococcales and Methanococcales isolated from around the globe but are restricted to these two orders. All members of the family share 7 core genes but employ different integration and replication strategies. Phylogenetic analysis of the core genes and CRISPR spacer distribution suggest that plasmids of the pT26-2 family evolved with their hosts independently in Thermococcales and Methanococcales. Remarkably, core genes are conserved even in integrated plasmids that have lost replication genes and/or replication origins suggesting that they may be beneficial for their hosts. We hypothesise that the core proteins encode for a novel type of DNA transfer mechanism, explaining the widespread oceanic distribution of pT26-2-related plasmids.

Introduction

Mobile genetic elements (MGEs) are a crucial component of the living world, being the major vehicles for horizontal gene transfer (HGT) (Koonin and Wolf 2008), agents of genomic recombination (Cossu et al. 2017) and cradles of novel gene (Keller et al. 2009; Forterre and Gaïa 2016; Legendre et al. 2018). Whereas Archaea are much more closely related to Eukaryotes than to Bacteria in terms of fundamental molecular mechanisms (replication, transcription and translation), the set of MGEs (mobilome) infecting Archaea and Bacteria are strikingly similar and very different from those present in Eukaryotes (Forterre 2013; Forterre et al. 2014). It is unclear if the observed resemblance between the archaeal and bacterial mobilomes is a result of convergence due to the comparable chromosome structure and organization of archaeal and bacterial cells, or if it reflects widespread distribution of these MGEs by HGT between these two domains, or perhaps inheritance of a similar type of MGE present in the Last Universal Common Ancestor (LUCA). Further studies on the archaeal mobilome should possibly help to resolve this conundrum.

Most of the research on the archaeal mobilome has focused on a narrow range of model organisms, including Sulfolobales, Haloarchaeales, Thermococcales, and a few methanogens. Among them, plasmids and viruses from the order Thermococcales (comprised of the genera

Thermococcus, *Pyrococcus* and *Palaeococcus*) represent among the most hyperthermophilic MGEs known to date and have been studied in several laboratories (Forterre et al. 2014; Lossouarn et al. 2015; Wang et al. 2015). Extensive screening of extrachromosomal MGEs showed that between 30% and 40% of Thermococcales strains carry at least one MGE (Prieur et al. 2004). Two viruses, PAV1 from *Pyrococcus abyssi* (Geslin et al. 2007) and TPV1 from *Thermococcus prieurii* (Gorlas et al. 2012), and 19 plasmids have been isolated and characterized (Forterre et al. 2014; Lossouarn et al. 2015; Wang et al. 2015). Additionally, other MGEs have been detected in the course of genome sequencing projects, either as extrachromosomal or integrated plasmids (Fukui et al. 2005; Zivanovic et al. 2009; Vannier et al. 2011).

All plasmids from Thermococcales can be grouped into seven families based on their replication proteins (table 1). They will be named hereafter according to their prototype plasmids pGT5, pTN2, pT26-2, pTBMP1, pAMT11, pTP2, and pTN3 (Erauso et al. 1996; Geslin et al. 2007; Soler et al. 2010; Gonnet et al. 2011; Soler et al. 2011; Vannier et al. 2011; Gorlas et al. 2013; Forterre et al. 2014; Gaudin et al. 2014; Gill et al. 2014; Lossouarn et al. 2015; Kazlauskas et al. 2018). The two families with pGT5 and pTP2 as prototype correspond to small rolling-circle plasmids (Erauso et al. 1996; Gorlas et al. 2013). These two plasmid families have been used to construct *E. coli-Thermococcus* shuttle vectors for manipulation of Thermococcales (Lucas et al. 2002; Santangelo et al. 2008; Catchpole et al. 2018). The five other plasmid families probably replicate via a theta mode, although none of their replication mechanisms have been characterized biochemically (Forterre et al. 2014).

The pT26-2 family is particularly interesting, as these plasmids encode many transmembrane proteins and an AAA+ ATPase that could be involved in the formation of protein complexes involved in DNA transfer (Soler et al. 2010). The first member of the pT26-2 family, was isolated from *Thermococcus* sp. 26-2 collected in the East Pacific ocean (Lepage et al. 2004), and representatives have since been identified in other isolates. The pT26-2 plasmid family is composed of mid-sized MGEs ranging from 17 to 38 kb, that are found either as extrachromosomal or as integrated elements in the chromosome of the host. Comparative analysis of pT26-2-related plasmids indicated that the sequence can be divided in two parts: a highly conserved region which includes seven genes that are present in all related pT26-2 elements and are considered as “core genes”, and a variable region that includes both ORFans and genes putatively horizontaly aquired (Soler et al. 2010). The structure of one of the two largest core proteins was determined (t26-6p) (Keller et al. 2009) and this protein contains

several domains exhibiting novel folds not found in cellular proteins, supporting the idea that plasmids and/or viruses could be reservoirs of novel protein folds (Keller et al. 2009; Soler et al. 2010).

Interestingly, several MGEs related to pT26-2 were identified in organisms of the Methanococcales order (Soler et al. 2010; Soler et al. 2011), including plasmid pMEFER01 (22.2 kb) in a hyperthermophile *Methanocaldococcus fervens* (Soler et al. 2011) and several integrated MGEs in mesophilic *Methanococcus* species. Plasmid sharing between Thermococcales and Methanococcales was also reported for other families, e.g., pEXT9a-like plasmids (Krupovic et al. 2013). It is unclear if closely related plasmids are present in these two orders because they share the same biotope or because of their phylogenetic proximity. The Thermococcales are strictly anaerobic hyperthermophiles (optimal growth temperature >80°C) that are ubiquitous in hydrothermal vent systems (Zillig et al. 1983; Fiala and Stetter 1986; Takai et al. 2001). The majority of Thermococcales were isolated from marine geothermal environments, both shallow and deep hydrothermal vents, and a few strains were also isolated from continental oil reservoirs (at high temperature and salinity) (Ravin et al. 2009) and from fresh water terrestrial hot springs (Antranikian et al. 2017). The Methanococcales are also strictly anaerobic, but, in contrast to Thermococcales, they are not restricted to high temperature environments; *Methanocaldococcaceae* (*Methanocaldococcus* and *Methanotorris*) are hyperthermophiles and *Methanococcaceae* (*Methanococcus* and *Methanothermococcus*) are either hyperthermophiles or mesophiles (supplementary fig. S1). All members of these two families were isolated from aquatic environments and are capable of forming methane by reduction of CO_2 with H_2 (Albers and Siebers 2014).

Thermococcales and Methanococcales are close relatives phylogenetically, both belonging to the group I Euryarchaeota *sensu* Raymann *et al.* (Brochier-Armanet et al. 2011; Raymann et al. 2015), and even form sister groups in some analyses (Makarova et al. 2015). However, most analyses support the super-class Methanomada, which groups Methanococcales with other group I methanogens, i.e. Methanobacteriales and Methanopyrales (supplementary fig. S1) (Adam et al. 2017; Da Cunha et al. 2017). In recent years, several genome sequencing projects have dramatically increased the number of available archaeal genome sequences, including numerous genomes of Thermococcales, Methanococcales and other Euryarchaeal Group I and II species (Raymann et al. 2015). In particular, metagenomic analyses led to the identification of two new candidate archaeal orders, Methanofastidiosa and Theionarchaea, that branch as sister groups to Thermococcales in archaeal phyogenies, forming the super-class

Acherontia (Adam et al. 2017). These new orders are thus good candidates to detect new MGE related to those of Thermococcales.

The core proteins encoded by the first described pT26-2 plasmid strikingly had homologues only in related elements, raising challenging questions concerning the origin of this family. Considering the dramatic increase in the number of archaeal, bacterial and eukaryotic genomes available, we wished to update the search for pT26-2-related plasmids in the hope of expanding the number of known elements, which could shed light on their origins, functions and mechanism of transfer.

Here, we report the identification and analysis of 17 new members of the pT26-2 family. Surprisingly, even though pT26-2-related plasmids are widespread (both taxonomically and geographically) in Thermococcales and Methanococcales isolates, they seem to remain strictly limited to these two orders. We observed that genes of all episomal and integrated plasmids of the pT26-2 family are organized into several modules that can be exchanged by recombination with modules from other plasmids or viruses. We confirmed the existence of the 7 core genes that define this family and show that these core genes are conserved in MGE that have apparently lost the plasmid DNA replication protein and/or the origin of replication, suggesting that they may confer some selective advantage to their hosts. Our different phylogenetic analyses suggest that recent MGE transfers have occured between different Thermococcales, but not between the Thermococcales and Methanococcales. Moreover, these MGEs exhibit different integration strategies mediated by non-orthologous types of integrases. Although the origin of this family remains mysterious, the modular structure and remarkably broad distribution of the pT26-2-related plasmids across two archaeal orders provides a unique opportunity to study plasmid evolution.

Results

Identification of new integrated plasmids of the pT26-2 family and isolation of the first pT26-2-like plasmid from a *Pyrococcus* species

In order to expand the known representatives of the pT26-2 family, each of the seven previously identified core genes (Soler et al., 2010) were used as a query for homology searches in complete or partial archaeal, bacterial or eukaryotic genomes. All targeted regions, even if they encode few putative core genes, were carefully analysed. We determined their extremities by the identification of direct repeat sequences (att sites) resulting from the recombination reaction. We thus identified 17 new pT26-2-related plasmids integrated in Thermococcales and Methanococcales genomes, more than doubling the number of known elements (table 2). In

addition, we isolated and sequenced a new plasmid of this family, pGE2 (23,702 bp), from the strain GE2. This strain was isolated from deep-sea vent samples collected in 1989 in the North-Fidji basin (S-W Pacific) (Marteinsson et al. 1995). Preliminary characterization of the GE2 strain indicated that it belongs to the *P. abyssi* species (Erauso et al. 1993). In primary cultures of strain GE2, pGE2 co-existed both as an integrated copy and an episome (Gonnet, 2008 PhD thesis). The latter form was present in up to ~40 copies per chromosome (as estimated by qPCR) but it disappeared during subculturing. The pGE2 integration site in a tRNA gene was determined using an inverse PCR and sequencing strategy. Recently, the whole genome of *P. abyssi* strain GE2 was sequenced in the framework of a large project on comparative genomics of Thermococcales. The episomal form could not be retrieved from the assembly and by the read mapping analysis, and surprisingly, the integrated copy was found to be slightly smaller (21,837 bp) than the sequence of the episomal form established several years ago. The differences correspond to the presence of an insertion sequence (IS) of 1825 bp containing two genes, encoding a resolvase (cds 1; 163 amino acids) and a transposase (cds 2; 429 amino acids), respectively. This IS element belongs to the IS family IS*200*/IS*605* (Chandler and Mahillon 2002) and is identical to the element found in *P. abyssi* GE5^{T} (PAB2076/PAB2077) and was also detected in the *P. abyssi* GE2 genome, suggesting that the copy found in the episomal form of pGE2, originated from the host chromosome.

The distribution of the pT26-2 plasmid family is restricted to two euryarcheal orders: Thermococcales and Methanococcales.

Surprisingly, we could not find members of the pT26-2 family in any genome outside of the Thermococcales and Methanococcales orders, despite the recent discovery of several new archaeal lineages closely related to these species (Nobu et al. 2016; Lazar et al. 2017). The 29 members of the pT26-2 family are widespread within the two orders with 30% (12/39) of available Thermococcales genomes and 45% (10/22) of available Methanococcales genomes containing at least one pT26-2-related plasmid (fig. 1a.). Notably, *Methanococcus maripaludis* C7, *Thermococcus kodakarensis* and *Thermococcus barophilus* each contain two integrated pT26-2-related plasmids. In the cases of *T. barophilus* CH5 and *M. maripaludis* C7 genomes, the att sequences of one of the integrated plasmid are extremely mutated and can presumably no longer be used for the reverse reaction of excision. The mutated att sites are still similar enough to be detected for the TbaCH5_IP1 integrated plasmid in the *T. barophilus* genome but not for the MMC7V1 element in the *M. maripaludis* C7 genome. A similar situation is observed

for the plasmid integrated in the *M. maripaludis* X1 genome for which we could not detect the att sites.

In *T. kodakarensi*s, the att sites of the two pT26-2-like elements, TKV2 and TKV3, are unexpectedly mixed. TKV2 attL2 is not identical to TKV2 attR2 but to TKV3 attL3. Similarly, TKV2 attR2 is identical to TKV3 attR3 (supplementary fig. S2). This can be explained by an inversion between the two integrated plasmids (supplementary fig. S2). We analysed the illumina read mapping coverage over the TKV2 and TKV3 elements and at the limits of the integrated plasmids compared to that observed for the rest of the *T. kodakarensis* genome. We observed no read mapping defect at the limits of the integrated elements, confirming that this inversion does not result from an assembly problem. Interestingly, a back-inversion between TKV2 and TKV3 was previously detected experimentally in a subpopulation (<10%) of *T. kodakarensis* TS559 cells (Gehring et al. 2017) which restored their excision potentiality. This back-inversion was asymmetrical and led to the gain or loss of 2 kb (4 ORFs) in TKV3 and TKV2, respectively. The reassembled elements of this subpopulation consequently have the potential to be mobile.

In order to identify strains which likely encountered pT26-2-related plasmids in the past but no longer encode these elements, we searched for CRISPR spacers against sequences of all pT26-2-related plasmid in the CRISPRdb database (Grissa et al. 2007). We only found pT26-2-related spacers in the genomes of Thermococcales and Methanococcales, confirming that these elements have a restricted host range (fig. 1a). Notably, we did not detect any CRISPR spacers against pT26-2-related plasmids in the MAGs of *Methanofastidiosa* and *Theionarchaea*, which are sister groups of Thermococcales (Adam et al. 2017). Among the available genomes, 30% of the Thermococcales and 15% of the Methanococcales genomes contain a CRISPR spacer against pT26-2-related plasmids. When combined with the data on plasmid distribution, 56% of Thermococcales and 55% of Methanococcales have either a resident pT26-2-related plasmid or a CRISPR spacer against them (fig. 1a), indicating that more than half of Thermococcales and Methanococcales strains have encountered a pT26-2-related plasmid at least once during their evolution. In addition, we observed that 5 of the 12 Thermococcales isolates containing a pT26-2-related plasmid also contain a spacer against a different pT26-2-related plasmid. For example, in the genomes of *T. kodakarensis* and *T. guaymasensis,* the CRISPR loci contain spacers against the pT26-2-related plasmids TliDSM11113_IP1 and TceDSM17994_IP1, respectively, which are found in integrated form in the genomes of *T. litoralis* and *T. celericresence,* respectively. Intriguingly, we were unable to find CRISPR spacers against Methanococcales-encoded pT26-2-related plasmids in

Thermococcales genomes, and *vice versa*. Such observations seem to indicate that individual plasmids are not able to colonize hosts from both taxonomic orders, but rather show order-specific infectivity. We therefore surmise that pT26-2 plasmid transfer between Thermococcales and Methanococcales is no longer possible, in contrast to the recent transfer from Thermococcales to Methanococcales suggested for a pEXT-9 like plasmid (Krupovic et al. 2013). Together these results confirm that the pT26-2 family is widespread and mobile within the Thermococcales and the Methanococcales, but remains limited to these two orders.

pT26-2 MGEs are globally widespread

It seems possible that pT26-2-related plasmids are over-represented in Thermococcales and Methanococcales isolates due to the isolation of these strains from a limited number of geographical areas. In order to estimate a potential sampling bias, we performed a biogeographic analysis by comparing the isolation sites of Methanococcales and Thermococcales genomes available from NCBI (fig. 1b). Except for a handful of *Thermococcus* species, all strains were isolated from various marine environments, particularly deep-sea hydrothermal vents located along oceanic ridges (in the Atlantic, Pacific and Indian oceans) or from volcanic back-arcs in the Mediterranean Sea. The available genomes originate mostly from strains isolated in the northern hemisphere. In addition, the six *Methanococcus maripaludis* strains (out of seven) whose genomes contains a pT26-2-related plasmid were isolated from neighbouring sites in the Gulf of Mexico (light green region in the fig. 1b.) indicating that the infectivity of these plasmids in Methanococcales could be overestimated. In several cases, *Thermococcus, Pyrococcus, Paleococcus, Methanocaldococcus*, and *Methanotorris* strains were isolated from the same deep-sea hydrothermal sites, such as the East Pacific Ocean ridge (fig. 1b), confirming that these two orders can share the same habitat.

To analyse plasmid distribution, we mapped information about the presence of pT26-2-related plasmids and the presence of CRISPR spacers against pT26-2-related plasmids on to the biogeographic analysis (fig. 1b). This revealed that pT26-2-related plasmids are integrated in the genomes of hosts isolated from all different sampling regions. It is thus clear that pT26-2-related plasmids are abundant and widespread in Thermococcales and Methanococcales all over the world.

No HGT observed between Thermococcales and Methanococcales

It was previously suggested that the sequences of pT26-2-related plasmids can be divided into two regions: a highly conserved region of twelve genes which includes the seven

core genes present in all pT26-2-related plasmids known at that time (t26-5p, 6p, 7p, 11p, 13p, 14p and 15p); and a variable region that includes singleton ORFans and genes of various origins (Soler et al., 2010). The seven core genes are contiguous in all integrated and episomal elements, and belong to a highly conserved region of twelve genes with conserved synteny (from *t26-4p* to *t26-15p*) (fig. 2, fig. 3). The 12 proteins of the highly conserved region do not share significant similarity with any other proteins in public databases outside of the pT26-2 family (as detectable by HMMER or BLASTP). Nine of these proteins contain at least one (and up to five) putative transmembrane domains. In some of these transmembrane proteins, additionnal domains were also detected, e.g. a SH3 domain in the protein t26-13p, a carboxypeptidase regulatory domain in the protein t26-5p and a carbohydrate binding domain in the protein t26-10p. Notably, carbohydrate binding domains are present in some viral capsid proteins, allowing the recognition of host cellular surfaces (Krupovic and Koonin 2017). Interestingly, Phyre-2 analyses point to strong structural similarities between the t26-14p-like core protein and several ATPases from the AAA+ superfamily: the HerA-like hexameric DNA translocase VirB4 of type IV secretion systems of conjugative plasmids, or the genome packaging ATPase B204 from the *Sulfolobus* Turreted Icosahedral Virus 2 (STIV2). Such structural conservation suggests that these proteins could be involved in the translocation of DNA through membranes.

To determine if the 7 previously identified “core genes” are present in the 29 members of the family, we analysed the conservation between all their proteins by the Reciprocal Best Hit (RBH) approach. In addition, the number of core genes was also determined using SiLiX (Miele et al. 2011), a program developed to cluster homologous proteins into families. Both approaches confirmed that 7 protein families are conserved among pT26-2-related plasmids, and correspond to the core genes. In addition, the comparative analysis of pT26-2-related plasmids by RBH analysis (fig. 4, supplementary fig. S3) reveals the presence of two distinct subgroups, one containing the related plasmids identified in Thermococcales, and the other, those in Methanococcales. This observation again suggests that pT26-2 and related elements have not been recently transferred between the two archaeal orders, and have co–evolved with theirs hosts.

To further test this hypothesis, we compared the individual and the concatenated phylogenetic trees obtained with the 7 core proteins. The single and the concatenated phylogenetic trees obtained were all congruent, with Thermococcales and the Methanococcales forming two separated monophyletic groups with internal phylogenies rather similar to the host phylogenies. In detail, we observed a clear co-evolution of the pT26-2-related plasmids within

the Methanococcales (fig. 5), but we also observed putative horizontal transfers between different genera of Thermococcales that are also evidenced in the network analysis (fig. 4). In the phylogenetic tree based on the concatenation of the 7 core proteins, the two monophyletic groups of pT26-2-related plasmids infecting either Thermococcales or Methanococcales were separated by a long branch (fig. 5), clearly indicating the absence of recent HGT between the two groups.

Functional modules encoded by non-core genes

The SiLiX analysis of all proteins encoded by pT26-2-related plasmids led to the classification of the 902 putative proteins into 356 families (supplementary file 1). A vast majority of pT26-2-encoded proteins (696 proteins) belong to the variable regions (supplementary file 1). These regions are in fact hypervariable, since among the 356 identified proteins families, 309 correspond to proteins present in less than three pT26-2-related plasmids, and 244 correspond to singletons. These results are in agreement with the usual observation than MGEs are rich in genes with no homologous sequence in the databases. Most proteins present in variable regions have small size, this is especially the case for singletons, suggesting that these proteins are either false ORFans, putative protogenes, or new genes that recently originated *de novo* from proto-genes (supplementary fig. S4). The evolutionary scenario whereby novel protein-coding genes could randomly emerge from intergenic regions (Keese and Gibbs 2006), then evolve to resemble protein-coding genes was first observed in several eukaryotic genomes (Levine et al. 2006; Heinen and Staubach 2009; Toll-riera et al. 2009; Donoghue et al. 2011; Carvunis et al. 2012). This hypothesis was proposed to be extend to giant viruses (Forterre and Gaïa 2016). Recently this scenario was supported by the comparative genomic analysis of the pan-genome of the different pandoraviruses (Legendre et al. 2018), where novel protein-coding genes correspond to clade-specific and strain-specific genes (Legendre et al. 2018). Similarly, in our comparative anlysis, small proteins present in two to four closely related pT26-2-related plasmids are good candidates to be new genes (supplementary fig. S4). The few large genes present in the variable regions encode integrases (FamAll_0015, 24 and 70 in supplementary file 1) and putative replication proteins (FamAll_0034 and 103 in supplementary file 1). The diversity of modules involved in these two mechanisms will be discussed in more detail thereafter.

Identification of DNA replication modules

Putative DNA replication proteins

We identified three different types of genes encoding putative replication (Rep) proteins among pT26-2-related plasmids. The first group is only present in four MGEs from Thermococcales and corresponds to the putative replication protein t26-22p previously detected in pT26-2 (Soler et al. 2010). The four proteins are composed of a large central P-loop NTPase domain framed by two short N- and C-terminal domains that have no detectable sequence similarities with other proteins in databases. A homologous central P-loop NTPase domain is present in primase/helicase Rep proteins encoded by pTIK4 and pORA1 plasmids of *Sulfolobus neozealandicus* (Greve et al. 2005). For these two *Sulfolobus* Rep proteins, the central domain is associated in the N-terminal with a PrimPol domain that exhibits primase and polymerase activity (Lipps et al. 2004) (supplementary fig. S5). Although we could not detect the classical signature of PrimPol primases in the four Thermococcales proteins, the presence of the central P-loop NTPase domain suggests that t26-22p is a novel type of Rep proteins with helicase activities fused to an additional domain of unknown function, possibly corresponding to a novel type of primase. The presence of the Rep protein t26-22p on only four Thermococcales pT26-2-related plasmids (fig. 2, table 3) makes it an unlikely ancestral replication module.

The most represented Rep protein identified in pT26-2-related plasmids corresponds to a minichromosome maintenance (MCM) replicative 5' to 3' helicase which is found in both Thermococcales and Methanococcales elements (fig. 2, fig. 3, table 3). These proteins are also encoded by MGEs from other families in Thermococcales, such as the plasmid pTN3 (Gaudin et al., 2014), the virus TPV1 (Gorlas et al., 2013), the virus-like TKV1 (Fukui et al. 2005) and by MGEs from other archaeal lineages (Krupovič, Gribaldo, et al. 2010; Krupovic et al. 2019). Moreover, one or several genes encoding MCM proteins are present in all archaeal genomes and correspond to chromosomal replicative helicases (Raymann et al., 2014).

It has been previously shown that viral/plasmidic archaeal MCMs were recruited several times independently from their hosts during archaeal evolution (Krupovič, Gribaldo, et al. 2010). Confirming this result, MCM encoded by MGEs from Thermococcales and Methanococcales branch as sister groups of MCMs from their respective cellular hosts in a phylogenetic tree. In addition, as previously observed in Methanococcales (Krupovič, Gribaldo, et al. 2010), the MCM history is complex (supplementary fig. S6). In order to improve the phylogenetic signal, we performed two analyses focusing either on Thermococcales or Methanococcales. In Thermococcales, MCMs encoded by MGEs (including those of the pT26-2 family) form a monophyletic group. Notably, the two MCMs encoded by elements of the pT26-2 family (PspNA2_IP1 and TbaCH5_IP2) do not cluster together (supplementary fig.

370 S7). Additionally, the MCM encoded by the virus TPV1 branches between different groups of
371 plasmidic MCMs. These results indicates that transfer of the *mcm* gene between MGE and their
372 hosts took place early in Thermococcales evolution, before the separation between
373 *Thermococcus* and *Pyrococcus* genera. Furthermore, many transfers of *mcm* genes have
374 occurred between different MGEs including plasmids and viruses.
375 The Methanococcales MCM phylogenetic analysis showed a more complex
376 evolutionary history (supplementary fig. S8). Here, the chromosomal MCMs correspond to two
377 monophyletic groups, designated MCM1 and MCM2 (Walters and Chong 2010). The
378 phylogeny of each MCM group reflect the host evolutionary history. Our phylogeny thus
379 confirms the duplication of the MCM gene before the last Methanococcales ancestor (Krupovič,
380 Gribaldo, et al. 2010; Walters and Chong 2010). In contrast to Thermococcales, MCMs encoded
381 by Methanococcales MGEs are mixed phylogenetically with the chromosomal MCMs, and
382 cluster in three different groups (supplementary fig. S8). One of these groups (group I) branched
383 between Methanococcales and the Methanobacteriales outgroup, another (group II) branched
384 as sister group to cellular MCM1, whereas the third one (group III) branched within the MCM2
385 clade. Notably, MCMs encoded by pT26-2-related plasmids are again not monophyletic but
386 belong to either group 2 or 3 (supplementary fig. S8). This MCM phylogeny indicates that
387 exchange of *mcm* genes has occurred more frequently in Methanococcales than in
388 Thermococcales In particular, the MCMs encoded by the *Methanococcus maripaludis* pT26-2-
389 related plasmids is sister group to the chromosomal MCM2 encoded by all *Methanococcus*
390 *maripaludis* strains. This indicated that the *mcm* transfer likely occurred rather recently.
391 Interestingly, the MCM encoded by the *Methanocaldococcus* plasmid pMEFER01 belongs to
392 group II - a basal group, suggesting that this replication module predates the divergence of
393 cellular MCM2 protein and the last Methanococcales ancestor. The observed basal position of
394 some pT26-2-encoded MCM in both Thermococcales and Methanococcales led us to
395 hypothesis that the MCM could correspond to the ancestral replication protein of the pT26-2
396 family.
397 Finally, sequence similarity searches seeded with pT26-2-encoded proteins showed that
398 four pT26-2-related plasmids in Methanococcales encode a new family of distantly related
399 MCM-like proteins (fig. 3, table 3), previously identified in bacterial (Mir-Sanchis et al. 2016)
400 and thaumarchaeal (PMID: 30773816) MGEs. Notably, the bacterial MGEs also encode a
401 serine recombinase downstream of the MCM-like replication gene, responsible for the
402 integration activity. The proximity of the replication and integration modules was proposed to
403 facilitate replication after excision, enhancing transfer efficiency (Mir-Sanchis et al. 2016). A

similar gene layout is observed in the pT26-2-related plasmids, where the MCM-like gene is located next to a tyrosine recombinase gene. Such organisation is not observed in the pT26-2-related plasmids encoding the classical MCM replication protein. As the presence of this replication protein is restricted to four MGEs, it probably does not correspond to the ancestral replication protein, suggesting a more recent acquisition in Methanococcaceae.

Overall, we were able to identify a putative Rep proteins in 15 out of 29 pT26-2-related MGEs, including the three episomal plasmids (table 3). The absence of a putative Rep protein in a particular MGE can be due to the presence of a novel type of Rep protein or due to the fact that these MGEs have lost the ability to replicate autonomously. In agreement with this second hypothesis we noticed that a PCNA is encoded separately in two pT26-2 related plasmid and in combination with the MCM in one additional element, such gene combination also observed in some haloarchaeal viruses (Mizuno et al. 2019). The second hypothesis is also supported for several of these MGEs by the fact that they do not encode large proteins of unknown function in their variable regions, and/or have no detectable replication origins (see below). Taken together, these analyses reveal a complex evolutionary scenario for the replication module of the pT26-2-related plasmids, with several replacements that could correspond to new gene aquisitions. This high frequency of replication module replacement could partially compensate the observed tendancy of pT26-2-related plasmids to lose the replication protein following integration.

Origins of replication

A putative replication origin (ori) was predicted by cumulative GC skew analysis for the plasmid pT26-2 between the *t26-20p* and *t26-21p* genes (Soler et al. 2010). Replication origins are usually AT rich regions of low stability that contain multiple direct and inverted repeated sequences (Sun et al. 2006; Krupovic et al. 2013). To predict putative *ori* for each pT26-2-related plasmids, we used two complementary methods: (1) we repeated a GC-skew analysis as was performed for the original plasmid pT26-2 (Soler et al. 2010) and (2) looked for repeat-rich regions by dotplot analysis. Together these methods allowed us to identity a putative *ori* for 24 of the 29 pT26-2-related elements (supplementary table1, supplementary fig. S9). Around half of these *ori* regions were identified by both methods independently, although for two elements, the two methods gave two different *ori* locations (supplementary table1, supplementary fig. S9). The majority of the remaning putative *ori* were predicted by GC-skew analysis, and for 5 elements, we could not detect any putative *ori* with either method (supplementary table1, supplementary fig. S9).

Most predicted *ori* are located in intergenic regions or in regions containing multiple small open reading frames which are potentially non-translated (fig. 2, fig. 3). Comparative analysis of the identified putative *ori* does not reveal any conserved consensus sequence. Given the low conservation of the non-core region, it is difficult to say whether *ori* location is conserved between the different elements - even for the closely related elements in *Methanococcus maripaludis*, the putative ori location is variable (fig. 2, fig. 3).

Overall, both the location and sequence of putative *ori* in pT26-2-related plasmids seem extremely variable. However, we still observed a linkage between the putative *ori* and the replication protein; in most cases (12/15) the *ori* was located nearby the gene encoding one of the three types of putative Rep proteins.

Thermococcales and Methanococcales pT26-2-related plasmids present two different integration strategies

We found that all pT26-2-related plasmids encode an integrase belonging to the tyrosine recombinase superfamily, even if for some of them the integrase gene was not annotated. So far, several families of site-specific integrases encoded by viruses and plasmids have been described in Archaea (She et al. 2004; Erauso et al. 2006; Cossu et al. 2017; Wang et al. 2018). They are divided in two major types based on the strategy of integration (She et al. 2004): for type-I integrases, recombination of the circular element with the host chromosome leads to division of the integrase gene into two fragments, a longer Int(C) fragment and a shorter Int(N) fragment; in contrast, the type-II integrases maintain an intact integrase-encoding gene after recombination.

The previously described Thermococcales and Methanococcales pT26-2-related plasmids encoded integrase of both type-I and type-II (Soler et al., 2010). In order to determine if this observation was still valid for our extended dataset, we carried out a clustering analysis based on pairwise protein similarity. Beside integrases encoded by the pT26-2-related plasmids, the analysed dataset included known archaeal integrases from different families, and putative integrases that show sequence similarity to the pT26-2 integrase (Int^{pT26-2}). Thermococcales Int^{pT26-2} were reconstituted from partitioned chromosomal Int(C) and Int(N) fragments. Clustering analysis was performed using SiLiX with a minimum threshold of 25% identity over 40% of the protein (Miele et al. 2011). The result confirmed that the pT26-2-encoded integrases in Thermococcales and in Methanococcales correspond to two different types, as they are not connected to each other (supplementary fig. S10). Rather, the integrases of the pT26-2 MGEs infecting Thermococcales were connected to the SSV-integrase family (type-I) and the

Methanococcales pT26-2 integrases were connected to XerC proteins, and less stringently, to the SNJ2 and pNOB8 integrases (type-II).

We then performed a phylogenetic analysis using the pNOB8, XerC/D, SNJ2, SSV, pTN3 and pT26-2-related integrases. This phylogeny confirmed the results of the network analysis, with Thermococcales integrases forming a sister group to pTN3 integrases and the Methanococcales integrases forming a sister group to SNJ2 integrases (supplementary fig. S11). The phylogenetic analysis also highlights several cases of integrases exchanges between different type of MGEs (supplementary fig. S11). For example, the integrases of *Methanococcus maripaludis* MMC7V2 and other *Methanococcus maripaludis* pT26-2-related plasmids are closely related to the integrase of a *Methanococcus vannielli* provirus, suggesting an integrase exchange between the pT26-2-related plasmids and the virus.

For all site-specific integrases, the recombination sites (att sites) correspond to identical sequences on the MGE (attP) and the host chromosome (attB). Close examination of the att sites of the pT26-2 family MGEs and of their integrase-encoding genes confirmed the two different recombination strategies used by these plasmids. In MGEs of Thermococcales, the attP site is located inside the coding sequence of the integrase gene, as expected for type I integrases. The att sites in Methanococcales are often located close to the integrase gene as expected for type II integrases. However, in some case, the att site can be located further away from the *int* gene; for instance, in *Methanococcus maripaludis* elements, the *int* gene is positioned in the middle of the integrated element. In addition, we also observed in Methanococcales, the tandem integration of multiple MGE next to the same tRNA confirming the previously observed existence of integration hotspots (Krupovič, Forterre, et al. 2010).

For both type I and II integrases, the attB sites are usually located at the 5’ or 3’ regions of tRNA genes (Faraco et al. 1989; She et al. 2004). All attB sites of pT26-2 family MGEs overlap with the 3’ end of a tRNA gene (fig. 6, table 3), most often including the anti-codon sequence. As observed for other integrase families, the tRNA genes are not disrupted by the integration event (Schleper et al. 1992). In Thermococcales, attB sites are present in a wide variety of tRNA genes, including Arg, Thr, Gly, Val, Glu, Tyr, and Ala tRNA genes (supplementary fig. S12a, table 3). In contrast, all attB sites in Methanococcales correspond to Ser tRNA genes, with the single exception of a Leu tRNA gene for MVV1. In both Thermococcales and Methanococcales, the tRNA target and att sites are conserved between integrases of closely related MGEs (supplementary fig. S12b, table 3). One notable exception is the integrases of TKV2 and TKV3 whose specificity was mixed during the inversion event discussed above (supplementary fig. S1).

The alignment of the different att sites from Thermococcales and Methanococales pT26-2-related plasmids showed that an imperfect palindromic sequence is conserved among the different att sites (fig. 6a). This sequence corresponds to the two T-stems of the T-arm in the 3' region of the tRNA (fig. 6b-c). The att sites of Thermococcales MGEs average 58 nt in length, similar to that previously observed with SSV1 and pTN3 integrase (Schleper et al. 1992; Cossu et al. 2017). The att sites are longer in Methanococcales MGEs, accounting for the presence of a variable loop in the tRNA-Leu and tRNA-Ser recognized by their integrases (fig. 6). Our results indicate that Thermococcales pT26-2 integrases present a high diversity of att sites, with no clear preferential target tRNA. In contrast, the att sites of Methanococcales pT26-2 integrases are presently limited to tRNAs containing the additional variable loop (Ser-tRNA and Leu-tRNA).

Discussion

The family of archaeal plasmids epitomised by the element pT26-2 was first described following the isolation of *Thermococcus* sp. 26-2 (Soler et al. 2010). Here, we expand our knowledge of this plasmid family by identifying new integrated pT26-2-related plasmids in Thermococcales and Methanococcales genomes and describe the first episomal pT26-2 family member present in a *Pyrococcus* strain, namely pGE2 from *Pyrococcus* sp. GE2.

All MGEs of this family are formed by the association of a variable region that often includes the Rep protein and a putative replication origin and a conserved region, the "core module", rich in genes encoding several putative membrane proteins and an ATPase that could be involved in DNA transfer . The core module include 7 genes that are conserved in all elements and can be used to define the pT26-2 family. Phylogenetic analyses of these pT26-2 core proteins reveals a well supported bipartition of Thermococcales and Methanococcales, suggesting that these MGEs have evolved independently in each taxonomic group after the separation of the two lineages, and were never transferred between members of the two orders.

The hypothesis of independent plasmid evolution in Thermococcales and Methanococcales is further supported by several observations. 1) All CRISPR spacers directed against MGEs present in strains of one order (Thermococcales or Methanococcales) are specific for MGE detected in this order; 2) Phylogeny of the Rep proteins shared between pT26-2 MGEs of these two orders (MCM) also show a clear-cut separation between them, and non-MCM Rep proteins are specific either for Thermococcales (t26-22p-like protein) or Methanococcales (the distanly related MCM-like protein and PCNA); 3) Thermococcales and Methanococcales

MGEs are characterized by different types of integrases (type I and II, respectively) and different integration specificities.

Two observations suggest that ancestral pT26-2-related MGEs were already present in the last common ancestor of both Thermococcales and Methanococcales. Firstly, these MGEs and CRISPR spacers which target them are widespread in both orders, and secondly, our phylogenetic analysis indicates that the core proteins have co-evolved with their hosts without inter-order transfers. It seems reasonable to hypothesize that an ancestral pT26-2-related plasmid was already infecting the last common ancestor of Thermococcales and Methanococcales and diverged after their separation. If Methanococcales are not a sister group to the Thermococcales, but rather to other group I methanogens (Adams et al., 2017, Da Cunha et al., 2017), a parallel loss of pT26-2-like elements from Methanobacteriales and Methanopyrales would have to be evoked. This could be explained by the appearance of a unique cell wall consisting of pseudomurein in Methanobacteriales and Methanopyrales (Steenbakkers et al. 2006; Visweswaran et al. 2010). In an alternative hypothesis, ancestors of Thermococcales and of Methanococcales were infected independently by two ancestral elements sharing the same core genes.

A few years ago, a clear case of horizontal plasmid transfer between *Thermococcus* and *Methanocaldococcus* was described based on comparative genomics of the pMETVU01 and pAMT7 plasmids from *Methanocaldococcus vulcanius* M7 and *Thermococcus* sp. AMT7, respectively (Krupovic et al. 2013). In contrast, our work suggests that the highly conserved region of pT26-2-related plasmids does not mediate the HGT between Methanococcales and Thermococcales. Despite the abundance of pT26-2-like MGEs within these taxonomic groups, their mobility appears to be prohibited at an inter-order scale. We could identify few cases of HGT between different members of the Thermococcales, including HGT between different *Pyrococcus* and *Thermococcus* strains (different genera). Similar to the core module, the integration and replication modules of MGE of the pT26-2 family have also evolved within the order boundaries. Nevertheless, they exhibit a more complex evolutionary history with many apparent exchanges with MGEs from other families of the same order. It was originally reported that DNA exchange was preferentially observed within 'DNA vehicles' of the same type (chromosome, plasmid or virus) (Halary et al. 2009). However, in the last decade, data suggesting a strong evolutionary connection between plasmids and viruses have accumulated. Mobile genetic elements have a modular organisation, and each module can follow its own evolutionary history by recombinational exhange with other MGE and/or their host genome (Guérillot et al. 2013; Iranzo et al. 2016). Here our analysis shows that the integration and

replication modules of pT26-2-related plasmids have been exchanged with the host chromosome and with other integrated elements, some of which have been identified as viruses sharing the same host. As previously noticed for Methanococales and Sulfolobales (Krupovič and Bamford 2008; Redder et al. 2009; Krupovič, Gribaldo, et al. 2010), and more recently for Thaumarchaeota (Krupovic et al. 2019), we also observed multiple tandem integration events in Methanococcales. Perhaps these tandem integrated MGEs could be excised together leading to the production of new patchwork element with new module combinations, as it has been proposed for the evolution of Sulfolobales fuselloviruses (Redder et al. 2009).

Taken together our results lead us to propose an evolutionary model for the pT26-2 family (fig. 7), where the core module was already encoded by the ancestral pT26-2-related plasmid (ancestral-pT26-2) infecting the last common ancestor of Methanococococales and Thermococcales. This ancestral pT26-2 element probably contained a replication module, potentially an MCM helicase. We observed that during the evolution in the Methanococcales the pT26-2 element has replaced its replication protein with the host chromosomal MCM2, and with another kind of MCM-like replication protein from an unknown source. In some pT26-2-related plasmids of Thermococcales, this replication protein has been replaced by a t26-22p-like protein. For the integration module, we can formulate two evolutionary hypotheses: 1) An integrase could have been present in the ancestral pT26-2 and then replaced in the ancestral Methanococcales pT26-2 and/or the ancestral Thermococcales pT26-2; or 2) the integration module could have been absent in the ancestral-pT26-2 and then acquired twice independently in the ancestral Methanococcales pT26-2 and ancestral Thermococcales pT26-2.

Surprisingly, we did not detect any pT26-2-related plasmids in any other archaeal phylum or in Bacteria. In particular, they are not present in Metagenomic Assembled Genomes (MAGs) of Theinoarchaea and Methanofastidiosa, which are closely related to Thermococcales in most recent phylogenetic analyses (Adams et al., 2017). However, only a limited number of partial MAGs are presentlty available for these two new candidate orders (Nobu et al. 2016; Lazar et al. 2017) and their future exploration might reveal new MGEs related to those of Thermococcales. Moreover, as our knowledge on the diversity of archaea and their mobilome is rapidly increasing, one can expect that pT26-2-related plasmids will be identified in other archaeal lineages. In the meantime, the origin of the core module proteins remains a mystery.

We noticed a tendency of pT26-2-related plasmid to loose the replication ability upon integration in both host orders, though more strongly so in Thermococcales. The latter observation could be linked to the integration mechanism employed by Thermococcales which

appears to be suicidal. The excision of Themococcales pT26-2-related plasmids after integration seems more difficult since the integrase gene is split during the integration process. Interestingly, the core module is still strictly conserved in several integrated MGEs that have lost their Rep protein and/or their putative replication origin. This could reflect some selective advantage that favours the conservation of this core module. In 2013, genetic studies of mutants lacking each of the four MGEs integrated in the genome of *T. kodakarensis*, showed that deletion of TKV2 and TKV3 (the two pT26-2-related plasmids) negatively effects growth (Tagashira et al. 2013). This suggests that pT26-2-related plasmids stimulate cell growth, at least under laboratory conditions (Tagashira et al. 2013). Following integration and the inactivation of mobility, the core module could still be used as a gene transfer agent, as a new kind of secretion system, or as a kind of interaction system between different cells. Two species containing integrated pT26-2-related plasmids, *T. gammatolerans* and *P. horikoshii,* were shown to produce membrane vesicles containing cellular DNA (Soler et al. 2008). Moreover, the DNA within the vesicles of *T. gammatolerans* is remarkably resistant to DNase treatment and thermodenaturation (Soler et al. 2008). Other plasmids in Thermococcales species, such as the pTN3, have been shown to use membrane vesicles to transfer horizontally (Gaudin et al. 2014), thus the promotion of such transfer via a plasmid-encoded mechanism in these organisms seems a reasonable possibility.

The core module of pT26-2-related MGEs encodes several predicted transmembrane proteins that may be involved in the formation of a DNA transfer complex. It was first thought that pT26-2 MGEs could represent a new kind of viral genome (or derived from such elements) (Soler et al. 2010). However, no genes encoding putative capsid proteins could be detected in pT26-2 or related elements, and no virions could be observed in cultures of the species containing such elements (Soler et al. 2010). Interestingly, the genetic organization of pT26-2 is comparable to that of the infectious plasmid pR1SE recently isolated from vesicles of the halophilic archaeon *Halorubrum lacusprofundi* R1SE, albeit without any direct sequence conservation (Erdmann et al. 2017). Similar to the conserved region of pT26-2 family, the conserved region of plasmid pR1SE includes several genes in the same orientation that encode putative transmembrane proteins with 1 to 5 transmembrane domains (Erdmann et al. 2017). In addition, pR1SE encodes a protein similar to the bacterial pili-assembly ATPase. It was experimentally shown that pR1SE can promote its own transfer between cells via extracellular vesicles which contain plasmid encoded proteins (Erdmann et al. 2017). Such infectious plasmid vesicles fostered by plasmid-encoded proteins appear to mimick virions, and were

proposed to be called 'plasmidions' (Forterre et al. 2017). Similarly to pR1SE, we hypothetize that pT26-2-related elements could facilitate their own transfer via a plasmidion-type mechanism. Plasmidion protection could facilitate passive transport compared to naked plasmid DNA under the harsh environmental conditions that prevail in Thermococcales habitats. This protection may help to explain the global geographic distribution of pT26-2-related elements, similarly to marine viruses where capsid proteins protect DNA during transport along oceanic currents (Brum et al. 2015). In the future, we would like to revive integrated pT26-2-related plasmids and raises the question of the transfer mechanism of the pT26-2 family. In addition we would like to tackle the outstanding question of the function of the core membrane protein and of the ATPase in transfer mechanism.

Material and Methods

Isolation, sequencing and detection of integrated copies of pGE2 in GE2.

Plasmid pGE2 was isolated from a 50 ml culture of *P. abyssi* strain GE2 in late exponential growth phase, using a modified alkaline-lysis method as previously described (Erauso et al. 1996). A shotgun plasmid library of clones of pGE2 was constructed in pUC18 vector and sequenced from both ends as described previously (Gonnet et al. 2011). The complete plasmid sequence was deposited to the GenBank under the following accession numbers: XXX.

The detection of integrated copies was performed by Southern blot of total DNA, the detailed procedure was previously described (Gonnet et al. 2011). A complete pGE2 genome probe was generated by digesting about 500 ng of pGE2 plasmid DNA with HindIII plus EcoRI and the mix of the fragments obtained was labeled with alkaline phosphatase by the AlkPhos Direct System kit (GE Healthcare, UK). A second probe targeting the IS element identified in pGE2 sequence was generated by using a PCR product obtained with specific primers (supplementary table2) and labelled as reported above.

pGE2 plasmid copy number

The pGE2 copy number was estimated using a real time quantitative PCR-based method (Lee et al. 2006; Providenti et al. 2006). The detailed procedure and calculation were as previously described (Gonnet et al. 2011). Two set of primers were tested for pGE2 specific primers targeting respectively the CDS7 (UVRD Helicase) and the CDS29 (putative Rep) and one pair, Arc344F-Uni516R, targeting the 16S rRNA gene of GE2 (sequences of the primers are given in the supplementary table2). The average copy number of pGE2 was essentially the same using either CDS7 or CDS29 primer pair.

675
676 **pT26-2 family update**
677 In order to identify new members of the pT26-2 family (Soler et al. 2010), each of the seven
678 previously identified core genes (t26-5p, 6p, 7p, 11p, 13p, 14p and 15p) were used as query for
679 homology search in complete archaeal genomes, using SynTax (Oberto 2013) a web server
680 linking protein conservation and synteny. In addition, we also screened by BLASTP search the
681 NCBI non-redundant protein database to access to plasmids and non-complete genomes. All
682 DNA regions, even if they encode few putative core genes, were carefully analysed because
683 these regions could correspond to remnant MGE.
684
685 **Limits detection and integrase extraction**
686 In Thermococcales, integrase genes were identified by homology with the integrase of the
687 plasmid pT26-2 (Int^{pT26-2}). Genomes were searched by tblastn using as query the N-ter or C-ter
688 region of Int^{pT26-2} and the subsequently identified integrases. Attachment (att) sites were
689 identified as identical region on the border of integrase N-ter and C-ter coding regions. Up to
690 three mismatches were accepted in the middle of the att site to take into account sequence
691 degenerescence after integration. N-ter coding sequences were defined from a start codon (ATG
692 or GTG) to the last non att codon. C-ter coding sequences were defined from the first att codon
693 to a stop codon. Complete integrase genes were reconstructed by adjoining N-ter and C-ter
694 coding regions. The N-ter and C-ter region did not have matching open reading frames only for
695 TKV2 and TKV3.
696 In Methanococcales, no Int^{pT26-2} homologs were identified by tblastn. Att sites were identified
697 as identical sequences in proximity to the detected core genes. Up to three mismatches were
698 accepted in the middle of the att site to take into account sequence degenerescence after
699 integration. Annotated integrase genes were located in between the att sites. Additional
700 integrases were searched by tblastn with annotated ones as query.
701
702 **Read Mapping and control of the inversion in the *T. kodakarensis* genome.**
703 As the inversion affecting the TKV2, TKV3 orientation observed in the *T. kodakarensis* could
704 be the result of an assembly problem, we mapped reads obtain from our laboratory strain against
705 the NCBI available genome, using Bowtie 2 (Langmead and Salzberg 2012). After mapping,
706 we observed a 500X coverage along the integrated elements and at both limits so this results
707 confirmed that this inversion do not results from an assembly mistake. The same approaches
708 was made on the *Pyrococcus abyssi GE2* genome and on the *Themococcus 26-2* genome in

order to detect a higher proportion of read mapping in the region corresponding to integrated pT26-2-related plasmid, unfortunately no clear higher coverage could have been observed.

Host specificity determination

The host specificity was determined by the presence of a CRISPRspacer against each pT26-2 element in the archaeal genomes present in the CRISPRdb database (Grissa et al. 2007) (http://crispr.i2bc.paris-saclay.fr/). At the time of the analysis the database contained 232 complete archaeal genomes, including 27 and 15 genomes of Thermococcales and Methanococcales respectively at the time of the analysis. The nucleotide sequences of all identified pT26-2-related plasmids were compared by blastn approach against the spacer-sequences in the CRISPR database.

Orthologous protein identification

For each pT26-2-related plasmid the encoded proteins were extracted. In order to identify orthologous proteins, we used Reciprocal Best Hits (RBH) a common strategy used in comparative genomics. Basicaly, a RBH is found when the proteins encoded by two genes, each in a different MGE, find each other as the best scoring match. NCBI's BLAST is the software most usually used for the sequence comparisons necessary to finding RBHs. The protein sequence comparisons were performed using the NCBI's BLAST version 2.2.28+ , every BLAST score was normalized to the alignment of query and hit proteins to themselves. Proteins showing normalized bi-directional BLASTs > 30% were considered orthologous as recommended by Lerat et al. (Lerat et al. 2003). Then we tested the impact of the selection of each different pT26-2-related plasmid as a pivot MGE on the core protein number size (protein present in 80% of the tested pT26-2-related plasmid). The comparative analysis showed that the number of “core genes” is affected by the pivot selection, and varies from 7 to 9. If TbaCH5_IP1 is selected as a pivot, the number of core gene falls to 3 reflecting the remnant state of this integrated element (table 3).

Silix Network

All-against-all blastp analyses were performed on all encoded pT26-2 integrases with the addition of related integrases found in Thermococcales and Methanococcales and of an integrase and a part of the dataset recently used for the analysis of the SNJ2 integrase family (Wang et al. 2018). The all-against-all integrases BlastP results were grouped using the SiLiX (for *SIngle LInkage Clustering of Sequences*) package v1.2.8 (http://lbbe.univ-lyon1.fr/

SiLiX)(Miele et al. 2011). This approach for the clustering of homologous sequences, is based on single transitive links with alignment coverage constraints. Several different criteria can be used separately or in combination to infer homology separately (percentage of identity, alignment score or E-value, alignment coverage). For this integrase dataset, we used the additional thresholds of 25% and 60% for the identity percentage and the query coverage, respectively. The network was visualized using igraph package from R (https://igraph.org/). In order to find densely connected communities in a graph via random walks, we used the cluster_walktrap function of the igraph package.

Synteny conservation

Synteny conservation among pT26-2-related plasmids was preliminary analysed using SynTax a web server linking protein conservation and synteny in complete archaeal genome. Then the synteny conservation among integrated and non-integrated related pT26-2 element was confirmed using easyFig a Python application for the comparison of genomic loci based on side-by-side visualization of BLAST results (Sullivan et al. 2011).

Putative protein function

As the blastp comparision is not sufficient to succeed to infer putative fonctions to the Core proteins. All Core proteins sequences were analysed with Phyre2 (Kelley et al. 2015). Phyre2 is a suite of tools available on the web to predict and analyze protein structure, this tool compare the given sequences to a Hidden Markov Model HMM database of known structures.

Replication origin prediction and module analysis

The replication origin was determined by two complementary methods: (1) GC-skews were drawn where the replication origin correspond to peaks and (2) dotplots were implemented with Gepard where the replication origin correspond to repeats-rich area.

To determine the origin of the MCM proteins encoded by pT26-2-related plasmids, we performed phylogenetic analyses using the core MCM helicases predicted by Raymann *et al*. (Raymann et al. 2014) and additional MCM helicases encoded by various MGEs identified in Thermococcales and Methanococcales genomes (Krupovič, Gribaldo, et al. 2010), that belong to pT26-2-related plasmids or not. In order to focus on the Thermococcales and Methanococcales MCM histories, we made two separated phylogenetic analyses using the *Theioarchaea* and the *Methanofastidiosa* or the Methanobacteriales as an outgroup respectively (supplementary fig. S7, supplementary fig. S8).

Alignments and trimming and phylogenetic analysis

Each alignment used for phylogenetic analyses was performed using MAFFT v7 with default settings (Katoh and Standley 2013) and trimmed with BMGE (Criscuolo and Gribaldo 2010) with a BLOSUM30 matrix, and the -b 1 parameter.

For the Maximum Likelihood (ML) analysis IQ-TREE v1.6 (http://www.iqtree.org/) was used with the best model as suggested by the best model selection option (Wong et al. 2017). Branch robustness was estimated with the nonparametric bootstrap procedure (100 replicates), or with SH-like approximate likelihood ratio test (Guindon et al. 2010) and the ultrafast bootstrap approximation (1,000 replicates) (Chernomor et al. 2017).

Funding and Acknowledgments

This work is supported by an European Research Council (ERC) grant from the European Union's Seventh Framework Program (FP/2007-2013)/ Project EVOMOBIL-ERC Grant Agreement no. 340440. CB is supported by Ecole Normale Supérieure de Lyon. MG was supported by allocation de recherche doctorale de la région Bretagne. GE was supported by grants from the EU Project PYRED QLK3-CT-2001-01676. We are grateful to Mart Krupovic for his comments.

Tables

Table 1. List of Thermococcales plasmid families

Plasmid family (type plasmid)	Replication mode	Related MGE	Size	References
pTN2	θ	PAV1	8.5-13kb	(Geslin et al. 2007; Soler et al. 2010; Krupovic et al. 2013; Gill et al. 2014; Kazlauskas et al. 2018)
pTBMP1	θ	-	55.5kb	(Vannier et al. 2011)
pAMT11	θ	TKV1	18.3-20.5kb	Gonnet et al. 2011
pT26-2	θ	TKV2, TKV3	17-38kb	(Soler et al. 2010)
pTN3	θ	TKV4	13.8-20.2kb	(Gonnet et al. 2011; Gaudin et al. 2014; Cossu et al. 2017)
pGT5	RC	-	3.4kb	(Erauso et al. 1996)
pTP2	RC	-	2kb	(Gorlas et al. 2013)

θ theta mode, RC rolling-circle

798

799 Table 2. List of pT26-2-related plasmids.

Element	Host	Integration location	Att length	state	Access	Reference
pT26-2	*Themococcus sp* 26-2	1..21566	51	free/integrated	295126597	Soler 2010
TKV2	*Thermococcus kodakarensis* KOD1	320075..347187	48/53	integrated	AP006878.1	Keller 2009
TKV3	*Thermococcus kodakarensis* KOD1	499284..526865	48/53	integrated	AP006878.1	Fukui 2005
TguDSM11113_IP1	*Thermococcus guayamensis* DSM11113	153065..178766	46	integrated	CPU007140.1	This analysis
TliDSM5473_IP1	*Thermococcus litoralis* DSM 5473	500722..523246	44	integrated	CP006670.1	This analysis
TbaCH5_IP1	*Thermococcus barophilus* CH5	770185..788746	44	integrated	CP013050.1	This analysis
TbaCH5_IP2	*Thermococcus barophilus* CH5	2013643..2038136	48	integrated	CP013050.1	This analysis
TspJCM11816_IP1	*Thermococcus sp.* JCM11816	162578..186018	49	integrated	Ga0128353_102	This analysis
TGV1	*Thermococcus gammatolerans* EJ3	621669..642462	50	integrated	CP001398.1	Keller 2009
TceDSM17994_IP1	*Thermococcus celericrescens* DSM17994	15770..43341	129	integrated	NZ_LLYW01000013	This analysis
PchGC74_IP1	*Pyrococcus chitonophagus* GC74	1137169.. 1159084	46	integrated	NZ_CP015193	This analysis
PkuNCB100_IP1	*Pyrorcoccus kukulkanii sp.* NCB100	456321..486708	102	integrated	CP010835.1	This analysis
PHV1	*Pyrococcus horikoshii* OT3	1061525..1083228	47	integrated	BA000001.2	Keller 2009
PyaCH1_IP16	*Pyrococcus yayanosii* CH1	1238312..1255830	46	integrated	CP002779	This analysis
PspNA2_IP1	*Pyrococcus sp.* NA2	1199678..1221811	47	integrated	CP002670	This analysis
pGE2 = PabGE2_IP1	*Pyrococcus abyssi* GE2	1467989..1488841	48	free/integrated	-	This analysis
MMC6V1	*Methanococcus maripaludis* C6	358..48565	56	integrated	NC_009975	Keller 2009
MMC7V1	*Methanococcus maripaludis* C7*	no detectable limits	-	Integrated	NC_009637	Keller 2009
MMC7V2	*Methanococcus maripaludis* C7	1436513..1469347	56	integrated	NC_009637	Keller 2009
MMPV1= MmaS2_IP	*Methanococcus maripaludis* S2	735195..773477	53	integrated	NC_005791	Keller 2009
MmaKA1_IP1	*Methanococcus maripaludis* KA1	466296..491741	54	integrated	AP011526	This analysis

MmaOS7_IP1	*Methanococcus maripaludis* OS7	45126..475878	54	integrated	AP011528	This analysis
MmaC5_IP1	*Methanococcus maripaludis* C5*	no detectable limits	-	remnant		This analysis
MmaX1_IP1	*Methanococcus maripaludis* X1	no detectable limits	-	integrated	340623184	This analysis
MVV1	*Methanococcus voltae* A3	1715487..1742050	102	integrated	NC_014222	Keller 2009
MthDSM2095_IP1	*Methanothermococcus thermolithotrophicus* DSM2095	1..21165	54	-	NZ_AQXV01000029	This analysis
MigKol5_IP1	*Methanotorris igneus* Kol5	500181..524602	58	integrated	NC_015562	This analysis
MspFS406-22_IP1	Methanocaldococcus sp. FS406-22	1092561..1123012	54	integrated	NC_013887	This analysis
pMEFER01	Methanocaldococcus fervens AG86	1..22190	57	free	NC_013157	Soler 2011

800

801 Table 3. Conserved features among the pT26-2 family.

Name	Relative Core size in BDBH	Replication	Ori rep	Integrase Type	Target tRNA
PspNA2_IP1	7	MCM	1	Type-I	tRNA-Val
TbaCH5_IP1	3	-	Not found	Type-I	tRNA-Val
TceDSMA7994_IP1	6	T26-22p-like	1	Type-I	tRNA-Thr
PchCG74_IP1	7	-	1	Type-I	tRNA-Gly
PHV1	8	-	1	Type-I	tRNA-Ala
PyaCH1_IP16	9	-	1	Type-I	tRNA-Gly
TbaCH5_IP2	9	MCM	1	Type-I	tRNA-Tyr
PabGE2_IP2	7	T26-22p-like	1	Type-I	tRNA-Ala
PkuNCB100_IP1	7	-	1	Type-I	tRNA-Ala
TliDSM5473_IP1	8	-	Not found	Type-I	tRNA-Gly
pT26-2	8	T26-22p-like	1	Type-I	tRNA-Arg
TguDSM11113_IP1	9	-	1	Type-I	tRNA-Arg
TGV1	9	-	1	Type-I	tRNA-Arg
TKV3	8	T26-22p-like	1	Type-I	tRNA-Arg
TKV2	8	-	1	Type-I	tRNA-Glu
TspJCM11816_IP1	7	-	1	Type-I	tRNA-Arg
MmaOS7_IP1	9	MCM-like	1	Type-II	tRNA-Ser
MmaKA1_IP1	8	MCM-like	2	Type-II	tRNA-Ser
MmaX1_IP1	9	MCM	Not found	Type-II	-
MMC7V1	7	MCM	Not found	Type-II	tRNA-Ser
MmaS2_IP	8	MCM	1	Type-II	tRNA-Ser
MMC6V1	9	MCM	1	Type-II	tRNA-Ser
MMC7V2	8	MCM-like	1	Type-II	tRNA-Ser
MmaC5_IP1	2	-	Not found	-	-
MVV1	7	-	2	Type-II	tRNA-Leu
MthDSM2095_IP1	8	MCM-like	1	Type-II	tRNA-Ser
pMEFER01	9	MCM	1	Type-II	-
MspFS406-22_IP1	8	-	1	Type-II	tRNA-Ser
MigKol5_IP1	8	-	1	Type-II	tRNA-Ser

802

803

804

805

806

Figures captions

Fig. 1. Biogeography of the Thermococcales and Methanococcales isolation sites of the NCBI available genomes. **a.** Barplot indicating the number of Methanococcales and Thermococcales isolates**.** For both orders, we also indicate the number of isolate containing a spacer against a pT26-2 related element, or containing a pT26-2 related element or containing both **b.** The isolation sites corresponding to Methanococcales and Thermococcales are indicated on the world map by red and blue dots respectively. Six major regions have been also indicated by different cloud on the world map *East Pacific Ocean Ridge, Gulf of Mexico, North Atlantic Ridge, Vulcano island, North West Pacific Ocean Ridges, Oceania*. For each region the number of isolate is indicated with a pie chart and the presence of pT26-2-related plasmid or a spacer against theses are indicated using the same colour code than in the a panel.

Fig. 2. Comparison of Thermococcales pT26-2-related plasmids.
In this schematic representation the CORE genes and the integrase genes are indicated in green and orange respectively. The different genes encoding for putative replication protein are indicated with different shade of purple. The result of conservation between two related pT26-2 by tblastx is indicated with several shade of blue based on the protein identity percentage. The schematic phylogenetic tree in the left correspond to a part of the phylogenetic tree obtain with the concatenantion of the core proteins in Fig. 5.

Fig. 3. Comparison of Methanococcales pT26-2-related plasmids.
In this schematic representation the CORE genes and the integrase genes are indicated in green and orange respectively. The location of the putative replication origin is indicated on the plasmid with a purple circle. The different genes encoding for putative replication protein are indicated with different shade of purple. The result of conservation between two pT26-2-related plasmid by tblastx is indicated with several shade of blue based on the protein identity percentage. The schematic phylogenetic tree in the left correspond to a part of the phylogenetic tree obtain with the concatenantion of the core proteins in Fig. 5.

Fig. 4. Network view of pT26-2 elements conservation. Results of Bidirectional Best-Hit are represented as a network. The line thickness is related to the number of conserved genes between two elements. In addition for pT26-2 related plasmid they are colored depending of their host genera in several kind of green for Methanococcales and two kind of blue for Thermococcales. This network analysis suggests that pT26-2 and related elements are not transferred between the two orders, and have co–evolved with theirs host. This network show that some pT26-2 related plasmids shared genes with archaeal viruses, or other unknown kind of archaeal MGEs.

Fig. 5. Maximum Likelihood tree of the concatenated CORE proteins. The isolation region is indicated by a colored square. The scale-bars represent the average number of substitutions per site. Values at nodes represent support calculated by nonparametric bootstrap (out of 100).

Fig. 6. att site variability

A. The att sites correspond to the 3' terminus of the tRNA genes. On the alignment, the anti-codon is framed. The consensus sequences among Thermococcales and among Methanococcales att sites are highlighted in color. Long sequences were only partially presented.

B. and C. The att sites are displayed on the structure of the targeted tRNA for Thermococcales and Methanococcales, respectively. Circles represent tRNA nucleotides. Red circles correspond to the anticodon. Squares represent att sites nucleotides downstream of the tRNA gene. Black nucleotides are present in all att sites, darker grey in more than 77% and lighter grey in more than 33%.

Fig. 7. The evolutive model of the pT26-2 family. In this schematic representation the core module and the integrase module are indicated in green and orange respectively. The different replication modules are indicated with different shade of purple.

Figures

Fig. 1.

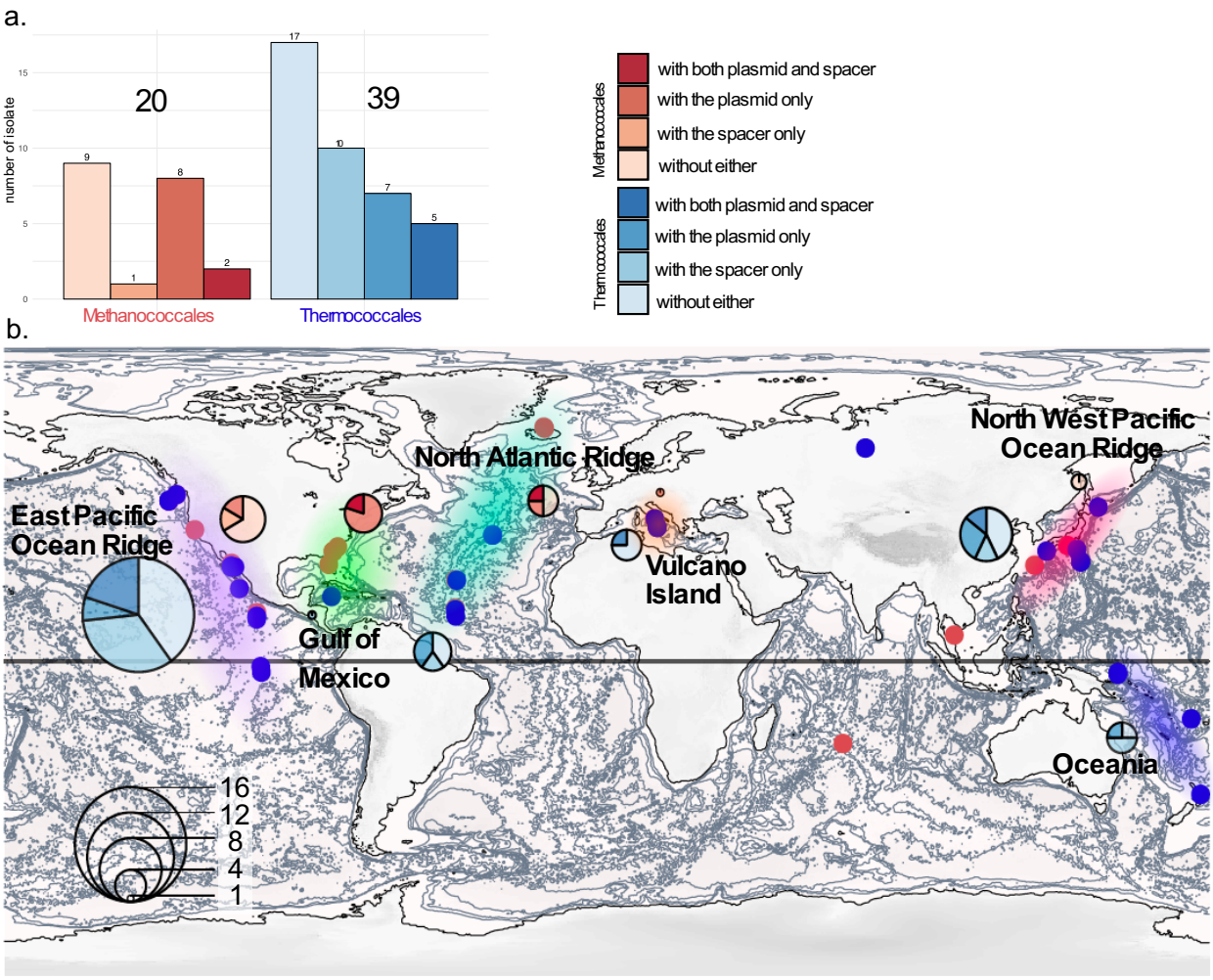


Fig. 2.

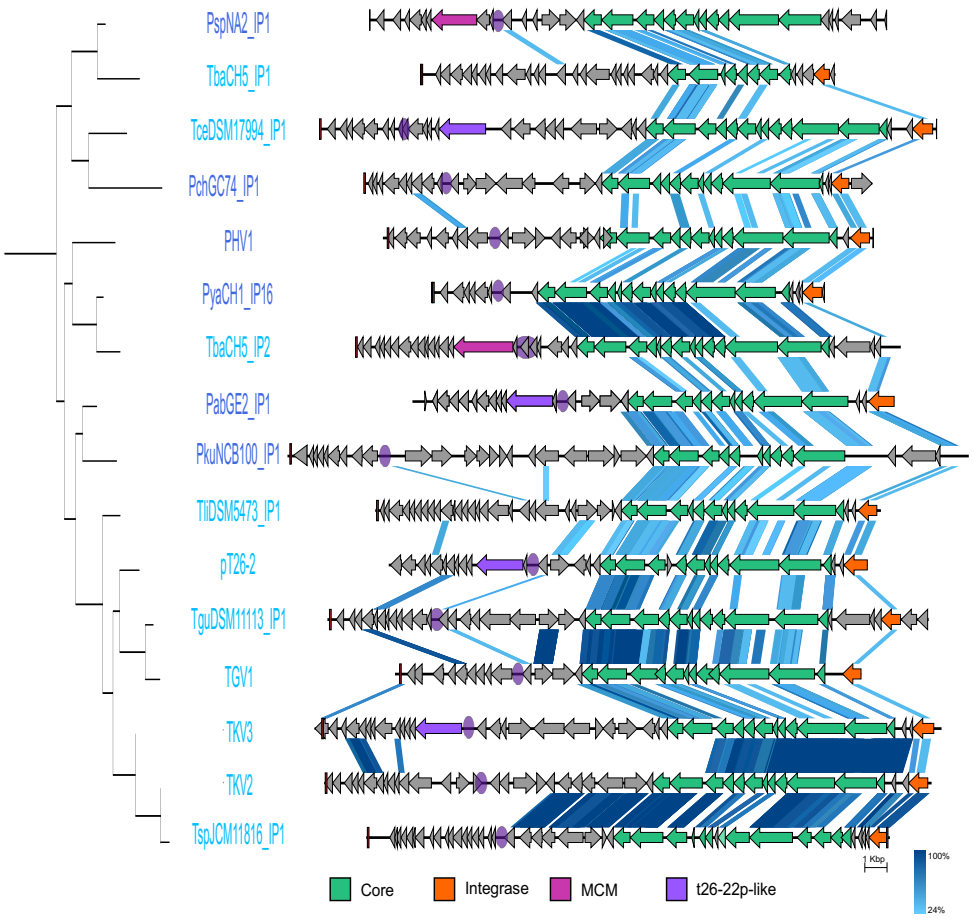

PspNA2_IP1
TbaCH5_IP1
TceDSM17994_IP1
PchGC74_IP1
PHV1
PyaCH1_IP16
TbaCH5_IP2
PabGE2_IP1
PkuNCB100_IP1
TliDSM5473_IP1
pT26-2
TguDSM11113_IP1
TGV1
TKV3
TKV2
TspJCM11816_IP1
1 Kbp
100%
24%
Core
Integrase
MCM
t26-22p-like

Fig. 3.

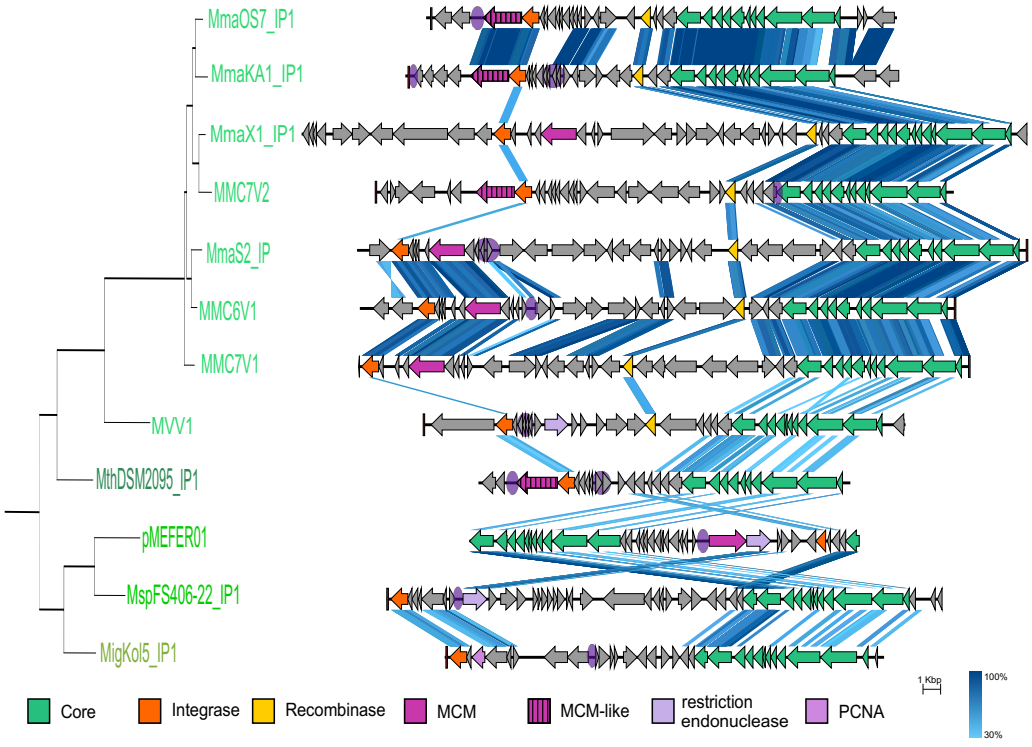

MmaOS7_IP1
MmaKA1_IP1
MmaX1_IP1
MMC7V2
MmaS2_IP
MMC6V1
MMC7V1
MVV1
MthDSM2095_IP1
pMEFER01
MspFS406-22_IP1
MigKol5_IP1
1 Kbp
100%
30%
Core
Integrase
Recombinase
MCM
MCM-like
restriction endonuclease
PCNA

Fig. 4.

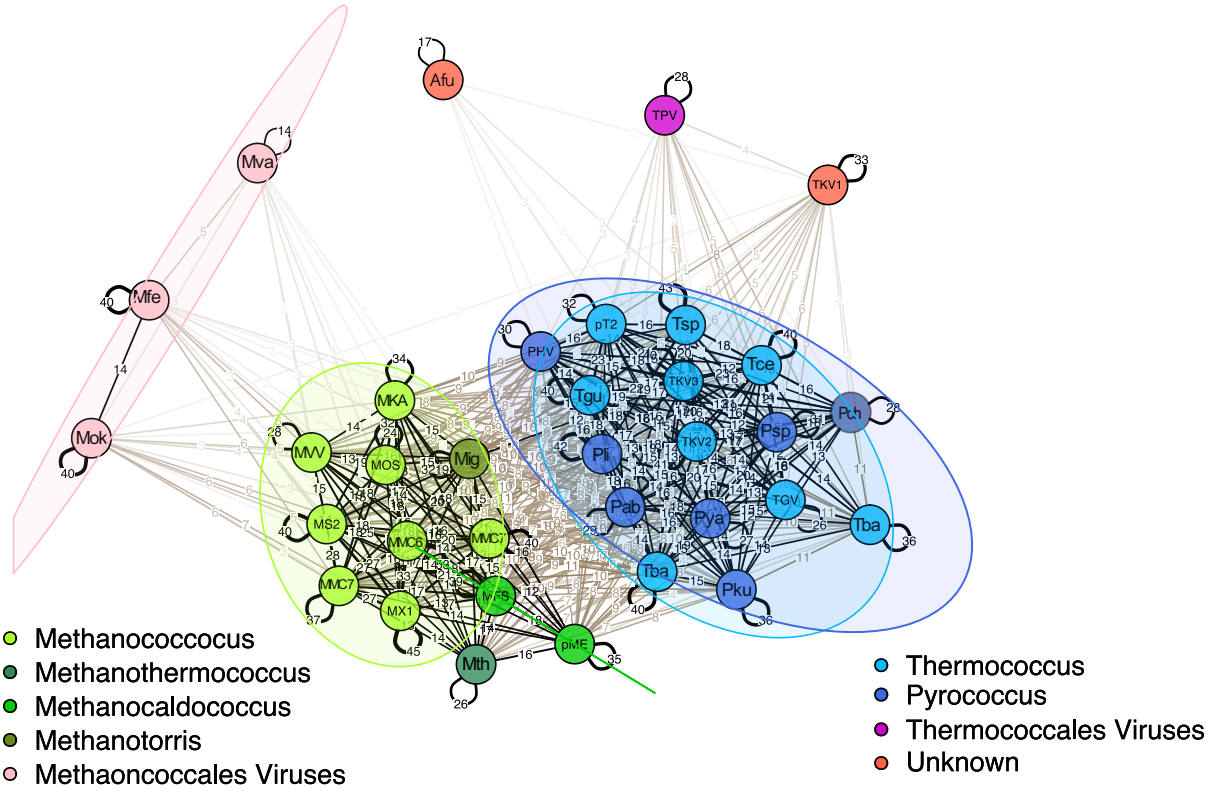

Afu
TPV
TKV1
Mva
Mfe
Mok
MKA
MMV
MOS
Mig
MS2
MMC6
MMC7
MMC7
MX1
MFS
Mth
pME
pT2
Tsp
PHV
Tce
TKV3
Tgu
Pch
Psp
Pli
TKV2
Pab
TGV
Pya
Tba
Tba
Pku
Methanococcocus
Methanothermococcus
Methanocaldococcus
Methanotorris
Methaoncoccales Viruses
Thermococcus
Pyrococcus
Thermococcales Viruses
Unknown

Fig. 5.

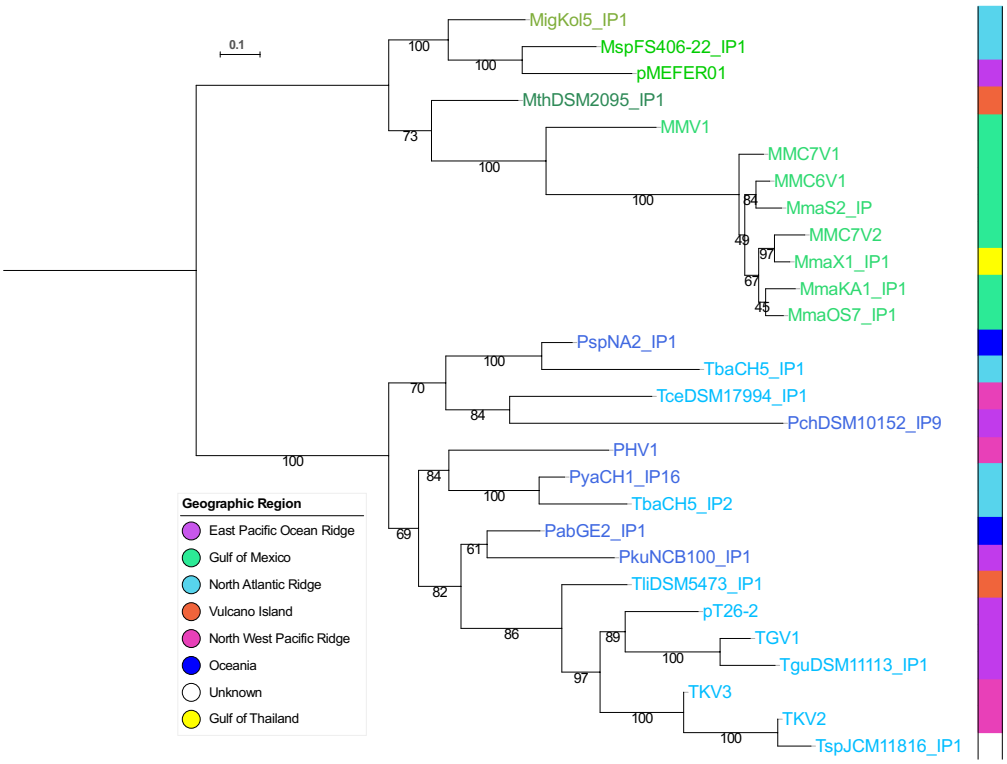

0.1
MigKol5_IP1
MspFS406-22_IP1
pMEFER01
MthDSM2095_IP1
MMV1
MMC7V1
MMC6V1
MmaS2_IP
MMC7V2
MmaX1_IP1
MmaKA1_IP1
MmaOS7_IP1
PspNA2_IP1
TbaCH5_IP1
TceDSM17994_IP1
PchDSM10152_IP9
PHV1
PyaCH1_IP16
TbaCH5_IP2
PabGE2_IP1
PkuNCB100_IP1
TliDSM5473_IP1
pT26-2
TGV1
TguDSM11113_IP1
TKV3
TKV2
TspJCM11816_IP1
Geographic Region
East Pacific Ocean Ridge
Gulf of Mexico
North Atlantic Ridge
Vulcano Island
North West Pacific Ridge
Oceania
Unknown
Gulf of Thailand

Fig. 6.

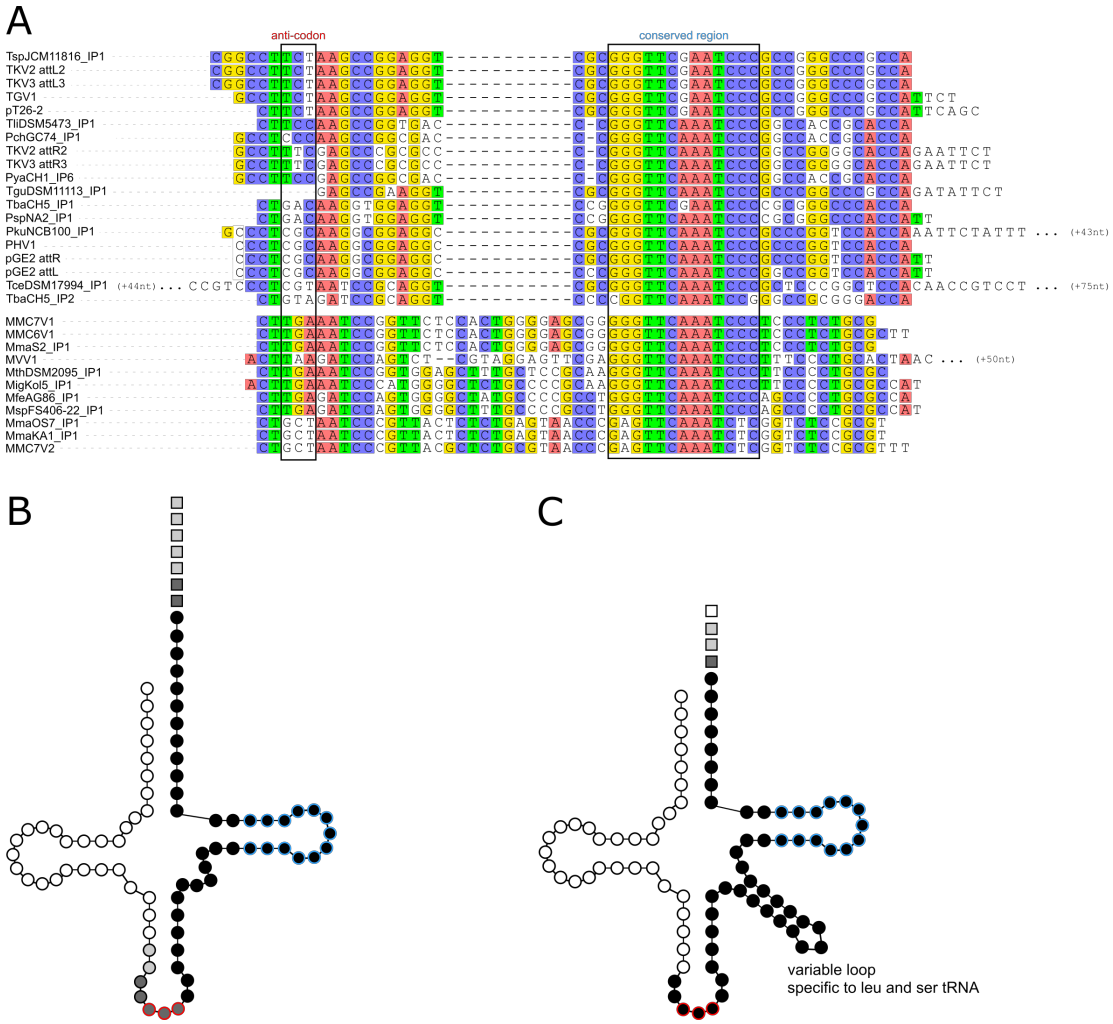
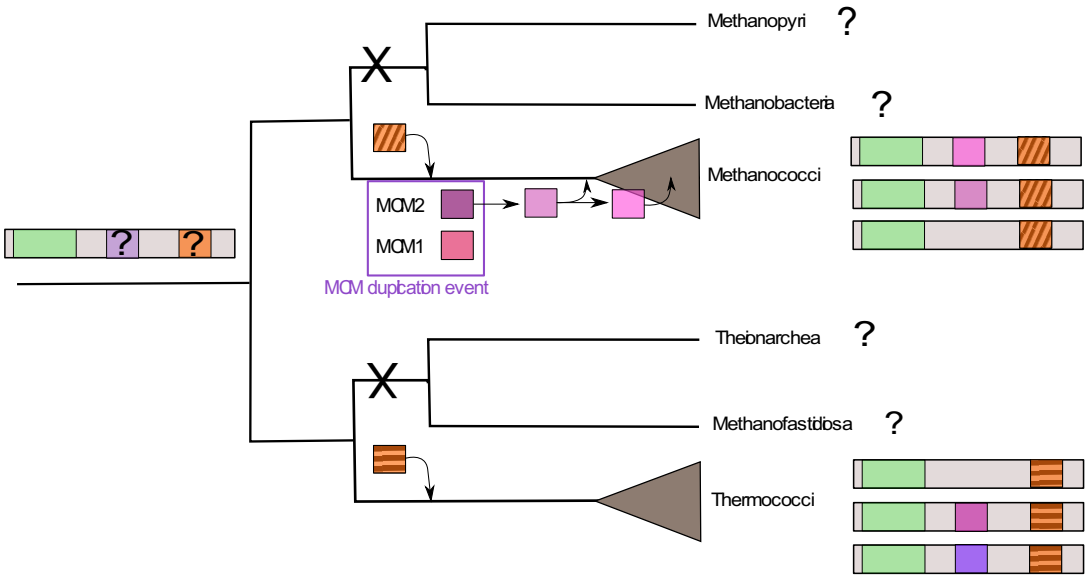

A
anti-codon
conserved region
TspJCM11816_IP1
TKV2 attL2
TKV3 attL3
TGV1
pT26-2
TliDSM5473_IP1
PchGC74_IP1
TKV2 attR2
TKV3 attR3
PyaCH1_IP6
TguDSM11113_IP1
TbaCH5_IP1
PspNA2_IP1
PkuNCB100_IP1
PHV1
pGE2 attR
pGE2 attL
TceDSM17994_IP1
TbaCH5_IP2
MMC7V1
MMC6V1
MmaS2_IP1
MVV1
MthDSM2095_IP1
MigKol5_IP1
MfeAG86_IP1
MspFS406-22_IP1
MmaOS7_IP1
MmaKA1_IP1
MMC7V2
B
C
variable loop
specific to leu and ser tRNA

Fig. 7.


Methanopyri
Methanobacteria
Methanococci
MCM2
MCM1
MCM duplication event
Thermoplasmatales
Methanofastidiosa
Thermococci

References

Adam PS, Borrel G, Brochier-armanet C. 2017. The growing tree of Archaea : new perspectives on their diversity , evolution and ecology. ISME J. [Internet]:1–19. Available from: http://dx.doi.org/10.1038/ismej.2017.122

Albers SV, Siebers B. 2014. The Prokaryotes: Other Major Lineages of Bacteria and The Archaea.

Antranikian G, Suleiman M, Schäfers C, Adams MWW, Bartolucci S, Blamey JM, Kåre N, Elizaveta B, Osmolovskaya B, Milton S, et al. 2017. Diversity of bacteria and archaea from two shallow marine hydrothermal vents from Vulcano Island. Extremophiles 21:733–742.

Brochier-Armanet C, Forterre P, Gribaldo S. 2011. Phylogeny and evolution of the Archaea: One hundred genomes later. Curr. Opin. Microbiol. 14:274–281.

Brum JR, Cesar Ignacio-Espinoza J, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S, Cruaud C, De Vargas C, Gasol JM, et al. 2015. Patterns and ecological drivers of ocean viral communities. Science (80-.). 348.

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

Catchpole R, Gorlas A, Oberto J, Forterre P. 2018. A series of new E . coli – Thermococcus shuttle vectors compatible with previously existing vectors. Extremophiles [Internet]. Available from: https://doi.org/10.1007/s00792-018-1019-6

Chandler M, Mahillon J. 2002. Insertion Sequences Revisited. In: Mobile DNA II. American Society of Microbiology. p. 305–366. Available from: http://www.asmscience.org/content/book/10.1128/9781555817954.chap15

Chernomor O, Minh BQ, Hoang DT, Vinh LS, von Haeseler A. 2017. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. 35:518–522.

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

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

Da Cunha V, Gaia M, Gadelle D, Nasir A, Forterre P. 2017. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. PLoS Genet. 13.

Donoghue MTA, Keshavaiah C, Swamidatta SH, Spillane C. 2011. Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. BMC Evol. Biol. 11:1–23.

Erauso G, Marsin S, Benbouzid-Rollet N, Baucher MF, Barbeyron T, Zivanovic Y, Prieur D, Forterre P. 1996. Sequence of plasmid pGT5 from the archaeon Pyrococcus abyssi: Evidence for rolling-circle replication in a hyperthermophile. J. Bacteriol. 178:3232–3237.

Erauso G, Reysenbach A, Godfroy A, Meunier J, Crump B, Partensky F, Baross JA, Marteinsson V,

Barbier G, Pace NR, et al. 1993. Pyrococcus abyssi sp. nov., a new hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent. Arch. Microbiol. 160:338–349.

Erauso G, Stedman KM, van den Werken HJG, Zillig W, van der Oost J. 2006. Two novel conjugative plasmids from a single strain of Sulfolobus. Microbiology 152:1951–1968.

Erdmann S, Tschitschko B, Zhong L, Raftery MJ, Cavicchioli R. 2017. Specialized Membrane Vesicles To Disseminate and. Nat. Microbiol.:0–1.

Faraco JH, Morrison NA, Baker A, Shine J, Frossard PM. 1989. Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. Nucleic Acids Res. 17:94043.

Fiala G, Stetter KO. 1986. Pyrococcus furiosus sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C. Arch. Microbiol. 145:56–61.

Forterre P. 2013. The common ancestor of archaea and eukarya was not an archaeon. Archaea 2013.

Forterre P, Da Cunha V, Catchpole R. 2017. Plasmid vesicles mimicking virions. Nat. Microbiol. 2:1340–1341.

Forterre P, Gaïa M. 2016. Giant viruses and the origin of modern eukaryotes. Curr. Opin. Microbiol. 31:44–49.

Forterre P, Krupovic M, Raymann K, Soler N. 2014. Plasmids from Euryarchaeota. Microbiol. Spectr. 2:PLAS-0027-2014.

Fukui T, Atomi H, Kanai T, Matsumi R, Fujiwara S, Imanaka T. 2005. Complete genome sequence of the hyperthermophilic archaeon Thermococcus kodakaraensis KOD1 and comparison with Pyrococcus genomes. Genome Res. 15:352–363.

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

Gehring AM, Astling DP, Matsumi R, Burkhart BW, Kelman Z, Reeve JN, Jones KL, Santangelo TJ. 2017. Genome replication in Thermococcus kodakarensis independent of Cdc6 and an origin of replication. Front. Microbiol. 8:1–10.

Geslin C, Gaillard M, Flament D, Rouault K, Le Romancer M, Prieur D, Erauso G. 2007. Analysis of the first genome of a hyperthermophilic marine virus-like particle, PAV1, isolated from Pyrococcus abyssi. J. Bacteriol. 189:4510–4519.

Gill S, Krupovic M, Desnoues N, Béguin P, Sezonov G, Forterre P. 2014. A highly divergent archaeo-eukaryotic primase from the Thermococcus nautilus plasmid, pTN2. Nucleic Acids Res. 42:3707–3719.

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

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

Gorlas A, Krupovic M, Forterre P, Geslin C. 2013. Living side by side with a virus: Characterization of two novel plasmids from Thermococcus prieurii, a host for the spindle-shaped virus TPV1. Appl. Environ. Microbiol. 79:3822–3828.

Greve B, Jensen S, Phan H, Brügger K, Zillig W, She Q, Garrett RA. 2005. Novel RepA-MCM proteins encoded in plasmids pTAU4, pORA1 and pTIK4 from Sulfolobus neozealandicus. Archaea 1:319–325.

Grissa I, Vergnaud G, Pourcel C. 2007. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics 8:172.

Guérillot R, Cunha V Da, Sauvage E, Bouchier C, Glaser P. 2013. Modular evolution of TnGBSs, a new family of integrative and conjugative elements associating insertion sequence transposition, plasmid replication, and conjugation for their spreading. J. Bacteriol. 195:1979–1990.

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

Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E. 2009. Network analyses structure genetic diversity in independent genetic worlds. Proc. Natl. Acad. Sci. 107:127–132.

Heinen TJAJ, Staubach F. 2009. Emergence of a New Gene from an Intergenic Region.

Iranzo J, Krupovic M, Koonin E V. 2016. The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. 7:1–21.

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

Kazlauskas D, Sezonov G, Charpin N, Venclovas Č, Forterre P, Krupovic M. 2018. Novel Families of Archaeo-Eukaryotic Primases Associated with Mobile Genetic Elements of Bacteria and Archaea. J. Mol. Biol. 430:737–750.

Keese PK, Gibbs A. 2006. Origins of genes: “big bang” or continuous creation? Proc. Natl. Acad. Sci. 89:9489–9493.

Keller J, Leulliot N, Soler N, Collinet B, Vincentelli R, Forterre P, Van Tilbeurgh H. 2009. A protein encoded by a new family of mobile elements from Euryarchaea exhibits three domains with novel folds. Protein Sci. 18:825–838.

Kelley LA, Mezulis S, Yates C, Wass M, Sternberg M. 2015. The Phyre2 web portal for protein modelling, prediction, and analysis. Nat. Protoc. [Internet] 10:845–858. Available from: http://dx.doi.org/10.1038/nprot.2015-053

Koonin E V., Wolf YI. 2008. Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 36:6688–6719.

Krupovič M, Bamford DH. 2008. Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota. Virology 375:292–300.

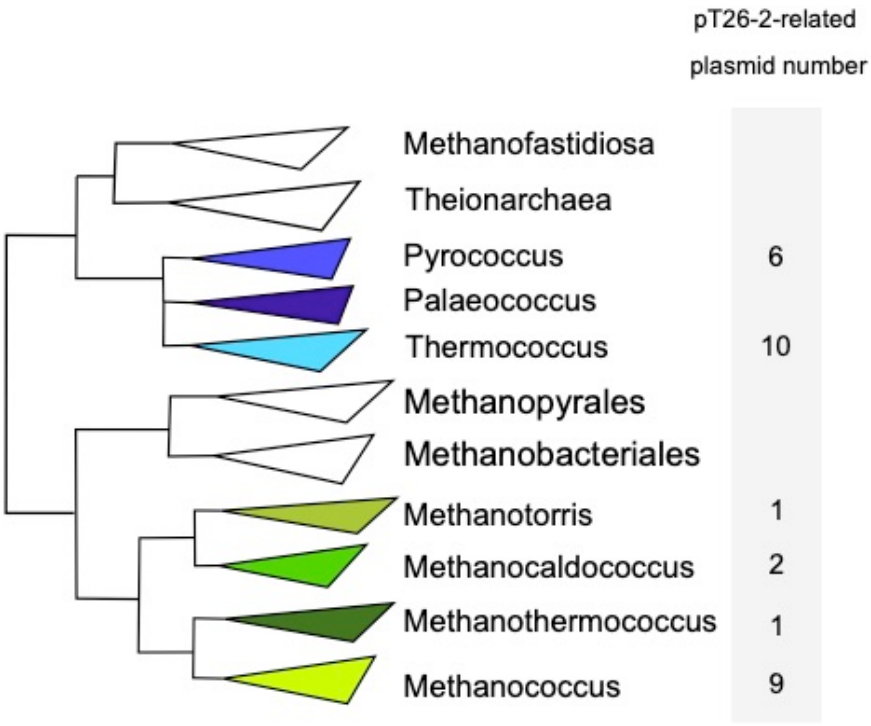
Krupovič M, Forterre P, Bamford DH. 2010. Comparative Analysis of the Mosaic Genomes of Tailed

1024 Archaeal Viruses and Proviruses Suggests Common Themes for Virion Architecture and
1025 Assembly with Tailed Viruses of Bacteria. J. Mol. Biol. 397:144–160.
1026 Krupovic M, Gonnet M, Hania W Ben, Forterre P, Erauso G. 2013. Insights into Dynamics of Mobile
1027 Genetic Elements in Hyperthermophilic Environments from Five New Thermococcus Plasmids.
1028 PLoS One 8:1–10.
1029 Krupovič M, Gribaldo S, Bamford DH, Forterre P. 2010. The evolutionary history of archaeal MCM
1030 helicases: A case study of vertical evolution combined with Hitchhiking of mobile genetic
1031 elements. Mol. Biol. Evol. 27:2716–2732.
1032 Krupovic M, Koonin E V. 2017. Multiple origins of viral capsid proteins from cellular ancestors.
1033 :2401–2410.
1034 Krupovic M, Makarova KS, Wolf YI, Medvedeva S, Prangishvili D, Forterre P, Koonin E V. 2019.
1035 Integrated Mobile Genetic Elements in Thaumarchaeota. Environ. Microbiol.:1–23.
1036 Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.
1037 Lazar CS, Baker BJ, Seitz KW, Teske AP. 2017. Genomic reconstruction of multiple lineages of
1038 uncultured benthic archaea suggests distinct biogeochemical roles and ecological niches. ISME J.
1039 11:1118–1129.
1040 Lee C, Kim J, Shin SG, Hwang S. 2006. Absolute and relative QPCR quantification of plasmid copy
1041 number in Escherichia coli. J. Biotechnol. 123:273–280.
1042 Legendre M, Fabre E, Poirot O, Jeudy S, Lartigue A, Alempic J-MM, Beucher L, Philippe N, Bertaux
1043 L, Christo-Foroux E, et al. 2018. Diversity and evolution of the emerging Pandoraviridae family.
1044 Nat. Commun. [Internet] 9. Available from: http://dx.doi.org/10.1038/s41467-018-04698-4
1045 Lepage E, Marguet E, Geslin C, Matte-Tailliez O, Zillig W, Forterre P, Tailliez P. 2004. Molecular
1046 diversity of new Thermococcalesisolates from a single area of hydrothermal deep-sea vents as
1047 revealed by randomly amplified polymorphic DNA fingerprinting and 16S rRNA gene sequence
1048 analysis. Appl. Environ. Microbiol. 70:1277–1286.
1049 Lerat E, Daubin V, Moran NA. 2003. From gene trees to organismal phylogeny in prokaryotes: The
1050 case of the γ-Proteobacteria. PLoS Biol. 1:101–109.
1051 Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding
1052 DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression.
1053 Proc. Natl. Acad. Sci. 103:9935–9939.
1054 Lipps G, Weinzierl AO, Von Scheven G, Buchen C, Cramer P. 2004. Structure of a bifunctional DNA
1055 primase-polymerase. Nat. Struct. Mol. Biol. 11:157–162.
1056 Lossouarn J, Dupont S, Gorlas A, Mercier C, Bienvenu N, Marguet E, Forterre P, Geslin C. 2015. An
1057 abyssal mobilome: Viruses, plasmids and vesicles from deep-sea hydrothermal vents. Res.
1058 Microbiol. 166:742–752.
1059 Lucas S, Toffin L, Zivanovic Y, Charlier D, Forterre P, Prieur D. 2002. Construction of a Shuttle
1060 Vector for , and Spheroplast Transformation of , the Hyperthermophilic Archaeon Pyrococcus

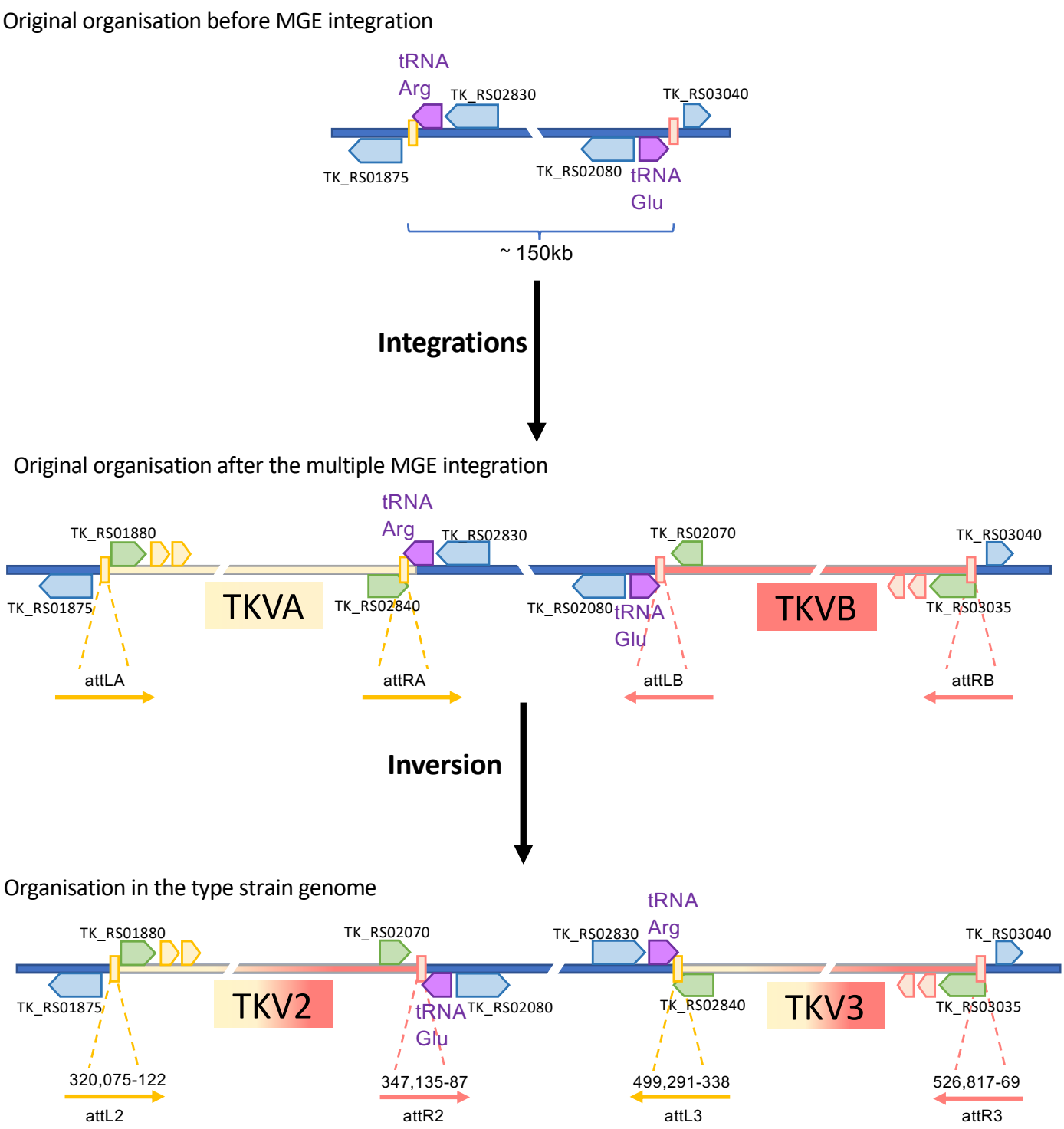
abyssi. Appl. Environ. Microbiol. 68:5528–5536.
Makarova KS, Wolf YI, Koonin E V. 2015. Archaeal Clusters of Orthologous Genes (arCOGs): An Update and Application for Analysis of Shared Features between Thermococcales, Methanococcales, and Methanobacteriales. Life (Basel, Switzerland) 5:818–840.
Marteinsson VT, Watrin L, Prieur D, Caprais C, Raguenes G, Erauso G. 1995. Phenotypic Characterization, DNA Similarities, and Protein Profiles of Twenty Sulfur-Metabolizing Hyperthermophilic Anaerobic Archaea Isolated from Hydrothermal Vents in the Southwestern Pacific Ocean. Int. J. Syst. Bacteriol. 45:623–632.
Miele V, Penel S, Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 12.
Mir-Sanchis I, Roman CA, Misiura A, Pigli YZ, Boyle-Vavra S, Rice PA. 2016. Staphylococcal SCCmec elements encode an active MCM-like helicase and thus may be replicative. Nat. Struct. Mol. Biol. [Internet] 23:891–898. Available from: http://dx.doi.org/10.1038/nsmb.3286
Mizuno CM, Prajapati B, Lucas-Staat S, Sime-Ngando T, Forterre P, Bamford DH, Prangishvili D, Krupovic M, Oksanen HM. 2019. Novel haloarchaeal viruses from Lake Retba infecting Haloferax and Halorubrum species. Environ. Microbiol.
Nobu MK, Narihiro T, Kuroda K, Mei R, Liu WT. 2016. Chasing the elusive Euryarchaeota class WSA2: Genomes reveal a uniquely fastidious methyl-reducing methanogen. ISME J. 10:2478–2487.
Oberto J. 2013. SyntTax : a web server linking synteny to prokaryotic taxonomy. BMC Bioinformatics 14:1471–2105.
Prieur D, Erauso G, Geslin C, Lucas S, Gaillard M, Bidault a, Mattenet a-C, Rouault K, Flament D, Forterre P, et al. 2004. Genetic elements of Thermococcales. Biochem. Soc. Trans. 32:184–187.
Providenti MA, O'Brien JM, Ewing RJ, Paterson ES, Smith ML. 2006. The copy-number of plasmids and other genetic elements can be determined by SYBR-Green-based quantitative real-time PCR. J. Microbiol. Methods 65:476–487.
Ravin N V., Beletsky A V., Mardanov A V., Skryabin KG, Svetlitchnyi VA, Bonch-Osmolovskaya EA, Miroshnichenko ML. 2009. Metabolic Versatility and Indigenous Origin of the Archaeon Thermococcus sibiricus, Isolated from a Siberian Oil Reservoir, as Revealed by Genome Analysis. Appl. Environ. Microbiol. 75:4580–4588.
Raymann K, Brochier-Armanet C, Gribaldo S. 2015. The two-domain tree of life is linked to a new root for the Archaea. Proc. Natl. Acad. Sci. U. S. A. 112:6670–6675.
Raymann K, Forterre P, Brochier-Armanet C, Gribaldo S. 2014. Global phylogenomic analysis disentangles the complex evolutionary history of DNA replication in archaea. Genome Biol. Evol. 6:192–212.
Redder P, Peng X, Brügger K, Shah SA, Roesch F, Greve B, She Q, Schleper C, Forterre P, Garrett RA, et al. 2009. Four newly isolated fuselloviruses from extreme geothermal environments

1098 reveal unusual morphologies and a possible interviral recombination mechanism. Environ.
1099 Microbiol. 11:2849–2862.
1100 Santangelo TJ, Čuboňová L, Reeve JN. 2008. Shuttle vector expression in Thermococcus
1101 kodakaraensis: Contributions of cis elements to protein synthesis in a hyperthermophilic
1102 archaeon. Appl. Environ. Microbiol. 74:3099–3104.
1103 Schleper C, Kubo K, Zillig W. 1992. The particle SSV1 from the extremely thermophilic archaeon
1104 Sulfolobus is a virus: demonstration of infectivity and of transfection with viral DNA. Proc. Natl.
1105 Acad. Sci. [Internet] 89:7645–7649. Available from:
1106 http://www.pnas.org/cgi/doi/10.1073/pnas.89.16.7645
1107 She Q, Shen B, Chen L. 2004. Archaeal integrases and mechanisms of gene capture. Biochem. Soc.
1108 Trans. 32:222–226.
1109 Soler N, Gaudin M, Marguet E, Forterre P. 2011. Plasmids, viruses and virus-like membrane vesicles
1110 from Thermococcales. Biochem. Soc. Trans. 39:36–44.
1111 Soler N, Marguet E, Cortez D, Desnoues N, Keller J, van Tilbeurgh H, Sezonov G, Forterre P. 2010.
1112 Two novel families of plasmids from hyperthermophilic archaea encoding new families of
1113 replication proteins. Nucleic Acids Res. 38:5088–5104.
1114 Soler N, Marguet E, Verbavatz JM, Forterre P. 2008. Virus-like vesicles and extracellular DNA
1115 produced by hyperthermophilic archaea of the order Thermococcales. Res. Microbiol. 159:390–
1116 399.
1117 Steenbakkers PJM, Geerts WJ, Ayman-Oz NA, Keltjens JT. 2006. Identification of pseudomurein cell
1118 wall binding domains. Mol. Microbiol. 62:1618–1630.
1119 Sullivan MJ, Petty NK, Beatson SA. 2011. Easyfig: A genome comparison visualizer. Bioinformatics
1120 27:1009–1010.
1121 Sun C, Zhou M, Li Y, Xiang H. 2006. Molecular characterization of the minimal replicon and the
1122 unidirectional theta replication of pSCM201 in extremely halophilic archaea. J. Bacteriol.
1123 188:8136–8144.
1124 Tagashira K, Fukuda W, Matsubara M, Kanai T, Atomi H, Imanaka T. 2013. Genetic studies on the
1125 virus-like regions in the genome of hyperthermophilic archaeon, Thermococcus kodakarensis.
1126 Extremophiles 17:153–160.
1127 Takai K, Komatsu T, Inagaki F, Horikoshi K. 2001. Distribution of Archaea in a Black Smoker
1128 Chimney Structure. Appl. Environ. Microbiol. 67:3618–3629.
1129 Toll-riera M, Bosch N, Castelo R, Armengol L, Estivill X, Alba MM. 2009. Origin of Primate Orphan
1130 Genes : A Comparative Genomics Approach. Mol. Biol. Evol. 26:603–612.
1131 Vannier P, Marteinsson VT, Fridjonsson OH, Oger P, Jebbar M, Copernic PN, Plouzane F-. 2011.
1132 Complete Genome Sequence of the Hyperthermophilic , Piezophilic ,Heterotrophic, and
1133 Carboxydotrophic Archaeon Thermococcus barophilus MP. J. Bacteriol. 193:1481–1482.
1134 Visweswaran GRR, Dijkstra BW, Kok J. 2010. Two Major Archaeal Pseudomurein

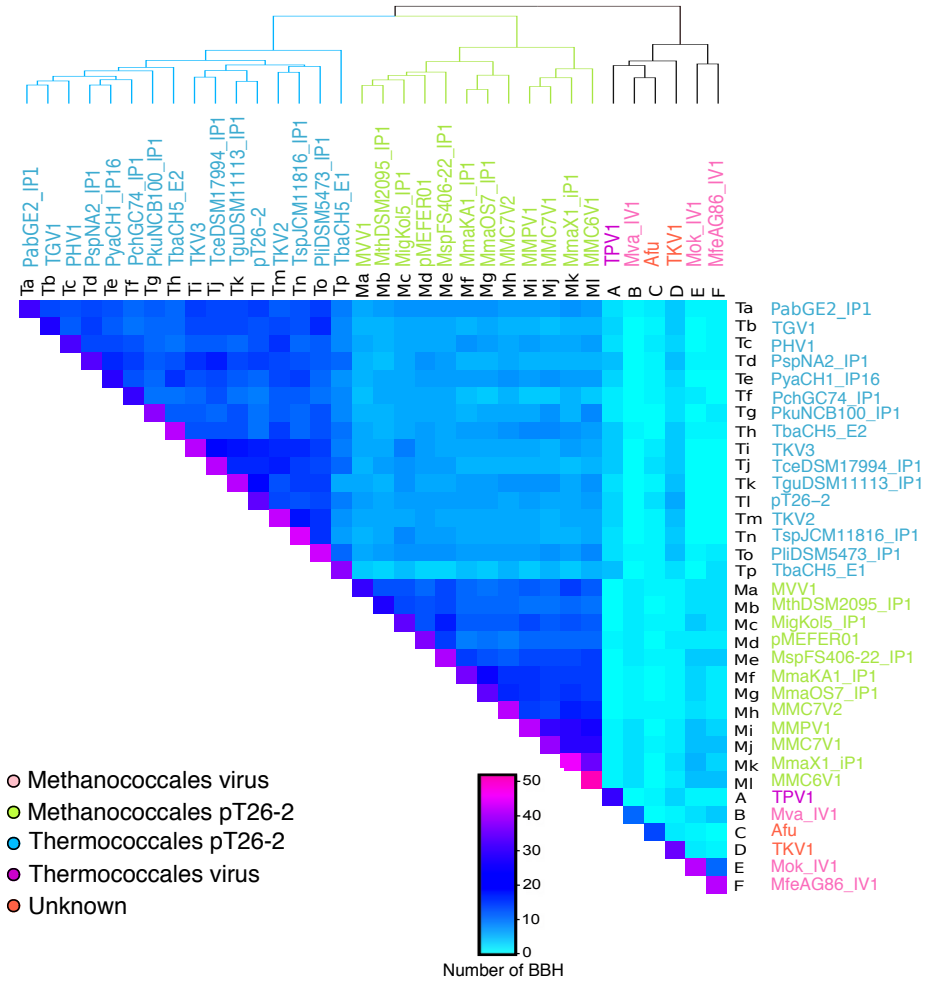
1135 Endoisopeptidases : PeiW and PeiP. Archaea 2010.
1136 Walters AD, Chong JPJ. 2010. An archaeal order with multiple minichromosome maintenance genes.
1137 Microbiology 156:1405–1414.
1138 Wang H, Peng N, Shah S a, Huang L, She Q. 2015. Archaeal Extrachromosomal Genetic Elements.
1139 Microbiol. Mol. Biol. Rev. 79:117–152.
1140 Wang J, Liu Y, Liu Y, Du K, Xu S, Wang Y, Krupovic M, Chen X. 2018. A novel family of tyrosine
1141 integrases encoded by the temperate pleolipovirus SNJ2. Nucleic Acids Res.:1–16.
1142 Wong TKF, Jermiin LS, Minh BQ, Kalyaanamoorthy S, von Haeseler A. 2017. ModelFinder: fast
1143 model selection for accurate phylogenetic estimates. Nat. Methods [Internet] 14:587–589.
1144 Available from: http://dx.doi.org/10.1038/nmeth.4285
1145 Zillig W, Holz I, Janekovic D, Schäfer W, Reiter WD. 1983. The Archaebacterium Thermococcus
1146 celer Represents, a Novel Genus within the Thermophilic Branch of the Archaebacteria. Syst.
1147 Appl. Microbiol. 4:88–94.
1148 Zivanovic Y, Armengaud J, Lagorce A, Leplat C, Guérin P, Dutertre M, Anthouard V, Forterre P,
1149 Wincker P, Confalonieri F. 2009. Genome analysis and genome-wide proteomics of
1150 Thermococcus gammatolerans , the most radioresistant organism known amongst the Archaea.
1151 Genome Biol. 10.
1152



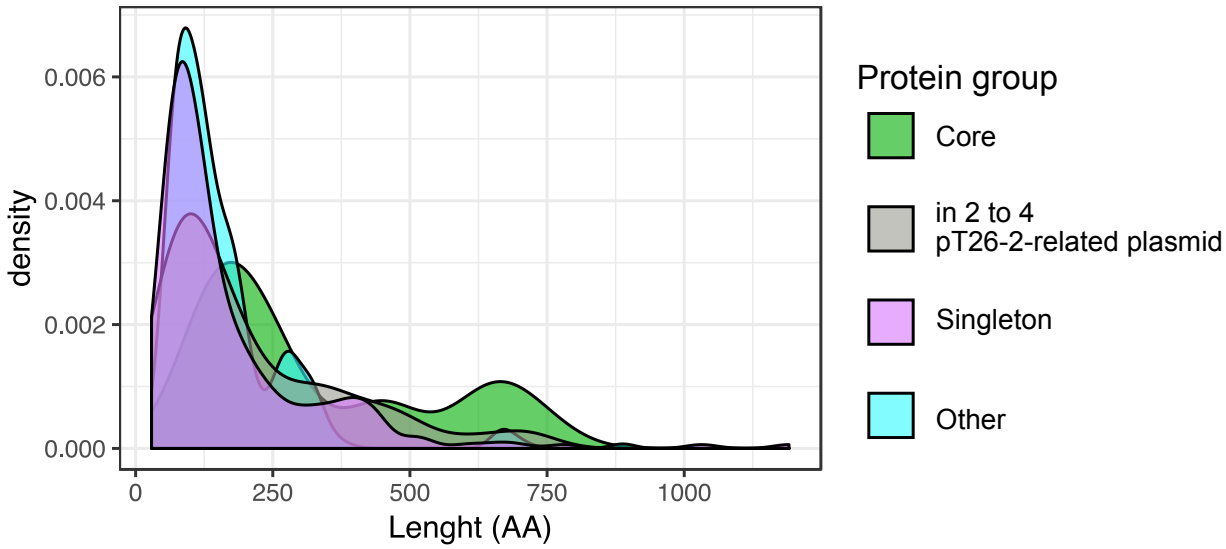
Supplementary Figure S1. Schematic representation of the phylogenetic relation between the Thermococcales, the Methanococcales and their closest archaeal relatives. In this schema the different genus of the Thermococcales and the Methanococcales were coloured in blue and green respectively



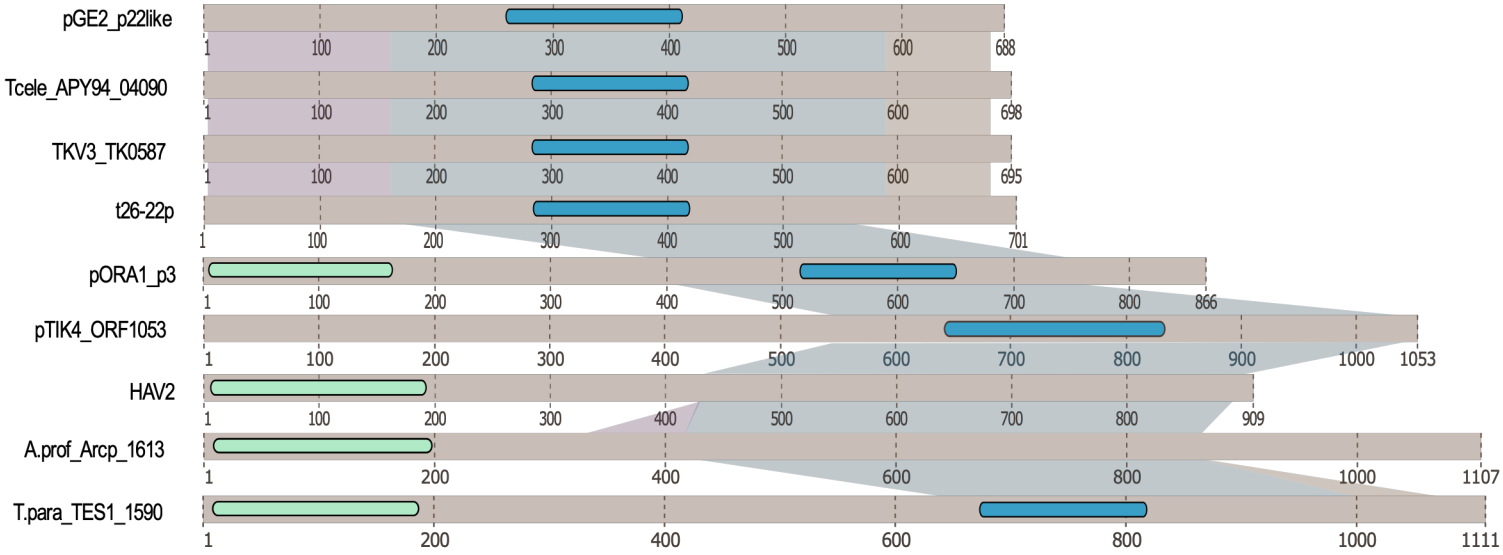
Supplementary Figure S2. TKV2 and TKV3 inactivation by large DNA inversion in *T. kodakarensis* genome.



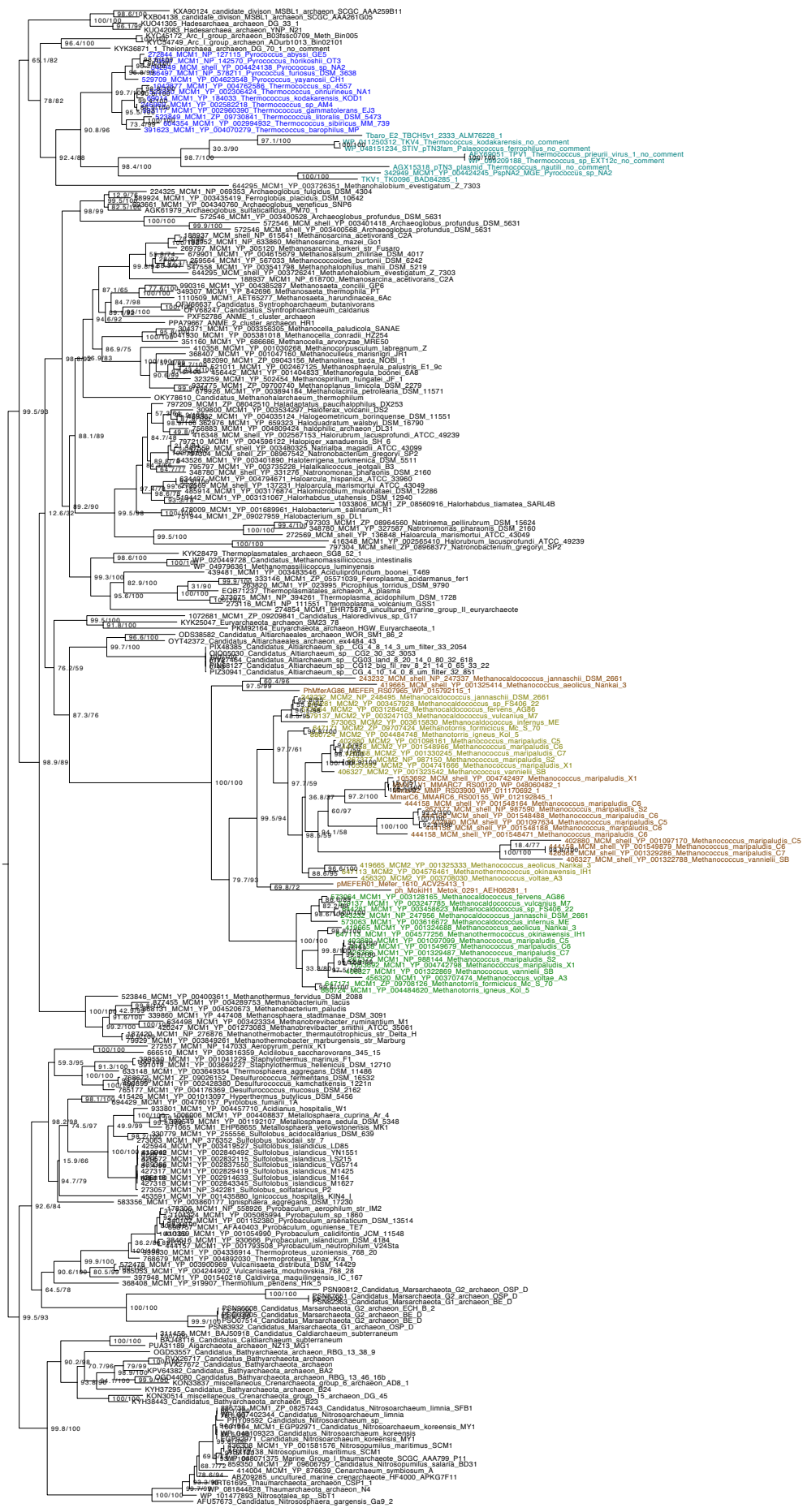
Supplementary Figure S3. Heatmap view of pT26-2-related plasmid conservation and their connections with viruses. Results of Bidirectional Best-Hit are represented as a heatmap. In the heatmap the number of conserved proteins is indicated in the scale. The heatmap analysis revealed two distinct groups, one containing the elements in Thermococcales and the other those of Methanococcales. The heatmap also show that some of the pT26-2-related plasmids shared genes with viruses.



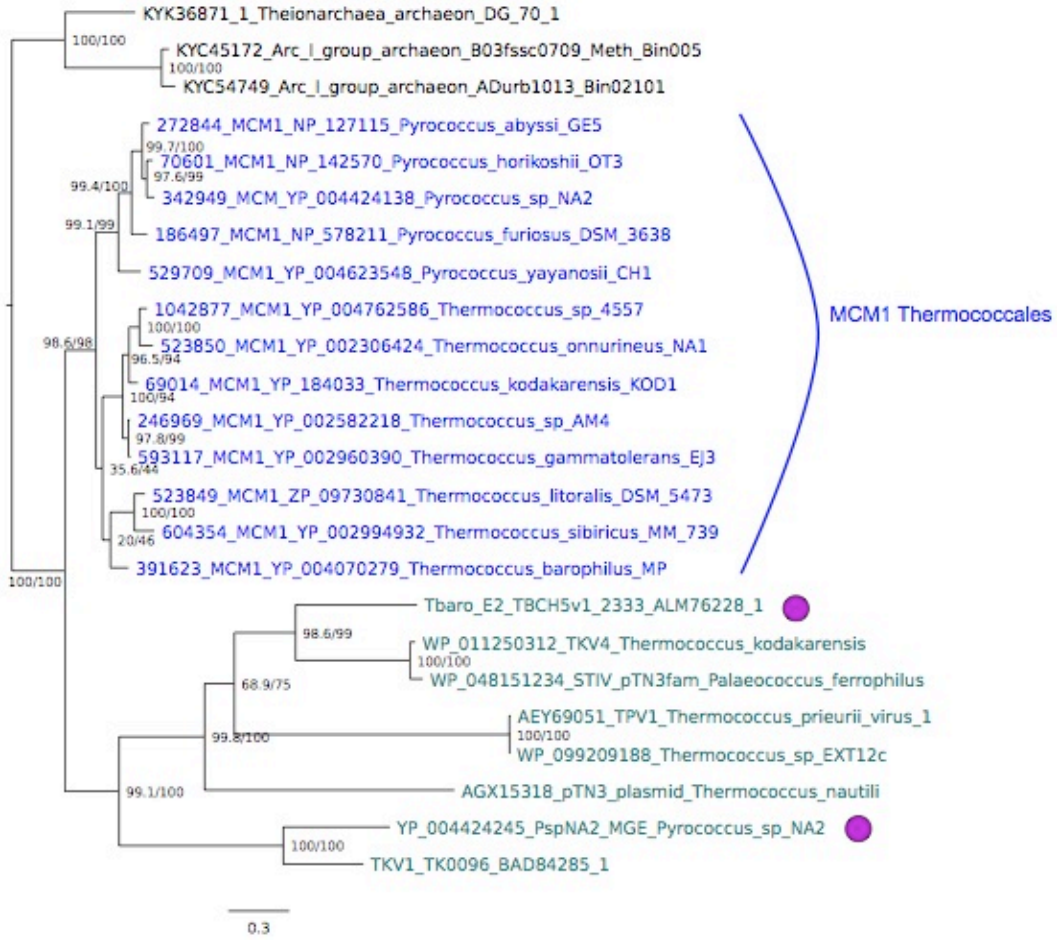
Supplementary Figure S4. Size of pT26-2-related plasmid encoded proteins. The protein set were divided in 4 different groups. For each group the density of protein is indicated at the different protein size.



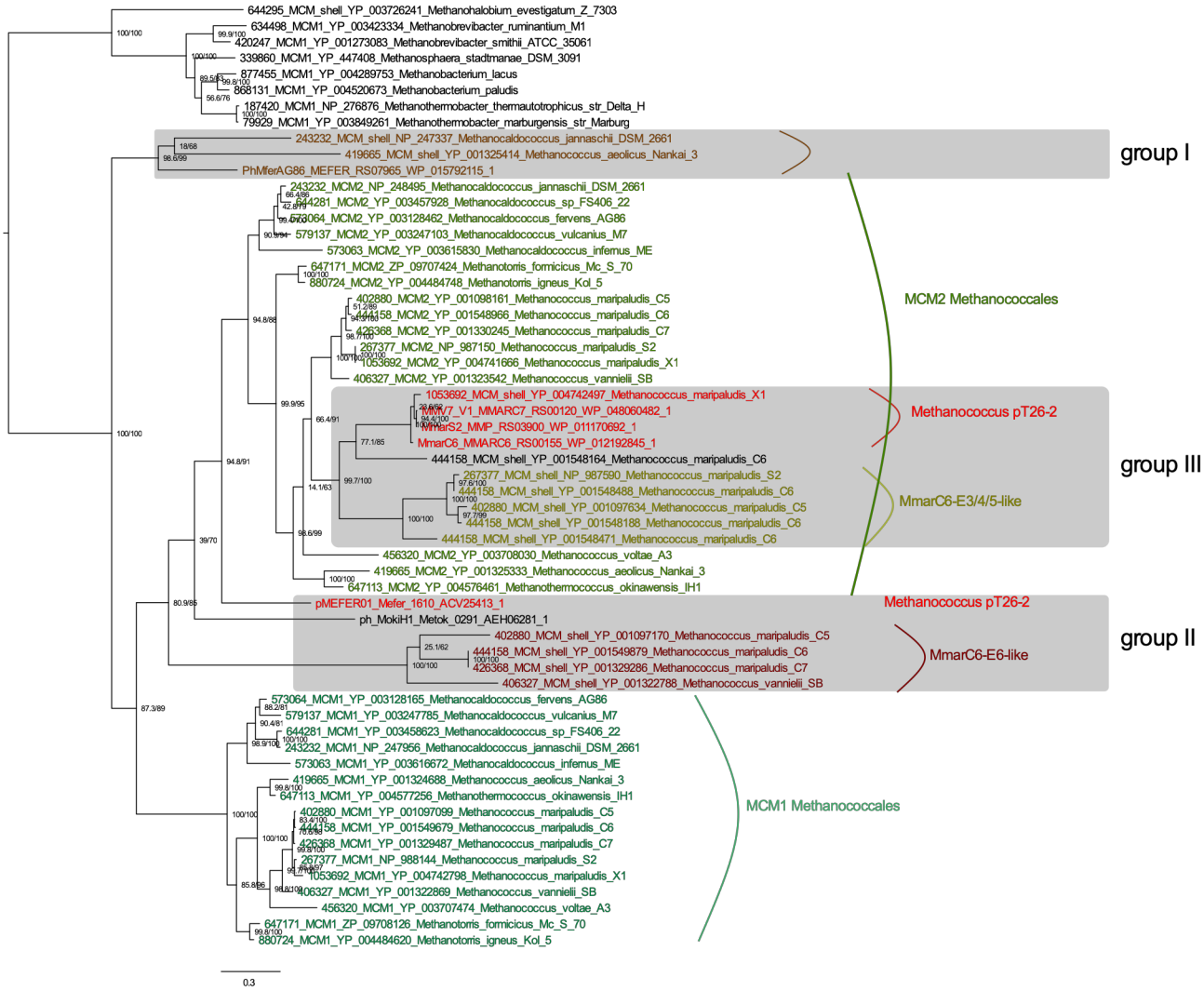
Supplementary Figure S5. Organisation of the pT26-2 putative replication protein and its comparison with other replication proteins. In this schematic representation the Primpol and the P-loop NTPase domain are indicated in green and blue respectively.



Supplementary Figure S6. MCM history within the Archaea. Maximum Likelihood phylogeny of the MCM protein within the Archaea. The Thermococcales cellular MCM are indicated in darkblue, the MCM encoded by Thermococcales MGEs are indicated in lightblue. The two Methanococcales chromosomic MCM are indicated with two different green. The Methanococcales MGEs encoded MCM are indicated in brown. The scale-bar represent the average number of substitutions per site. Values at nodes represent support calculated by aLRT and ultrafast bootstrap approximation (1,000 replicates).



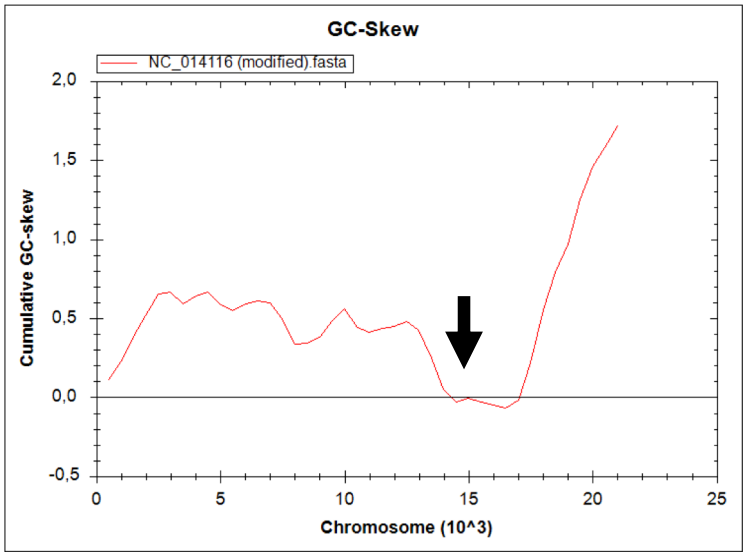
Supplementary Figure S7. MCM history within *Thermococcales.* Maximum Likelihood phylogeny of the MCM protein within Thermococcales using the *Theionarchaea* and the *Methanofastidiosa* as an out group. The Thermococcales Core MCM are indicated in dark blue, the MCM encoded by MGEs are indicated in and the purple dots correspond to pT26-2 related elements. The scale-bars represent the average number of substitutions per site. Values at nodes represent support calculated by aLRT and ultrafast bootstrap approximation (1,000 replicates).

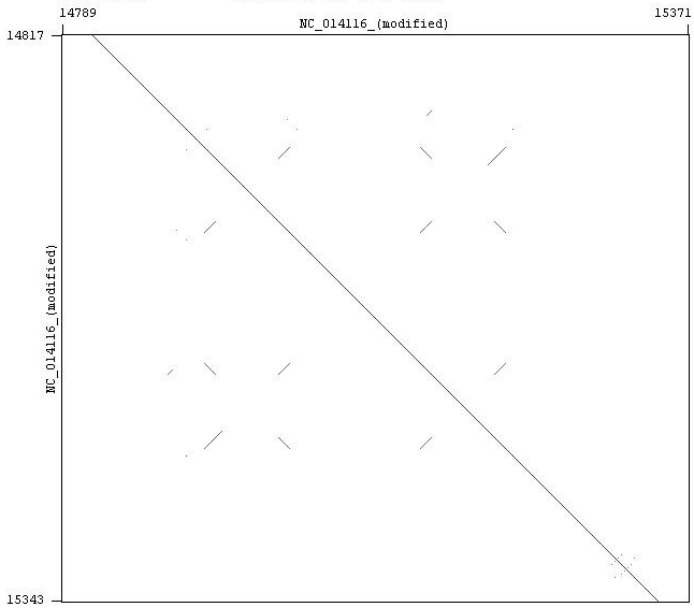


Supplementary Figure S8. MCM history within *Methanococcales.* Maximum Likelihood phylogeny of the MCM protein within Methanococcales using the Methanobacteriales as an out group. The two Methanococcales Core MCM are indicated in dark green by the names MCM1 and MCM2. the MCM encoded by MGEs are indicated with several other colours and are all within three groups highlighted in grey. The MCM encoded by Methanococcales pT26-2 related elements are indicated in red. The scale-bars represent the average number of substitutions per site. Values at nodes represent support calculated by aLRT and ultrafast bootstrap approximation (1,000 replicates).

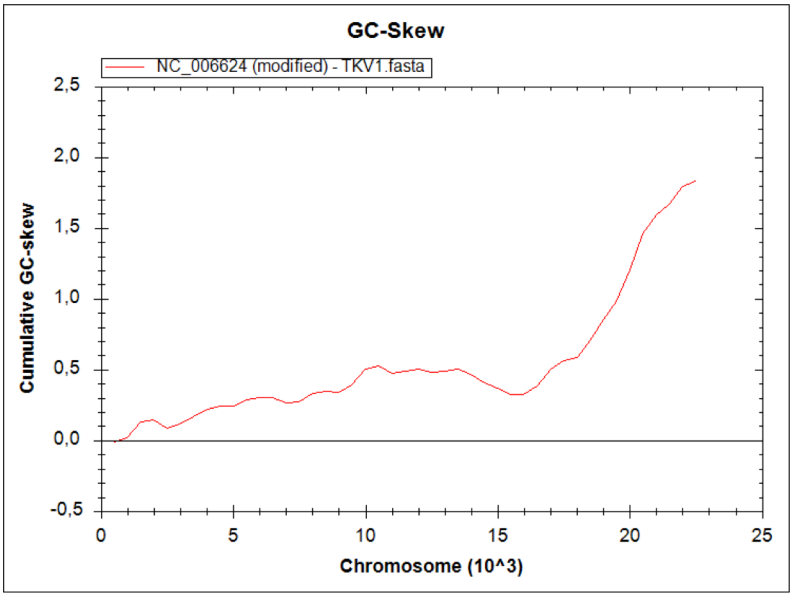
Supplementary Figure S9. Putative Replication origin prediction by GC-skew (left) and dotplot (right) in pT26-2 related elements. For integrated elements, the zero coordinate correspond to the first nucleotide of attL. Dotplots were drawn with Gepard.

pT26-2 plasmid

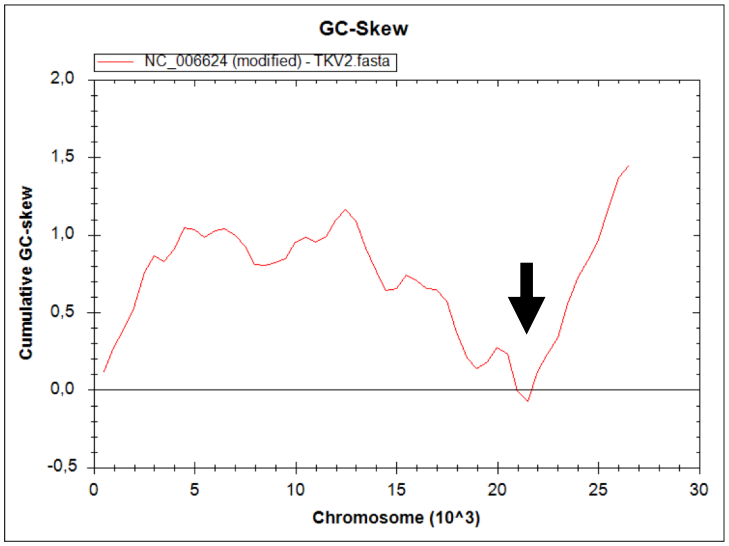


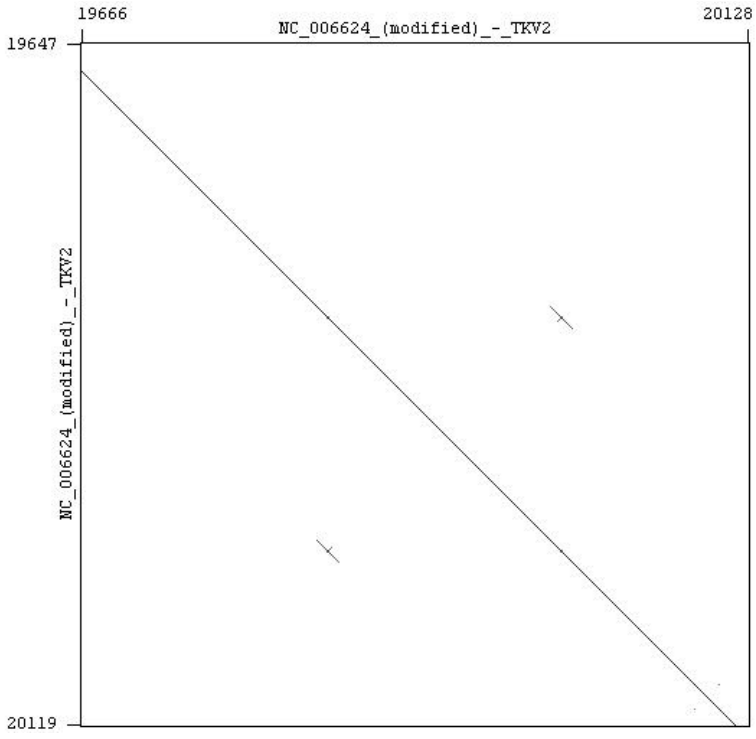


Thermococcus kodakarensis TKV1

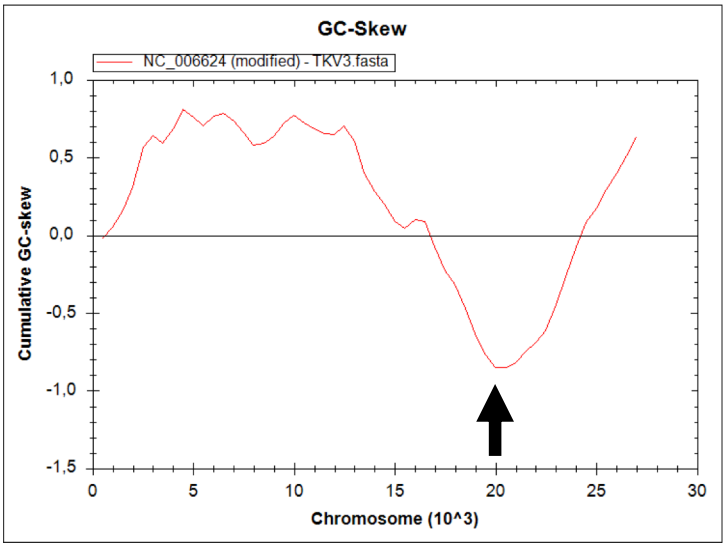


Thermococcus kodakarensis TKV2

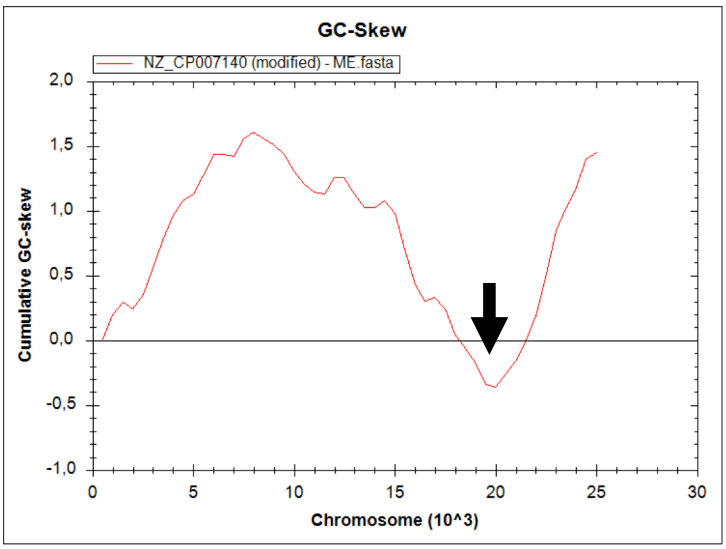




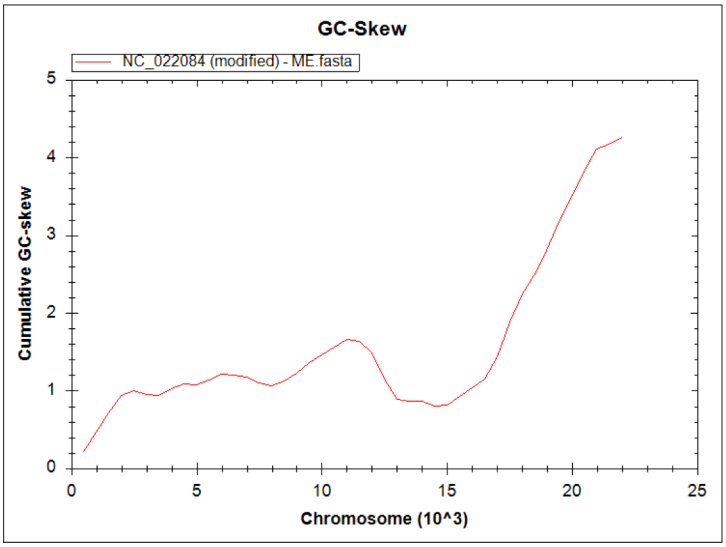
Thermococcus kodakarensis TKV3



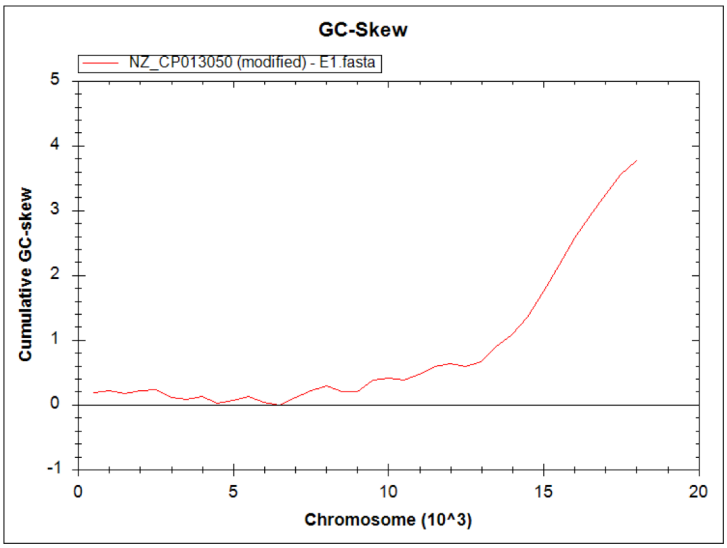
Thermococcus guaymasensis



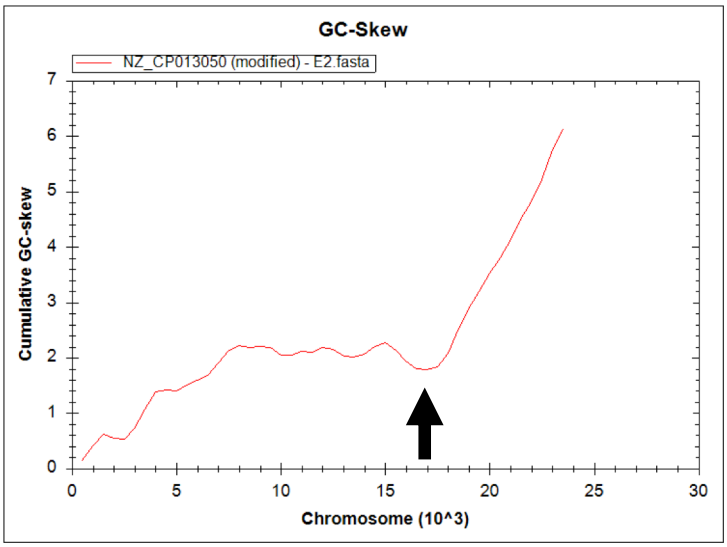
Thermococcus litoralis



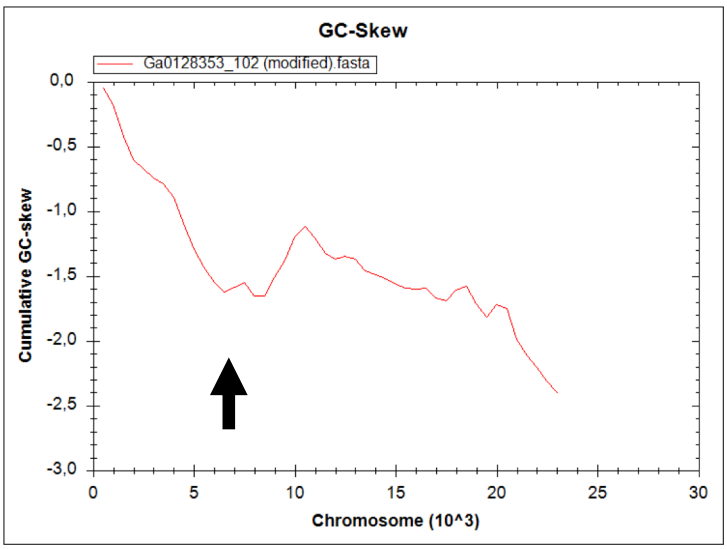
Thermococcus barophilus E1

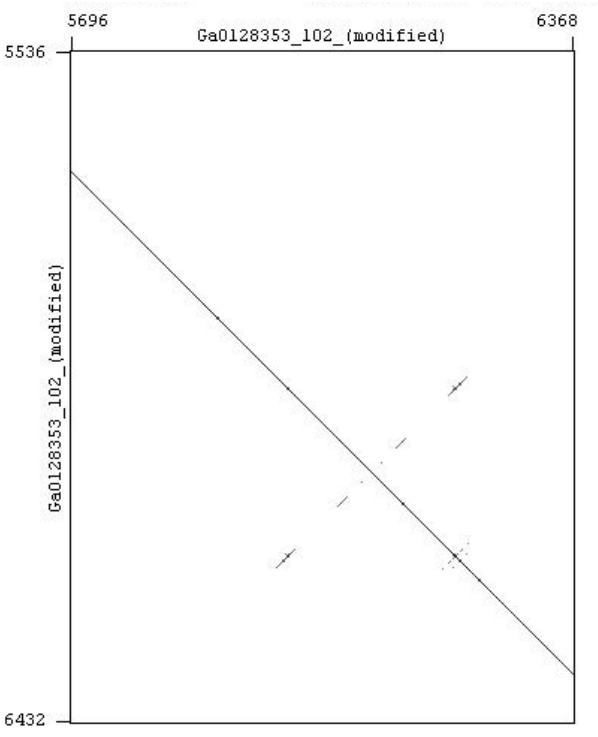


Thermococcus barophilus E2

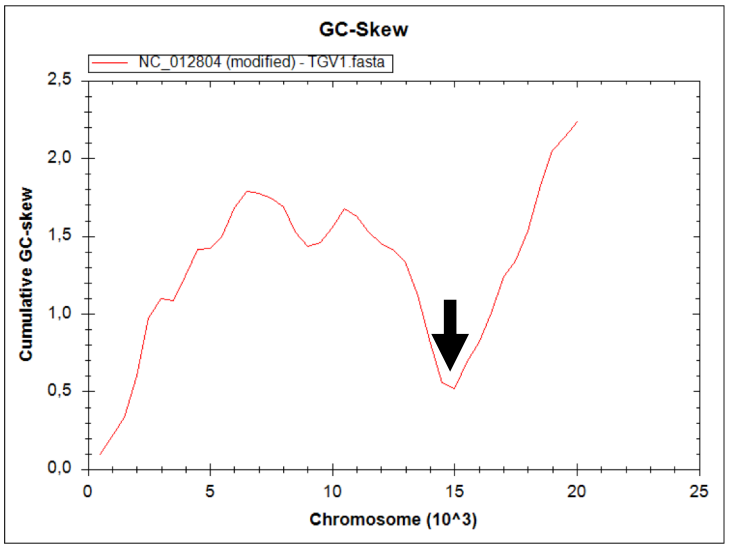


Thermococcus sp. JCM11816

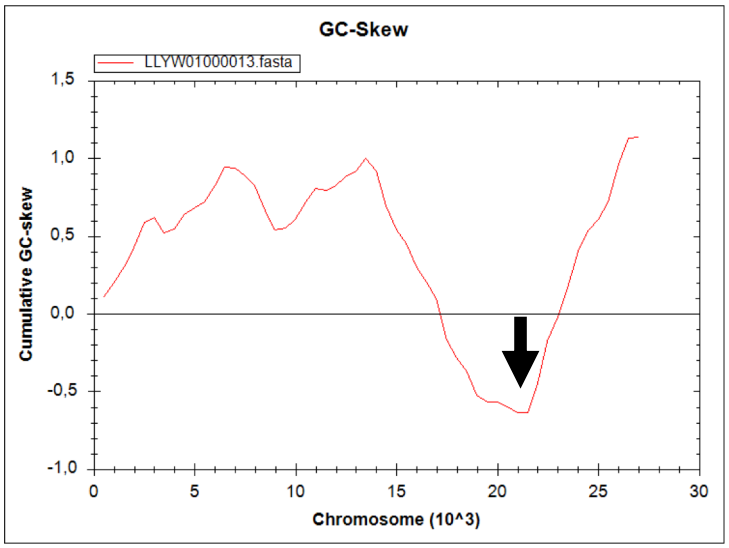


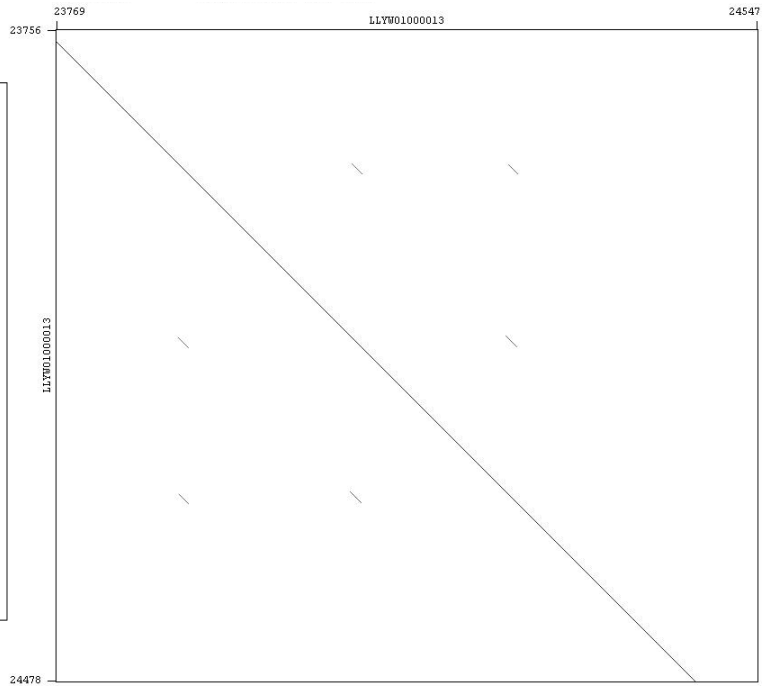


Thermococcus gammatolerans TGV1

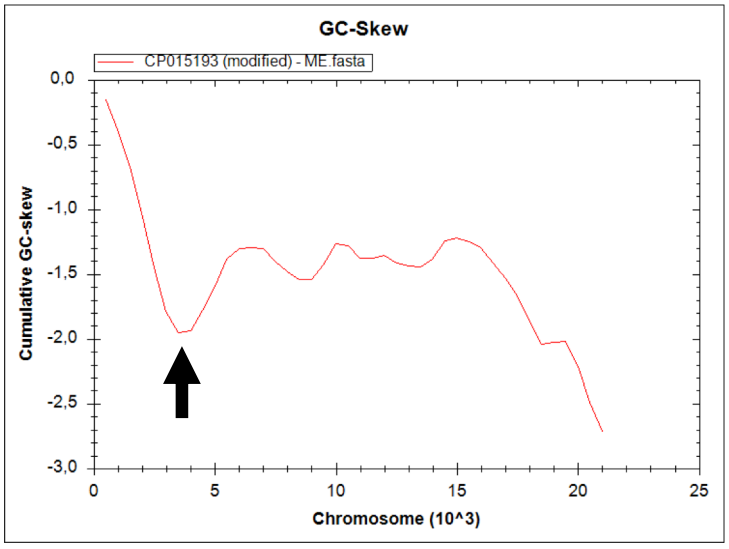


Thermococcus celericrescens

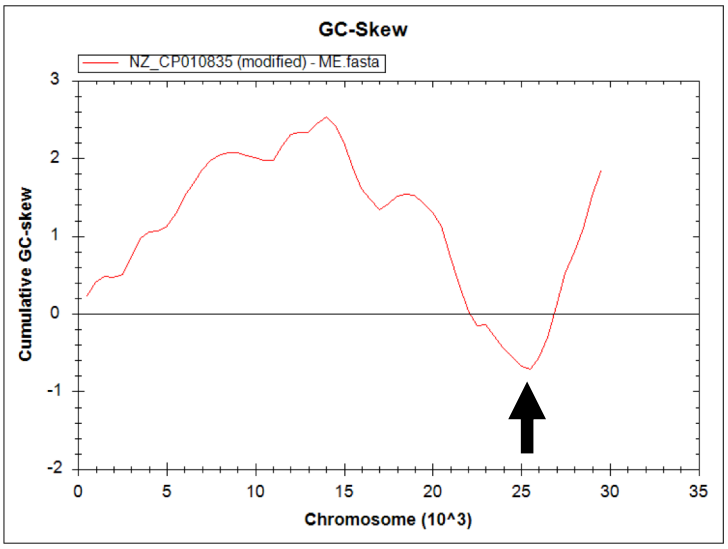




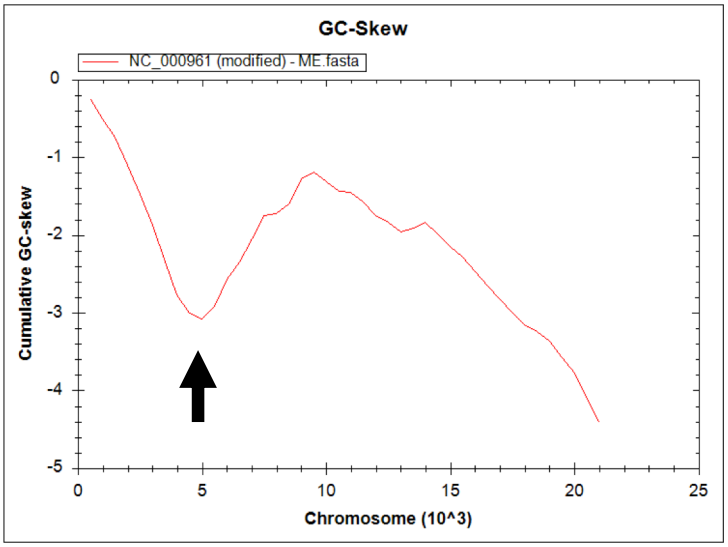
Pyrococcus chitonophagus



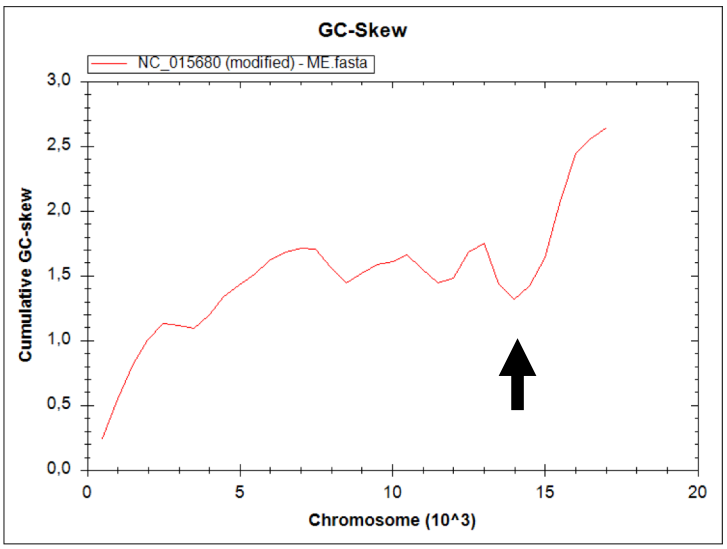
Pyrococcus kukulkanii

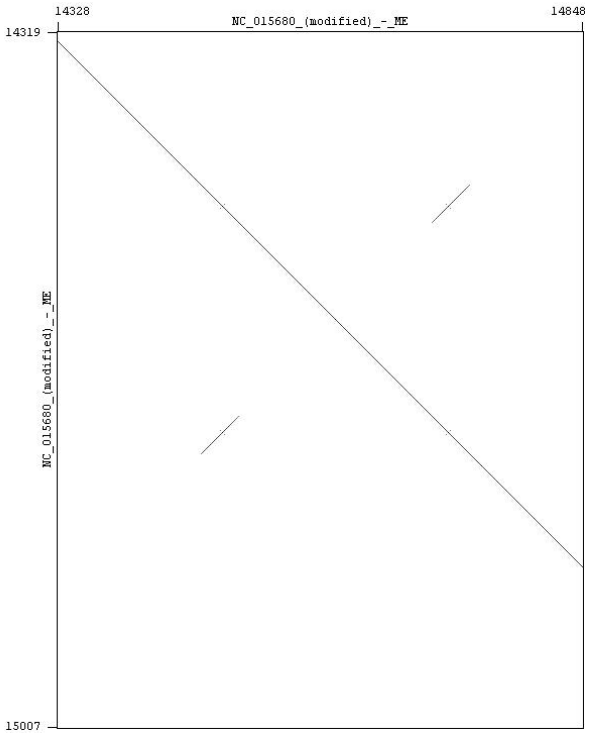


Pyrococcus korikoshii

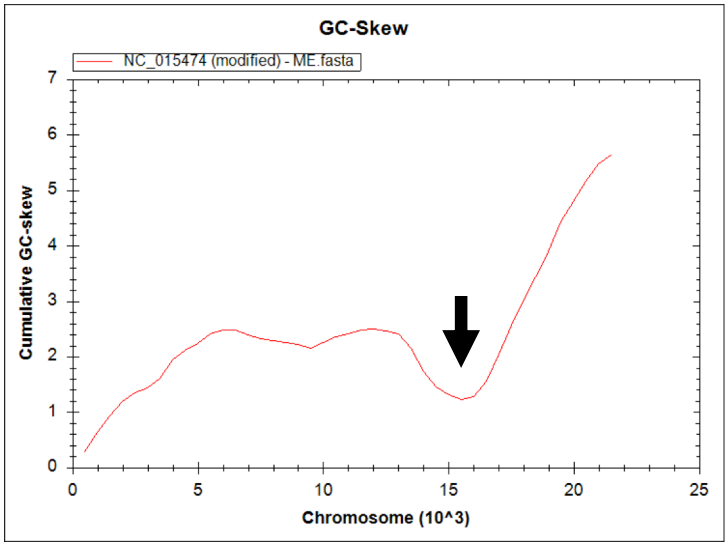


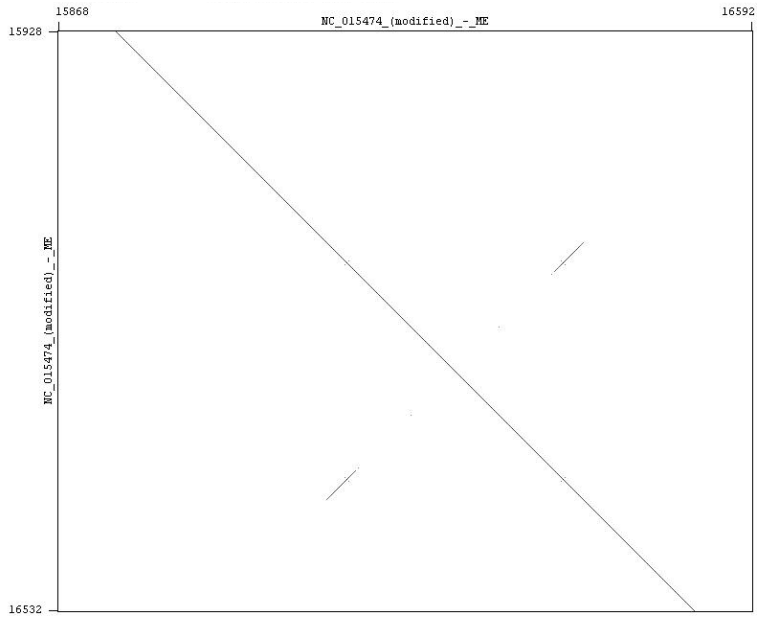
Pyrococcus kukulkanii



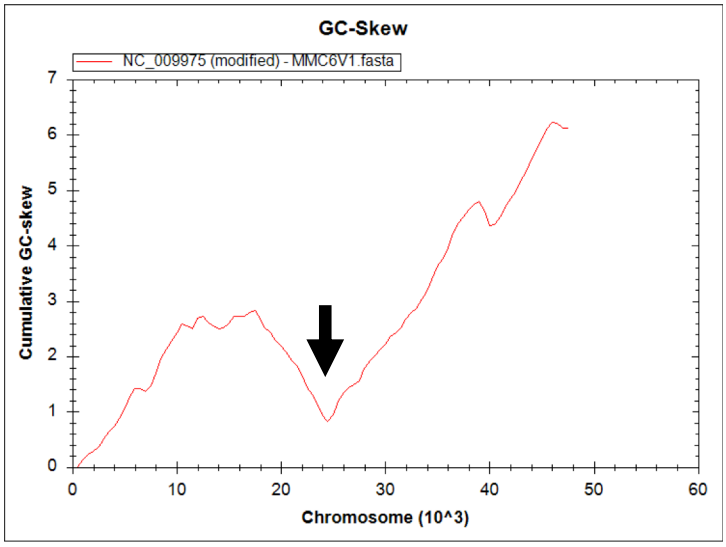


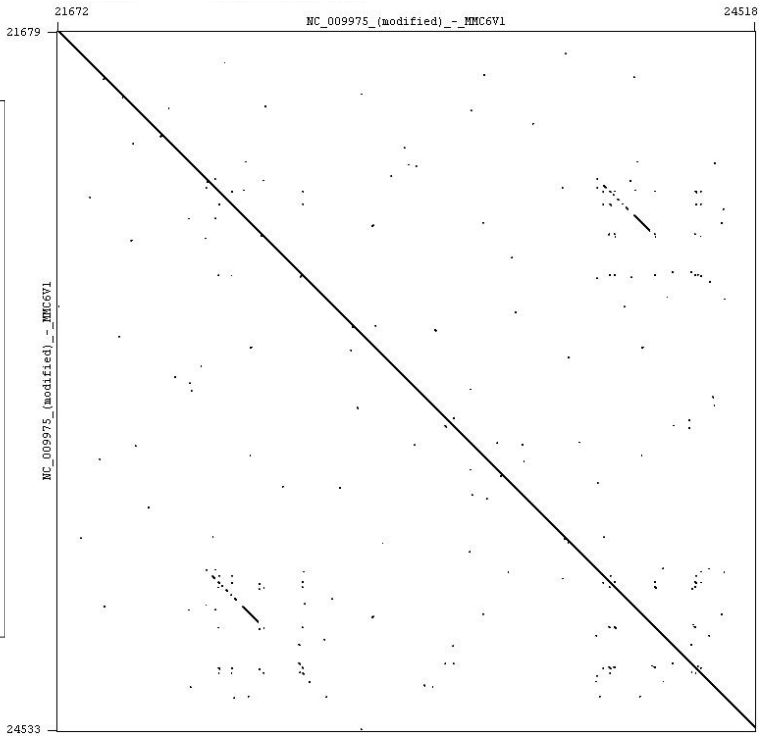
Pyrococcus sp. NA2



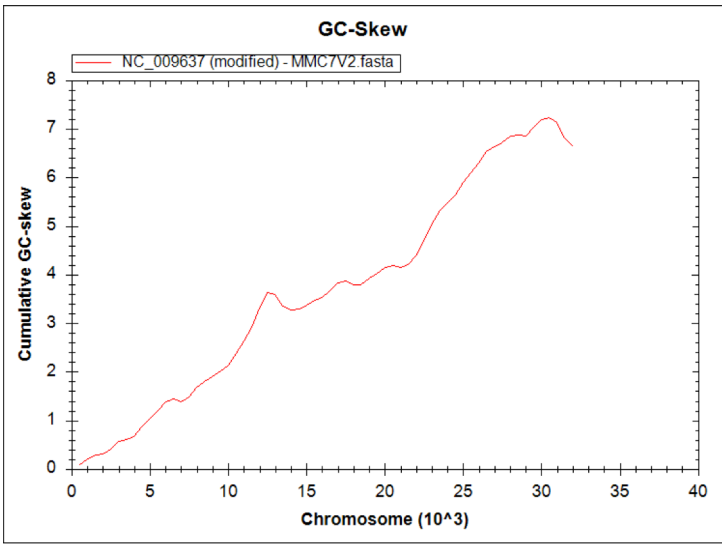


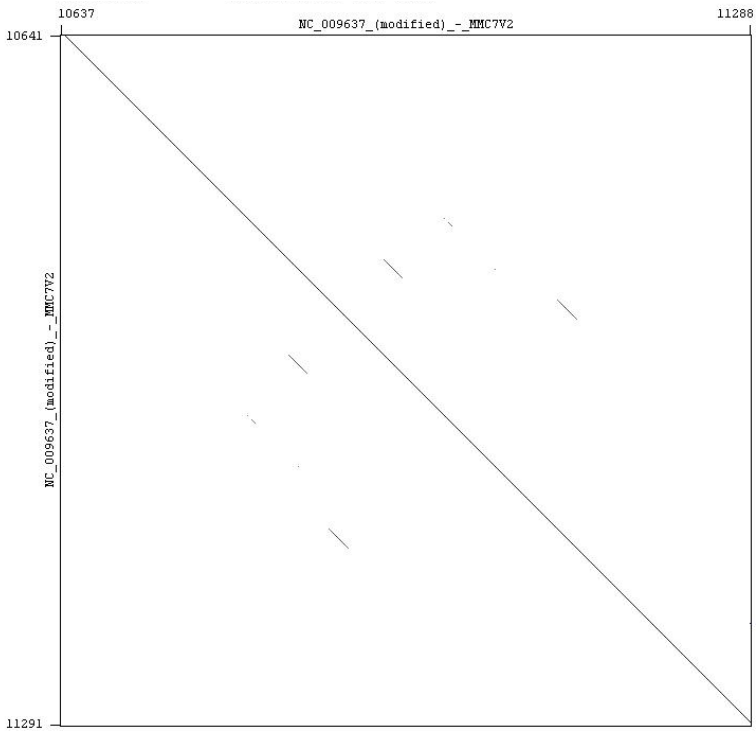
Methanococcus maripaludis MMC6V1



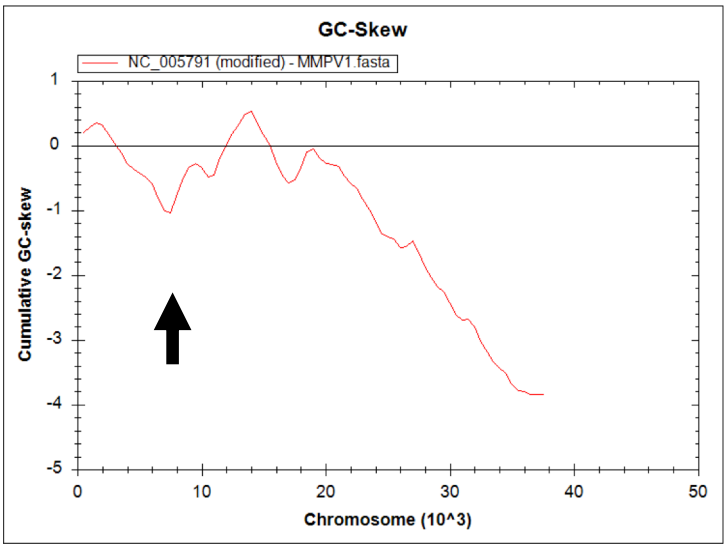


Methanococcus maripaludis MMC7V2

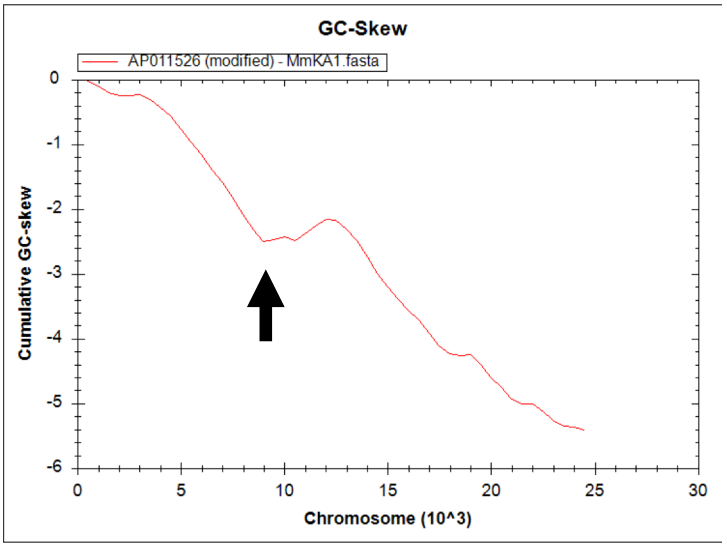


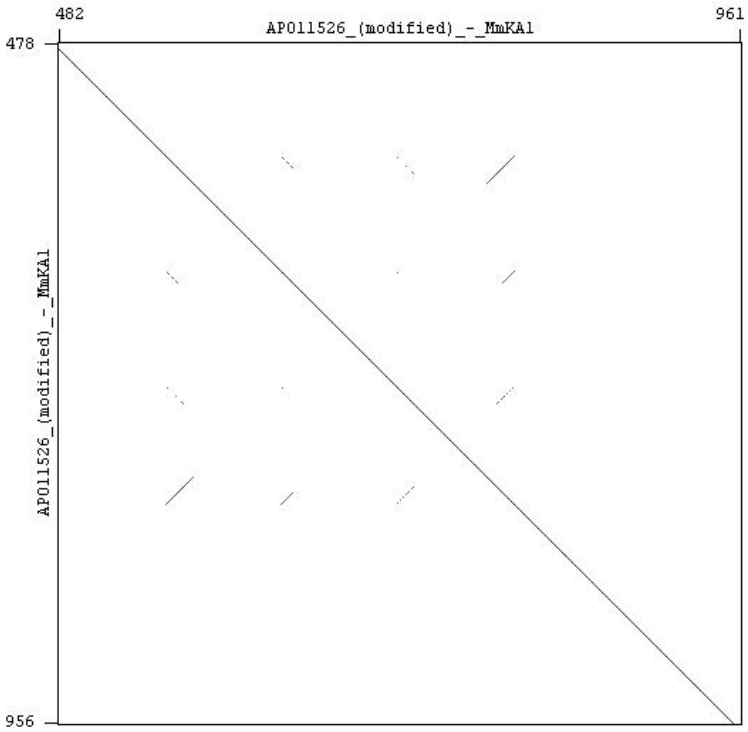


Methanococcus maripaludis S2 MMPV1

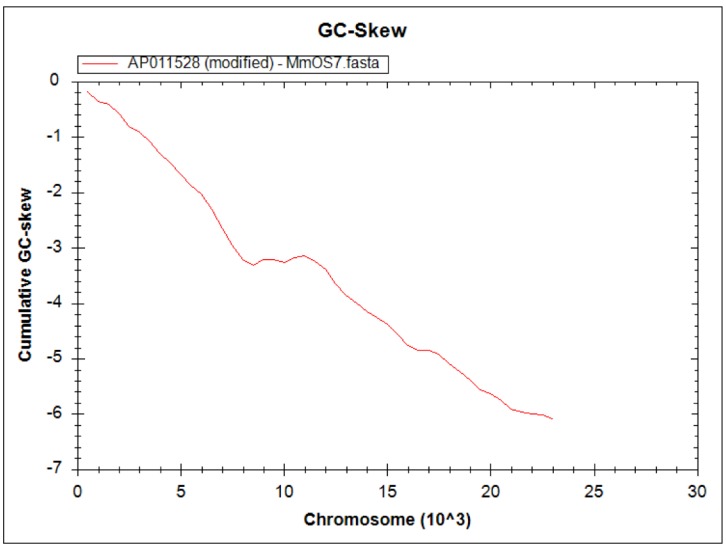


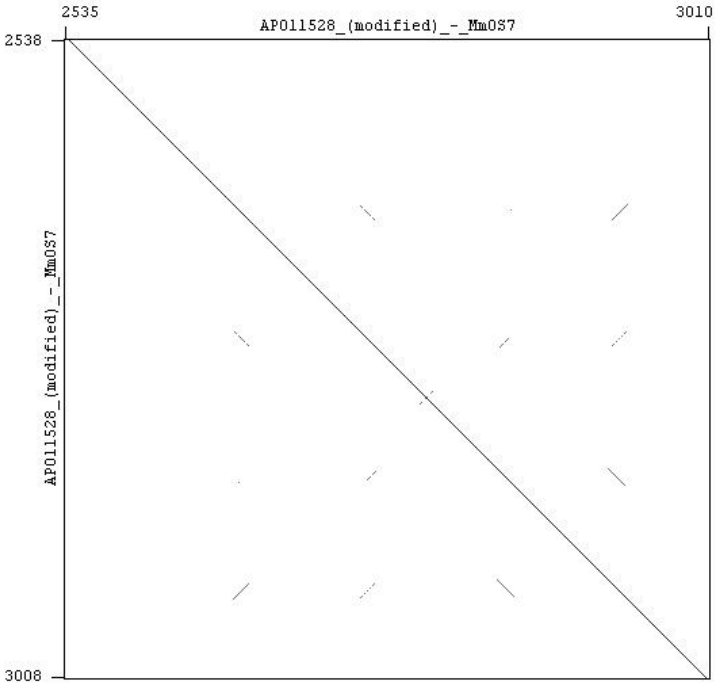
Methanococcus maripaludis KA1



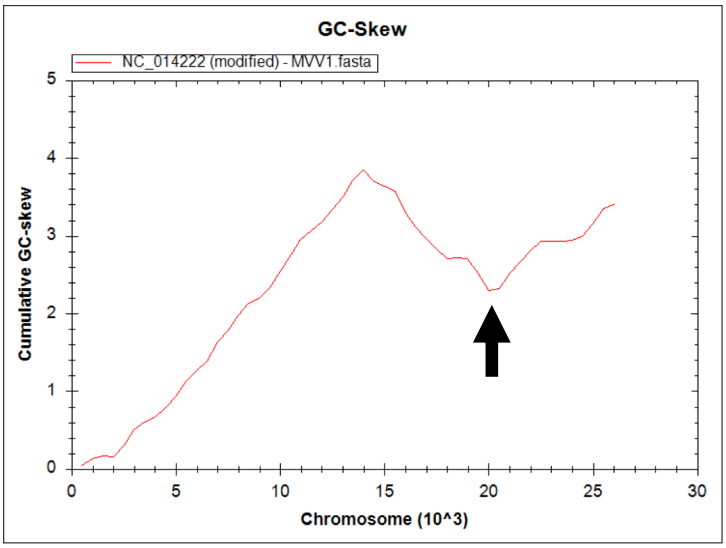


Methanococcus maripaludis OS7

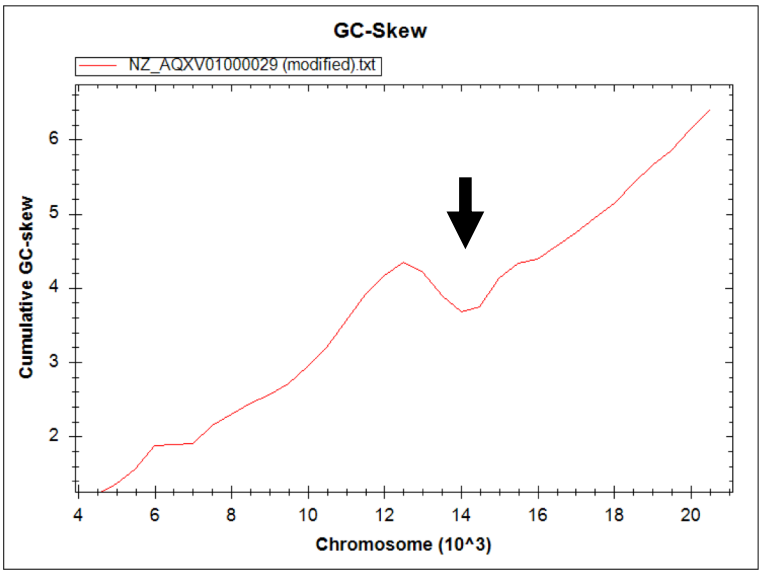


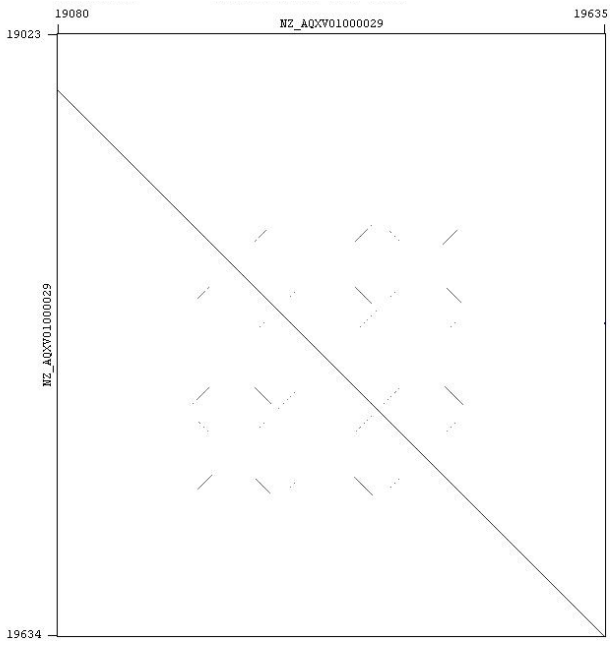


Methanococcus voltae MVV1

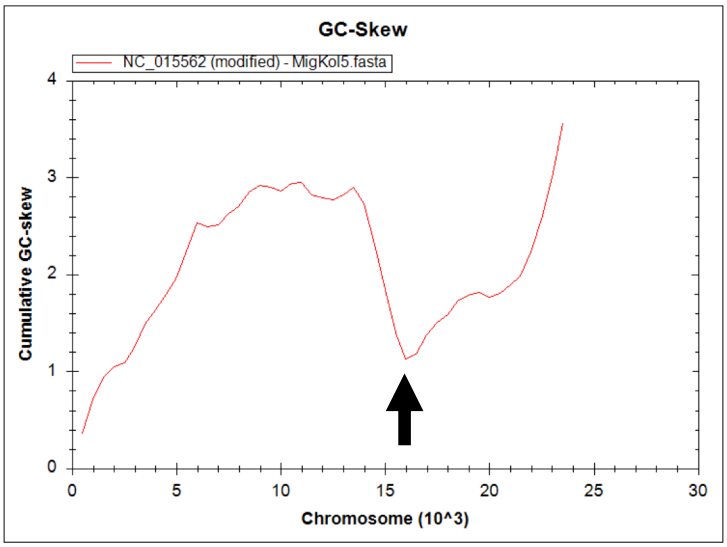


Methanothermococcus thermolithotrophicus

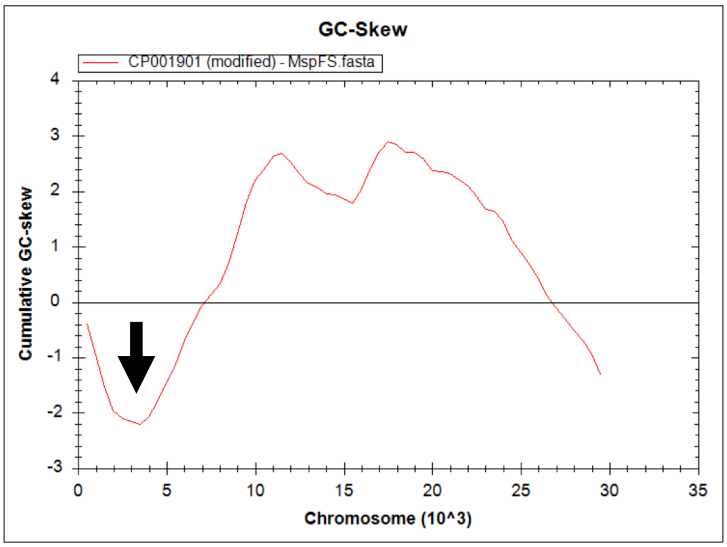


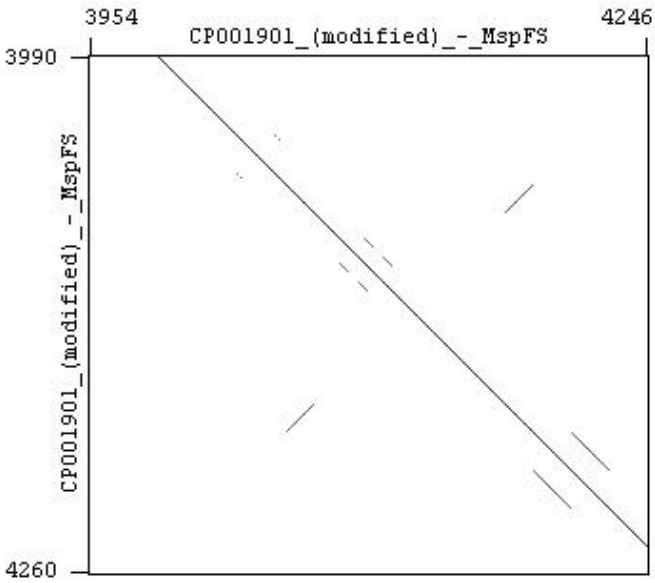


Methanotorris igneus

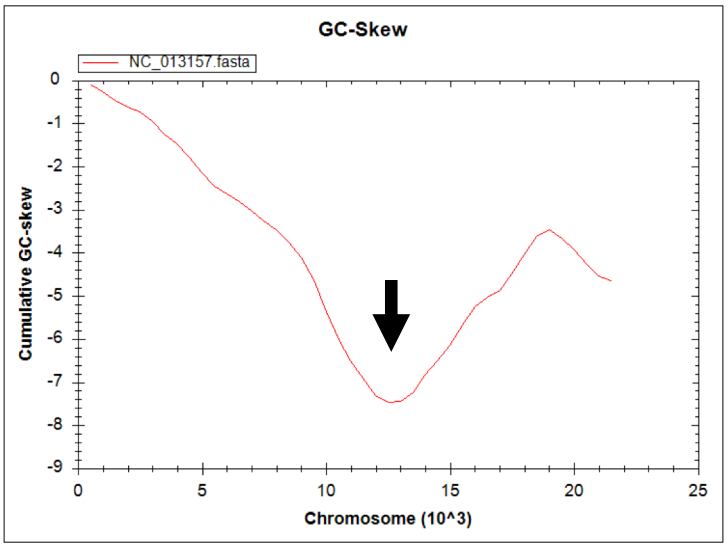


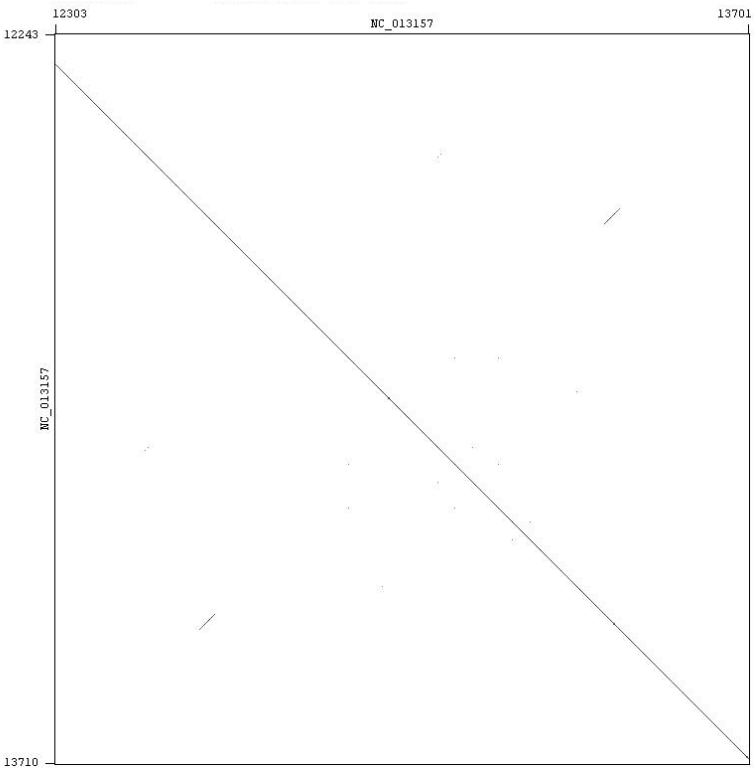
Methanocaldococcus sp. FS406-22

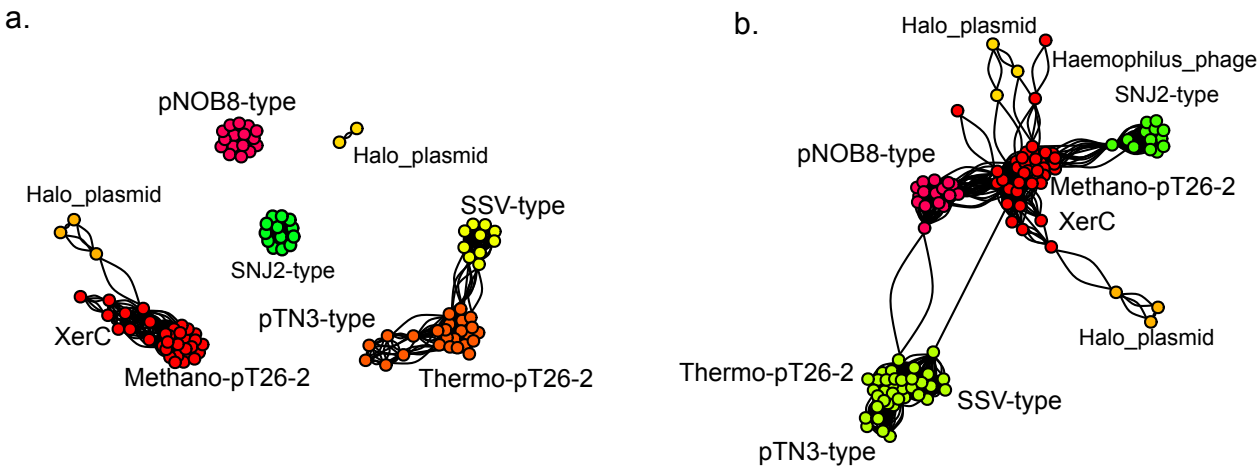




pMEFER01

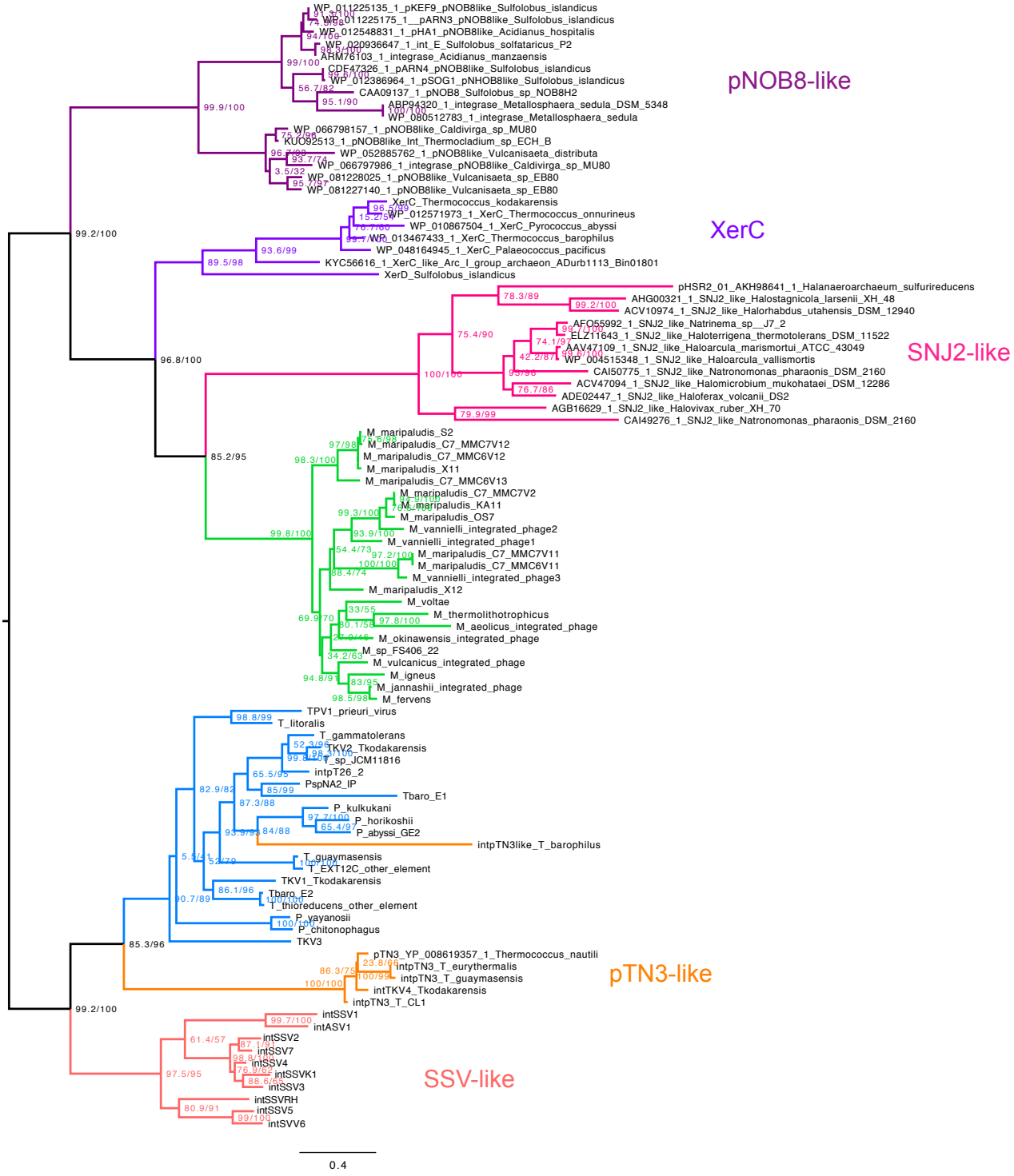




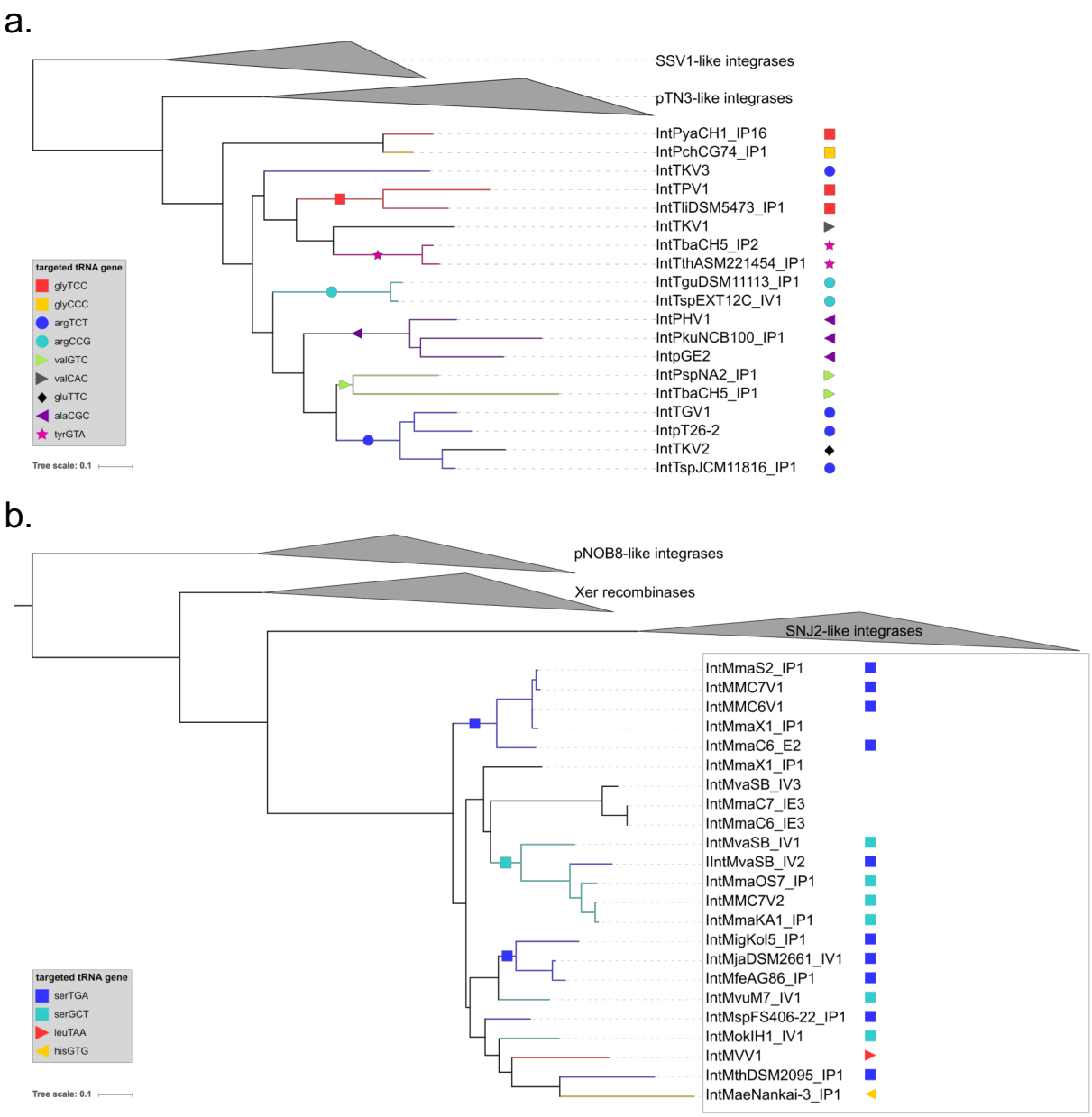


Supplementary Figure S10. Integrase similarity network view and corresponding family attribution.

The results of all integrases versus all integrase protein by BlastP (expect >0.001) are represented as a network with two additional different criteria in the limits of similarity in **a** the similarity is >25% among 65% of the protein, and in **b** >25% among 40% of the protein.



Supplementary Figure S11. Two integrases family encoded by pT26-2 MGEs
Maximum Likelihood phylogeny of thepT26-2 encoded integrases protein using several other known integrases family as an outgroup.



Supplementary Figure S12. Integrase phylogeny and targeted tRNA gene for *Thermococcales* (top) and *Methanococcales* (bottom).
Maximum Likelihood tree of integrases proteins (see uncollapsed version in supplementary fig. S11). The tRNA gene in which the att site is located is indicated by coloured symbols. tRNA corresponding to the same amino-acid but with different anti-codon are indicated by the same symbol but varying colour. Thermococcales integrases target 9 different tRNA genes while Methanococcales integrases target 4 different tRNA genes. Most parsimonious ancestral tRNA gene target is indicated when possible. Integrase homologs found in integrated element that differ from pT26-2 are included.

Host	element name	ori localisation	methodology
Themococcus sp 26-2	pT26-2	between t26-20p and t26-21p	GC-skew and dotplot
Thermococcus kodakarensis KOD1	TKV1	**Not found**	
Thermococcus kodakarensis KOD1	TKV2	from TK_RS02005 to TK_RS02015	GC-skew and dotplot
Thermococcus kodakarensis KOD1	TKV3	inside TK_RS02910	GC-skew
Thermococcus guayamensis DSM11113		from X802_RS00960 to X802_RS00965	GC-skew
Thermococcus litoralis DSM 5473		**Not found**	
Thermococcus barophilus CH5	E1	**Not found**	
Thermococcus barophilus CH5	E2	from TBCH5V1_RS11215 to TBCH5V1_RS11225	GC-skew
Thermococcus sp. JCM11816 *CONTIG 00002*		between Ga0128353_102299 and Ga0128353_102300	GC-skew and dotplot
Thermococcus gammatolerans EJ3	TGV1	from TGAM_RS03400 to TGAM_RS03405	GC-skew
Thermococcus celericrescens DSM17994 CONTIG 013		from the end of APY94_04115 to APY94_04120 and downstream	GC-skew and dotplot
Pyrococcus chitoniphagus GC74		between A3L04_06365 and A3L04_06370	GC-skew
Pyrorcoccus kukulkanii sp. NCB100		between TQ32_RS02635 and TQ32_RS02645	GC-skew
Pyrococcus horikoshii OT3	PHV1	between PH_RS05530 and PH_RS05540	GC-skew
Pyrococcus yayanosii CH1		between PYCH_RS07010 and PYCH_RS07015	GC-skew and dotplot
Pyrococcus sp. NA2		immediatly downstream of PNA2_RS06700	GC-skew and dotplot
Pyrococcus abyssi GE2	pGE2	In the intergenic region two ORFs upstream the putative replication protein	GC-skew and dotplot
Methanococcus maripaludis C6	MMC6V1	between MMARC6_RS00125 and MMARC6_RS00130	GC-skew and dotplot
Methanococcus maripaludis C7	MMC7V1	**Not found**	
Methanococcus maripaludis C7	MMC7V2	from MMARC7_RS07665 to MMARC7_RS07670	dotplot
Methanococcus maripaludis S2	MMPV1	between MMP_RS03905 and MMP_RS08880	GC-skew
Methanococcus maripaludis KA1		Dotplot: between tRNAser and MMKA1_04820 (or GC-skew between MMKA1_04910 and MMKA1_04970)	
Methanococcus maripaludis OS7		between MMOS7_04780 and MMOS7_04790	dotplot
Methanococcus maripaludis C5		**Not found**	
Methanococcus maripaludis X1		**Not found**	
Methanococcus voltae A3	MVV1	between MVOL_RS08870 and downstream of MVOL_RS07785	GC-skew
Methanothermococcus thermolithotrophicus		Dotplot: between F555_RS08905 and F555_RS0101670 GC skew: from F555_RS0101630 to F555_RS0101635	
Methanotorris igneus Kol5		between METIG_RS02530 and METIG_RS02535	GC-skew
Methanocaldococcus sp. FS406-22		between MFS40622_1105 and MFS40622_1106	GC-skew and dotplot
Methanocaldococcus fervens AG86	pMEFER01	between MEFER_RS08100 and MEFER_RS08105	GC-skew and dotplot

Supplementary Table 2. List of primers

Name	Sequence (5'->3')	Ref
SP-ISC913		This study
ASP-ISC913		This study
SP-pGE2-CDS7	ATGAATACCGGAGTGTTCCTGAAGC	This study
ASP-pGE2-CDS7	AACGATGGCGTAACTTACGGTAAGA	This study
SP-pGE2-CDS29	TTGCTGCGTTTAGAATTAGCTCGTT	This study
ASP-pGE2-CDS29	TGGGTTGGGAGTACACCATAAAGAA	This study
Arc344F	ACGGGGYGCAGCAGGCGCGA	Raskin et al., 1994
Uni516R	GTDTTACCGCGGCKGCTGRCA	Takai &Horikoshi 2000

Raskin L, Stromley JM, Rittmann BE, Stahl DA (1994) Group-specific 16S rRNA hybridization probes to describe natural communities of methanogens. Appl Environ Microbiol 60: 1232–1240.
Takai & Horikoshi 2000 AEM 66 : 5066-5072

Characterization of the integrases of Methanococcales

To elucidate the concomitant presence of the two types of integrases for the same pT26-2 plasmidic backbone and in the same oceanic hydrothermal environment, we decided to characterize one integrase of each type. This study is detailed in Part 3 for the Thermococcales integrases. Here, we will present the preliminary characterization of the classical Methanococcales integrases.

Classical integrases were extensively characterized in bacteria (Duyne, 2015; Escudero et al., 2016; Landy, 2015) or in eukaryotes (Sadowski, 1995). However, in archaea, only one classical integrase has been characterized *in vivo* so far: the integrase from the halophilic pleolipovirus SNJ2 (Wang et al., 2018). In addition, tyrosine recombinases XerA were characterized *in vitro* for *Thermoplasma acidophilum* (Jo et al., 2017) and for *Pyrococcus abyssi* (Cortez et al., 2010). Our identification of a family of integrases in Methanococcales presented an opportunity to characterize, for the first time, the *in vitro* activity of a classical archaeal integrase.

We identified 23 Methanococcales candidate integrases in Article 2 suitable for a biochemical analysis. They all present the consensus catalytic motif R...HxxR...Y (Figure 28) that corresponds to the one previously identified in bacteria and differs from those identified in archaea (R...KxxR...Y or R...YxxR...Y) (She et al., 2004). All the identified integrases are integrated in the host chromosome. They might be subjected to a relaxed purifying selection compared to episomal copies leading to the accumulation of deleterious mutations (Bobay et al., 2014). We therefore decided to characterize two integrases to maximize the chance of *in vitro* activity. We chose the two integrases from *Methanocaldococcus* hosts, which have the growth conditions closest to Thermococcales, i.e. the integrase identified in *Methanocaldococcus fervens* and the integrase identified in *Methanocaldococcus sp.* 406-22.

We constructed two *E. coli* expression vectors to overproduce each integrase. We then purified the integrases using affinity chromatography and size exclusion chromatography (cf. Material and methods page 291). In the size exclusion chromatography, both integrases predominantly behaved as monomers (Figure 29). The integrase Cre is also monomeric in absence of DNA (Ghosh et al., 2007) and the tyrosine recombinase XerA from *P. abyssi* is monomeric at low protein concentrations (Serre et al., 2013). On the contrary, the bacterial HP1 integrase and the archaeal SSV2 integrase are tetrameric in solution in absence of DNA (Hakimi and Scocca, 1996; Zhan et al., 2015). The integrase oligomerization in absence of DNA seems to be variable among classical integrases and among hyperthermophilic archaeal integrases.

The recombinational activity of the two purified integrases was assayed *in vitro* on a plasmid substrate presenting one copy of the attachment site (Figure 30.C). The attachment site of the two integrases corresponds to the 3' half of $tRNA^{Arg}$ genes and varies by 2 nucleotides (Figure 30.A-B). For the integrase from *Methanocaldococcus fervens*, the activity assay was not conclusive and reaction conditions should be further optimized (data not shown). The integrase from *Methanocaldococcus sp. FS406-22* catalyzed the formation of open circular plasmid dimer at 65°C (Figure 30.D). It is now necessary to control the activity on non-specific substrates and of a catalytic tyrosine mutant. Nevertheless, this integrase seems to be active *in vitro* and is a good candidate for further analysis. It would for example be interesting to test the other directionalities of reaction to determine whether the integrase can catalyze all three reactions without additional co-factors.

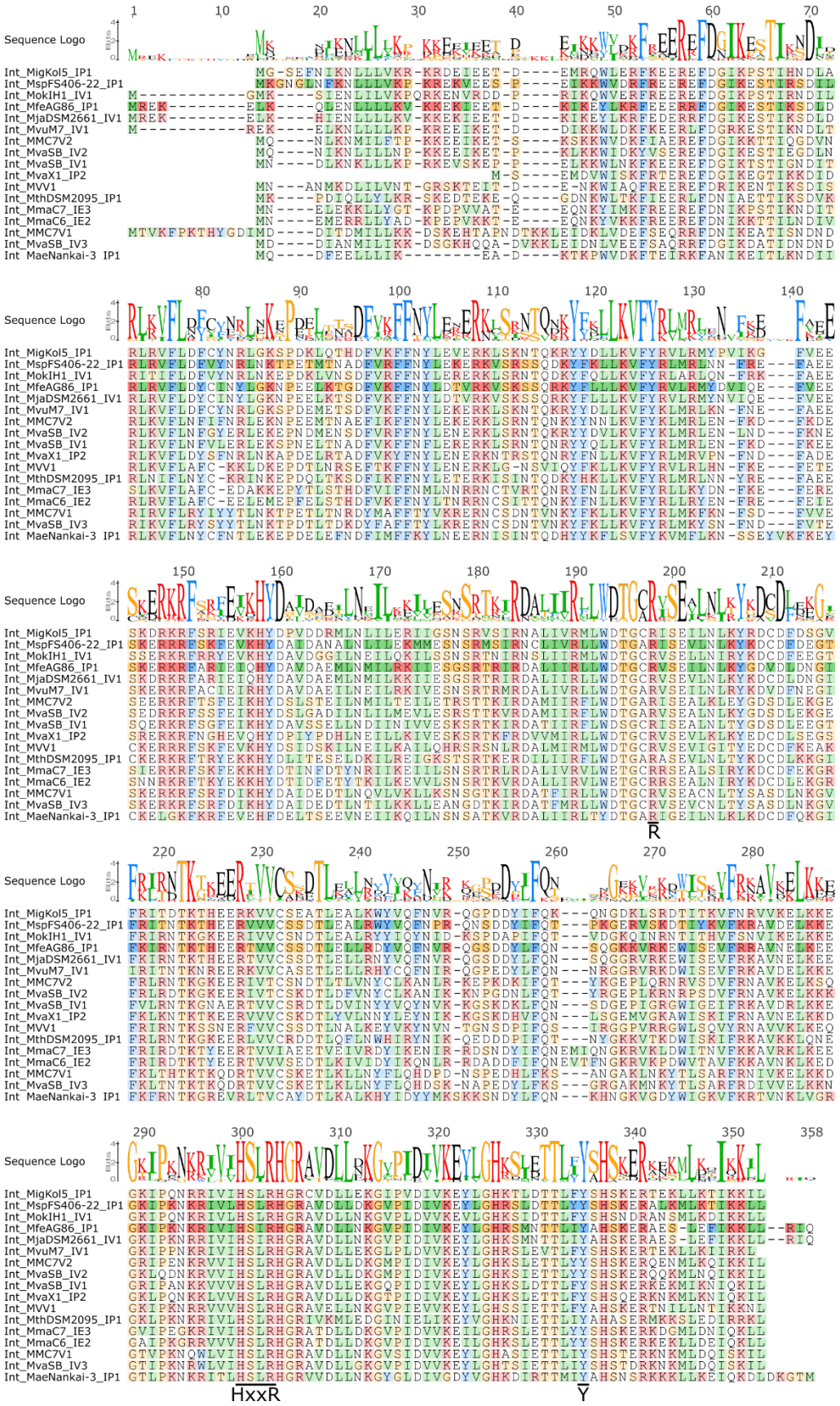


Figure 28. MAFFT alignment of the integrases identified in pT26-2 like elements of Methanococcales. The catalytic motif R…HxxR…Y is indicated. Several redundant integrases from Methanococcus maripaludis strains were removed from the alignment.

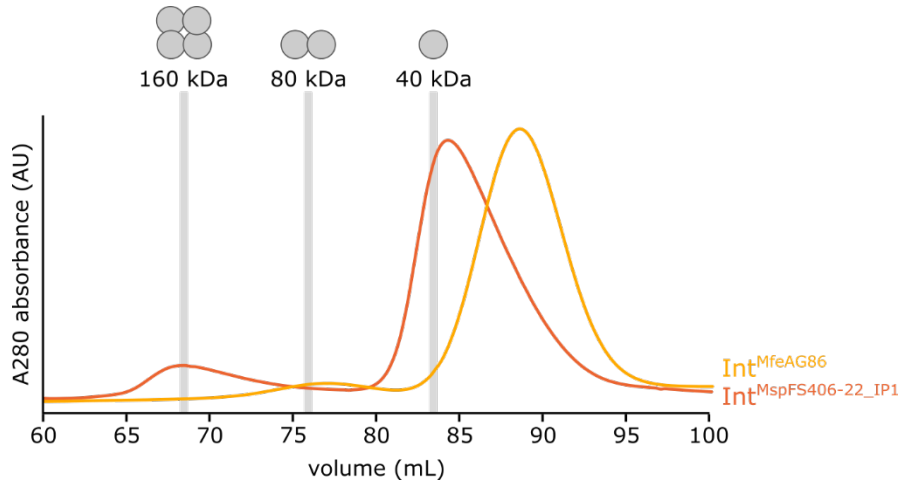


Figure 29. Size exclusion chromatography profile for the purification of the integrases of *Methanocaldococcus sp. 406-22* and *Methanocaldococcus fervens* AG86. Protein extract were loaded on a HiLoad 16/600 Superdex 200 size exclusion column. The buffer was 1M KCl, 40 mM Tris-HCl pH=8, 5 mM B-mercaptoethanol and 10% glycerol. The elution volumes for the different oligomer were calculated using a calibration curve obtained in another buffer and are indicated on the plot.

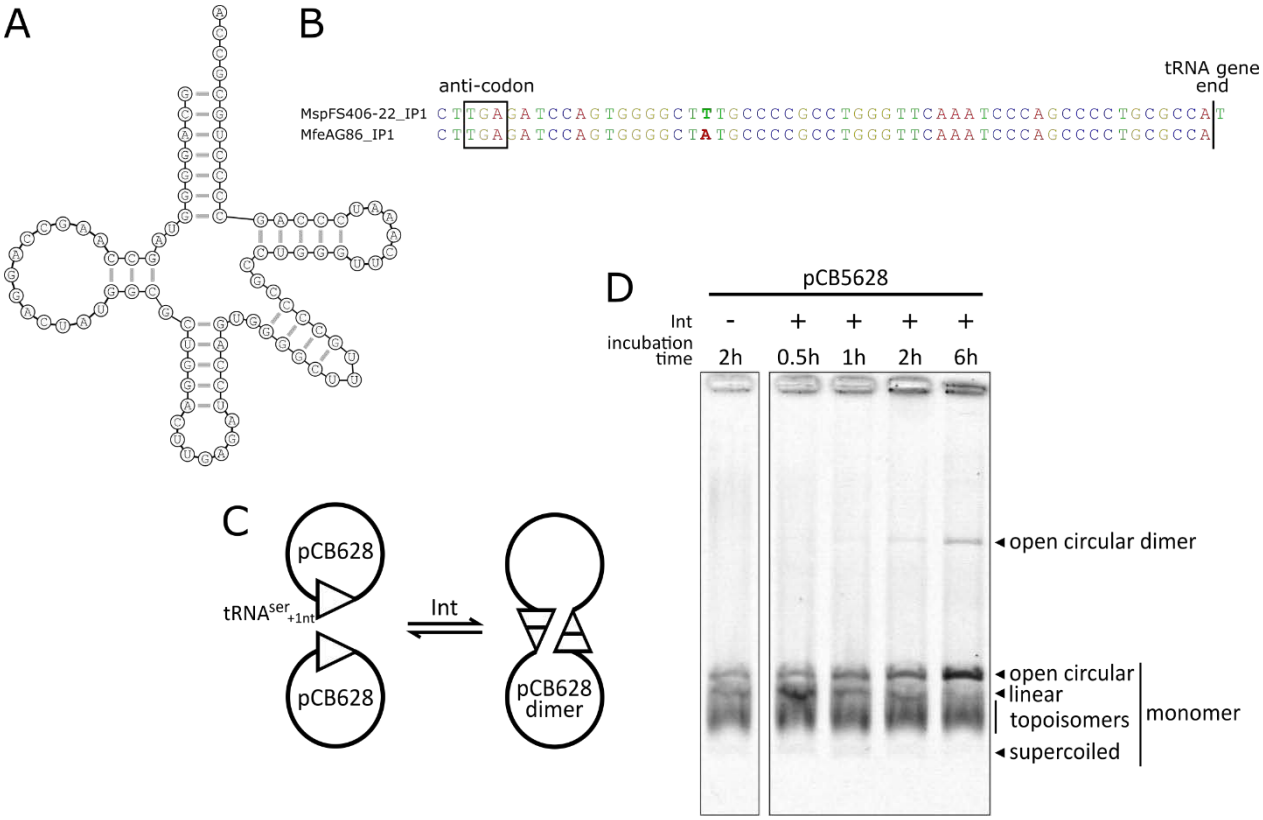


Figure 30. The integrase from the integrated element MspFS406-22_IP1 can catalyze site-specific recombination. A. Structure of the tRNA^{Ser} whose gene is targeted for the integration. B. Comparison of the attachement site with the one of the integrated element MfeAG86_IP1. C Recombination between two att sites (triangles) carried by two identical plasmids pCB628 produce plasmid dimers. D. Plasmid pCB628 was incubated with the purified Integrase at 65°C, treated with proteinase K and separated on a gel.

Part 3. The activity and evolution of Thermococcales suicidal integrases

Article 3. Pervasive suicidal integrases in archaea

To elucidate the concomitant presence of two integrase types for the same pT26-2 plasmidic backbone and in the same oceanic hydrothermal environment, we decided to characterize one integrase of each type. The analysis is presented above for the classical Methanococcales integrases. Here, we present the characterization of suicidal Thermococcales integrases. Among the candidates integrases carried by pT26-2-like elements, two are carried by a free mobile elements and therefore more likely to be active: the integrase Int^{pT26-2} from plasmid pT26-2 and the integrase from plasmid pGE2. We decided to purify and characterize the integrase from the type plasmid pT26-2. We also analyzed the diversity and evolutionary history of a dataset of Int^{pT26-2} related integrases.

In this work, we evidenced biochemical activities and evolutionary patterns that concurred to explain the sustainability of the suicidal activity. Firstly, the integrase Int^{pT26-2} can catalyze all three direction of recombination *in vitro* in the absence of additional protein co-factor, as the integrase Int^{pTN3}. It seems that suicidal integrase avoid the utilization of protein recombination directionality factor (RDF) by controlling the directionality of reaction through suicidal integration. The integration module is consequently very compact with the integrase gene coding for the recombinase, the control of the reaction directionality and the recombination site. It can therefore easily be acquired by mobile elements. Secondly, we uncovered a pattern of protein evolution where integrated mobile genetic elements exchange integrase moieties. Gene fragmentation upon integration can therefore be considered as an evolutionary strategy facilitating modular protein evolution and specificity exchange.

The previously characterized integrase Int^{pTN3} is related to Int^{pT26-2}. Both integrate *Thermococcus* host that inhabit hyperthermophilic hydrothermal environments. Int^{pTN3} presents a second catalytic activity of homologous recombination that was not identified for Int^{pT26-2}. The two integrases are active at high temperature and Int^{pT26-2} optimal temperature is even higher than Int^{pTN3}. This excludes high temperature as a cause for Int^{pTN3} dual activity. Moreover, Int^{pTN3} modeled structure exhibits supplementary loops that are absent from Int^{pT26-2}. Int^{pT26-2} can accordingly be viewed as a natural deletion mutant of the supplementary loops, suggesting their importance for the second catalytic activity.

Pervasive suicidal integrases in hyperthermophilic archaea

Catherine BADEL[1], Violette DA CUNHA[1], Patrick FORTERRE[1,2] & Jacques OBERTO[1]

[1]Institute for Integrative Biology of the Cell (I2BC), Microbiology Department, CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette cedex, France

[2]Institut Pasteur, Unité de Biologie Moléculaire du Gène chez les Extrêmophiles, Département de Microbiologie, 25 rue du Docteur Roux, 75015 Paris, France

Abstract

Mobile genetic elements are modular and evolve by recruiting modules advantageous for their survival and propagation. Integration modules are particularly beneficial as they allow addition of MGE genomes as part of their host chromosome. The integration module for the Thermococcales members of the pT26-2 plasmid family comprises only a single gene encoding a site-specific tyrosine recombinase of the SSV-family. These integrases carry a recombination site within their open reading frame and have been termed suicidal due to the disruption of their gene upon MGE integration in the host genome. The suicidal nature of these enzymes is in sharp contrast with the high prevalence and multiples recruitments of these integrases in hyperthermophilic archaea. In this work, we have investigated the biochemical properties of Int^{pT26-2}, the prominent member of this family of integrases and observed its unprecedented site-specific recombination activity at near-boiling water temperature. The reconstruction of the evolution history of this family of integrases revealed their capacity to efficiently exchange recombination specificity and allowed to propose a model explaining the mechanisms of target site switching. The biochemical, genomic and phylogenetic data presented here concurred in explaining the prevalence and pervasiveness of this family of integrases in the most hyperthermophilic living organisms.

Introduction

The maintenance and propagation of mobile genetic elements (MGEs) such as plasmids and viruses impose the infection of a suitable cellular host and the deployment of appropriate strategies (1). Brute force mechanisms such as high copy number grant MGE inheritance into daughter cells after cell division (2). More refined toxin-antitoxin systems ensure MGE maintenance by relentlessly killing hosts trying to eliminate them (3). These mechanisms prove a burden for the host cells which develop effective countermeasures such as CRISPR or restriction modification systems (4,5). Alternatively, to favor their maintenance, MGEs alleviate their physiological cost for the host (6). For example, an efficient MGE partitioning allows propagation with a low copy number (7,8). Some MGEs even carry functions that present an advantage for the host such as resistance genes that increase the fitness of the symbiont in the presence of antibiotics (6). Contrastingly, some MGEs have adopted a different potent survival strategy. They have acquired the capacity to integrate their DNA at a particular location

of the cellular chromosome without overly altering both genetic programs, using a mechanism known as site-specific recombination (9-11). By disguising their genome as part of the host chromosome, these MGEs succeed in lowering their negative impact on the host metabolism and in bypassing defense mechanisms. This improved cellular acceptance ensures MGE maintenance and vertical propagation. The reverse reaction of excision regenerates the MGE in its independently replicating form (12) which can infect other host cells. The bi-stable mechanism of site-specific integration/excision is orchestrated by MGE-encoded enzymes belonging to the serine- or tyrosine-recombinases (9). Tyrosine recombinases constitute the most widespread site-specific recombinases and are ubiquitous in the three domains of life. The enzymatic properties of tyrosine integrases have been investigated for decades (11,13,14). They typically recognize short identical DNA sequences present simultaneously on the MGE DNA and on its host chromosome. According to the phage Lambda/*Escherichia coli* paradigm, these sequences have been termed attB (for attachment Bacteria) and attP (for attachment Phage) (11). Integrases catalyze site-specific recombination between these sequences using a timely orchestrated mechanism consisting of two sequentially integrase-generated single-strand cuts in the two att sequences followed by strand-migration and religation (9). As a result, the exact MGE DNA is integrated into the host chromosome and bordered by attL and attR sequences that are hybrid of attB and attP. The site-specific recombination between attL and attR, known as excision, regenerates perfectly intact MGE and host chromosomes. The recombination reaction requires in some cases additional protein partners called recombination directionality factors (RDFs) that regulate the directionality of the reaction (15). Interestingly, integrases sharing very similar enzymatic properties have been identified in the three domains of life. Bacterial and eukaryotic tyrosine recombinases have been extensively studied (11,16-18). In contrast, very few archaeal integrases have been fully characterized (19-21). Archaea possess peculiar suicidal integrases whose attP site resides within the integrase-coding gene (22). Upon integration, the integrase-coding gene is split into two inactive moieties int(N) and int(C) on each side of the integrated MGE.

Integrases are found in geothermal environments where the most hyperthermophilic organisms belong to the Euryarchaeal Thermococcales order (23-25). Plasmid pTN3 (26) from *Thermococcus nautili* 30-1 (27,28) encodes Int^{pTN3}, the only suicidal integrase that has been characterized in Thermococcales so far (19). In addition to *bona fide* site-specific recombination properties, Int^{pTN3} promotes efficient recombination between short stretches of identical sequence resulting in massive genomic inversions (19). In the view of the peculiar recombination activities of Int^{pTN3}, it was of great importance to characterize additional suicidal integrase from Thermococcales. By carefully examining the distribution of MGE integrases, it appeared that a particular class of integrases was prevalent in hyperthermophilic archaea (this manuscript, Article 2). To better understand the reasons of this success, we have reconstituted in this work the evolution history of this class of integrases and characterized the enzymatic properties of its prominent member, the integrase of plasmid pT26-2 hosted by *Thermococcus* sp. 26-2 (29).

Results

We reported recently the broad distribution in hyperthermophilic Archaea of MGEs constituting the pT26-2 plasmid family (this manuscript, Article 2). These plasmids and integrated elements alternatively encode two types of integrase: a classical type II integrase following the Lambda Int paradigm in Methanococcales and a suicidal type I integrase in Thermococcales (29,30) (this manuscript, Article 2). This sharp dichotomic distribution and the widespread of suicidal integrases in Thermococcales prompted us to investigate their biochemical and enzymatic properties. Most of the identified Thermococcales integrases were poor candidates for biochemical analysis as they are encoded by inactive genes. Most suicidal integrase genes are indeed present in the host genome in a disrupted form resulting from their integration (30). Moreover, inactive gene moieties are similar to pseudogenes and bound to rapidly accumulate deleterious mutations (31) or can originate from distinct genes as discussed below. To avoid the risk of degeneration for the candidate for biochemical analysis, we selected an integrase expressed by a replicative plasmid. Only two freely replicating elements were described for this class so far. The abundance of specific CRISPR spacers directed against pT26-2-related elements could explain the relative instability of their replicative plasmidic form (this manuscript, Article 2). Among the two plasmids, we selected the candidate integrase encoded by plasmid pT26-2 carried by *Thermococcus* sp. 26-2 (29). Genomic sequence analysis revealed the presence of an integrated pT26-2 copy in *T.* sp. 26-2 chromosome. The DNA sequence of the reconstituted chromosome copy of the integrase gene was identical to the plasmid version. By DNA sequence comparison between the plasmid sequence and the extremities of the integrated copy, we identified the attachment sites of plasmid pT26-2 (Fig. 1B). The attP site corresponded to a portion of the integrase coding gene as expected for suicidal integrases. The chromosomal attachment site (attB) was found in a gene coding for a $tRNA^{Arg}$(TCT). The identification of these sequences allowed to reconstitute the integration scenario of plasmid pT26-2 into its host chromosome (Fig. 1A). Precisely, the attB site extends form 2 nt upstream the anticodon sequence to 6nt downstream of the gene.

Int^{pT26-2} can catalyze all three canonical site-specific recombination activities

To investigate the enzymatic properties of this integrase, we over-produced in *Escherichia coli* Strep-tagged versions of Int^{pT26-2} and of the Int^{pT26-2}Y327F variant where the catalytic tyrosine is substituted by a phenylalanine. We purified these enzymes and tested their *in vitro* recombination activities using an extended array of synthetic substrates. The most straightforward assay to rapidly assert the activity of purified Int^{pT26-2} was an *in vitro* integration reaction. Integrase-catalyzed recombination between two att sites carried by two identical supercoiled plasmids was monitored through the formation of plasmid dimers as described in (19) (Fig. 2A). It was previously reported for another suicidal integrase that the minimal att site defined by a strict attB/attP sequence identity was not sufficient for *in vitro* recombination activity (19). We therefore defined our synthetic att recombination site as the entire sequence of the $tRNA^{Arg}$(TCT) gene followed by 6 nucleotides downstream carried by plasmid pCB568 (Fig. 2A). This reaction generated the formation of plasmid dimers which could be readily identified by electrophoretic migration while reactions with longer incubation time accumulated higher-order multimers by cumulative integration (Fig. 2B). Int^{pT26-2} was thus capable of efficiently catalyzing site-specific integration *in vitro*. In a similar reaction, the mutated variant Int^{pT26-2}Y327F failed to produce detectable pCB568 plasmid dimers, confirming the catalytic role of tyrosine 327 (Fig. 2B). Interestingly, only wild-type Int^{pT26-2} was capable of producing topoisomers with supercoiled templates

devoid of att site such as pUC18. This suggests that Int^{pT26-2} could perform the first step of recombination, i.e. non-specific single-strand cleavage, followed by re-ligation on non-specific substrate, leading to the formation of a topoisomer ladder (Fig. 2B). The capacity of Int^{pT26-2} to promote excision was assayed in a recombination reaction using supercoiled plasmid pCB596 carrying two att sites in direct orientation followed by endonuclease restriction. This excision reaction effectively produced two smaller plasmids each containing a single att site (Fig. 1C). The substrate and the excised products were easily discriminated via their respective restriction pattern (Fig. 1D). This reaction demonstrated the capacity of Int^{pT26-2} to promote site-specific excision *in vitro*. The incubation with Int^{pT26-2} Y327F did no yield the excised products (Fig. 1D) confirming the implication of the catalytic tyrosine in the excision activity. Intra-molecular DNA inversion constituted the third canonical site-specific recombination reaction. We assayed Int^{pT26-2} inversion activity on supercoiled plasmid pCB598 containing two att sites in opposite orientation in a recombination reaction followed by endonuclease restriction. In the presence of Int^{pT26-2}, the assay produced the inversion of the sequence delimited by the att sites (Fig. 1E). Differential restriction patterns allowed to easily discriminate between the template and the inverted products (Fig. 1F). As for the integration and excision activities, the Int^{pT26-2}Y327F variant could not catalyze the site-specific inversion. All three positive recombination assays demonstrated that Int^{pT26-2} is a fully functional tyrosine recombinase able to catalyze efficient site-specific integration, excision and inversion *in vitro* in the absence of additional co-factors.

The recombinase activity of Int^{pT26-2} is independent of substrate DNA topology

The reported activity of several integrases demonstrated a high or mandatory requirement for negatively supercoiled templates (32,33). In hyperthermophilic archaeal cells, the DNA topological state is still conjectural even if some reports favor a relaxed chromosome (34). Additionally, in the context of hyperthermophilic enzymes, substrates could be differentially affected by the high temperature of the reactions depending on their topological state. Relaxed substrates could for instance adopt inactive conformations. Since all recombination assays described above were performed on supercoiled substrates, we wondered if Int^{pT26-2} could efficiently recombine templates in other topological forms. We compared Int^{pT26-2} integration and inversion capacity on supercoiled plasmids, relaxed templates and linear DNA fragments (Fig. 3 and Fig. S1). Inverted and integrated products were obtained after Int^{pT26-2} incubation with all three topological forms at 75°C. Int^{pT26-2} could therefore catalyze recombination at high temperature on a variety of topological forms. We quantified Int^{pT26-2} enzymatic activity on the three different topological forms for two replicates of the inversion assay (Fig. 3B). The results indicated that Int^{pT26-2} recombined both supercoiled and linear DNA with the same efficiently whereas the activity was slightly lower on relaxed templates.

Int^{pT26-2} att site extremities are not precisely delimited

Contrary to other integrases of pT26-2-related elements, the Int^{pT26-2} attachment site extends 6 nt downstream of the tRNA^{ARG} gene (this manuscript, Article 2) (Fig. 4AB). As all the previous assays were performed with recombining sites comprising the 6 nucleotides, we wondered if shorter sites would be equally proficient in site-specific recombination reactions (Fig. 4A). We implemented a recombination assay for shorter sites as follows. A constant 2106 bp long linear fragment carrying a full length tRNA^{Arg} gene was incubated in equimolar ratio with a nested set of shorter linear fragments whose sequence was tested for recombination (Fig. 4C). This test was designed in such a way that only

reactions using recombination-proficient templates would generate two linear fragments of intermediate size. Similar experiments have been used to strictly delimit the DNA segment required by other archaeal site-specific recombinases (19,21). Surprisingly, this test did not provide clear-cut limits for the Int^{pT26-2} att site. At the 5' end, we observed a progressive reduction in recombination efficiency: the segments L56, L55 and L54 are positive for recombination while sequence L53 is weakly active (Fig. 4D). At the 3' end, a wide range of sequences exhibited a barely detectable gradient of reduced recombination. The observed trend also strongly suggested that the attP site detected *in silico* (L51) is not recombination-proficient *in vitro*. From all the tested sequences, it appeared that Int^{pT26-2} retained partial activity over a remarkably wide range of recombination site extensions.

IntpT26-2 is active at near boiling water temperature

The natural hosts of the pT26-2 plasmid class belong to Methanococcales and Thermococcales which constitute some of the most hyperthermophilic organisms known to date. In the laboratory where it is readily cultivable, the optimal growth temperature of *Thermococcus* sp. 26-2 amounts to 85°C. The particular distribution of plasmid pT26-2 raised the question whether the Int^{pT26-2} recombinase activity was optimized for, and restricted to, high temperatures. All *in vitro* Int^{pT26-2} activity assays described above were performed at a near optimal 75°C which was the highest documented temperature for *in vitro* site-specific recombination. The optimal reported temperature for other hyperthermophilic recombinases never exceeded 65°C (19,21,35). It was therefore of great interest to test whether the Int^{pT26-2} integrase would be able to catalyze recombination reactions at yet higher temperatures. We performed the inversion assay described in Figure 1E across a wide range of incubation temperatures from 60°C to 99°C (Fig. 5). Interestingly, Int^{pT26-2} was able to efficiently catalyze site-specific recombination across the whole range of temperature tested (Fig. 5A). The maximal amount of recombination product was obtained between 75°C and 80°C (Fig. 5A). Interestingly, the inverted product was still observed at 99°C, which was the highest temperature we could assay at atmospheric pressure. It is to be noted that template DNA was decaying with the increasing temperature, probably due to thermal degradation during the 30 min incubation. To take degradation into account, we quantified the substrate/product ratio in 3 replicate experiments which demonstrated an optimal inversion rate between 80°C and 85 °C (Fig. 5B). It appeared that the difference between apparent and real *in vitro* Int^{pT26-2} optimal recombination temperatures was due to DNA degradation at high temperatures.

A cluster of related integrases is prevalent in hyperthermophilic Euryarchaea

Our data indicated that Int^{pT26-2} could catalyze all three canonical site-specific recombination reactions and that it was active at the highest demonstrated temperature *in vitro*. In a recent work, we reported that the integrase is encoded by 15 pT26-2-related elements of hyperthermophilic Thermococcales (this manuscript, Article 2). Given this temperature range and wide distribution among Thermococcales, we wondered whether we would detect Int^{pT26-2} homologs in other hyperthermophilic organisms. A similarity search in the protein databases could not be implemented because suicidal integrases are often misannotated due to their fragmentation upon integration. We therefore used a similarity search in translated nucleotide databases. This led to the identification of 73 integrases, including 54 in Thermococcales, 14 in Archaeoglobales and 5 in Methanosarcinales (Table S1). Of these, 20 were already published and 53 were newly identified, including 34 in our

genome collection (to be published elsewhere). These integrases originating from hyperthermophilic Thermococcales and Archaeoglobales and mesophilic Methanosarcinales constituted Dataset 1 (Table 1). To decipher the evolutionary relationship between these integrases, we built two similarity networks including Dataset 1 and all previously known suicidal integrases (Fig. 6A). We first applied a low similarity threshold of 25% over 60% of the protein length to take into account the suicidal integrase decay after integration (Fig. 6A). Random walk clustering indicated that all Dataset 1 integrases belonged to the same cluster and diverged from pTN3-like integrases and SSV-like integrases. Using a more refined similarity threshold to analyze the structure of Dataset 1, it appeared that the Thermococcales integrases and 8 Archaeoglobales integrases were more related, and constituted Dataset 2 (Fig. 6B & Table 1). Random walk clustering among Dataset 2 indicated a stronger relationship between the 54 Thermococcales integrases constituting Dataset 3 (Table 1). The Archaeoglobales integrases linked to IntpT26-2 in the second network were defined as Dataset 4 (Fig. 6B & Table 1). The integrases of Dataset 2 were present in 30% of closed Thermococcales chromosomes (15/51) and in 50% of closed Archaeoglobales chromosomes (4/8). Several genomes even contained several copies of these integrases, up to 5 for *Archaeoglobus profundus* DSM5631. Remarkably, we did not detect any multiple or tandem integration events at the same chromosomal site (Table 1). The very high prevalence rate designated the integrases of Dataset 2 as a very efficient and successful for those chromosomes. Additionally, the integrases from Dataset 2 were only present in organisms with optimal growth temperature above 75°C (Table S1). This selectivity for hyperthermophily is probably facilitated by an efficient activity at high temperatures as we evidenced for $Int^{pT26\text{-}2}$.

Constrained choice of integrases among highly variable hyperthermophilic MGEs

The more stringent network analysis clearly restricted the distribution of $Int^{pT26\text{-}2}$ homologs from Dataset 2 to hyperthermophilic Euryarchaeota. In order to trace the evolutionary history of these integrases, we identified all the MGEs encoding these enzymes. Most of them were integrated elements except virus TPV1 (36) and plasmids pT26-2 and pGE2 (29) (this manuscript, Article 2) and the newly identified plasmid pIRI06c infecting *Thermococcus* IRI06c. The size of the elements was highly variable, from 13 to 38 kb in Thermococcales (20 to 66 ORFs) and from 8 to 38 kb (12 to 50 ORFs) in Archaeoglobales (Table S1). The integrated elements from Thermococcales were ranked on the basis of core protein homologues. We identified 25 pT26-2-related elements that encoded at least 6 of the 7 core genes (this manuscript, Article 2), 8 fuselloviruses that encoded a major capsid protein (MCP) (37) and 8 pAMT11-related elements that encoded at least 3 of the common proteins shared by pAMT11 and TKV1 (38) (Fig. S2 and Table S1). In addition, 6 integrated elements were unrelated to any known plasmid of the WASPS plasmid database (in preparation). Among them, pIRI42c_IE1, T29-3_IE1 and TAMTc94_IE1 encoded common proteins (Fig. S2) and presumably correspond to a new MGE family. In Archaeoglobales, the integrated elements did not match with any known plasmid of the WASPS database. In all cases, the majority of the annotated proteins corresponded to unknown functions while the known functions were predominantly related to replication and transcriptional regulation. Furthermore, toxin anti-toxin systems, Cas proteins and nucleases were also encoded by the integrated elements. Finally, an additional tyrosine recombinase was also encoded by some elements. Overall, a wide range of different elements, plasmids and viruses, recruited integrases from Dataset 2, further underlining the pervasiveness of these recombinases.

Assessing the diversity of Int^{pT26-2}-related hyperthermophilic suicidal integrases by phylogenetic analysis

To investigate the relationships between the more closely related integrases of Dataset 2, we performed a phylogenetic tree (Fig. 7). Based on the network analysis results, we rooted the tree between the Thermococcales integrases (Dataset 3) and the 8 Archaeoglobales integrases from Dataset 4. Thermococcales distal branches were well resolved even if Archaeoglobales and basal Thermococcales were poorly supported (Fig. 7). Dataset 4 comprises very divergent Archaeoglobales integrases. Their corresponding long branches in the phylogenetic tree (Fig. 7) did not permit to infer evolutionary relationships. The Thermococcales integrases of Dataset 3 presented a mixture of closely related and divergent enzymes. This variability distribution provided the opportunity to study integrase evolution at different scales.

Thermococcales integrases are frequently exchanged between mobile elements

In the light of the phylogenetic tree we were able to approach various evolutionary aspects of Dataset 3. Precisely, we explored the variability and evolution of target specificity. We also investigated whether integrases presented host selectivity whether they followed their MGE phylogeny. Precisely, two phylogenetic histories could be invoked to explain the wide distribution of Thermococcales integrases that we observed among the various types of elements such as fuselloviruses, plasmids pT26-2 and pAMT11 and unidentified MGEs could be explained by: (i) the congruence of MGE and associated integrase phylogenies indicating that theses enzymes diverged from a single common ancestor whilst they co-evolved with the rest of the mobile element or (ii) the exchange of integrases between different MGE types. Strikingly, a very similar integrase (94% mean pairwise similarity) was found in the genomes of very distinct mobiles elements: in fuselloviruses (TspEXT12C_IV1 and TAMTc70_IV1), in a pT26-2-like integrated plasmid (TguDSM11113_IP1) and in unidentified integrated elements (T29-3_IE1 and TAMTc94_IE1) (Fig. S3). Such high similarity values indicated that these integrase genes were recently exchanged between these integrated elements. However we could not trace the directionality of the exchange due to the lack of bootstrap support. As another example, the pAMT11-related plasmid family presumably captured integrases from Dataset 3 at least twice independently, in TpIRI06c_IP1 and TprCol3_IP1 (Fig. 7). For the other integrases encoded by pAMT11-like elements, the evolutionary history could not be recovered due to the absence of bootstrap support for the basal branches. Interestingly, the plasmid pAMT11 described originally did not encode an integrase (38), suggesting corresponding gene loss in this particular plasmid or independent integrase gene acquisitions in pAMT11-related elements identified in this study. Module exchange between MGEs is a well-known process (39-41) and here we demonstrated that closely related integrases are encoded by different types of mobile elements (Fig. 7). In the case of this integrase family, the frequency of genetic exchange or acquisition highlighted the selective advantage provided by Int^{pT26-2}-related integrases to their respective MGE.

Thermococcales integrases are not species-specific

Our phylogenetic analysis indicated clearly that integrase and host chromosome phylogenies are not congruent (Fig. 7). Distinct Thermococcales genera such as *Thermococcus barophilus* CH5 and *Pyrococcus* sp. NA2 harbored very closely related integrases whereas the distant integrases of elements TKV1, 2 and 3 were found in a single species, *Thermococcus kodakarensis.* However, a limited

Pyrococcus genus specificity was observed for the integrases of the elements pGE2, PkuNCB100_IP1 and PHV1. Overall, the integrases from Dataset 3 seemed to be capable of pervading all Thermococcales, without species specificity.

Reduced att site selectivity suggests a biological function

In a recent work, we reported that the chromosomal attachment site for 19 published integrases consisted of the 3' end of various genes encoding tRNAs without supplementary loop (this manuscript, Article 2). This general rule still applied to the integrases of Dataset 1 with the exceptions of the integrases from the Archaeoglobales elements AprDSM5631_IE1 and AveSNP6_IE1 where that att site was located at the 3' and 5' end of tRNA genes, respectively, with a supplementary loop (Fig. S4). The 54 Thermococcales integrases from Dataset 3 used 14 different tRNA genes for integration whereas the 14 Archaeoglobales integrases from Dataset 4 used 9 different tRNA genes reflecting a flexibility in their integration specificity (Table S1). In *Thermococcus kodakarensis*, the mean pairwise identity between all tRNAs with no supplementary loop and no intron (35/47 tRNA genes) amounted to 73% in the portion after the anticodon but only 62% before the anticodon (Datafile 2AB). For the integrases of Dataset 1, targeting the 3'-end of tRNA genes might constitute an advantageous strategy to easily change specificity due to higher similarity between potential att targets. The cleavage site for the suicidal integrase of Sulfolobales virus SSV1 was identified at the extremities of the anti-codon loop (42). For Dataset 1, the att site mostly encompassed the anti-codon, one anti-codon arm, the T stem-loop and the amino-acid binding site (Fig. S4) consistently with a cleavage site at the extremities of the anti-codon loop. However, 9 att sites did not overlap the anti-codon sequence and 8 of them corresponded to a group of closely related integrases present in various integrated elements which integrated into a $tRNA^{Arg}$(CCG) gene. Strikingly, the att site of TCIR10A_IP1 did not include the anti-codon arm. The absence of the anti-codon sequence in several att sites argued against a possible cleavage site at the extremities of the anti-codon loop. Overall, the att site extended downstream of the tRNA gene for less than 40% of the integrases, representing less frequent cases than the 60% acknowledged previously (this manuscript, Article 2). For Thermococcales integrases of Dataset 3, a core sequence at the center of the att site was highly conserved: the portion corresponding to the T stem- loop presented a 90% mean pairwise identity between the att sites (Datafile 2D) the mean pairwise identity between full att sites was only 75% (Datafile 2C). As a general rule in all tRNAs, the T stem-loop is significantly more conserved than the remaining features. In *Thermococcus kodakarensis*, all T stem-loops share 85% similarity (Datafile 2G). The conserved core in the center of the att site might have a functional importance for the integrase as it could contains the cleavage and strand exchange positions. Very interestingly, the analysis of IntpT26-2 *in vitro* activity presented above revealed that as long as the core site was present, the requirement for specific arm sequence was less stringent. In these experiments, the last 10 nucleotides of the tRNA gene were not crucial to allow site-specific recombination. Consistently, we observed in Dataset 3 a variation in the extent of the att site for integrases recognizing the same tRNA gene target (Table S1, Fig. S4).

Promiscuous integration increases the pool of target sites

In the integrase datasets, we evidenced one case of non-specific integration with the Thermococcales element TspEXT12c_IV1 integrating in a $tRNA^{Arg}$(CGC) gene (Fig. S5). The attL and attR sequences of this element presented a single A-G nucleotide mismatch at the tip of the tRNA T loop (Fig. S5A-C). Both the A and G alleles were found for $tRNA^{Arg}$(CGC) in Thermococcales (Datafile 3) therefore, ruling out sequencing errors or random mutations. Strikingly, the sequences corresponding to attL and attR were also present in the $tRNA^{Arg}$(TCG) in Thermococcales (Datafile 3). Two scenarios could explain the difference between the attL and attR sites for TspEXT12c_IV1 (Fig. S5B). Firstly, non-specific recombination could have occurred between an attP site with a G and an attB site with an A (Fig. S5B). This scenario is less likely since the attP site would then have been identical to the host $tRNA^{Arg}$(TCG) gene and site-specific recombination would have been more favorable with that alternative attB. Secondly, non-specific recombination could have occurred between an attP site with an A and an attB site with a G. In any case, a non-specific recombination occurred that allowed the integration into a new site. This possibility to catalyze less stringent site-specific recombination for this integrase family could explain the capacity to target different tRNA genes.

Independent evolution histories of protein and target site sequences elucidate the diversity of suicidal integrases

Suicidal integrases share as common characteristic to use part of their own gene as attP integration site. Therefore, the integrase protein sequence at the junction between the Int(N) and Int(C) moieties corresponds to the translation of the att site (Fig. 1). The phylogenetic analysis presented in Figure 7 showed that, as a general rule, closely related integrases targeted the same tRNA gene for integration. In a number of cases however, cognate integrases appeared to have switched specificity resulting in a substantial modification of their amino-acid sequence. The variety of integrases populating Dataset 3 allowed to investigate how integrases acquire new target specificities without impairing protein integrity. By comparing both DNA and protein sequenced we identified 5 different cases of specificity switch which deviate from the general rule (Fig. 8). The first case demonstrated the acquisition of different att sites, the $tRNA^{Val}$(CAC) and $tRNA^{Tyr}$(GTA) genes respectively, by the distant integrases of elements TKV1 and TthOGL-20P_IP1. These target sites were very likely acquired via two independent events (Fig. 8A & Fig. S6AB). The second case illustrated the recruitment of an identical $tRNA^{Val}$(TAC) att site by two phylogenetically distantly related integrases from Tsp_IP1 and TEXT15c_IE1 (Fig. 7 & Fig. 8B). The two att sites exhibited different lengths and three nucleotide mismatches giving rise to a different protein sequence in the corresponding segment (Fig. S6CD). The att site similarity presumably constituted a convergence due to the limited pool size of the possible tRNA genes for integration rather than an ancestral characteristic inherited from their common ancestor, explaining the variation in att site size and translation. In the third case, the closely related integrases of pT26-2 and TGV1 shared the same specificity for a $tRNA^{Arg}$(TCT) gene (Fig. 8C & Fig. S6EF). These proteins exhibited high amino-acid similarity (>70%) (Fig. 8C) as reflected by their proximity in the phylogenetic analysis (Fig. 7). However, the amino-acid sequences corresponding to their respective att site were strikingly different. This difference was caused by two translation frameshifts occurring immediately upstream and downstream the att site, accounting also for a slight difference in site length (Fig. S6EF). Surprisingly, in its phylogenetic clade, the $Int^{pT26\text{-}2}$ integrase was the only one exhibiting these frameshifts therefore suggesting a single att site acquisition for all the members of the clade. A similar situation of

frameshifting was observed in a fourth case for the integrases of PIRI42c_IE1 and TE10P11_IP1 even if it resulted in similar glycine and proline-rich sequences due to the high GC content of the att site (Fig. 8D & Fig. S6GH). Notably, these proteins and their respective gene exhibited differential sequence conservation upstream and downstream the att region, suggesting a hybrid origin of the two moieties. The fifth case illustrated further the recombinant nature of these enzymes. Integrases originating from two different phylogenetic clades and carried by TE15P30_IV1 and TpiCDGS_IP1 opted for att sites in the related $tRNA^{Gly}$(CCC) and $tRNA^{Gly}$(TTC) genes and the sequence similarities between these integrases mirrored those found in the fourth case. (Fig. 8E & Fig. S6IJ). The integration of TpiCDGS_IP1 in *Thermococcus piezophilus* CDGS further exposed the recombination mechanism involved in the evolution of the suicidal integrases. Contrarily to other integration events, the *in silico* reconstituted integrase genes of TpiCDGS_IP1 carried a frameshift mutation due to a missing nucleotide in the attachment site. The presence of this mutation was confirmed by sequence read mapping (kindly provided by the original author). Taken together with our *in vitro* data demonstrating $Int^{pT26\text{-}2}$ relaxed target recognition, the succession of cases presented above demonstrated the high frequency of specificity change in suicidal integrases and allowed to propose a molecular model for the evolution of these enzymes which will be discussed in the next section.

Discussion

We have identified 73 suicidal integrases related to the Thermococcales pT26-2 integrase (Dataset 1), including 53 newly described enzymes. These suicidal integrases were identified in Thermococcales and for the first time in Archaeoglobales and Methanosarcinales. A subset of these integrases, Dataset 2, comprised closely related enzymes which were recruited by a variety or mobile genetic elements. These integrases infected exclusively thermophilic archaea and were proven prevalent in these organisms. Suicidal integrases carry their DNA recombination site within their own coding sequence and efficient integration requires a compatible site on the host chromosome. This site-specific recombination causes the disruption -or suicide- of the integrase gene into two inactive stumps therefore preventing MGE excision. Strikingly, the irreversibly captive state of MGEs carrying suicidal integrases is difficult to reconcile with the ubiquitous presence of these MGEs as integrated elements or freely replicative plasmids. The popularity of these integrases must imply the conveyance of a selective advantage to the genetic element and/or host chromosome. The deep genomic analysis of this unprecedented dataset of related integrases and of their *in vitro* recombination activity provided highly plausible clues for their evolutionary success. Firstly, we demonstrated that one integrase from the dataset, $Int^{pT26\text{-}2}$, can outstandingly catalyze *in vitro* site-specific recombination at near-boiling water temperatures. Other hyperthermophilic site-specific recombinases have been characterized and their activity was assayed *in vitro* at a maximal temperature not exceeding 65°C (19,21,42,43). Additionally, they were encoded by self-replicating mobile elements infecting hosts with optimal growth temperatures of 85°C at the most. The integrases from the Dataset 2 are encoded by self-replicating mobile elements infecting hosts with much higher optimal growth temperature, up to 105°C as reported for *Pyrococcus kukulkanii NCB100* (25). Such integrases were therefore particularly well suited to efficiently catalyze integration in hyperthermophilic environments. In a previous report, we demonstrated that the hyperthermophilic site-specific integrase Int^{pTN3} exhibited the additional property to catalyze low sequence specificity recombination reactions with the same outcome as

homologous recombination events in vitro at 65°C (19). Here, we showed that Int^{pT26-2} did not carry the additional subdomains found in Int^{pTN3} and only carried out site-specific recombination reactions.

Secondly, typical integration modules such as in the bacteriophage Lambda are composed of the integrase gene, a separate att site and additional RDF genes to avoid spontaneous MGE excision (11). In contrast, the suicidal integrases IntpTN3 and IntSSV2 they were shown to promote *in vitro* excision without recombination directionality factors (19,21); the disruption of their gene upon integration officing as directionality regulator. This property was also confirmed for Int^{pT26-2} which was able to perform both *in vitro* integration and excision reactions with comparable efficiencies. The compactness of a single gene integration module constituted very likely a strong advantage for the pervasiveness of suicidal integrases among related organisms.

Thirdly, this integrase family allows MGE integration in a variety of chromosomal sites and in a wide range of archaeal organisms corresponding to three distinct taxonomic orders. These enzymes are uniquely resilient by efficiently switching target specificity. The comparison of all chromosomal attachment sites demonstrated that these integrases target the 3' end of various tRNA genes which corresponds to their most conserved region. In addition, the *in vitro* activity analysis of Int^{pT26-2}, the most prominent integrase of this dataset clearly showed a somewhat relaxed requirement for specific att site extremities. These two properties certainly contributed to the evolution of these enzyme but were not sufficient to explain the extensive target site exchange among closely related integrases (Fig. 7). One would expect that any abrupt att switching would lead to drastic changes in the protein sequence in the att site segment and these alterations could also extent further downstream due to frameshifting. It can be intuited that in both cases the resulting protein would lose its integrase function. Unexpectedly, in Dataset 2, integrase sequences corresponding to the att site diverged either due to different att sequences or to identical att sites translated in alternate frames. In the latter case, we observed frequent site size variation and the presence of indels bordering the att site. These changes, allowing the restoration of a sense reading frame for the C-terminal end of the protein were often found among closely related integrases and were compatible with our biochemical evidence of relaxed sequence requirement of att borders. In addition, the variability of protein sequence encompassing the att site was somewhat constrained by the extensive conservation of the 3' end of the target tRNA genes and its high GC content giving rise to proline or glycine-rich protein segments. Overall, it appeared clearly that protein sequence changes corresponding to the att site did not affect protein function, making specificity switching easier than anticipated.

The aforementioned results and a thorough genomic comparison of 54 chromosomal integration events from Dataset 1 permitted to propose an integrated molecular model to explain the prevalence and pervasiveness of suicidal integrases in hyperthermophilic organisms which is bound to the mechanism used for att target switching. The model is based on successive MGE infections and integrations in the same cellular host. Any integration episode would generate identical attL and attR sequences at its borders and disrupt the suicidal integrase gene (Fig. 1). Each of these att sites can be targeted by the same MGE in a second event of integration to produce a tandem integration reconstituting an intact copy of the integrase gene (Fig. 9A). This particular situation is prone to efficient excision catalyzed by the intact integrase and has not been observed even in the larger Dataset 1 nor for other suicidal enzymes (19). On the other hand, tandem integration has been observed for MGEs carrying non-suicidal integrases (this manuscript, Article 2) (44,45) as their excision

might be regulated by RDFs. The integration instability of tandem MGEs could also be used to generate new hybrid suicidal integrases as observed in the case of *Thermococcus piezophilus* TpiCDGS_IP1 (Fig 8E & Fig. S6IJ). The model we propose is backed by the following molecular mechanism. The tandem integration of two related MGEs carrying divergent integrases followed by homologous recombination released a hybrid plasmid presumably carrying a frameshift in the reconstituted hybrid integrase gene. A mirrored hybrid MGE remaining integrated in the chromosome and presenting two integrase gene moieties of different origin and different reading frame corresponded to the observed TpiCDGS_IP1 element (Fig. 9B). We have documented several cases of hybrid integrases displaying separate evolution histories in their Int(N) and Int(C) moieties (Fig. 8DE and Fig. S6 GHIJ). As discussed below, efficient genomic homologous recombinations between cognate integrated copies of MGEs in *Thermococcus kodakarensis* has been reported (46) and fully supports this model.

An alternative scenario could also account for the generation of hybrid suicidal integrases. It involved different integrases from Dataset 2 carried by MGEs often found inserted in different locations of the same host chromosome (30,47). Two cognate MGEs integrated in opposite orientations and sharing enough DNA similarity could undergo homologous recombination and generate chromosomal inversions events as reported for *Thermococcus kodakarensis* TKV2 and TKV3 elements (46) (Fig. 9C). Such an inversion would bring heterologous attL and attR sites and heterologous integrase moieties into the correct register. A new incoming MGE with a relaxed integrase specificity could excise these recombinant MGEs and generate hybrid integrases with modified target specificities (Fig. 9C).

Suicidal integrases were thought to leave only dead corpses behind. This work demonstrated on the contrary that by integrating, these enzymes generated a fertile bed of pseudogenes whose combinations created a wide array of new integrases able to efficiently target 14 different tRNA genes. Our data showed that this variability, a somewhat relaxed target specificity, a very compact integration module devoid of RDFs and an extreme thermostability very likely accounted for the prevalence and unique pervasiveness of this integrase family in hyperthermophilic archaea.

Materials and methods

Recombinant protein production and purification

The gene coding for the integrase of plasmid pT26-2 ($Int^{pT26\text{-}2}$, NCBI protein accession YP_003603594.1) was PCR amplified from pT26-2 plasmid DNA with primers pT26-2_F and _R (Table **). The forward primer added a sequence coding for the Strep-tag at the 5' end of the gene. The PCR product was then assembled in the linearized expression plasmid pET-26b(+) by Gibson assembly (NEB) and transformed into *Escherichia coli* strain XL1-Blue. The resulting plasmid pCB558 was verified by DNA sequencing. Plasmid pCB616 encoding the variant $Int^{pT26\text{-}2}$Y327F was constructed with the Q5® Site-Directed Mutagenesis Kit (NEB) using the primers pT26-2_Y327F_F and _R (Table **). *E. coli* Rosetta BL21 (DE3) carrying pCB558 or pCB616 was grown in LB medium to OD 0.5 and recombinant protein production was induced with 250 µM IPTG. $Int^{pT26\text{-}2}$ overproduction in *E. coli* was somewhat toxic. After 1.5 hour induction, cells were harvested and resuspended in the purification buffer (1 M KCl, 40 mM Tris HCl pH 8, 10 % glycerol and 5 mM β-mercaptoethanol) supplemented with a protease inhibitor cocktail (cOmplete™ ULTRA Tablets, EDTA-free, Roche). Cells were lysed by a pressure shock with a one shot cell disruptor (Constant Systems Ltd) and centrifuged at 4°C for 30 min at 18000 g. The supernatant

was recovered, heated at 65°C for 10 minutes, centrifuged at 5000 g for 15 min and filtered. The solution was then loaded on a 1 mL StrepTrap HP column (GE Healthcare). The STREP-tagged Int$^{pT26\text{-}2}$ and Int$^{pT26\text{-}2}$Y327F were eluted by the purification buffer supplemented with 2.5 mM d-desthiobiotin. The buffer of Int$^{pT26\text{-}2}$Y327F was depleted in d-desthiobiotin by buffer exchange with a Vivaspin® Centrifugal Concentrators (Sartorius). Int$^{pT26\text{-}2}$ was subsequently loaded on a HiLoad 16/600 75 prep grade column (GE Healthcare) for size exclusion chromatography and fractions containing the protein were concentrated with a Vivaspin® Centrifugal Concentrators (Sartorius). Protein solutions harvested at different steps of the purification were analyzed by SDS-PAGE (Fig. S7). The purified concentrated proteins contained the N-ter strep-tag and their concentration was determined by spectrophotometry.

Integrase substrates construction

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

Integrase substrates production

Supercoiled plasmids were extracted from *E. coli* XL1-Blue using NucleoSpin Plasmid (Macherey-Nagel) or NucleoBond Xtra Midi (Macherey-Nagel) accordingly to the manufacturer instructions. Relaxed pCB568 and pCB598 were obtained by Nt.BspQI digestion (NEB) followed by a column purification (NucleoSpin Gel and PCR clean-up, Macherey-Nagel). ScaI and PvuII pCB568 fragments were obtained by ScaI and PvuII digestion (FastDigest, ThermoFisher) followed by a gel purification of the desired fragment (NucleoSpin Gel and PCR clean-up, Macherey-Nagel). Linear pCB598 was obtained by ScaI digestion (FastDigest, ThermoFisher) followed by a gel purification (NucleoSpin Gel and PCR clean-up, Macherey-Nagel). A 2106 bp fragment of pCB568 was amplified by Phusion Polymerase (ThermoFisher) with the primers pUC1481-1503 and P30-REV followed by column purification (NucleoSpin Gel and PCR clean-up, Macherey-Nagel). The various fragments of 800 bp were amplified from the appropriate plasmid by Phusion Polymerase (ThermoFisher) with the primers pUC195-217 and pZE21_rev followed by column purification (NucleoSpin Gel and PCR clean-up, Macherey-Nagel).

In vitro integrase enzymatic assay

For *in vitro* enzymatic assays, 500 µg substrate DNA and 200 ng (240 nM) integrase were incubated for 1h at 75°C in 300 mM KCl, 7 mM Tris HCl pH 8, 0.4 % glycerol and 825 µM β-mercaptoethanol in a total volume of 20 µL unless otherwise indicated. In certain cases, two different substrates were mixed in an equimolar ratio for a total mass of 500 µg. For integration assays, reaction product were treated with proteinase K, separated by agarose gel electrophoresis at 50V and subsequently stained with ethidium bromide for visualization. For inversion and excision assays, reaction products were purified with the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel), digested with the appropriate restriction enzymes (Fast-digest, ThermoFisher) and separated by gel electrophoresis. Band intensity was quantified with ImageJ (48) on non-saturated gel pictures using 3 repetitions of the activity assay.

Integrase homologs and mobile elements detection

To detect proteins closely related to $Int^{pT26\text{-}2}$, a classical similarity search in the protein databases could not be implemented since SSV-type integrases are often mis-annotated due to their fragmentation after integration. Instead, we used tBLASTn with already known and subsequently detected Int(N) and Int(C) moieties as query. As subject sequence, we used the nr/nt nucleotide collection and our own collection of sequenced Thermococcales genomes (to be published elsewhere). We selected hits with an e-value lower than 1e-30 and then reconstituted the complete integrase protein as indicated in this manuscript, Article 2. We then delimited the integrated mobile element presenting the integrase gene by the direct duplication at its extremities (att site). GenBank files were extracted for each element from attL site to attR sites (Datafile 1). Mobile elements were assigned to a MGE family based on the presence of marker genes: the 7 core genes of the pT26-2 plasmid family (this manuscript, Article 2) and the Major Capsid protein (MCP) gene for the fuselloviruses (37). For the pAMT11 plasmid family, no marker gene was previously proposed. We used the three longer genes (ORF1 to 3) conserved between the two previously known members of the family pAMT11 and TKV1 (38).

Silix Network

All-against-all BLASTP analyses were performed on all the integrases of Dataset 1, a set of Sulfolobales integrases identified in free Fuselloviridae, and all available pTN3 integrases. The all-against-all integrases BlastP results were grouped using the SiLiX (for *Single Linkage Clustering of Sequences*) package v1.2.8 (http://lbbe.univ-lyon1.fr/SiLiX) (49). This approach for the clustering of homologous sequences, is based on single transitive links with alignment coverage constraints. Several different criteria can be used separately or in combination to infer homology separately (percentage of identity, alignment score or E-value, alignment coverage). For this integrase dataset, we used the additional thresholds of 25% or 35% for the identity percentage and 60% for the query coverage. The network was visualized using igraph package from R (https://igraph.org/). In order to find densely connected communities in a graph via random walks, we used the cluster_walktrap function of the igraph package.

Protein alignment, trimming and phylogenetic analysis

The alignment used for phylogenetic analyses was performed using MAFFT v7 with default settings (50) and trimmed with BMGE (51) with a BLOSUM30 matrix, and the -b 1 parameter. IQ-TREE v1.6 (http://www.iqtree.org/) (52) was used to calculate maximum likelihood (ML) trees with the best model as suggested by the best model selection option (53). Branch robustness was estimated with the nonparametric bootstrap procedure (100 replicates) or with the SH-like approximate likelihood ratio test (54) and the ultrafast bootstrap approximation (1,000 replicates) (55). The integrases phylogenetic tree shown in Figure 7 correspond to the tree obtained with the VT+F+I+G4 model on a matrix of 318 positions and with ultrafast bootstrap to indicate the tree robustness. The phylogenetic tree was shaped with the iTOL webtool (56).

Other bioinformatics analysis

Synteny maps were created using EasyFig (57). Pairwise alignments and att site alignments were performed with MUSCLE (58). Thermococcus kodakarensis tRNA genes were extracted with GtRNAdb (59).

Funding and acknowledgements

This work was funded by CNRS and the European Research Council under the European Union's Seventh Framework Program (FP/2007-2013)/Project EVOMOBIL - ERC Grant Agreement no. 340440 (PF). Catherine Badel is supported by 'Ecole Normale Supérieure de Lyon'. The authors wish to thank Philippe Oger for kindly providing *T. piezophilus* sequencing reads.

Tables

Table 1. Integrase datasets used in this study. The dataset are defined using the network analysis (Fig. 6A)

Integrase dataset	Integrase number	Integration in tandem number	Host phyla	Dataset description
1	73	0	Thermococcales, Archaeoglobales and Methanosarcinales	Larger dataset. All suicide integrases clustered with $Int^{pT26\text{-}2}$ in the network A
2	62	0	Thermococcales and Archaeoglobales	Subset of Dataset 2. Integrases connected in the network B.
3	54	0	Thermococcales	Subset of Dataset 2. Thermococcales integrases
4	8	0	Archaeoglobales	Subset of Dataset 3 Archaeoglobales integrases

Figures

Figure 1. Plasmid pT26-2 integration model. A. Plasmid pT26-2 was present as a freely replicating element and integrated in the chromosome of *Thermococcus 26-2*. The chromosomal attachment site (attB) corresponds to a the $tRNA^{Arg}$(CTC) gene (in grey). Upon integration, the integrase gene is split in two parts named int(C) and int(N). The catalytic tyrosine residue (*) is located in the int(C) part. **B.** Alignment of the pT26-2 attP (P) sequence with the attL (L) and attR (R) sequences from *Thermococcus 26-2*. The conserved sequence is the attachment site (att) that corresponds to the 3' end of a $tRNA^{Arg}$(TCT) gene. The anti-codon sequence is underlined. The att site continues downstream of the tRNA gene additional 6 nt underlined by a dotted line. The sequences of the integrase and tRNA genes are antiparallel.

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

Figure 3. Int^{pT26-2} inversion activity on different topological forms. The inversion assay presented in Figure 2C was used to test the topological forms that Int^{pT26-2} can use as substrate. **A.** Supercoiled, relaxed and ScaI-linearized pCB598 were incubated with purified Int^{pT26-2} (WT) at 75°C for 1h or 6h and digested with ScaI and XhoI. Digestion was incomplete to retain relaxed plasmids (dotted arrow) in addition to the linear form. **B.** The relative amounts of substrate and product molecules were quantified for each lane. The histogram represents the mean value of 2 replicates and crosses indicate individual values.

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

Figure 5. Temperature activity range of Int^{pT26-2}. The inversion assay presented in Figure 2C was used to test the temperature activity range of Int^{pT26-2}. **A.** Plasmid pCB598 was incubated with purified Int^{pT26-2} at different temperatures during 0.5h and digested with ScaI and XhoI. **B**. Relative amounts of substrate and product were calculated for each lane, in triplicate. The error bar represents a 95% confidence interval.

Figure 6. Archaeal suicidal integrases similarity network. All available archaeal suicidal integrases were analyzed through a similarity network. Each dot corresponds to a protein. Links between two proteins refer to a BLASTP pairwise similarity >25% over 60% for the protein of **Panel A** and >35% over 60% for the protein of **Panel B**. A random walk algorithm was used for protein clustering. For both networks, proteins are colored depending on their clustering as indicated in the boxed legend. The star points to Int^{pT26-2}.The datasets defined in Table 1 are indicated.

Figure 7. Maximum likelihood phylogenic tree of the integrases from Dataset 2. Branch values represent the posterior probability. Branches supported by both the posterior probability and ultrafast bootstrap (>95%) are indicated by a black dot. The integrated element classification is color-indicated indicated when known. The individual tRNA gene used for integration are indicated as well as the anti-codon sequence. The scale bar represents the average number of substitutions per site.

Figure 8. **Independent evolution of integrases and their target sites**. For suicidal integrases, the att site is located inside the gene coding for the integrase and is therefore translated along with the integrase. Different cases illustrating the independent evolution of the integrases of Dataset 3 and their respective target sites are summarized here. Gene sequences (DNA) or integrase protein sequences (proteins) were aligned. Mean pairwise similarity over the Int(N), att or Int(C) regions is indicated by a color scale. High similarities (>70%) are indicated in dark blue. Lower similarities (<70%) are indicated in light blue. The 70% cutoff was selected because it corresponds to the similarity between the closely related integrases from elements TGV1 and pT26-2. The phylogenetic distance (d) between proteins is calculated in the same units as in Figure 7. **A.** General case: completely divergent integrases at the DNA and protein levels. **B.** Two divergent integrases sharing the same att site but translated in different frames. **C.** The integrases are closely related as indicated by their similar gene and protein sequences. The same att sequence is translated in a different frame. **D.** The two integrases are closely related at their Int(C) as indicated by their similar gene and protein sequences but with divergent Int(N) segments. Similarly to C, the att sequence translation is different between the two proteins, due to a frameshift. **E.** The two integrases are closely related at their Int(N) as indicated by their similar gene and protein sequences but with divergent Int(C) segments. The att site is translated in a different frame. Complete att site and protein alignments are available in Figure S6.

Figure 9. Events leading to the formation of hybrid integrases. A. Tandem insertion of the same MGE in the same tRNA gene target reconstituting a functional integrase gene able to excise the element. Identical tandem insertions have never been observed. **B.** A first MGE integration event generated an attR site with a single nucleotide deletion as compared to the original tRNA^{Gly} gene (red dot) (Fig. S6.I-J). The second integration event involved a related MGE but with a more distant integrase. This integration generates an inactive integrase gene due to frameshifting. Homologous recombination

between related MGE backbones could have excised a hybrid plasmid leading to the situation observed for the integrated TpiCDGS_IP1. The Int(N) and Int(C) segments of its integrase have a different evolution history and cannot be assembled due to a mirrored frameshift in the att region. **C.** Multiple MGE integration events at separate chromosomal locations and in inverted orientation can give rise to a large genomic inversion by homologous recombination between related MGE backbones as reported (46). This inversion generates hybrid MGEs which could excise by the means of a compatible integrase provided in trans via superinfection.

Supplementary data

Table S1. Integrase list.

Table S2. Plasmids used in this work

Table S3. Oligonucleotides used in this work

Figure S1. Int^{pT26-2} integration activity on different topological forms. **A.** Recombination between two att sites (triangles) carried by two identical plasmids pCB568, either supercoiled or relaxed, produce plasmid dimers that are supercoiled and relaxed respectively. Recombination between two att sites carried by two linear fragments produce two chimeric linear fragments. **B.** Plasmid pCB598 DNA in supercoiled or relaxed form or as and linear ScaI and PvuII fragments was incubated with purified Int^{pT26-2} (WT) at 75°C for 1h or 6h and treated with proteinase K.

Figure S2. Integrated elements synteny map. The synteny between all the genetic elements encoding integrases from the dataset 3 (Thermococcales) is visualized with the software Easyfig (57). Each arrow represents an open reading frame (ORF). The ORFs encoding the core proteins of each MGE family are indicated in a specific color: green for pT26-2 elements, red for fuselloviruses and yellow for pAMT11 elements. The integrase fragments are colored in blue. Pairwise similarity calculated using TBLASTX is indicated in a shade of greys.

Figure S3. Synteny map of 5 elements encoding closely related integrases (in blue). TspEXT12c_IV1 and TAMTc70_IV1 are fuselloviruses and encode a major capsid protein (red). TguDSM11113_IP1 is pT26-2 related element as indicated by the presence of the pT26-2 family 7 core proteins (green). T29-3_IE1 and TAMTc11_IE1 are unrelated to any other know MGE. Pairwise similarity calculated using TBLASTX is indicated in a shade of greys.

Figure S4. Att sites sequences. A. Archaeoglobales and Methanosarcinales elements. **B.** Thermococcales elements. The sequences corresponding to relevant tRNA structures are indicated. The anti-codon is framed. In A, the star indicates elements included in Dataset 2. **C.** Nucleotide identity in the Thermococcales att site alignment.

Figure S5. TspEXT12c_IV1 non-specific integration. **A.** TspEXT12c_IV1 attL and attR sequences are not identical. This difference is not due to a sequencing error or mutation as both attL and attR sequences are found in Thermococcales tRNAs. **B.** Two scenarios of non-specific integration can explain the difference between the attL and attR sites. The two alternative ancestral chromosomes are found in Thermococcales and we can therefore not discriminate between the scenarios. **C.** The leaf-like structure of the $tRNA^{Arg}$(CCG) from *Thermococcus sp.* EXT12c is represented. The difference between the attL and attR sequence corresponds to the tip of the T loop. The att site nucleotides are indicated in red.

Figure S6. Nucleotidic att sites alignments (A, C, E, G and I) **and corresponding integrase protein alignment** (B, D, F, H and J). In the att alignments, the mismatched nucleotides are indicated in bold. In the protein alignments, the translation of the att site sequence is boxed in blue.

Figure S7. $Int^{pT26\text{-}2}$ and $Int^{pT26\text{-}2}$Y327F purification. A. At each step of $Int^{pT26\text{-}2}$ purification, a small volume was sampled and analyzed by SDS-PAGE. The gels are presented in chronological order. The bolded eluate of the size exclusion chromatography were retained for the next purification steps. **B.** At each step of $Int^{pT26\text{-}2}$Y327F purification, a small volume was sampled and analyzed by SDS-PAGE. The gels are presented in chronological order. The bolded eluate of the affinity chromatography were retained for the next purification steps. **C.** At the end of both purification, concentrated $Int^{pT26\text{-}2}$ and $Int^{pT26\text{-}2}$Y327F were analyzed by SDS-PAGE. All presented SDS-PAGE gels are composed of 12% acrylamide.

Datafile 1. GenBank files of all the genetic elements encoding an integrase from Dataset 1. The file can be downloaded from https://archaea.i2bc.paris-saclay.fr/Datafile1.zip

Datafile 2. Pairwise similarity tables. A. Pairwise identity between the sequences before the anti-codon of all tRNAs with no supplementary loop and no intron (35/47 tRNA genes) from *Thermococcus kodakarensis*. **B.** Pairwise identity between the sequences after the anti-codon of all tRNAs with no supplementary loop and no intron (35/47 tRNA genes) from *Thermococcus kodakarensis*. **C.** Pairwise identity between the entire att sites from Dataset 3 (Thermococcales integrases). **D.** Pairwise identity between the portion corresponding to the T stem-loop from the att sites from Dataset 3 (Thermococcales integrases). **E.** Pairwise identity between the portion corresponding to the T stem-loop from the att sites of Dataset 3 against Dataset 4. **F.** Pairwise identity between the segments corresponding to the T stem-loop from the att sites Dataset 3 against the att sites of the integrases included in Dataset 1 but excluded from the dataset 2. **G.** Pairwise identity between the T stem-loop region of all tRNAs (47/47 tRNA genes) from *Thermococcus kodakarensis.* The file can be downloaded from https://archaea.i2bc.paris-saclay.fr/Datafile2.xlsx

Datafile 3. List of $tRNA^{Arg}$(TCG) and $tRNA^{Arg}$(CGC) from Thermococcales. The file can be downloaded from https://archaea.i2bc.paris-saclay.fr/Datafile3.txt

References

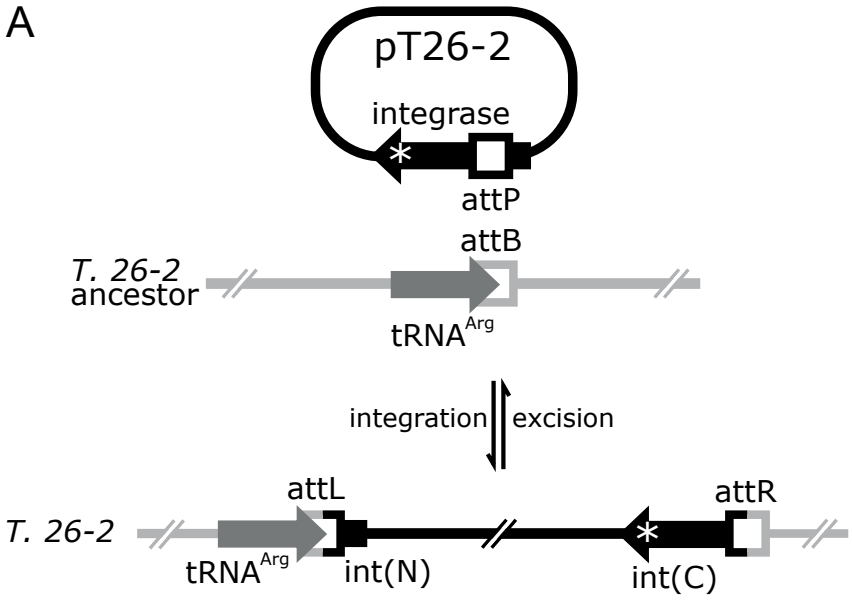
1. Hulter, N., Ilhan, J., Wein, T., Kadibalban, A.S., Hammerschmidt, K. and Dagan, T. (2017) An evolutionary perspective on plasmid lifestyle modes. *Current opinion in microbiology*, **38**, 74-80.
2. Million-Weaver, S. and Camps, M. (2014) Mechanisms of plasmid segregation: have multicopy plasmids been overlooked? *Plasmid*, **75**, 27-36.
3. Harms, A., Brodersen, D.E., Mitarai, N. and Gerdes, K. (2018) Toxins, Targets, and Triggers: An Overview of Toxin-Antitoxin Biology. *Mol Cell*, **70**, 768-784.
4. Arber, W. and Dussoix, D. (1962) Host specificity of DNA produced by Escherichia coli. I. Host controlled modification of bacteriophage lambda. *Journal of molecular biology*, **5**, 18-36.
5. Hille, F., Richter, H., Wong, S.P., Bratovic, M., Ressel, S. and Charpentier, E. (2018) The Biology of CRISPR-Cas: Backward and Forward. *Cell*, **172**, 1239-1259.
6. Carroll, A.C. and Wong, A. (2018) Plasmid persistence: costs, benefits, and the plasmid paradox. *Canadian journal of microbiology*, **64**, 293-304.
7. Gerdes, K., Howard, M. and Szardenings, F. (2010) Pushing and pulling in prokaryotic DNA segregation. *Cell*, **141**, 927-942.
8. Nordstrom, K. (2006) Plasmid R1--replication and its control. *Plasmid*, **55**, 1-26.
9. Grindley, N.D.F., Whiteson, K.L. and Rice, P.A. (2006) Mechanisms of site-specific recombination. *Annual review of biochemistry*, **75**, 567-605.
10. Landy, A. (1989) Dynamic, structural, and regulatory aspects of lambda site-specific recombination. *Annual review of biochemistry*, **58**, 913-949.
11. Landy, A. (2015) The lambda Integrase Site-specific Recombination Pathway. *Microbiol Spectr*, **3**, MDNA3-0051-2014.
12. Gandon, S. (2016) Why Be Temperate: Lessons from Bacteriophage lambda. *Trends in microbiology*, **24**, 356-365.
13. Guo, F., Gopaul, D.N. and Van Duyne, G.D. (1999) Asymmetric DNA bending in the Cre-loxP site-specific recombination synapse. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 7143-7148.
14. Chen, J.W., Lee, J. and Jayaram, M. (1992) DNA Cleavage in Trans by the Active-Site Tyrosine during Flp Recombination - Switching Protein Partners before Exchanging Strands. *Cell*, **69**, 647-658.
15. Lewis, J.A. and Hatfull, G.F. (2001) Control of directionality in integrase-mediated recombination: examination of recombination directionality factors (RDFs) including Xis and Cox proteins. *Nucleic acids research*, **29**, 2205-2216.
16. Van Duyne, G.D. (2015) Cre Recombinase. *Microbiol Spectr*, **3**, MDNA3-0014-2014.
17. Dorman, C.J. and Bogue, M.M. (2016) The interplay between DNA topology and accessory factors in site-specific recombination in bacteria and their bacteriophages. *Science progress*, **99**, 420-437.
18. Jayaram, M., Ma, C.H., Kachroo, A.H., Rowley, P.A., Guga, P., Fan, H.F. and Voziyanov, Y. (2015) An Overview of Tyrosine Site-specific Recombination: From an Flp Perspective. *Microbiol Spectr*, **3**.
19. Cossu, M., Badel, C., Catchpole, R., Gadelle, D., Marguet, E., Barbe, V., Forterre, P. and Oberto, J. (2017) Flipping chromosomes in deep-sea archaea. *PLoS genetics*, **13**, e1006847.
20. Wang, J., Liu, Y., Liu, Y., Du, K., Xu, S., Wang, Y., Krupovic, M. and Chen, X. (2018) A novel family of tyrosine integrases encoded by the temperate pleolipovirus SNJ2. *Nucleic acids research*, **46**, 2521-2536.
21. Zhan, Z.Y., Zhou, J. and Huang, L. (2015) Site-Specific Recombination by SSV2 Integrase: Substrate Requirement and Domain Functions. *Journal of virology*, **89**, 10934-10944.
22. She, Q., Peng, X., Zillig, W. and Garrett, R.A. (2001) Gene capture in archaeal chromosomes. *Nature*, **409**, 478.
23. Adam, P.S., Borrel, G., Brochier-Armanet, C. and Gribaldo, S. (2017) The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. *The ISME journal*, **11**, 2407-2425.

24. Schut, G.J., Lipscomb, G.L., Han, Y., Notey, J.S., Kelly, R.M., and Adams, M.M.W. (2014) In E. Rosenberg, E. F. D., S. Lory, E. Stackebrandt, and F. Thompson, (ed.), *The Prokaryotes*. Springer Berlin Heidelberg, pp. 363–383.
25. Callac, N., Oger, P., Lesongeur, F., Rattray, J.E., Vannier, P., Michoud, G., Beauverger, M., Gayet, N., Rouxel, O., Jebbar, M. *et al.* (2016) Pyrococcus kukulkanii sp. nov., a hyperthermophilic, piezophilic archaeon isolated from a deep-sea hydrothermal vent. *International journal of systematic and evolutionary microbiology*, **66**, 3142-3149.
26. Gaudin, M., Krupovic, M., Marguet, E., Gauliard, E., Cvirkaite-Krupovic, V., Le Cam, E., Oberto, J. and Forterre, P. (2013) Extracellular membrane vesicles harbouring viral genomes. *Environmental microbiology*, **16**, 1167-1175.
27. Gorlas, A., Croce, O., Oberto, J., Gauliard, E., Forterre, P. and Marguet, E. (2014) Thermococcus nautili sp. nov., a hyperthermophilic archaeon isolated from a hydrothermal deep sea vent (East Pacific Ridge). *International journal of systematic and evolutionary microbiology*, **64**, 1802-1810.
28. Oberto, J., Gaudin, M., Cossu, M., Gorlas, A., Slesarev, A., Marguet, E. and Forterre, P. (2014) Genome Sequence of a Hyperthermophilic Archaeon, Thermococcus nautili 30-1, That Produces Viral Vesicles. *Genome announcements*, **2**, e00243-00214.
29. Soler, N., Marguet, E., Cortez, D., Desnoues, N., Keller, J., van Tilbeurgh, H., Sezonov, G. and Forterre, P. (2010) Two novel families of plasmids from hyperthermophilic archaea encoding new families of replication proteins. *Nucleic acids research*, **38**, 5088-5104.
30. She, Q., Chen, B. and Chen, L. (2004) Archaeal integrases and mechanisms of gene capture. *Biochemical Society transactions*, **32**, 222-226.
31. Liu, Y., Harrison, P.M., Kunin, V. and Gerstein, M. (2004) Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome biology*, **5**, R64.
32. Mizuuchi, K., Gellert, M. and Nash, H.A. (1978) Involement of supertwisted DNA in integrative recombination of bacteriophage lambda. *Journal of molecular biology*, **121**, 375-392.
33. Reed, R.R. (1981) Transposon-mediated site-specific recombination: a defined in vitro system. *Cell*, **25**, 713-719.
34. Lopez-Garcia, P. and Forterre, P. (1997) DNA topology in hyperthermophilic archaea: reference states and their variation with growth phase, growth temperature, and temperature stresses. *Molecular microbiology*, **23**, 1267-1279.
35. Cortez, D., Quevillon-Cheruel, S., Gribaldo, S., Desnoues, N., Sezonov, G., Forterre, P. and Serre, M.C. (2010) Evidence for a Xer/dif system for chromosome resolution in archaea. *PLoS genetics*, **6**, e1001166.
36. Gorlas, A., Koonin, E.V., Bienvenu, N., Prieur, D. and Geslin, C. (2012) TPV1, the first virus isolated from the hyperthermophilic genus Thermococcus. *Environmental microbiology*, **14**, 503-516.
37. Krupovic, M., Quemin, E.R., Bamford, D.H., Forterre, P. and Prangishvili, D. (2014) Unification of the globally distributed spindle-shaped viruses of the archaea. *Journal of virology*, **88**, 2354-2358.
38. Gonnet, M., Erauso, G., Prieur, D. and Le Romancer, M. (2011) pAMT11, a novel plasmid isolated from a Thermococcus sp. strain closely related to the virus-like integrated element TKV1 of the Thermococcus kodakaraensis genome. *Research in microbiology*, **162**, 132-143.
39. Hendrix, R.W., Lawrence, J.G., Hatfull, G.F. and Casjens, S. (2000) The origins and ongoing evolution of viruses. *Trends in microbiology*, **8**, 504-508.
40. Iranzo, J., Koonin, E.V., Prangishvili, D. and Krupovic, M. (2016) Bipartite Network Analysis of the Archaeal Virosphere: Evolutionary Connections between Viruses and Capsidless Mobile Elements. *Journal of virology*, **90**, 11043-11055.
41. Oberto, J., Sloan, S.B. and Weisberg, R.A. (1994) A segment of the phage HK022 chromosome is a mosaic of other lambdoid chromosomes. *Nucleic acids research*, **22**, 354-356.
42. Serre, M.C., Letzelter, C., Garel, J.R. and Duguet, M. (2002) Cleavage properties of an archaeal site-specific recombinase, the SSV1 integrase. *Journal of Biological Chemistry*, **277**, 16758-16767.

43. Jo, M., Murayama, Y., Tsutsui, Y. and Iwasaki, H. (2017) In vitro site-specific recombination mediated by the tyrosine recombinase XerA of Thermoplasma acidophilum. *Genes to cells : devoted to molecular & cellular mechanisms*, **22**, 646-661.
44. Krupovic, M., Forterre, P. and Bamford, D.H. (2010) Comparative analysis of the mosaic genomes of tailed archaeal viruses and proviruses suggests common themes for virion architecture and assembly with tailed viruses of bacteria. *Journal of molecular biology*, **397**, 144-160.
45. Krupovic, M., Makarova, K.S., Wolf, Y.I., Medvedeva, S., Prangishvili, D., Forterre, P. and Koonin, E.V. (2019) Integrated mobile genetic elements in Thaumarchaeota. *Environmental microbiology*.
46. Gehring, A.M., Astling, D.P., Matsumi, R., Burkhart, B.W., Kelman, Z., Reeve, J.N., Jones, K.L. and Santangelo, T.J. (2017) Genome Replication in Thermococcus kodakarensis Independent of Cdc6 and an Origin of Replication. *Frontiers in microbiology*, **8**, 2084.
47. Wang, Y., Duan, Z., Zhu, H., Guo, X., Wang, Z., Zhou, J., She, Q. and Huang, L. (2007) A novel Sulfolobus non-conjugative extrachromosomal genetic element capable of integration into the host genome and spreading in the presence of a fusellovirus. *Virology*, **363**, 124-133.
48. Schneider, C.A., Rasband, W.S. and Eliceiri, K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nature methods*, **9**, 671-675.
49. Miele, V., Penel, S. and Duret, L. (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC bioinformatics*, **12**, 116.
50. Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, **30**, 772-780.
51. Criscuolo, A. and Gribaldo, S. (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC evolutionary biology*, **10**, 210.
52. Nguyen, L.T., Schmidt, H.A., von Haeseler, A. and Minh, B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, **32**, 268-274.
53. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. and Jermiin, L.S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, **14**, 587-589.
54. Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, **59**, 307-321.
55. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q. and Vinh, L.S. (2018) UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular biology and evolution*, **35**, 518-522.
56. Letunic, I. and Bork, P. (2019) Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic acids research*.
57. Sullivan, M.J., Petty, N.K. and Beatson, S.A. (2011) Easyfig: a genome comparison visualizer. *Bioinformatics*, **27**, 1009-1010.
58. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, **32**, 1792-1797.
59. Chan, P.P. and Lowe, T.M. (2016) GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic acids research*, **44**, D184-189.

Figure 1

A


pT26-2
integrase
attP
attB
T. 26-2
ancestor
tRNA^Arg
integration
excision
attL
attR
T. 26-2
tRNA^Arg
int(N)
int(C)

B

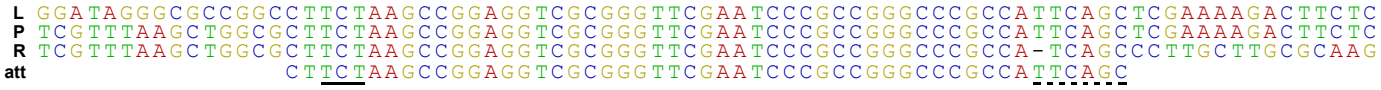

L GGATAGGGCGCCGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCCATTCAGCTCGAAAAGACTTCTC
P TCGTTTAAGCTGGCGCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCCATTCAGCTCGAAAAGACTTCTC
R TCGTTTAAGCTGGCGCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCCA-TCAGCCCTTGCTTGCGCAAG
att CTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCCATTCAGC

Figure 2

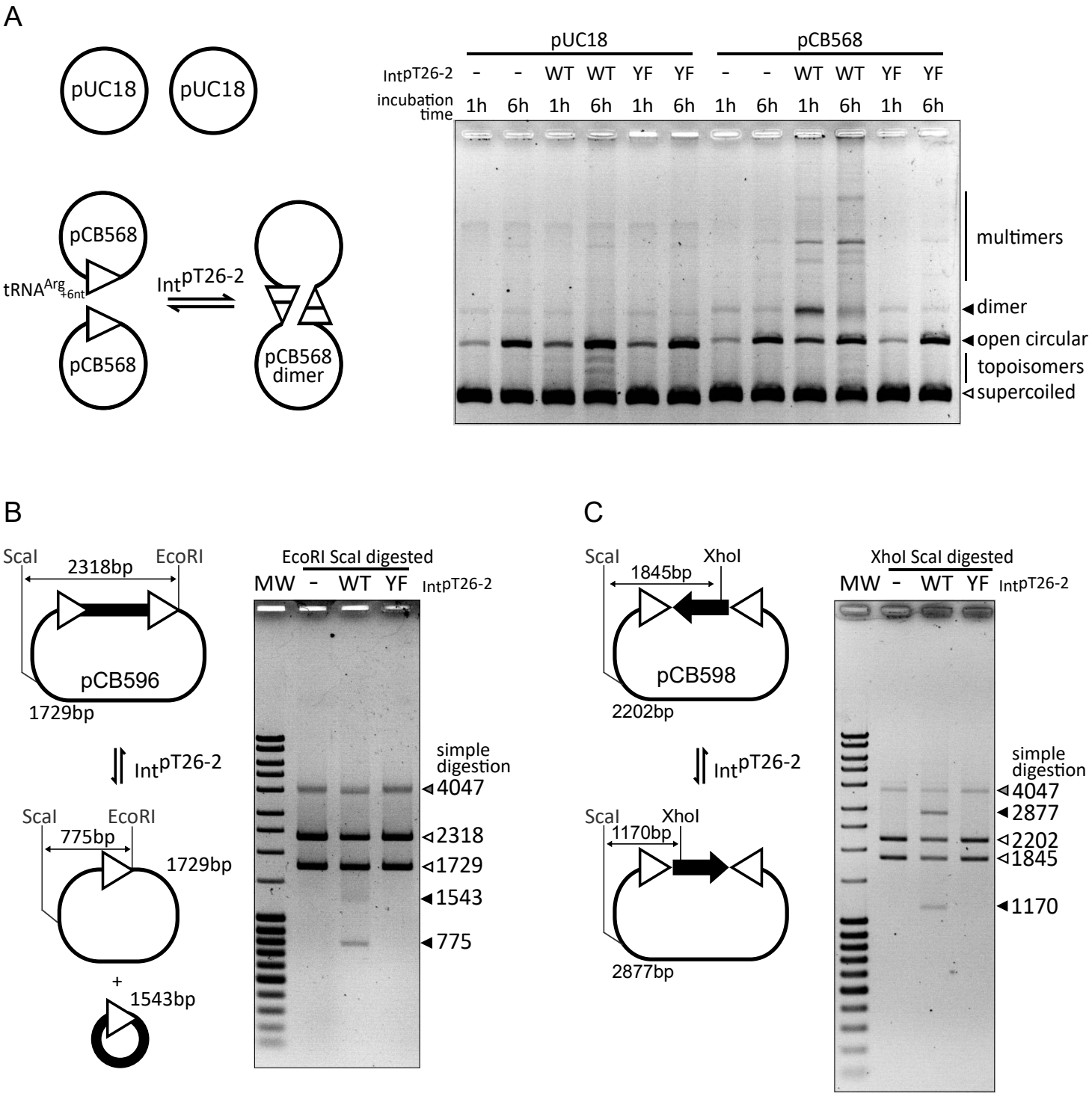
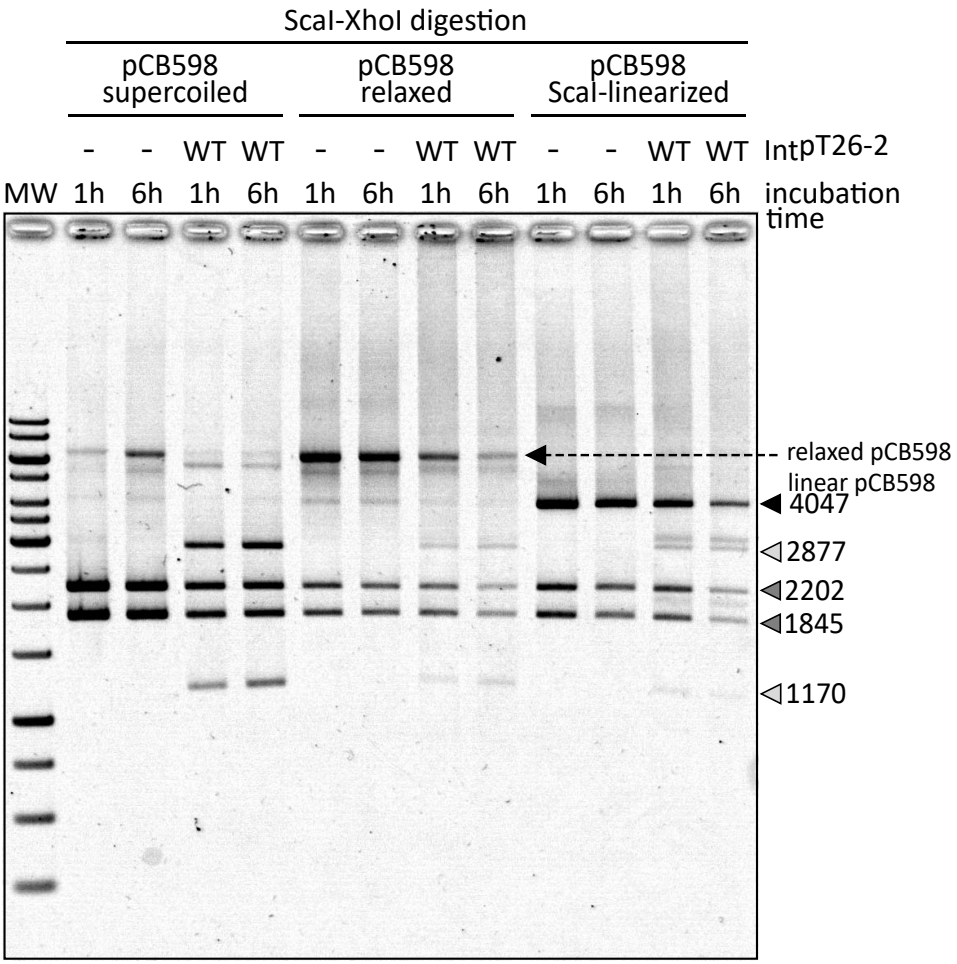

A
pUC18
pUC18
pCB568
tRNA$^{Arg}_{+6nt}$
Int^{pT26-2}
pCB568
pCB568
dimer
pUC18
pCB568
Int^{pT26-2} - - WT WT YF YF - - WT WT YF YF
incubation time 1h 6h 1h 6h 1h 6h 1h 6h 1h 6h 1h 6h
multimers
dimer
open circular
topoisomers
supercoiled
B
ScaI
2318bp
EcoRI
pCB596
1729bp
Int^{pT26-2}
ScaI
EcoRI
775bp
1729bp
+
1543bp
EcoRI ScaI digested
MW - WT YF Int^{pT26-2}
simple digestion
4047
2318
1729
1543
775
C
ScaI
1845bp
XhoI
pCB598
2202bp
Int^{pT26-2}
ScaI
XhoI
1170bp
2877bp
XhoI ScaI digested
MW - WT YF Int^{pT26-2}
simple digestion
4047
2877
2202
1845
1170

Figure 3

A


ScaI-XhoI digestion
pCB598 supercoiled
pCB598 relaxed
pCB598 ScaI-linearized
- - WT WT - - WT WT - - WT WT Int pT26-2
MW 1h 6h 1h 6h 1h 6h 1h 6h 1h 6h 1h 6h incubation time
relaxed pCB598
linear pCB598
4047
2877
2202
1845
1170

B

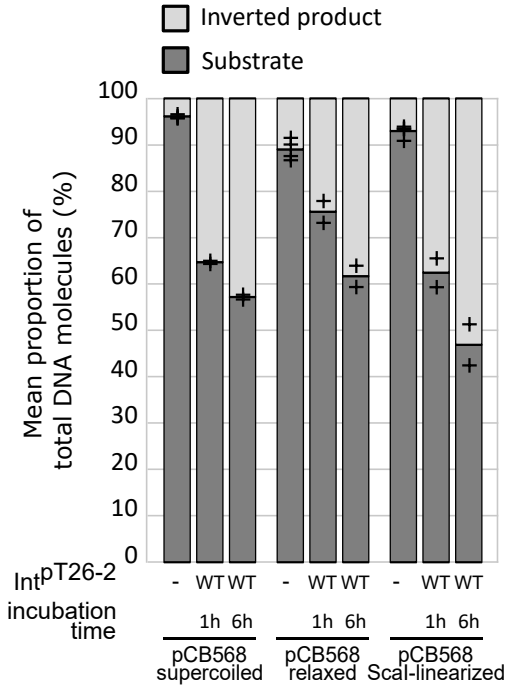

Inverted product
Substrate
Mean proportion of total DNA molecules (%)
100
90
80
70
60
50
40
30
20
10
0
Int pT26-2
- WT WT - WT WT - WT WT
incubation time
1h 6h 1h 6h 1h 6h
pCB568 supercoiled
pCB568 relaxed
pCB568 ScaI-linearized

Figure 4

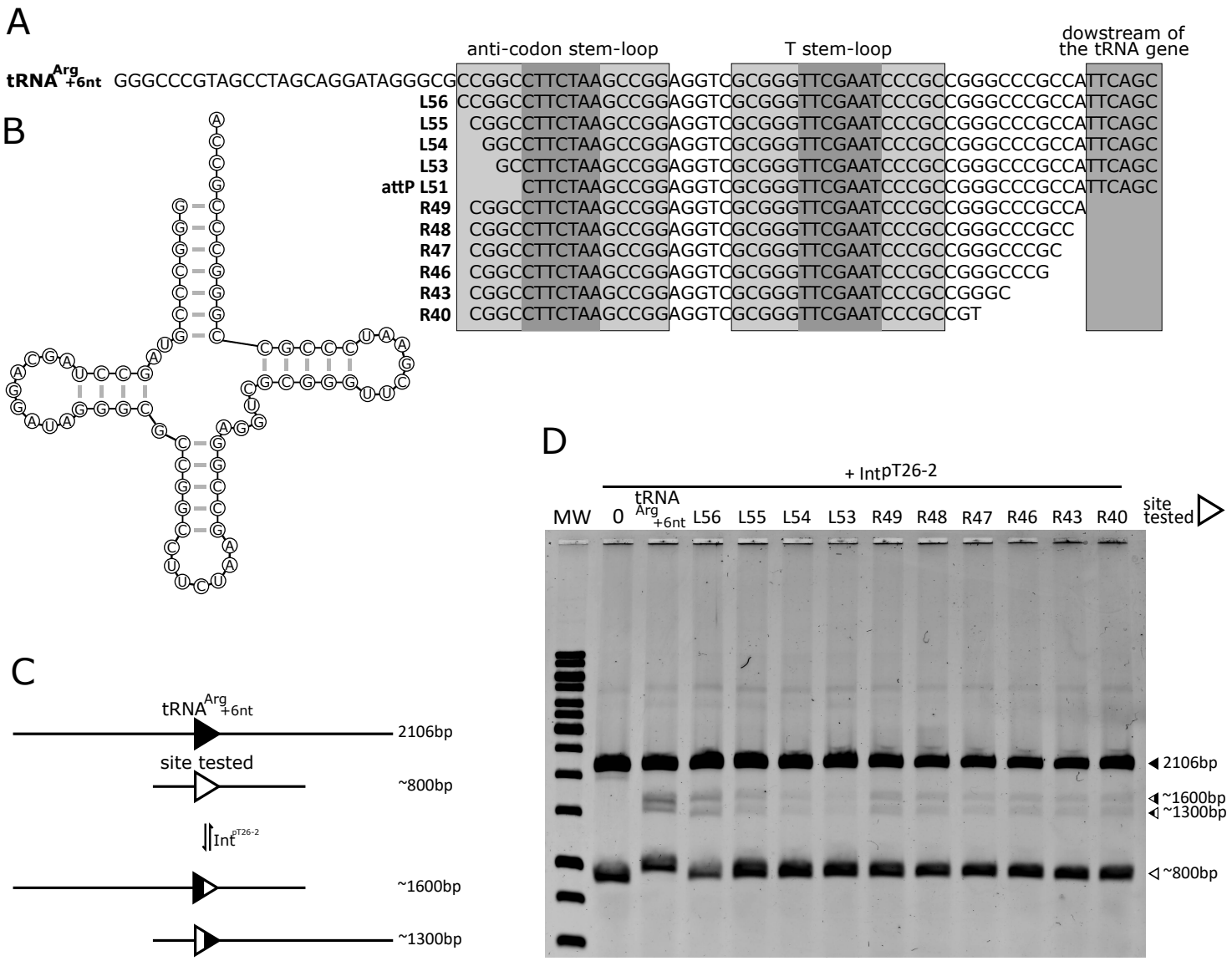

A
anti-codon stem-loop
T stem-loop
dowstream of the tRNA gene
tRNA Arg +6nt GGGCCCGTAGCCTAGCAGGATAGGGCGCCGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCCATTCAGC
L56 CCGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCCATTCAGC
L55 CGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCCATTCAGC
L54 GGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCCATTCAGC
L53 GCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCCATTCAGC
attP L51 CTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCCATTCAGC
R49 CGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCCA
R48 CGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCC
R47 CGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGC
R46 CGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCG
R43 CGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGC
R40 CGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGT
B
C
tRNA Arg +6nt
2106bp
site tested
~800bp
Int pT26-2
~1600bp
~1300bp
D
+ Int pT26-2
MW
0
tRNA Arg +6nt
L56
L55
L54
L53
R49
R48
R47
R46
R43
R40
site tested
2106bp
~1600bp
~1300bp
~800bp

Figure 5

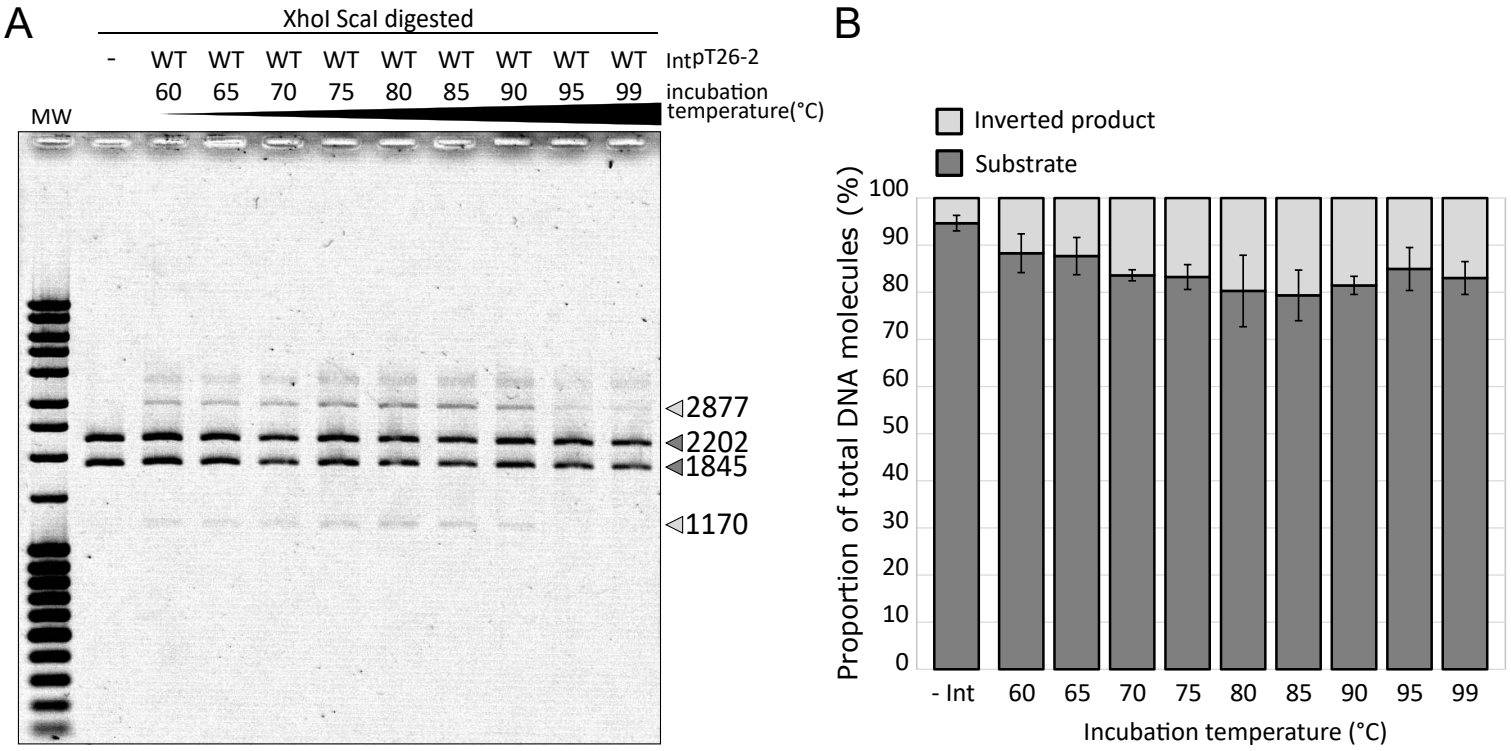
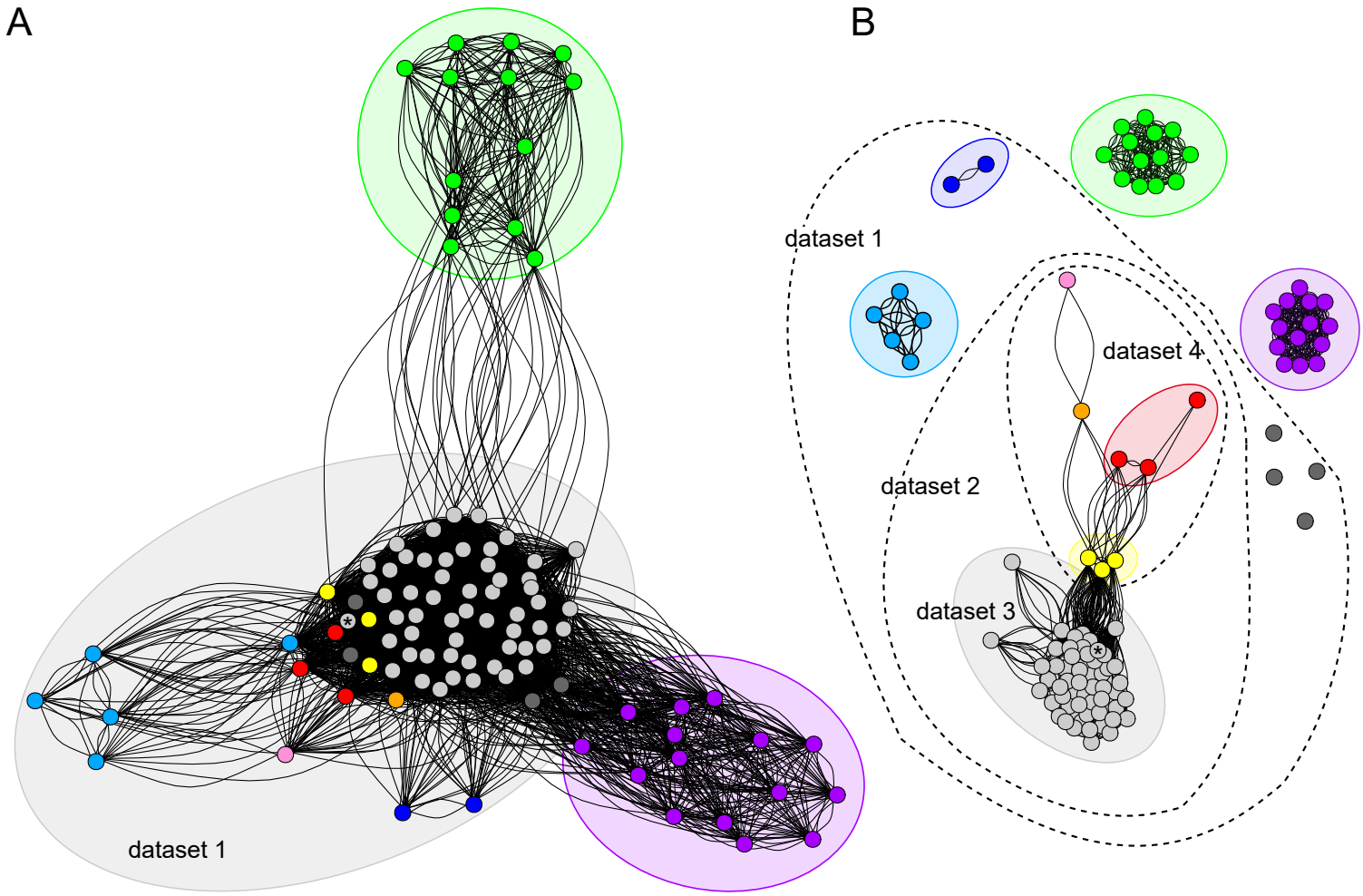

A
XhoI ScaI digested
- WT WT WT WT WT WT WT WT WT Int pT26-2
60 65 70 75 80 85 90 95 99 incubation temperature(°C)
MW
2877
2202
1845
1170
B
Inverted product
Substrate
Proportion of total DNA molecules (%)
100
90
80
70
60
50
40
30
20
10
0
- Int
60
65
70
75
80
85
90
95
99
Incubation temperature (°C)

Figure 6



- Thermococcales pT26-2-like integrases
- Integrases from AprDSM5631_IE2, _IE4 and _IE6
- Integrase from FplDSM10642_IE1
- Integrases from GacSBH6_IE2 and AprDSM5631_IE1 and _IE3
- Integrase from AfuDSM8774_IE1
- Integrases from AveSNP6_IE1 and _IE2, AprDSM5631_IE5 and GacSBH6_IE3
- Methanosarcinales integrases
- Integrases from GacSBH6_IE1 and Gah234T_IE1
- Thermococcales pTN3-like plasmid integrases
- Sulfolobales virus SSV integrases

Figure 7

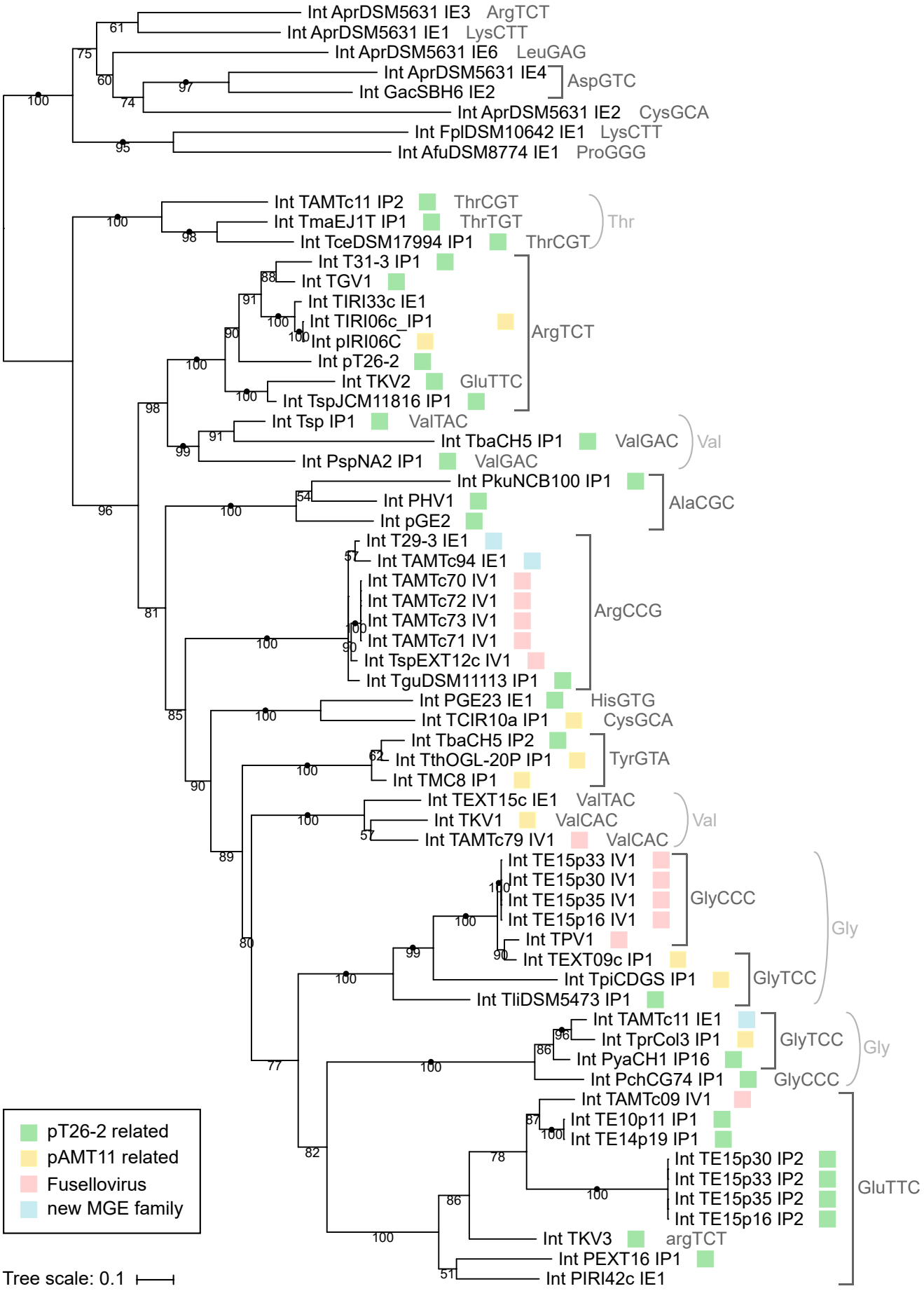

Int AprDSM5631 IE3 ArgTCT
Int AprDSM5631 IE1 LysCTT
Int AprDSM5631 IE6 LeuGAG
Int AprDSM5631 IE4
Int GacSBH6 IE2
AspGTC
Int AprDSM5631 IE2 CysGCA
Int FplDSM10642 IE1 LysCTT
Int AfuDSM8774 IE1 ProGGG
Int TAMTc11 IP2 ThrCGT
Int TmaEJ1T IP1 ThrTGT
Int TceDSM17994 IP1 ThrCGT
Thr
Int T31-3 IP1
Int TGV1
Int TIRI33c IE1
Int TIRI06c_IP1
Int pIRI06C
Int pT26-2
Int TKV2 GluTTC
Int TspJCM11816 IP1
ArgTCT
Int Tsp IP1 ValTAC
Int TbaCH5 IP1 ValGAC
Int PspNA2 IP1 ValGAC
Val
Int PkuNCB100 IP1
Int PHV1
Int pGE2
AlaCGC
Int T29-3 IE1
Int TAMTc94 IE1
Int TAMTc70 IV1
Int TAMTc72 IV1
Int TAMTc73 IV1
Int TAMTc71 IV1
Int TspEXT12c IV1
Int TguDSM11113 IP1
ArgCCG
Int PGE23 IE1 HisGTG
Int TCIR10a IP1 CysGCA
Int TbaCH5 IP2
Int TthOGL-20P IP1
Int TMC8 IP1
TyrGTA
Int TEXT15c IE1 ValTAC
Int TKV1 ValCAC
Int TAMTc79 IV1 ValCAC
Val
Int TE15p33 IV1
Int TE15p30 IV1
Int TE15p35 IV1
Int TE15p16 IV1
GlyCCC
Int TPV1
Int TEXT09c IP1
Int TpiCDGS IP1
Int TliDSM5473 IP1
GlyTCC
Gly
Int TAMTc11 IE1
Int TprCol3 IP1
Int PyaCH1 IP16
GlyTCC
Int PchCG74 IP1
GlyCCC
Gly
Int TAMTc09 IV1
Int TE10p11 IP1
Int TE14p19 IP1
Int TE15p30 IP2
Int TE15p33 IP2
Int TE15p35 IP2
Int TE15p16 IP2
GluTTC
Int TKV3 argTCT
Int PEXT16 IP1
Int PIRI42c IE1
pT26-2 related
pAMT11 related
Fusellovirus
new MGE family
Tree scale: 0.1

Figure 8

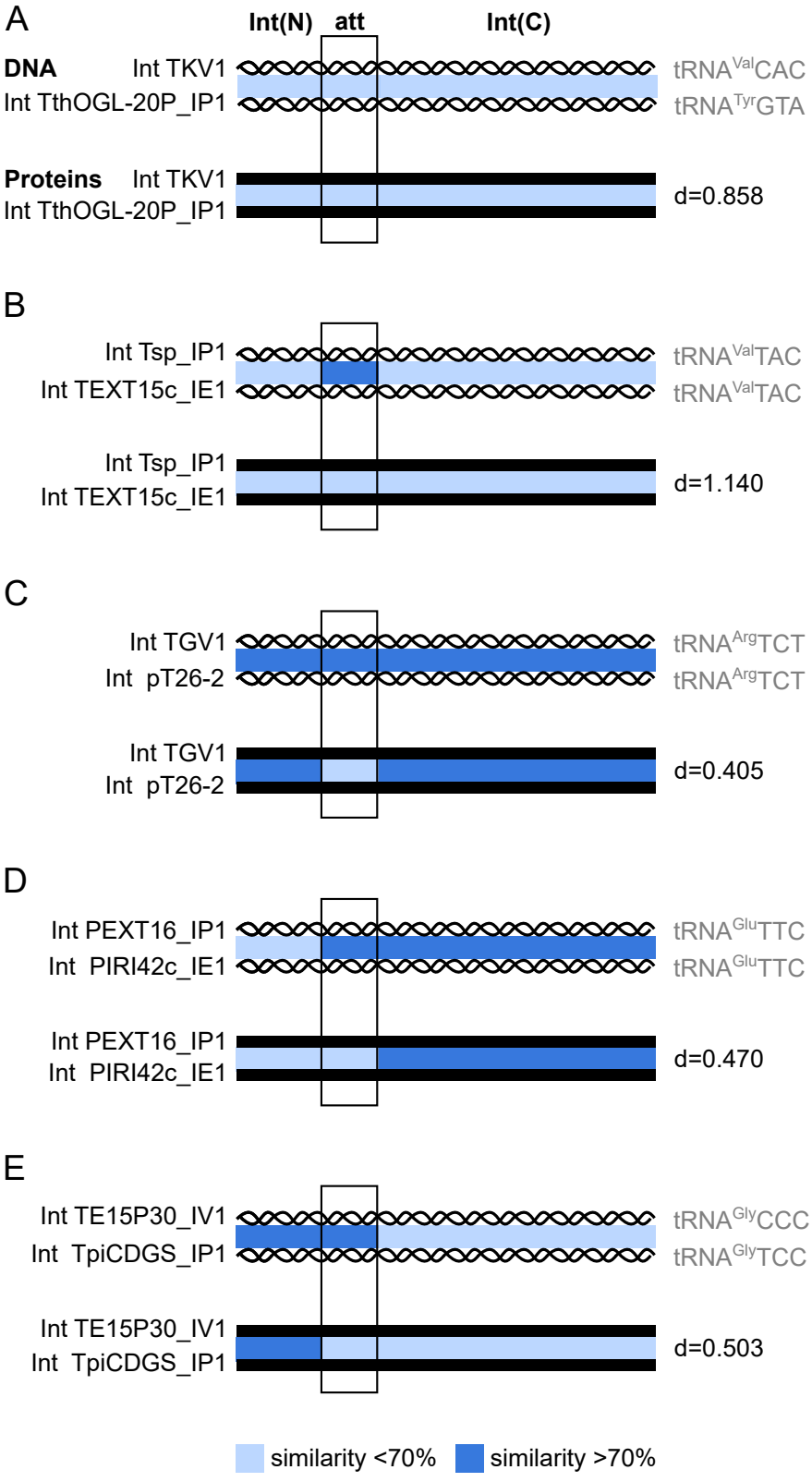

A
Int(N)
att
Int(C)
DNA
Int TKV1
Int TthOGL-20P_IP1
tRNA^ValCAC
tRNA^TyrGTA
Proteins
Int TKV1
Int TthOGL-20P_IP1
d=0.858
B
Int Tsp_IP1
Int TEXT15c_IE1
tRNA^ValTAC
tRNA^ValTAC
Int Tsp_IP1
Int TEXT15c_IE1
d=1.140
C
Int TGV1
Int pT26-2
tRNA^ArgTCT
tRNA^ArgTCT
Int TGV1
Int pT26-2
d=0.405
D
Int PEXT16_IP1
Int PIRI42c_IE1
tRNA^GluTTC
tRNA^GluTTC
Int PEXT16_IP1
Int PIRI42c_IE1
d=0.470
E
Int TE15P30_IV1
Int TpiCDGS_IP1
tRNA^GlyCCC
tRNA^GlyTCC
Int TE15P30_IV1
Int TpiCDGS_IP1
d=0.503
similarity <70%
similarity >70%

Figure 9

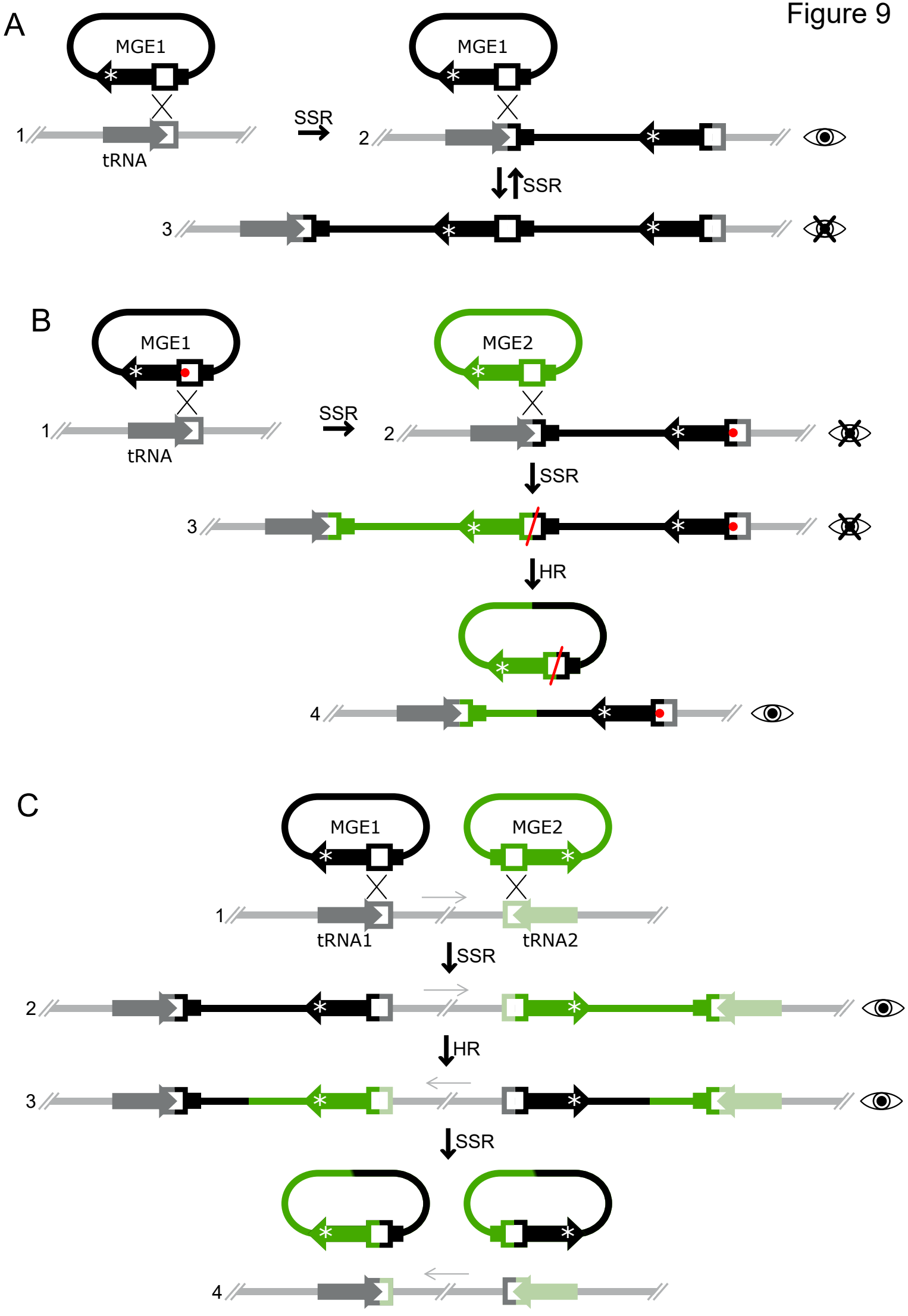

A
MGE1
tRNA
1
SSR
MGE1
2
SSR
3
B
MGE1
tRNA
1
SSR
MGE2
2
SSR
3
HR
4
C
MGE1
MGE2
tRNA1
tRNA2
1
SSR
2
HR
3
SSR
4

Table S1

element name	element type	element size (bp)	integration tRNA	host	optimal growth temperature	access
pGE2	pT26-2	20804	alaCGC	*Pyrococcus abyssi GE2*	ND	
PchCG74_IP1	pT26-2	21915	glyCCC	*Pyrococcus chitonophagus CG74*	85°C	NZ_CP015193.1
PEXT16_IP1	pT26-2	31892	gluTTC	*Pyrococcus EXT16*	ND	this study
PGE23_IE1	unknown	23334	hisGTG	*Pyrococcus GE23*	ND	this study
PHV1	pT26-2	21703	alaCGC	*Pyrococcus horikoshii OT3*	98°C	BA000001.2
PIRI42c_IE1	unknown	14352	gluTTC	*Pyrococcus IRI42C*	ND	this study
PkuNCB100_IP1	pT26-2	30387	alaCGC	*Pyrococcus kukulkanii NCB100*	105°C	CP010835
PspNA2_IP1	pT26-2	22133	valGTC	*Pyrococcus sp. NA2*	93°C	CP002670
PyaCH1_IP16	pT26-2	17518	glyTCC	*Pyrococcus yayanosii CH1*	98°C	CP002779
pT26-2	pT26-2	21610	argTCT	*Thermococcus 26-2*	ND	NC_014116.1
T29-3_IE1	group	12799	argCCG	*Thermococcus 29-3*	ND	this study
T31-3_IP1	pT26-2	25039	argTCT	*Thermococcus 31-3*	ND	this study
TAMTc09_IV1	TPV1	18832	gluTTC	*Thermococcus AMTc09*	ND	this study
TAMTc11_IP2	pT26-2	37864	thrCGT	*Thermococcus AMTc11*	ND	this study
TAMTc11_IE1	group	37864	glyTCC	*Thermococcus AMTc11*	ND	this study
TAMTc70_IV1	TPV1	23356	argCCG	*Thermococcus AMTc70*	ND	this study
TAMTc71_IV1	TPV1	23356	argCCG	*Thermococcus AMTc71*	ND	this study
TAMTc72_IV1	TPV1	23356	argCCG	*Thermococcus AMTc72*	ND	this study
TAMTc73_IV1	TPV1	23356	argCCG	*Thermococcus AMTc73*	ND	this study
TAMTc79_IV1	TPV1	22662	valCAC	*Thermococcus AMTc79*	ND	this study
TAMTc94_IE1	group	12899	argCCG	*Thermococcus AMTc94*	ND	this study
TbaCH5_IP1	pT26-2	18561	valGTC	*Thermococcus barophilus CH5*	ND	CP013050.1
TbaCH5_IP2	pT26-2	24493	tyrGTA	*Thermococcus barophilus CH5*	ND	CP013050.1
TceDSM17994_IP1	pT26-2	27699	thrCGT	*Thermococcus celericrescens DSM17994*	80°C	NZ_LLYW01000013
TCIR10a_IP1	pAMTc11	30801	cysGCA	*Thermococcus CIR10a*	ND	this study
TE10p11_IP1	pT26-2	26949	gluTTC	*Thermococcus E10p11*	ND	this study
TE14p19_IP1	pT26-2	26949	gluTTC	*Thermococcus E14P19*	ND	this study
TE15p16_IP2	pT26-2	24956	gluTTC	*Thermococcus E15P16*	ND	this study
TE15p16_IV1	TPV1	25057	glyCCC	*Thermococcus E15P16*	ND	this study
TE15p30_IP2	pT26-2	24956	gluTTC	*Thermococcus E15P30*	ND	this study
TE15p30_IV1	TPV1	25057	glyCCC	*Thermococcus E15P30*	ND	this study
TE15p33_IP2	pT26-2	24956	gluTTC	*Thermococcus E15p33*	ND	this study
TE15p33_IV1	TPV1	25057	glyCCC	*Thermococcus E15p33*	ND	this study
TE15p35_IP2	pT26-2	24956	gluTTC	*Thermococcus E15P35*	ND	this study
TE15p35_IV1	TPV1	25057	glyCCC	*Thermococcus E15P35*	ND	this study
TEXT09c_IP1	pAMTc11	28832	glyTCC	*Thermococcus EXT09c*	ND	this study
TEXT15c_IE1	unknown	18904	valTAC	*Thermococcus EXT15c*	ND	this study
TGV1	pT26-2	20793	argTCT	*Thermococcus gammatolerans EJ3*	88°C	CP001398
TguDSM11113_IP1	pT26-2	26701	argCCG	*Thermococcus guaymasensis DSM11113*	88°C	CPU007140.1
TIRI06c_IP1	pAMTc11	21855	argTCT	*Thermococcus IRI06C*	ND	this study
TIRI33c_IE1	unknown	17903	argTCT	*Thermococcus IRI33c*	ND	this study
TKV1	pAMTc11	23400	valCAC	*Thermococcus kodakarensis KOD1*	85°C	AP006878.1
TKV2	pT26-2	27112	gluTTC	*Thermococcus kodakarensis KOD1*	85°C	AP006878.1
TKV3	pT26-2	27581	argTCT	*Thermococcus kodakarensis KOD1*	85°C	AP006878.1
TliDSM5473_IP1	pT26-2	22524	glyTCC	*Thermococcus litoralis DSM5473*	88°C	CP006670.1
TmaEJ1T_IP1	pT26-2	24986	thrTGT	*Thermococcus marinus EJ1T*	88°C	this study
TMC8_IP1	pAMTc11	18910	tyrGTA	*Thermococcus MC8*	ND	this study
TpiCDGS_IP1	pAMTc11	18406	glyTCC	*Thermococcus piezophilus CDGS*	75°C	NZ_CP015520.1
TPV1	TPV1	21625	glyCCC	*Thermococcus prieuri chr1*	80°C	NZ_CP015193.1
TprCol3_IP1	pAMTc11	22198	glyTCC	*Thermococcus prieuri chr2*	80°C	this study
Tsp_IP1	pT26-2	23902	valTAC	*Thermococcus sp.*	ND	this study
TspEXT12c_IV1	TPV1	21307	argCCG	*Thermococcus sp. EXT12C*	85°C	NZ_LT900021.1
TspJCM11816_IP1	pT26-2	23440	argTCT	*Thermococcus sp. JCM11816*	85°C	Ga0128353_102
TthOGL-20P_IP1	pAMTc11	29542	tyrGTA	*Thermococcus thioreducens OGL-20P*	83-85°C	NZ_CP015105.1
AfuDSM8774_IE1	unknown	11760	proGGG	Archaeoglobus fulgidus DSM 8774	76°C	NZ_CP006577
AprDSM5631_IE1	unknown	10953	lysCTT	Archaeoglobus profundus DSM 5631	82°C	NC_013741.1
AprDSM5631_IE2	unknown	9715	cysGCA	Archaeoglobus profundus DSM 5631	82°C	NC_013741.1
AprDSM5631_IE3	unknown	7230	argTCT	Archaeoglobus profundus DSM 5631	82°C	NC_013741.1
AprDSM5631_IE4	unknown	10052	aspGTC	Archaeoglobus profundus DSM 5631	82°C	NC_013741.1
AprDSM5631_IE5	unknown	28635	gluCTC	Archaeoglobus profundus DSM 5631	82°C	NC_013741.1
AprDSM5631_IE6	unknown	38408	leuGAG	Archaeoglobus profundus DSM 5631	82°C	NC_013741.1
AveSNP6_IE1	unknown	13345	leuGAG	Archaeoglobus veneficus SNP6	80°C	NC_015320
AveSNP6_IE2	unknown	9698	serTGA	Archaeoglobus veneficus SNP6	80°C	NC_015320
FplDSM10642_IE1	unknown	8680	lysCTT	Ferroglobus placidus DSM 10642	85°C	NC_013849
GacSBH6_IE2	unknown	8255	aspGTC	Geoglobus acetivorans SBH6	81°C	NZ_CP009552
GacSBH6_IE3	unknown	29550	gluTTC	Geoglobus acetivorans SBH6	81°C	NZ_CP009552
GacSBH6_IE1	unknown	8657	pheGAA	Geoglobus acetivorans SBH6	81°C	NZ_CP009552
Gah234T_IE1	unknown	18313	pheGAA	Geoglobus ahangari 234T	88 °C	NZ_CP011267
MmeMM1_IE1	unknown	20170	valTAC	Methanococcoides methylutens MM1	30-35°C	NZ_CP009518
MbaCM1_IE1	unknown	19570	valTAC	Methanosarcina bakeri CM1	40°C	NZ_CP008746
MmaC16_IE1	unknown	15206	valTAC	Methanosarcina mazei C16	36°C	NZ_CP009514
MmaWWM610_IE1	unknown	15570	valTAC	Methanosarcina mazei WWM610	NI	NZ_CP009509
MspWWM596_IE1	unknown	15873	valCAC	Methanosarcina sp. WWM596	NI	NZ_CP009503

Table S2. Plasmids used in this work

Name	Backbone	Insertion	Selection
pCB558	pET-26b(+)	Wild-type $Int^{pT26\text{-}2}$ gene	KmR
pCB616	pET-26b(+)	$Int^{pT26\text{-}2}$Y327F gene	KmR
pUC18	-	-	ApR
pCB568	pUC18	$tRNA^{arg}$ gene and the 6 nucleotides downtreamfrom *T. sp. 26-2*	ApR
pUC4K	pUC18		ApR, KmR
pCB596	pUC18	Direct repetition of the pCB568 insert	ApR, KmR
pCB598	pUC18	Reverse repetition of the pCB568 insert	ApR, KmR
pCB590	pUC18	L56	ApR
pCB588	pUC18	L55	ApR
pCB586	pUC18	L54	ApR
pCB584	pUC18	L53	ApR
pCB602	pUC18	R49	ApR
pCB604	pUC18	R48	ApR
pCB630	pUC18	R47	ApR
pCB632	pUC18	R46	ApR
pCB636	pUC18	R43	ApR
pCB638	pUC18	R40	ApR

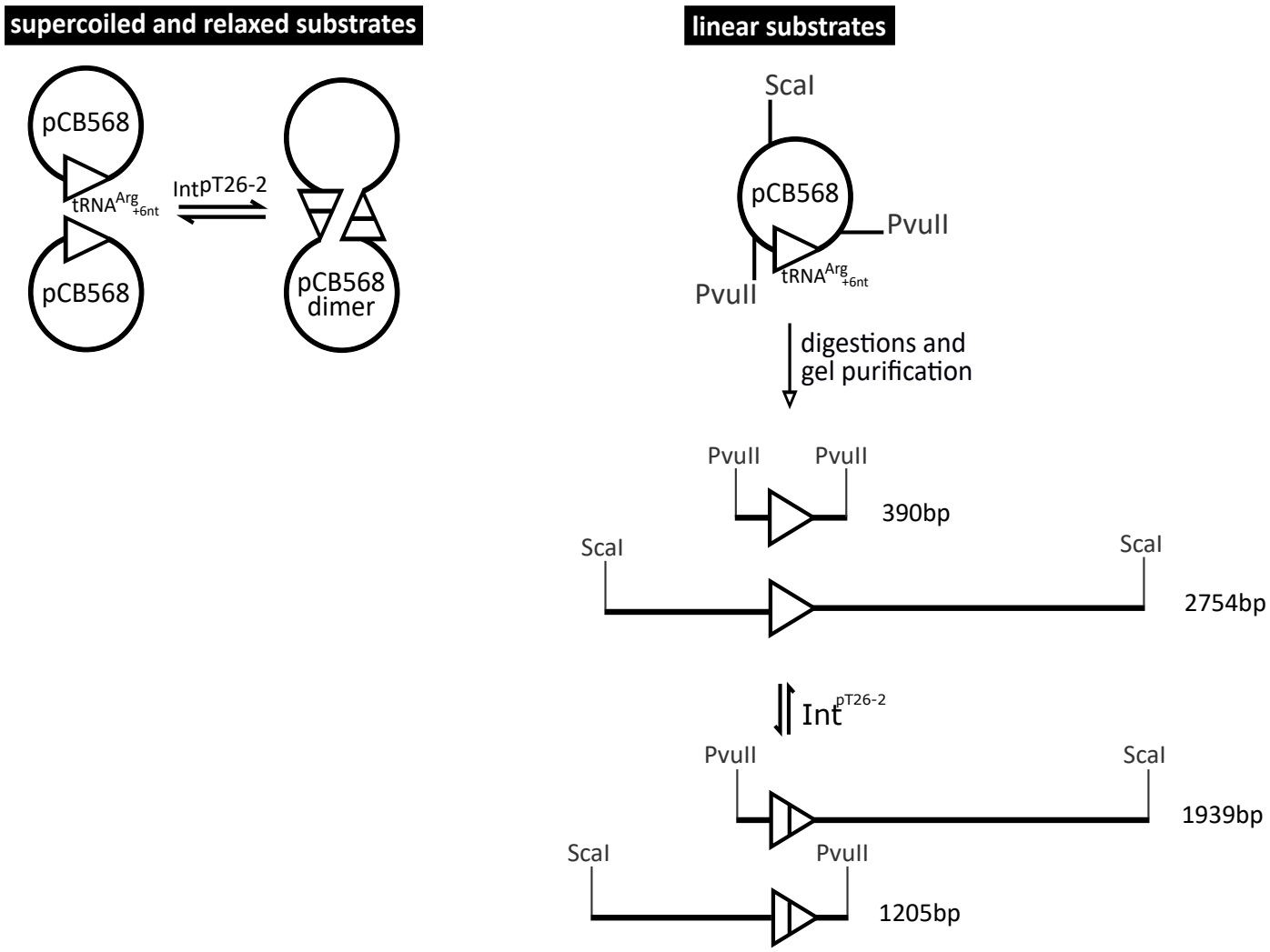
Table S3. Oligonucleotides used in this work

Name	PCR-Template	Sequence	Usage
pT26-2_F	pT26-2	TTAAGAAGGAGATATACATATGTGGAGCCACCCGCAGTTCGAAAAAGGCCGTAGTGGGCCGTC	Gibson assembly
pT26-2_R	pT26-2	CAGTGGTGGTGGTGGTGGTGCTCGAGTGCGGCCGCTCATTTTTCTAAAACCTTCCTCAAGTC	Gibson assembly
pT26-2_Y327F_F	558	TCTTGTTTAGAAAGTGCGTCGCA	Mutagenesis
pT26-2_Y327F_R	558	CTCTCCTGGCCGATGAAT	Mutagenesis
BamHI-tRNAarg+6-EcoRI_A	-	cgGGATCCGGGCCCGTAGCCTAGCAGGATAGGGCGCCGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCCATTCAGCGAATTCggtc	
BamHI-tRNAarg+6-EcoRI_B	-	gaccGAATTCGCTGAATGGCGGGCCCGGCGGGATTCGAACCCGCGACCTCCGGCTTAGAAGGCCGGCGCCCTATCCTGCTAGGCTACGGGCCCGGATCCcg	
BamHI-L56-coRI_A	-	gtacAAGCTTCCGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCCATTCAGCGAATTCggtc	
BamHI-L56-coRI_B	-	gaccGAATTCGCTGAATGGCGGGCCCGGCGGGATTCGAACCCGCGACCTCCGGCTTAGAAGGCCGGAAGCTTgtac	
BamHI-L55-coRI_A	-	gtacAAGCTTCGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCCATTCAGCGAATTCggtc	
BamHI-L55-coRI_B	-	gaccGAATTCGCTGAATGGCGGGCCCGGCGGGATTCGAACCCGCGACCTCCGGCTTAGAAGGCCGAAGCTTgtac	
BamHI-L53-coRI_A	-	gtacAAGCTTGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCCATTCAGCGAATTCggtc	
BamHI-L53-coRI_B	-	gaccGAATTCGCTGAATGGCGGGCCCGGCGGGATTCGAACCCGCGACCTCCGGCTTAGAAGGCAAGCTTgtac	
KanR-ex1	pUC4K	ggcccgccattcagcCGCTGAGGTCTGCCTCGT	Gibson assembly
KanR-ex2	pUC4K	gtgcggtatttcacaccgcaAAAGCCACGTTGTGTCTCAAAATC	Gibson assembly
tRNAarg+6-ex1	pCB568	attgtactgagagtgcaccaGGGCCCGTAGCCTAGCAG	Gibson assembly
tRNAarg+6-ex2	pCB568	attgtactgagagtgcaccaGGGCCCGTAGCCTAGCAG	Gibson assembly
KanR-inv1	pUC4K	ctaggctacgggcccCGCTGAGGTCTGCCTCGT	Gibson assembly

KanR-inv2	pUC4K	gtgcggtatttcacaccgcaAAAGCCACGTTGTGTCTCAAAATC	Gibson assembly
tRNAarg+6-inv1	pCB568	attgtactgagagtgcaccaGGGCCCGTAGCCTAGCAG	Gibson assembly
tRNAarg+6-inv2	pCB568	attgtactgagagtgcaccaGGGCCCGTAGCCTAGCAG	Gibson assembly
pUC18-H_FOR	pUC18	ATGCAAGCTTGGCACTGGCCG	
L54-pUC18_REV	pUC18	GATAAGCTTGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCCATTCAGCGAATTCGTAATCATGGTCATAGCTG	
R49-pUC18_REV	pUC18	GATAAGCTTCGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCCAGAATTCGTAATCATGGTCATAGCTG	
R48-pUC18_REV	pUC18	GATAAGCTTCGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCCGAATTCGTAATCATGGTCATAGCTG	
R47-pUC18_REV	pUC18	GATAAGCTTCGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGCGAATTCGTAATCATGGTCATAGCTG	
R46-pUC18_REV	pUC18	GATAAGCTTCGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCCCGGAATTCGTAATCATGGTCATAGCTG	
R43-pUC18_REV	pUC18	GATAAGCTTCGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGGGCGAATTCGTAATCATGGTCATAGCTG	
R40-pUC18_REV	pUC18	GATAAGCTTCGGCCTTCTAAGCCGGAGGTCGCGGGTTCGAATCCCGCCGTGAATTCGTAATCATGGTCATAGCTG	
pUC1481-1503	pUC18	GGCCGCAGTGTTATCACTCATGG	
P30-REV	pUC18	CTGCGCGTAATCTGCTGCTTGC	
pUC195-217	pUC18	GAAATACCGCACAGATGCGTAAG	
pZE21_rev	pUC18	GGGTTTCGCCACCTCTGACTTG	

Figure S1

A


supercoiled and relaxed substrates
pCB568
tRNA^Arg +6nt
Int^pT26-2
pCB568
pCB568 dimer
linear substrates
ScaI
pCB568
PvuII
tRNA^Arg +6nt
PvuII
digestions and gel purification
PvuII
PvuII
390bp
ScaI
ScaI
2754bp
Int^pT26-2
PvuII
ScaI
1939bp
ScaI
PvuII
1205bp

B

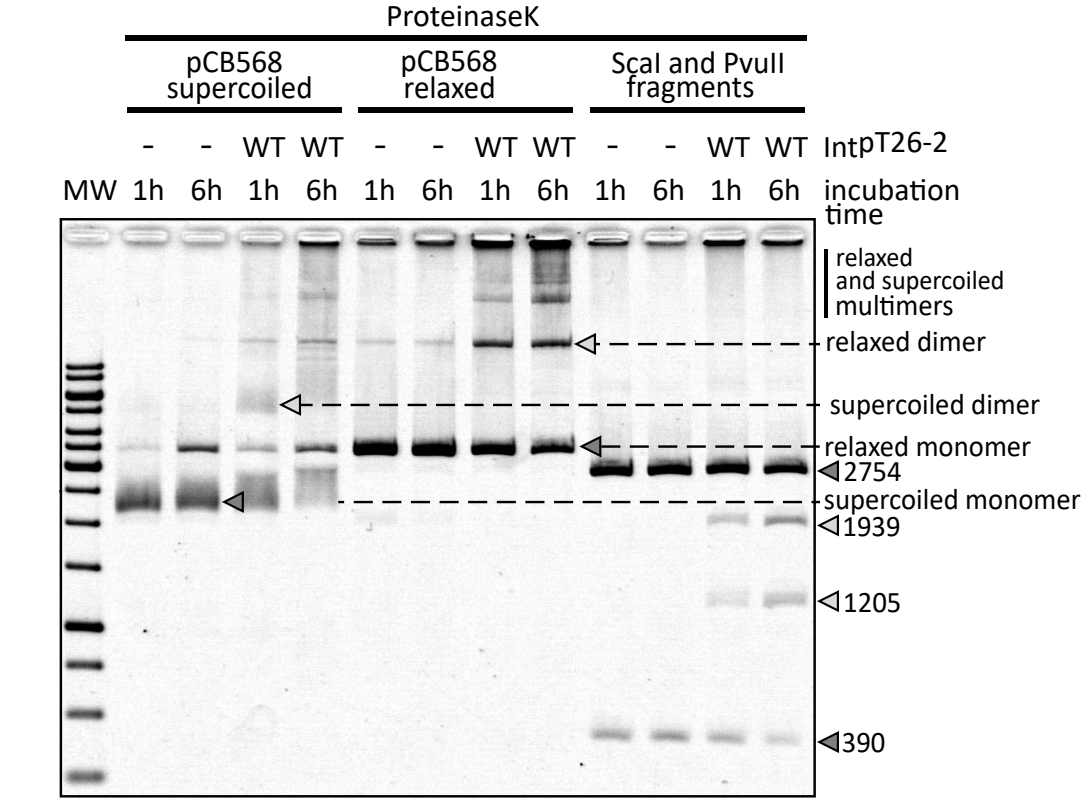

ProteinaseK
pCB568 supercoiled
pCB568 relaxed
ScaI and PvuII fragments
- - WT WT - - WT WT - - WT WT Int^pT26-2
MW 1h 6h 1h 6h 1h 6h 1h 6h 1h 6h 1h 6h incubation time
relaxed and supercoiled multimers
relaxed dimer
supercoiled dimer
relaxed monomer
2754
supercoiled monomer
1939
1205
390

Figure S2

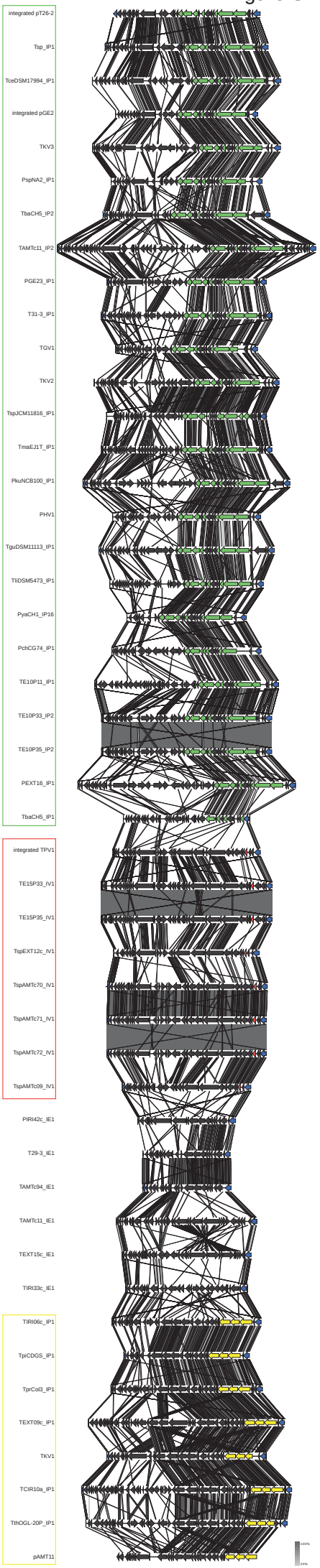

integrated pT26-2
Tsp_IP1
TceDSM17994_IP1
integrated pGE2
TKV3
PspNA2_IP1
TbaCH5_IP2
TAMTc11_IP2
PGE23_IP1
T31-3_IP1
TGV1
TKV2
TspJCM11816_IP1
TmaEJ1T_IP1
PkuNCB100_IP1
PHV1
TguDSM11113_IP1
TliDSM5473_IP1
PyaCH1_IP16
PchCG74_IP1
TE10P11_IP1
TE10P33_IP2
TE10P35_IP2
PEXT16_IP1
TbaCH5_IP1
integrated TPV1
TE15P33_IV1
TE15P35_IV1
TspEXT12c_IV1
TspAMTc70_IV1
TspAMTc71_IV1
TspAMTc72_IV1
TspAMTc09_IV1
PIRI42c_IE1
T29-3_IE1
TAMTc94_IE1
TAMTc11_IE1
TEXT15c_IE1
TIRI33c_IE1
TIRI06c_IP1
TpiCDGS_IP1
TprCol3_IP1
TEXT09c_IP1
TKV1
TCIR10a_IP1
TlhOGL-20P_IP1
pAMT11

Figure S3

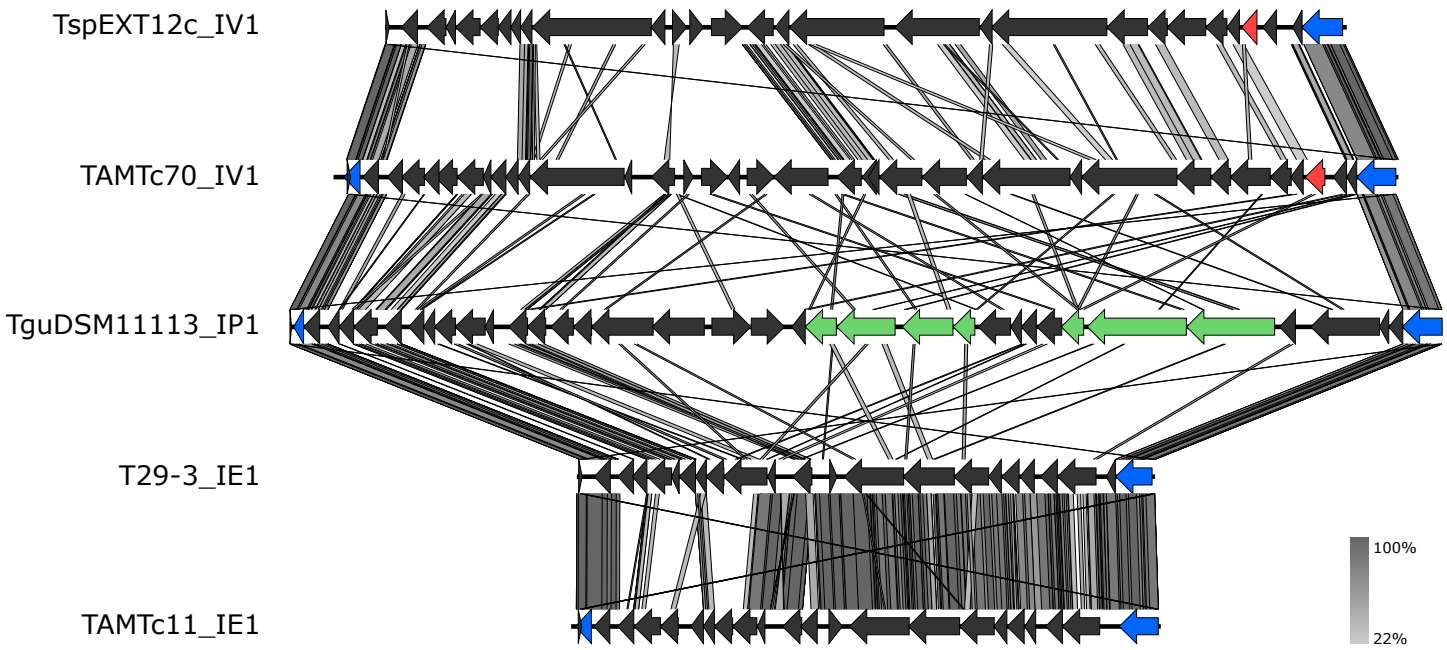

TspEXT12c_IV1
TAMTc70_IV1
TguDSM11113_IP1
T29-3_IE1
TAMTc11_IE1
100%
22%

Figure S4

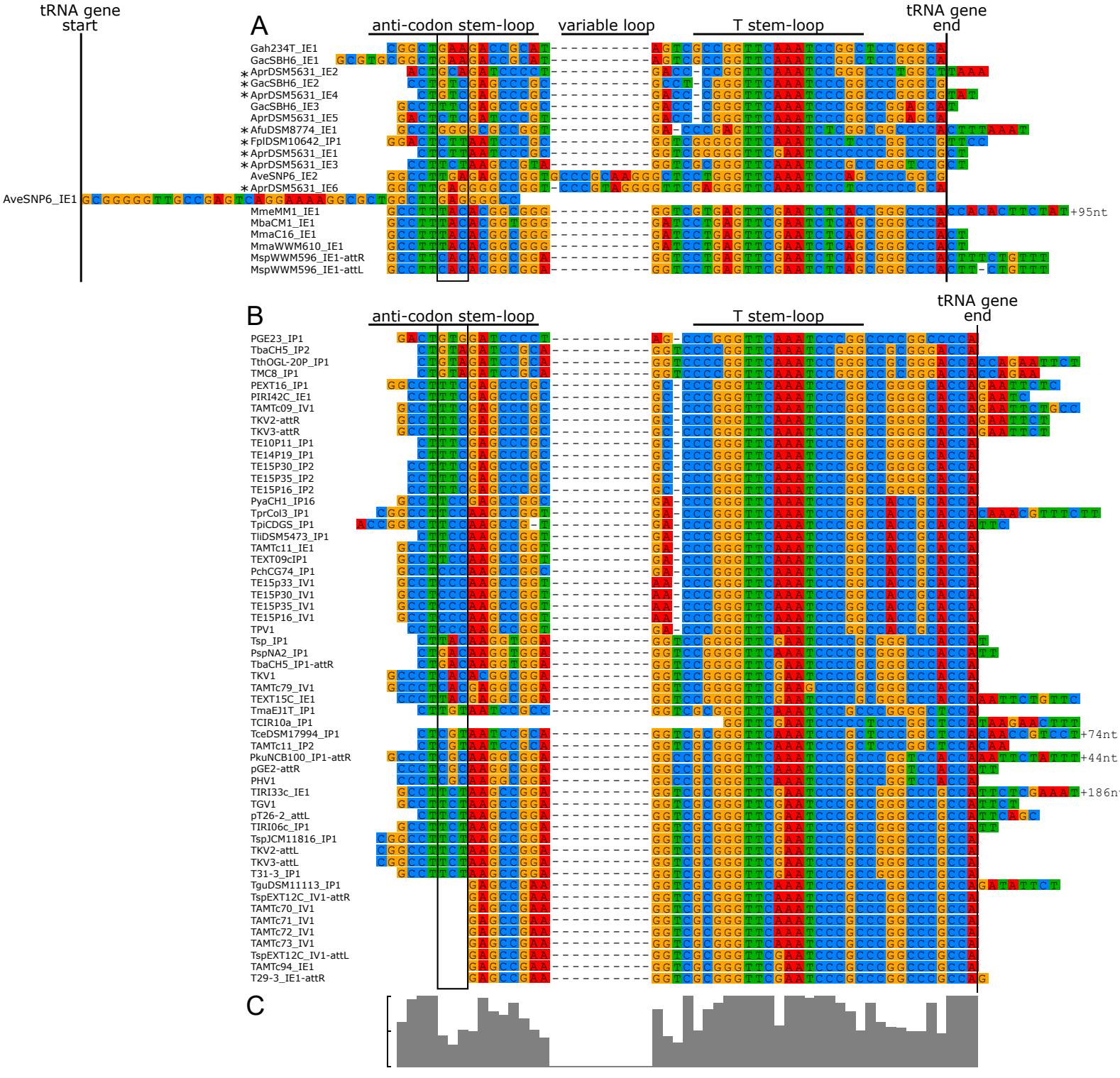
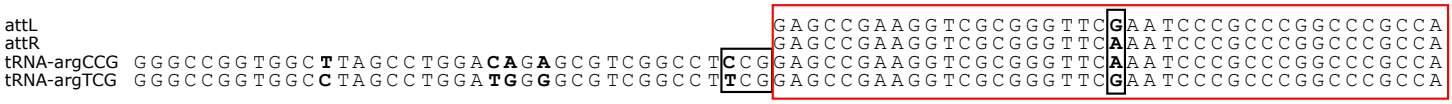
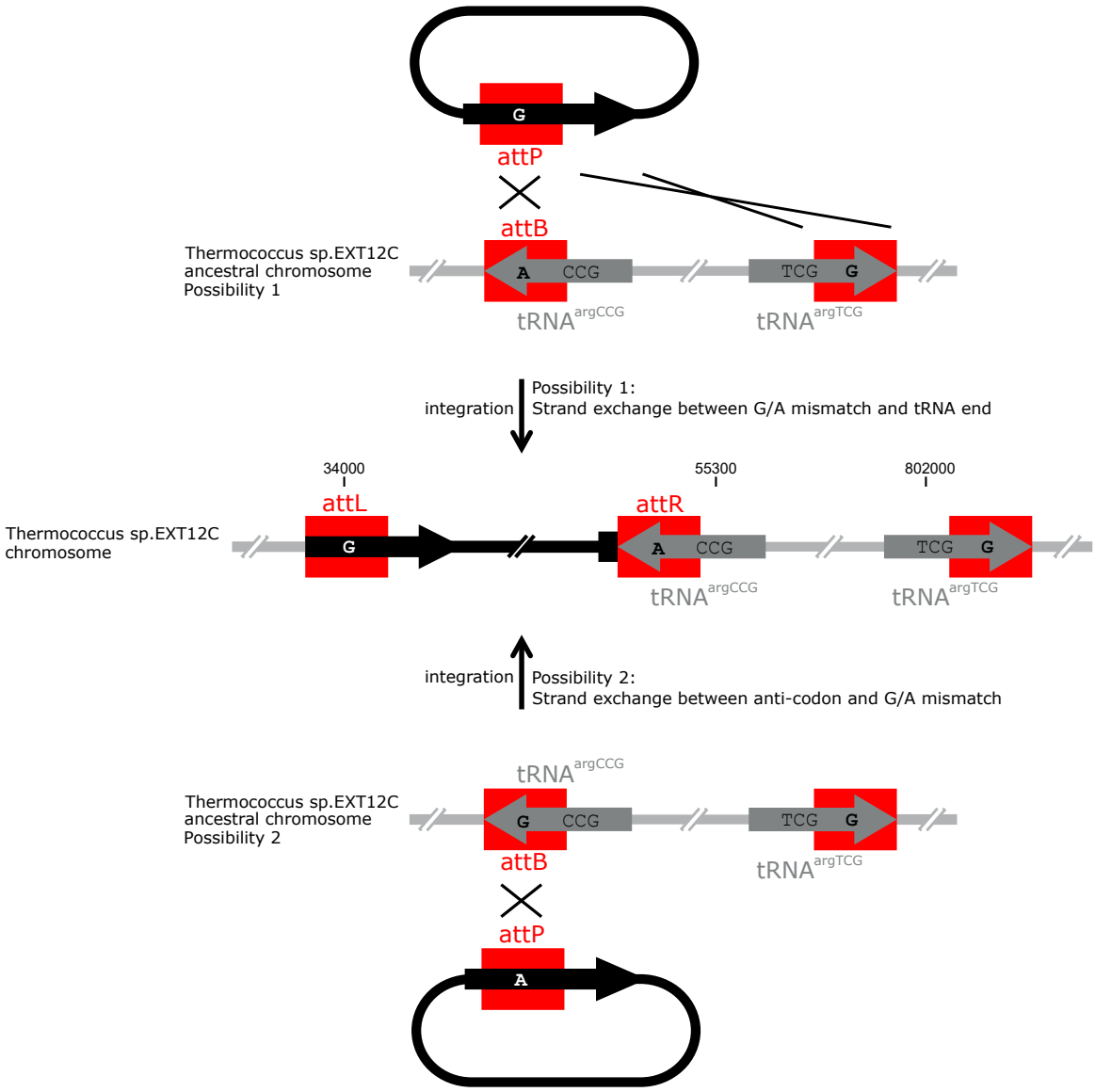

tRNA gene start
A
anti-codon stem-loop
variable loop
T stem-loop
tRNA gene end
Gah234T_IE1
GacSBH6_IE1
*AprDSM5631_IE2
*GacSBH6_IE2
*AprDSM5631_IE4
GacSBH6_IE3
AprDSM5631_IE5
*AfuDSM8774_IE1
*FplDSM10642_IP1
*AprDSM5631_IE1
*AprDSM5631_IE3
AveSNP6_IE2
*AprDSM5631_IE6
AveSNP6_IE1
MmeMM1_IE1
MbaCM1_IE1
MmaC16_IE1
MmaWWM610_IE1
MspWWM596_IE1-attR
MspWWM596_IE1-attL
+95nt
B
anti-codon stem-loop
T stem-loop
tRNA gene end
PGE23_IP1
TbaCH5_IP2
TthOGL-20P_IP1
TMC8_IP1
PEXT16_IP1
PIRI42C_IE1
TAMTc09_IV1
TKV2-attR
TKV3-attR
TE10P11_IP1
TE14P19_IP1
TE15P30_IP2
TE15P35_IP2
TE15P16_IP2
PyaCH1_IP16
TprCol3_IP1
TpiCDGS_IP1
TliDSM5473_IP1
TAMTc11_IE1
TEXT09cIP1
PchCG74_IP1
TE15p33_IV1
TE15P30_IV1
TE15P35_IV1
TE15P16_IV1
TPV1
Tsp_IP1
PspNA2_IP1
TbaCH5_IP1-attR
TKV1
TAMTc79_IV1
TEXT15C_IE1
TmaEJ1T_IP1
TCIR10a_IP1
TceDSM17994_IP1
TAMTc11_IP2
PkuNCB100_IP1-attR
pGE2-attR
PHV1
TIRI33c_IE1
TGV1
pT26-2_attL
TIRI06c_IP1
TspJCM11816_IP1
TKV2-attL
TKV3-attL
T31-3_IP1
TguDSM11113_IP1
TspEXT12C_IV1-attR
TAMTc70_IV1
TAMTc71_IV1
TAMTc72_IV1
TAMTc73_IV1
TspEXT12C_IV1-attL
TAMTc94_IE1
T29-3_IE1-attR
+74nt
+44nt
+186nt
C

Figure S5

A



B



C

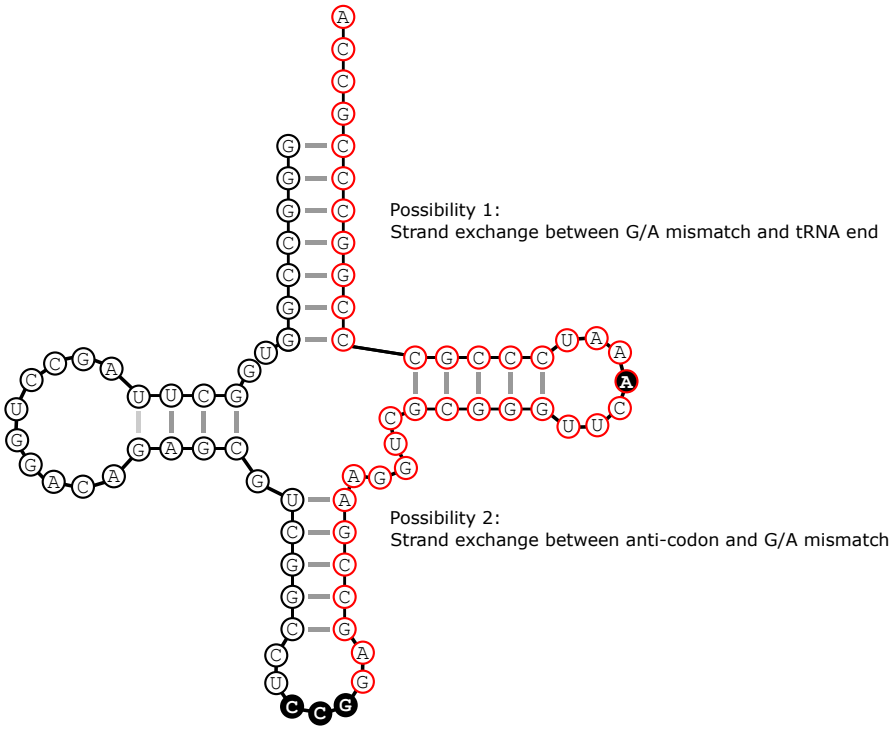


Figure S6

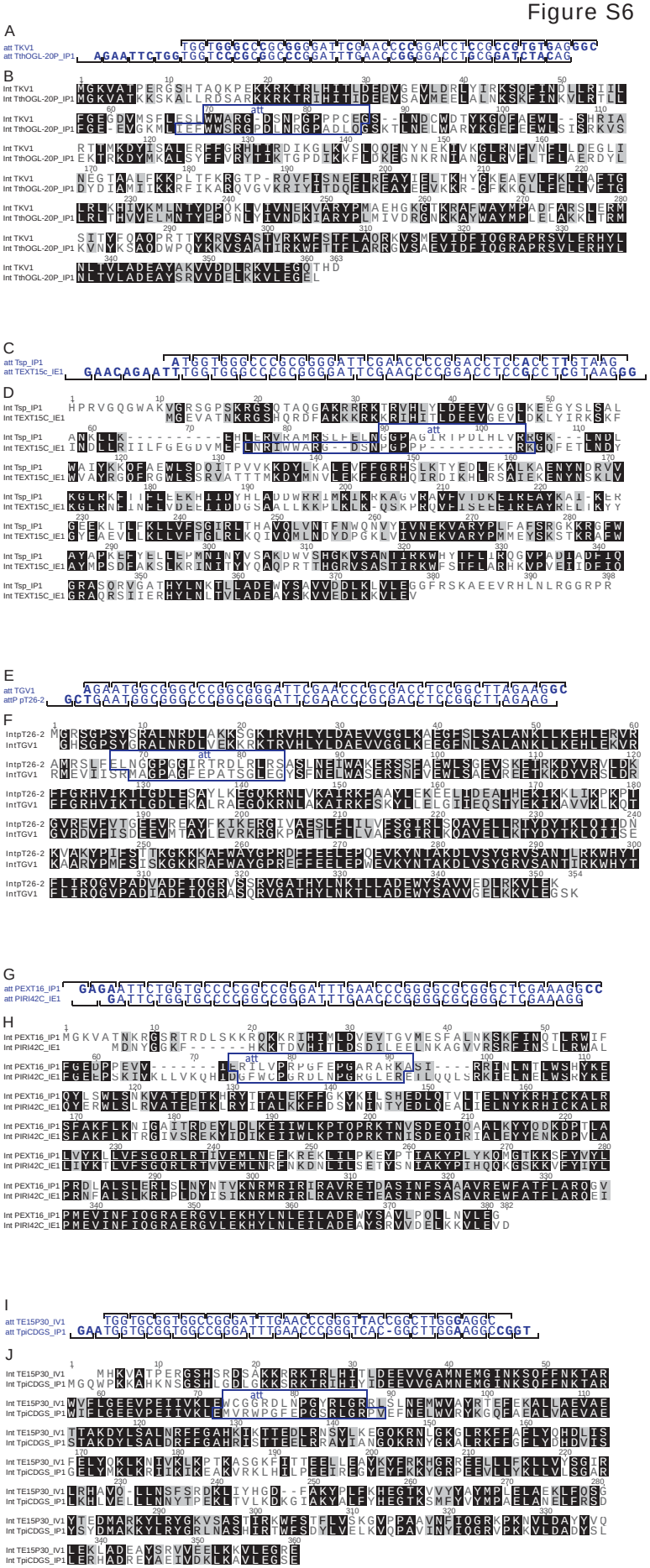
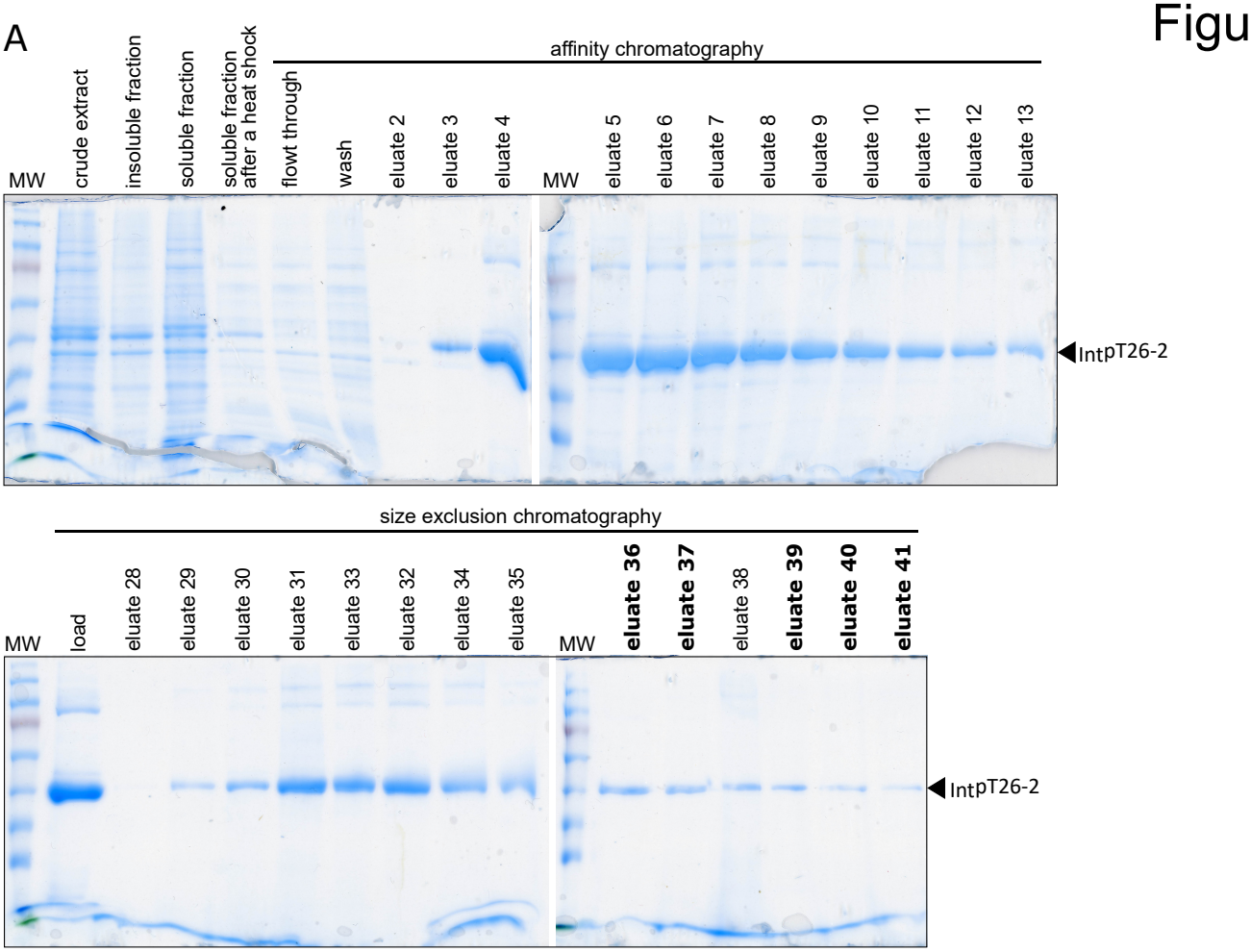
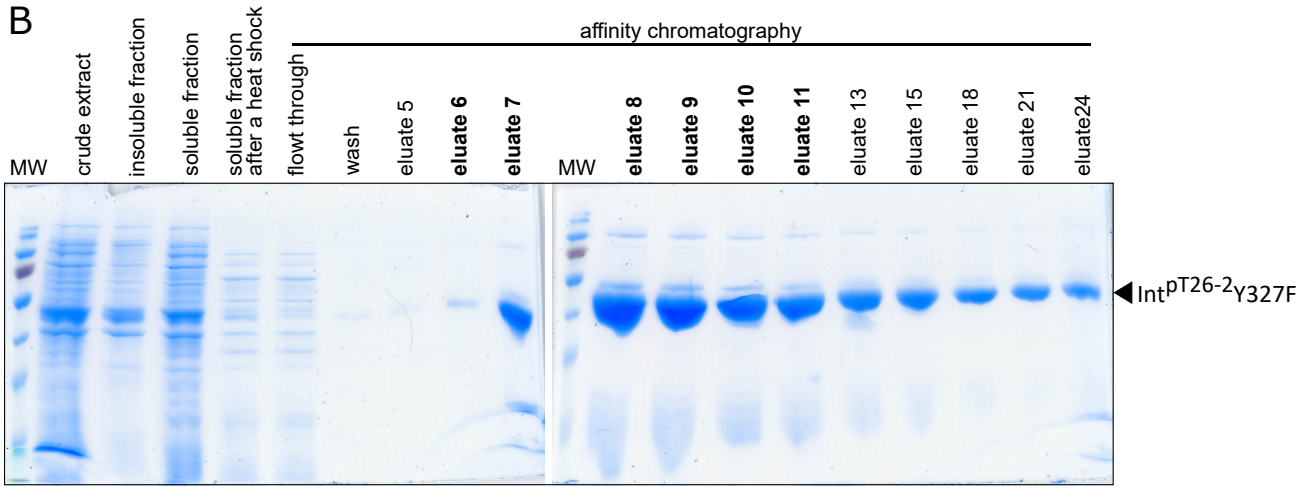
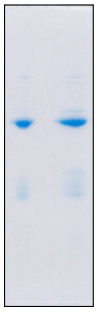


Figure S7


A
affinity chromatography
MW
crude extract
insoluble fraction
soluble fraction
soluble fraction after a heat shock
flowt through
wash
eluate 2
eluate 3
eluate 4
MW
eluate 5
eluate 6
eluate 7
eluate 8
eluate 9
eluate 10
eluate 11
eluate 12
eluate 13
IntpT26-2
size exclusion chromatography
MW
load
eluate 28
eluate 29
eluate 30
eluate 31
eluate 33
eluate 32
eluate 34
eluate 35
MW
eluate 36
eluate 37
eluate 38
eluate 39
eluate 40
eluate 41
IntpT26-2


B
affinity chromatography
MW
crude extract
insoluble fraction
soluble fraction
soluble fraction after a heat shock
flowt through
wash
eluate 5
eluate 6
eluate 7
MW
eluate 8
eluate 9
eluate 10
eluate 11
eluate 13
eluate 15
eluate 18
eluate 21
eluate24
IntpT26-2Y327F


C
IntpT26-2
IntpT26-2Y327F

The annotated Int^{TPV1} is a truncated and inactive mutant

Given the important diversity of the integrases related to Int$^{pT26\text{-}2}$, we wanted to characterize a second protein. Once again, we favored an integrase encoded by an episomal element that is less likely to accumulate deleterious mutations. We chose the integrase carried by the virus TPV1 (Int^{TPV1}) that is more distant from Int$^{pT26\text{-}2}$ than the integrases carried by the plasmids pGE2 and pIRI06c (Article 3).

We implemented the same strategy as for Int$^{pT26\text{-}2}$ for the production, the purification and the characterization of Int^{TPV1}. The site-specific activity was assayed in an inversion assay (Figure 31.A). After incubation with Int^{TPV1}, no restriction pattern corresponding to the inverted product was detected (Figure 31.B). The purified enzyme Int^{TPV1} cannot catalyze *in vitro* site-specific inversion in the tested conditions. Int^{TPV1} was also inactive *in vitro* for site-specific recombinations in various conditions and in integration and excision assays (data not shown). We can consider two explanations for the absence of activity. Either the conditions for activity were not found with for example a necessary co-factor or the purified protein truly is inactive. The first hypothesis is less probable since the closely related integrase Int$^{pT26\text{-}2}$ was active in similar conditions.

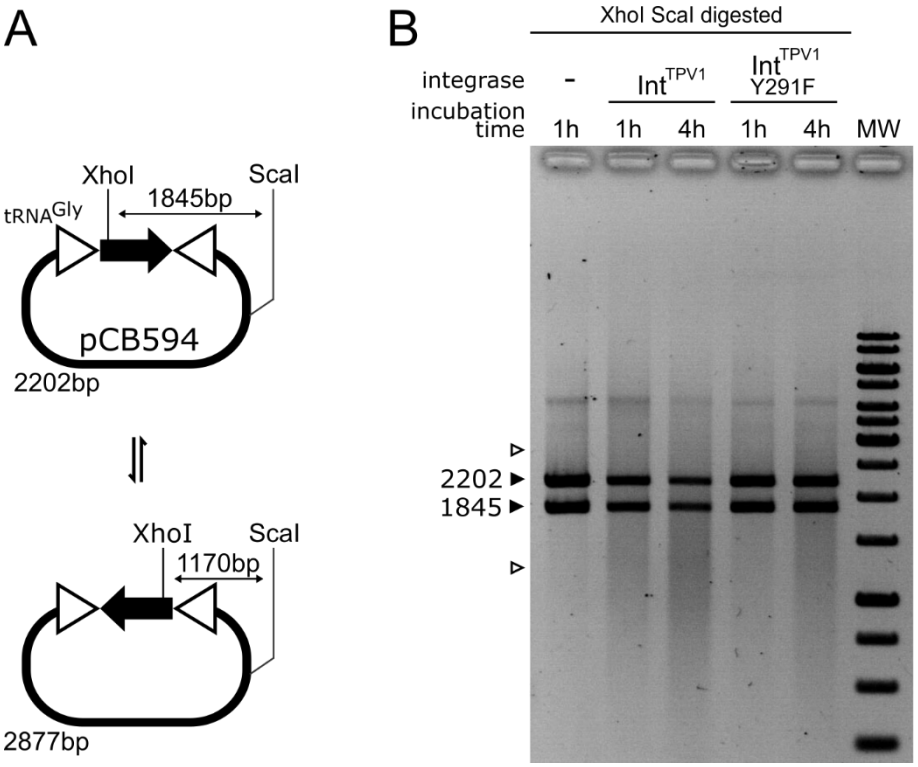


Figure 31. The integrase of the virus TPV1 presented no site-specific recombination activity *in vitro*. A. The site-specific activity is tested through an inversion assay. The substrate plasmid and inverted product can be distinguished after a restriction analysis. B. The plasmid pCB594 was incubated with the wild-type Int^{TPV1} or a tyrosine mutant Int^{TPV1}Y291F at 65°C and then digested with ScaI and XhoI. No inverted product was detected at the expected sizes (white arrows).

To better understand Int^{TPV1} absence of activity, we aligned its sequence with its most closely related integrases and with Int^{pT26-2} (Figure 32). Strikingly, Int^{TPV1} is shorter than all the other integrases at the N-terminal extremity. However, an alternative start-codon GTG creates a correct-length Int^{TPV1} that aligned perfectly with the other integrases. It seems that Int^{TPV1} is not active *in vitro* because it corresponds to a N-terminal truncated mutant.

The full-length Int^{TPV1} sequence was not annotated because it includes a stop-codon TGA (Figure 32). We confirmed that the stop codon is not a sequencing error by comparing the virus and the chromosome integrase copies that were sequenced independently ((Gorlas et al., 2012) and data not shown). The two copies are identical. The stop-codon therefore probably corresponds to a nonsense mutation. The integration of the virus with a mutated integrase in the chromosome is puzzling. One can speculate that enough stop-codon read-through happened that produced a full-length active integrase that catalyzed the site-specific integration.

The N-terminal part of suicidal integrases is involved in tetramer formation (Zhan et al., 2015). We therefore looked whether Int^{TPV1} was impaired in its oligomerization compared to Int^{pT26-2}. The proteins were loaded on a size exclusion column as part of the purification procedure (Figure 33). This allowed an estimation of a probable oligomerization state. Int^{pT26-2} behaved as a tetramer while Int^{TPV1} behaved as a monomer. This observation confirms the role of the N-terminal part of suicidal integrases in tetramer formation. Furthermore, since Int^{TPV1} was not active, tetramerization seems essential for activity. Finally, we can confidently assume that the shorter Int(C) fragment of Int^{TPV1} is not active *in vitro*. This further confirms the absence of activity for Int(C) fragments observed for the integrase Int^{TKV4} (Article 1) and contrary to the integrase Int^{SSV2} (Zhan et al., 2015)

Figure 32. MAFFT alignment of Int^{TPV1} with its most closely related integrases and Int^{pT26-2}. An alternative product of Int^{TPV1} is included that utilizes an alternative start codon but also includes a stop codon. The sequence translated from the att site is framed. The catalytic tyrosine is indicated by an arrow.

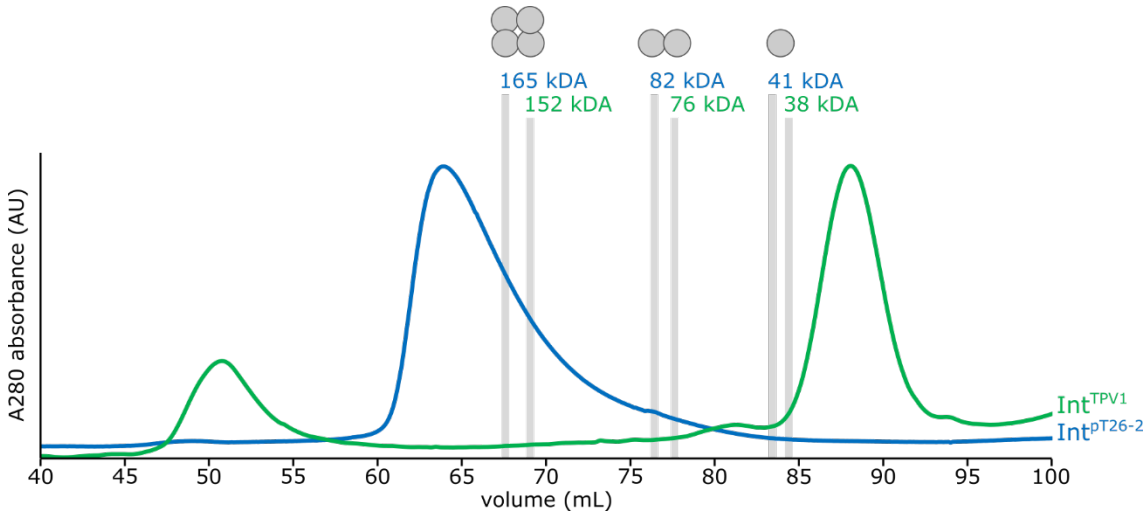


Figure 33. Size exclusion chromatography profile during the purification of the Int^{TPV1} and Int^{pT26-2}. Protein extract were loaded on a HiLoad 16/600 Superdex 200 size exclusion column. The buffer was 1M KCl, 40 mM Tris-HCl pH=8, 5 mM B-mercaptoethanol and 10% glycerol. The elution volume for the different oligomer were calculated using a calibration curve obtained in another buffer and are indicated on the plot.

Part 4. A new tool for mobile genetic element studies

Article 4. WASPS: Web Assisted Symbolic Plasmid Synteny

For our research on mobile genetic elements, some members of the BCA team and I faced the absence of comprehensive plasmid genome databases. We therefore joined forces and developed the WASPS Database (Web Assisted Symbolic Plasmid Synteny). This database comprises all the natural plasmids from the three domains of life and provides numerous functions for similarity searches of DNA and proteins. In addition, WASPS allows the drawing of accurate synteny maps, using pre-calculated orthologous clustering. The WASPS database is fully updated at regular intervals to ensure the highest standards in plasmid analysis.

The WASPS Database is fully functional and deployed at: https://archaea.i2bc.paris-saclay.fr/wasps/

WASPS: Web-Assisted Symbolic Plasmid Synteny Server

Catherine BADEL[1], Violette DA CUNHA[1], Ryan CATCHPOLE1[1], Patrick FORTERRE[1,2] & Jacques OBERTO[1,3]

[1]Institute for Integrative Biology of the Cell (I2BC), Microbiology Department, CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette cedex, France
[2]Institut Pasteur, Unité de Biologie Moléculaire du Gène chez les Extrêmophiles, Département de Microbiologie, 25 rue du Docteur Roux, 75015 Paris, France

Abstract

Motivation

Comparative plasmid genome analyses require complex tools, the manipulation of large numbers of sequences and constitute a daunting task for the wet bench experimentalist. Dedicated plasmid databases are sparse, only comprise bacterial plasmids and provide exclusively access to sequence similarity searches.

Results

We have developed WASPS (Web-Assisted Symbolic Plasmid Synteny), a web service granting protein and DNA sequence similarity searches against a database comprising all completely sequenced natural plasmids from bacterial, archaeal and eukaryal origin. This database pre-calculates orthologous protein clustering and enables WASPS to generate fully resolved plasmid synteny maps in real time using internal and user-provided DNA sequences.

Availability and implementation

WASPS queries befit all current browsers such as Firefox, Edge or Safari while the best functionality is achieved with Chrome. Internet Explorer is not supported. WASPS is freely accessible at https://archaea.i2bc.paris-saclay.fr/wasps/

Supplementary information

Supplementary data are available at Bioinformatics online.

Issue Section: SEQUENCE ANALYSIS

1. Introduction

The Darwinian evolution of genomes form the three domains of life is fast-tracked by the acquisition of traits through horizontal gene transfer (Cordaux and Batzer, 2009; Cossu, et al., 2017; Seth-Smith, et al., 2012). This process is mediated by plasmids and viruses defined as

[3] Corresponding author : jacques.oberto@i2bc.paris-saclay.fr

mobile genetic elements. In contact with their hosts, replicating plasmids in particular have been shown as remarkably plastic (Cury, et al., 2018). The comparative analysis of plasmid genomes constitutes a daunting task due to the lack of dedicated, plasmid-centric resources. Similarity searches in public databases of the National Centre for Biotechnology Information (NCBI) cannot be restricted exclusively to plasmid DNA or protein. Recently, plasmid assets have been developed proposing either a comprehensive manually curated bacterial plasmid list (Brooks, et al., 2019) or a bacterial plasmid database which can be interrogated using sequence similarity programs (Galata, et al., 2019). Bioinformatics tools providing access to a database of bacterial, archaeal and eukaryal plasmids are still missing at this time. In this work, we present the WASPS web service which in addition to similarity searches allows the real-time generation of fully resolved plasmid synteny maps. The evolutionary relationships between user-provided sequences and databases sequences can be inferred using these synteny maps. The WASPS Database is updated at regular intervals, comprises all completely sequenced natural plasmids from the three domains of life and boasts full pre-calculated orthologous gene clustering.

2. Materials and methods

WASPS consists of three modules:

The WASPS Database is a relational database containing all entries from the NCBI RefSeq plasmid repository encompassing all completely sequenced bacterial, archaeal and eukaryal plasmids of natural origin. At database creation, each protein-encoding gene is hierarchically linked to its plasmid of origin using GenBank keys. All corresponding protein sequences undergo orthologous clustering using UCLUST (Edgar, 2010) and obtain a centroid identifier. At this date, the WASPS Database contains 15,789 bacterial, 203 archaeal and 36 eukaryal plasmids.

The WASPS Webtool provides a user interface for the remote interrogation of the database and proposes 5 distinct features:

1) Text-based search in plasmid and proteins definitions and accession numbers.
2) Protein and DNA-based similarity searches with user-provided sequences using BLAST (Altschul, et al., 1990) and DIAMOND (Buchfink, et al., 2015) algorithms.
3) The drawing of synteny maps using a user-provided annotated DNA file in GenBank format. It proceeds by the extraction and matching of the corresponding protein sequences against WASPS centroids using BLAST or DIAMOND. The query sequence and related plasmids from the database are drawn using a consistent symbolic coloring.
4) The drawing of synteny maps using an unannotated raw or Fasta DNA file. It requires the initial six-frame translation of the query sequence using ATG, GTG or TTG as start codons and TAG, TAA or TGA as stop codons. All frames and related plasmids are drawn as above.
5) The similar drawing of plasmid synteny maps from the database according to orthologous clustering using WASPS plasmid and protein accession numbers.

Dedicated hyperlinks connect text and similarity searches with synteny queries. Wide genomic areas can be explored by panning and zooming directly within the browser. All synteny map genes grant access to relevant additional information and protein sequences using specific gestures in the web interface (Supplementary data). Maps can be exported as PNG or SVG.

The WASPS updater is fully automated and operates a pipeline in the background to refresh the database at fixed intervals (Supplementary data).

3. Results

To illustrate WASPS capabilities, we submitted for synteny analysis the unannotated Fasta sequence of the 13,015bp archaeal plasmid pTN2 (NC_014115.1) from *Thermococcus nautili* (Soler, et al., 2010). Using the option 'primary hits only' and DIAMOND searching algorithm, three hits were detected: plasmid pTN2, already existing in the database, pIRI48 from *Thermococcus* sp. IRI18 and pCIR10 from *Thermococcus* sp. CIR10. All three plasmids were selected to generate a plasmid synteny map (Fig. 1 & Supplementary data). Interestingly, WASPS synteny results for pTN2 resembled a prior analysis where the same *Thermococcus* plasmids were aligned and drawn manually (Krupovic, et al., 2013). In this representation, all plasmids shared orthologous *uvrD* genes whereas for the WASPS synteny pIRI48 gene YP_007195295.1 is not orthologous. To ascertain WASPS clustering quality, we assessed the orthology of the various *uvrD* genes using an alternative method. Protein sequence similarity among WASPS UvrD cluster members exceeded 26%, compatible with the proposed 30% orthology threshold (Lerat, et al., 2003). Strikingly, similarity between each member of the UvrD cluster and the translated YP_007195295.1 gene never exceeded 15.2% (Supplementary data). These results demonstrated the robustness of the WASPS synteny mapping.

In a second experiment, the predictive capabilities of WASPS were tested on NCBI sequence contigs. We discovered that QMOB01000129.1, a 10757bp contig assigned to a *Chloroflexi* bacterium metagenome (Dombrowski, et al., 2018) shared an excessive number of genes with archaeal plasmids (Supplementary data). We surmised that this particular contig corresponded to a low level of archaeal DNA contamination in the metagenomic sample. WASPS was therefore able to provide an effortless characterization of metagenomes.

Funding

This work was funded by CNRS and the European Research Council under the European Union's Seventh Framework Program (FP/2007-2013)/Project EVOMOBIL - ERC Grant Agreement no. 340440 (PF). Catherine Badel is supported by 'Ecole Normale Supérieure de Lyon'.

Figure 1. Synteny map of plasmid pTN2 from *T. nautili*. The first 6 tracks correspond to genes predicted in the Fasta query sequence in the 6 reading frames with corresponding E-values. The following tracks refer to plasmids from the WASPS database sharing synteny with the submitted sequence. Gene orthology is indicated by consistent graphic symbolism throughout the map. Genes in white color are singletons, devoid of ortholog in the database. As indicated, plasmids maps are accurately drawn to scale.

References

Altschul, S.F.*, et al.* (1990) Basic local alignment search tool, *Journal of molecular biology*, **215**, 403-410.

Brooks, L., Kaze, M. and Sistrom, M. (2019) A Curated, Comprehensive Database of Plasmid Sequences, *Microbiology resource announcements*, **8**.

Buchfink, B., Xie, C. and Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND, *Nature methods*, **12**, 59-60.

Cordaux, R. and Batzer, M.A. (2009) The impact of retrotransposons on human genome evolution, *Nature reviews. Genetics*, **10**, 691-703.

Cossu, M.*, et al.* (2017) Flipping chromosomes in deep-sea archaea, *PLoS genetics*, **13**, e1006847.

Cury, J.*, et al.* (2018) Host range and genetic plasticity explain the co-existence of integrative and extrachromosomal mobile genetic elements, *Molecular biology and evolution*.

Dombrowski, N., Teske, A.P. and Baker, B.J. (2018) Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments, *Nature communications*, **9**, 4999.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST, *Bioinformatics*, **26**, 2460-2461.

Galata, V.*, et al.* (2019) PLSDB: a resource of complete bacterial plasmids, *Nucleic acids research*, **47**, D195-D202.

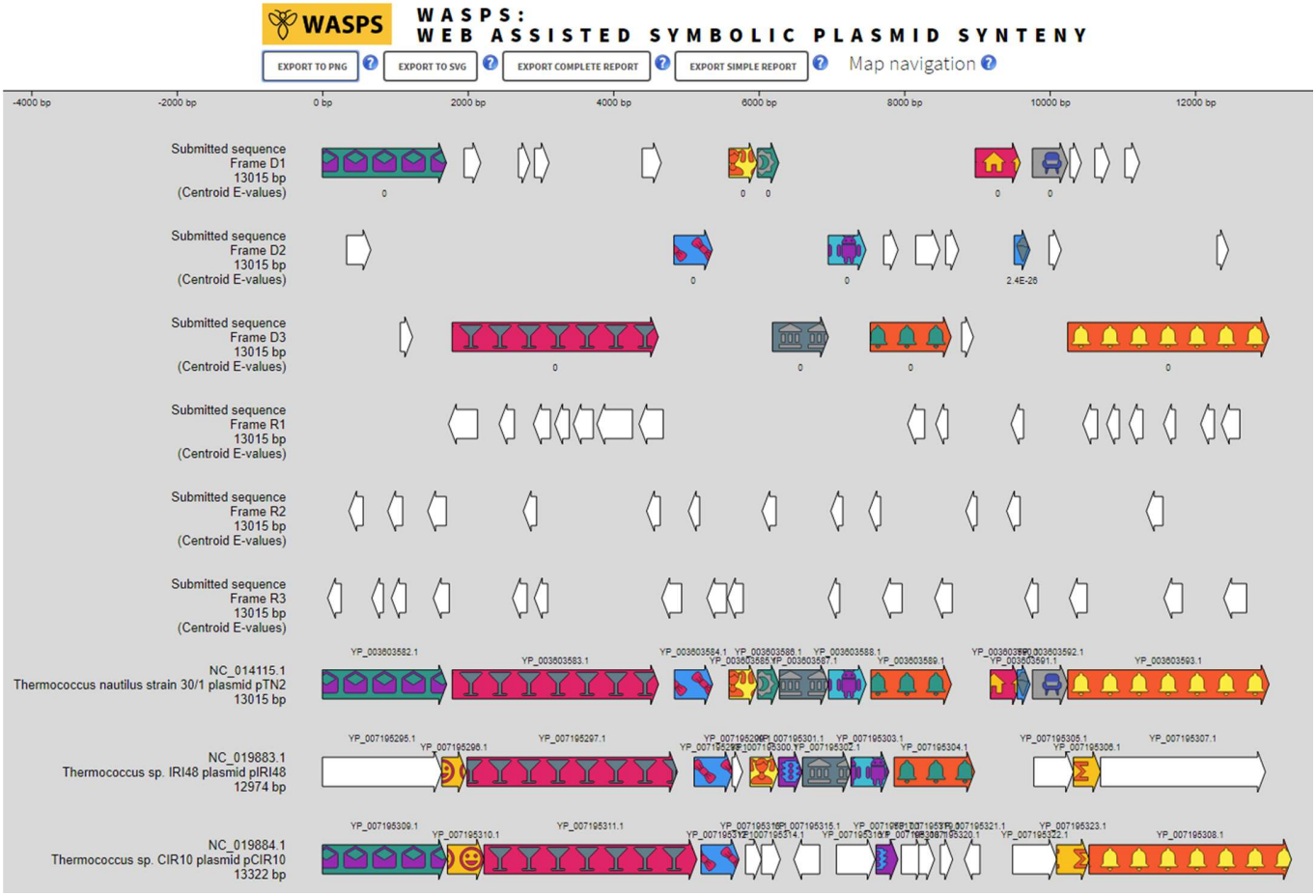
Krupovic, M.*, et al.* (2013) Insights into dynamics of mobile genetic elements in hyperthermophilic environments from five new Thermococcus plasmids, *PloS one*, **8**, e49044.

Lerat, E., Daubin, V. and Moran, N.A. (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria, *PLoS biology*, **1**, E19.

Seth-Smith, H.M.*, et al.* (2012) Structure, diversity, and mobility of the Salmonella pathogenicity island 7 family of integrative and conjugative elements within Enterobacteriaceae, *Journal of bacteriology*, **194**, 1494-1504.

Soler, N.*, et al.* (2010) Two novel families of plasmids from hyperthermophilic archaea encoding new families of replication proteins, *Nucleic acids research*, **38**, 5088-5104.

Figure 1


WASPS
WASPS:
WEB ASSISTED SYMBOLIC PLASMID SYNTENY
EXPORT TO PNG
EXPORT TO SVG
EXPORT COMPLETE REPORT
EXPORT SIMPLE REPORT
Map navigation
-4000 bp
-2000 bp
0 bp
2000 bp
4000 bp
6000 bp
8000 bp
10000 bp
12000 bp
Submitted sequence
Frame D1
13015 bp
(Centroid E-values)
Submitted sequence
Frame D2
13015 bp
(Centroid E-values)
Submitted sequence
Frame D3
13015 bp
(Centroid E-values)
Submitted sequence
Frame R1
13015 bp
(Centroid E-values)
Submitted sequence
Frame R2
13015 bp
(Centroid E-values)
Submitted sequence
Frame R3
13015 bp
(Centroid E-values)
NC_014115.1
Thermococcus nautilus strain 30/1 plasmid pTN2
13015 bp
NC_019883.1
Thermococcus sp. IRI48 plasmid pIRI48
12974 bp
NC_019884.1
Thermococcus sp. CIR10 plasmid pCIR10
13322 bp

Supplementary Data

WASPS: Web Assisted Symbolic Plasmid Synteny Server

Catherine BADEL[1], Violette DA CUNHA[1], Ryan CATCHPOLE1[1], Patrick FORTERRE[1,2] & Jacques OBERTO[1,3]

[1]Institute for Integrative Biology of the Cell (I2BC), Microbiology Department, CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette cedex, France

[2]Institut Pasteur, Unité de Biologie Moléculaire du Gène chez les Extrêmophiles, Département de Microbiologie, 25 rue du Docteur Roux, 75015 Paris, France

Contents

1. WASPS Database structure, generation and update pipeline

2. WASPS client-side user interface

3. Plasmid pTN2 synteny map and prediction quality assessment

4. Identification of NCBI contig QMOB01000129.1

5. Discussion

6. References

1. WASPS Database structure, generation and update pipeline

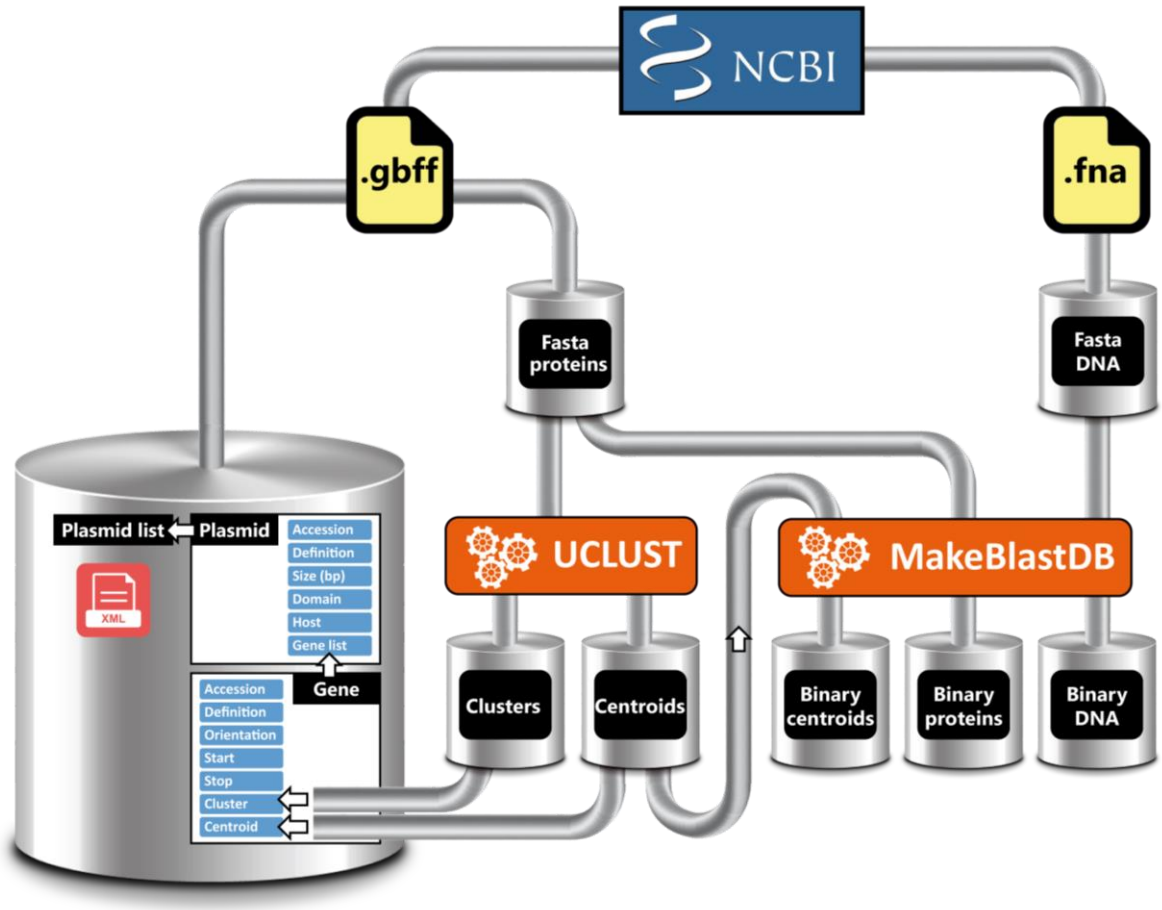
Natural plasmid data (RefSeq) is collected from the FTP site of the NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/plasmid) in binary form and processed locally on the server to generate the WASPS Database. At this stage, only the RefSeq plasmid releases are included in the WASPS Database for the reason that they constitute a non-redundant and well-annotated set of sequences, updated at regular intervals. The WASPS Database is updated by the WASPS Updater application in order to follow NCBI RefSeq refreshes. This process involves the addition of new entries which receive a two part GenBank identifier (accession.version). Updates of pre-existing entries keep the same accession number and will increment the version number by one unit. These 'accession.version' identifiers univocally describe both plasmid and gene entries. These identifiers are used to link the different data bins composing the WASPS relational database. The central part of the WASPS database consists of a single XML file containing a sequential list of plasmids exposing their relevant fields. Each plasmid contains a gene list field to store relevant genetic data (Suppl. Fig.1). Each protein in the WASPS database is therefore identified by double 'accession.version' under the format 'gene_accession.version=plasmid_accession.version'. Plasmid DNA sequences and protein sequences are stored in separate bins but intimately linked to the central XML using the 'accession.version' identifiers. All fields and DNA or protein sequences are extracted or

[3] Corresponding author : jacques.oberto@i2bc.paris-saclay.fr

parsed from the downloaded NCBI GenBank and DNA Fasta files. Protein orthology relationships are determined using UCLUST and injected appropriately in the XML file. The UCLUST orthology parameter used by WASPS amounts to 0.35 and is slightly lower than the recommended values (Edgar, 2010). This particular value was chosen since it empirically corresponds to the orthology threshold of 30% similarity proposed by (Lerat, et al., 2003)(see also Section 3, below) which is calculated as follows for proteins A and B:

$$BLAST_{bits(A\times B)} \times \frac{100}{BLAST_{bits(A\times A)}} \geq 30 \qquad \text{(Equation 1)}$$

The cluster centroids calculated with UCLUST are collected separately into an additional bin. The text bins containing DNA, total proteins and centroid proteins sequences are then converted to binary format in order to be efficiently queried by DIAMOND, BlastN, BlastP, TBlastN or PsiBlast. The database compilation is optimized and fully automated to allow frequent updates and ensure exhaustiveness of the analyses. The WASPS Updater pipeline is shown in Supplemental Figure 1.

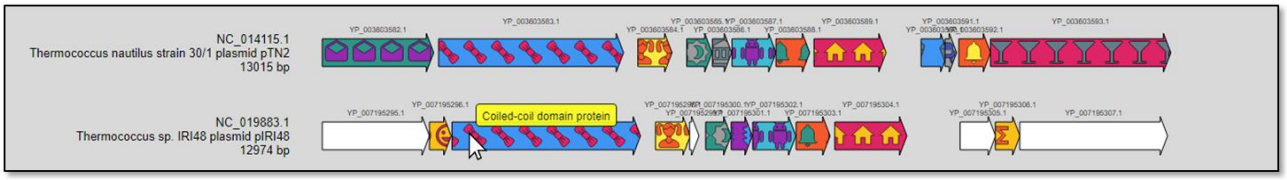


Supplemental figure 1. WASPS database structure and update pipeline. *The WASPS update process is fully automated and optimized for low CPU cycles. Due to its structure, the database is completely regenerated at each update to allow increased robustness and accuracy. Flux directionality is top-down except where noted. All metallic bins are qweryable in WASPS. Binary bins are used specifically by BLAST and DIAMOND.*

2. WASPS client-side user interface

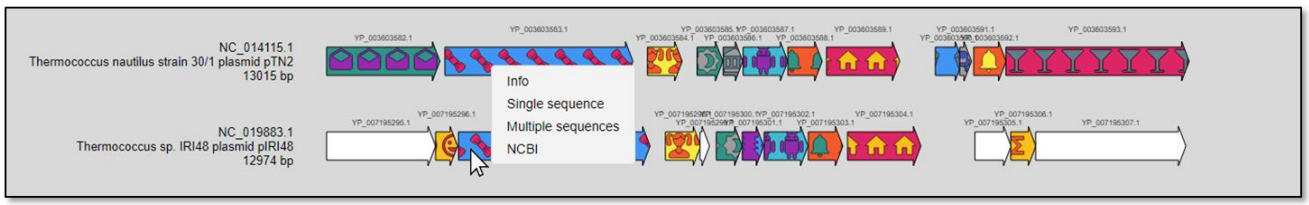
WASPS Synteny Map Interface has been developed to allow maximal user-interactivity. User interactivity is achieved by the means of a 'three button wheel mouse', a standard equipment for most modern desktop computers. Equivalent gestures are available for laptops, trackpads or touch screen devices and are provided by the respective operating systems. 2D synteny maps can be smoothly panned and zoomed directly in the web browser without requiring data transfer from or to the server.

- ***Pan.*** The Synteny Map Interface can be panned by holding down the left mouse button.
- ***Zoom.*** The Synteny Map Interface can be zoomed by rotating the mouse wheel.
- ***Hovering.*** Context-sensitive information is available for each displayed gene in the synteny maps. Mouse hovering on a specific gene will present its definition in a yellow tooltip (Suppl. Fig. 2).



Supplemental figure 2. *The hovering tooltip appears in yellow color.*

- ***Context menu.*** Right clicking on a specific gene will open a context menu with four options (Suppl. Fig. 3):
  i) Info : protein gene accession, plasmid accession, protein definition and protein cluster.
  ii) Single sequence : protein sequence of the highlighted gene in Fasta format.
  iii) Multiple sequences : all protein sequences of the WASPS cluster related to the highlighted gene.
  iv) NCBI : external link to the protein (in GenBank format) at the NCBI.



Supplemental figure 3. *The context menu appears in white background upon right mouse keypress.*

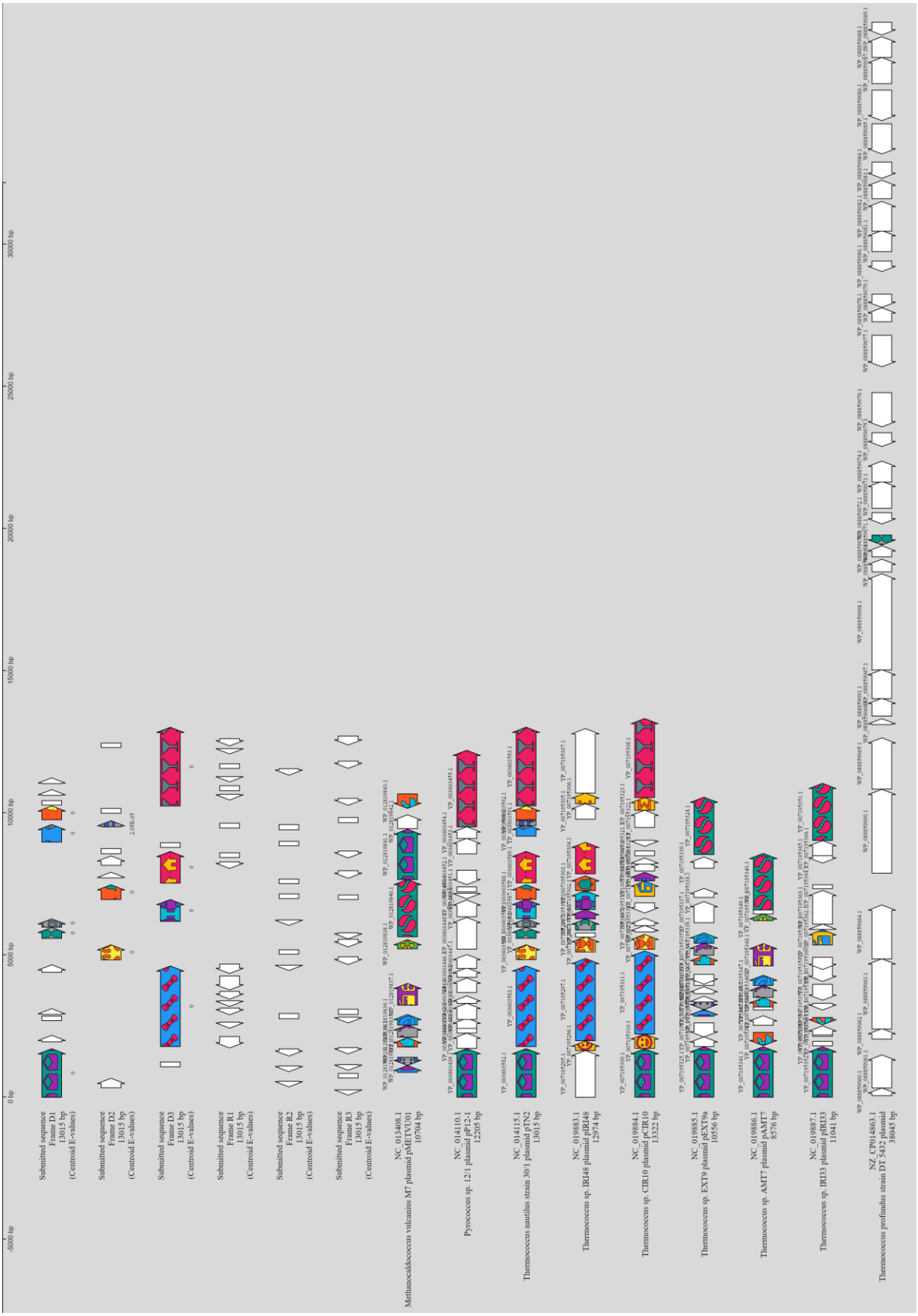
3. Plasmid pTN2 synteny map and assessment of the prediction quality

The sequence of plasmid pTN2 from *T. nautili* was submitted to the WASPS web service for synteny analysis using the parsimonious option 'Primary Hits Only' to limit the search to the best match for each protein using the fast DIAMOND algorithm. To obtain the results shown in Figure 1 (main article), WASPS internally translated the query plasmid sequence in the six possible reading frames using ATG, GTG or TTG as start codons and TAG, TAA or TGA as stop codons. All predicted open reading frames (ORFs) were retained without size limitation. Internal ORFS in the same reading frame of large predicted genes were not considered. The resulting 51 predicted proteins were then compared to the centroid database using DIAMOND. All matching translated ORFs presenting a DIAMOND E-value hit ≤ 10E-6 were

assigned the corresponding WASPS centroid cluster number and a unique symbolism composed of a SVG icon with foreground and background colors. Translated ORFs above this threshold were non-orthologous singletons and retained the default white color. Three related plasmids were detected as best hits in the WASPS database and drawn similarly using gene cluster numbers pre-calculated at database inception. A combination of graphic symbols and colors would grant the simultaneous display of a quasi-unlimited number or orthologous genes. This 'fully connected' pre-calculated clustering topology allowed for near-instantaneous generation of completely resolved plasmid synteny maps. The consistent ORF symbolism permitted immediate visual synteny analysis.

A second analysis using the same pTN2 plasmid data using the 'Extended cluster hits' option produced a deeper analysis and reported the same 3 plasmids found previously and 6 additional plasmids (Suppl. Fig. 4). In that case, the entire orthologous clusters corresponding to each best DIAMOND centroid hit were considered positive. All plasmids having at least one gene belonging to one of these clusters were retained and drawn according to the same 'fully connected' clustering topology.

We have tested the accuracy of the predictive capabilities of WASPS syntenies using an alternative method. All plasmids depicted in Figure 1 (main article) carry a leftmost gene annotated as 'UvrD'. However in the WASPS synteny map, the corresponding YP_007195295.1 gene (white color) carried by pIRI48 is not considered as part to the same orthologous cluster containing the UvrD genes of the other plasmids. Gene YP_007195295.1 clustered alone in the WASPS database. To investigate the nature of this discrepancy we computed orthologous relationships between gene YP_007195295.1 and the WASPS UvrD cluster composed of genes YP_007195309.1, YP_007195325.1, WP_012819841.1, YP_007195341.1, YP_003603582.1, YP_007195352.1 and YP_003603439.1. We used Equation 1 (see Section 1, above) for the calculations of the similarities between these 8 proteins and the results are tabulated in Supplemental Table 1. The UvrD intra-cluster similarities (orange) were always near or in excess of the 30% orthology threshold recommended by (Lerat, et al., 2003). Interestingly, similarity with YP_007195295.1 (blue) was significantly lower suggesting non-orthology on the basis of protein sequence. These results validated the threshold parameter chosen for UCLUST (see Section 1, above).



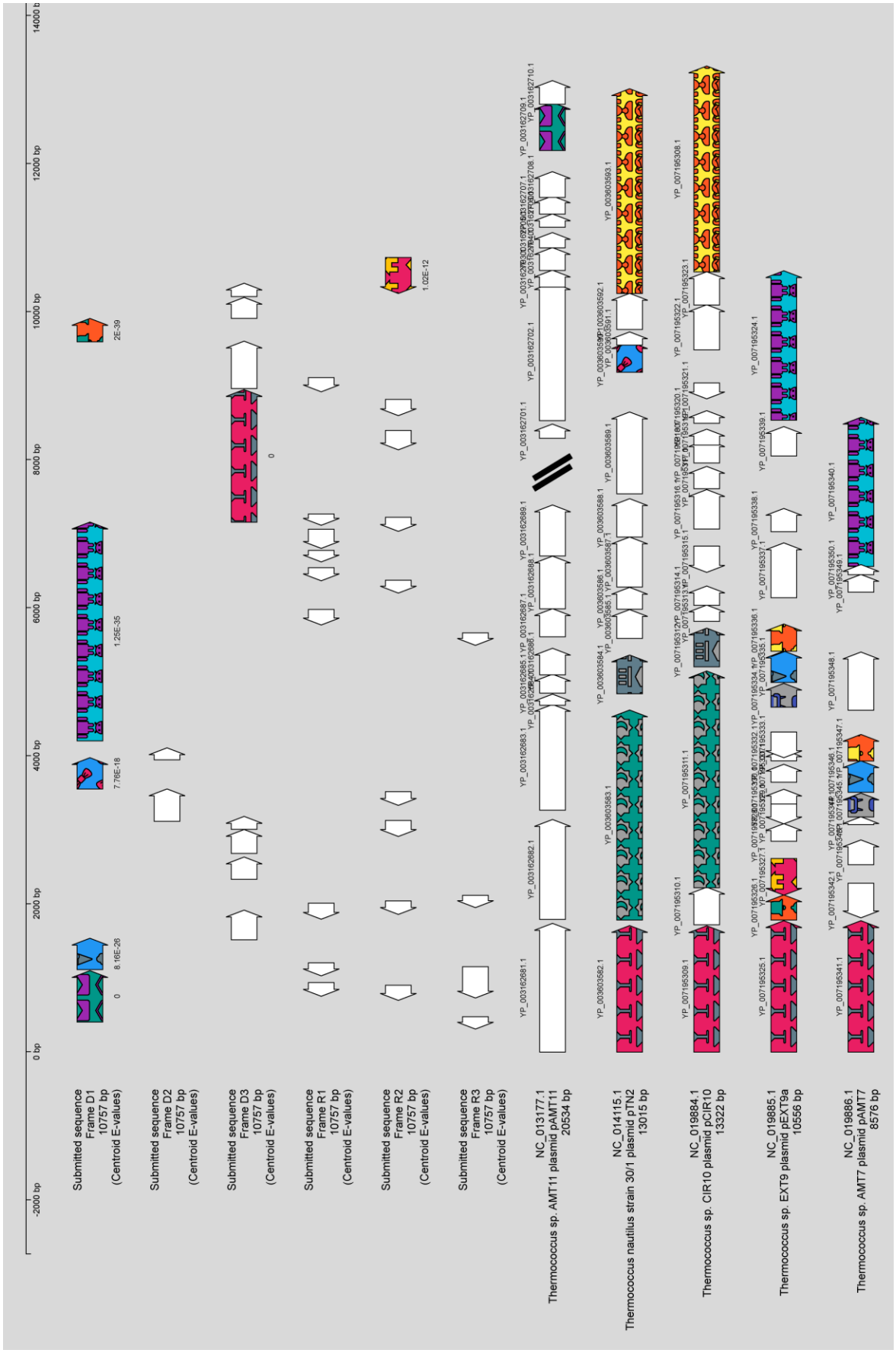
Supplemental figure 4. Extended pTN2 plasmid synteny*. Using the 'Extended cluster hits' option, the synteny analysis produced more extensive results, retrieving all 9 related plasmids from the WASPS database.*

Supplemental Table 1. Inter-cluster (blue) versus intra-cluster orthologous comparisons (orange)

Gene_Accession Plasmid_accession	YP_007195295.1 NC_019883.1	YP_007195309.1 NC_019884.1	YP_007195325.1 NC_019885.1	WP_012819841.1 NC_013408.1	YP_007195341.1 NC_019886.1	YP_003603582.1 NC_014115.1	YP_007195352.1 NC_019887.1	YP_003603439.1 NC_014110.1
YP_007195295.1 NC_019883.1	100	14.3	14.6	14.0	15.2	14.1	14.7	15.2
YP_007195309.1 NC_019884.1		100	27.6	26.7	27.2	74.1	26.5	82.7
YP_007195325.1 NC_019885.1			100	94.3	95.0	27.0	93.4	27.0
WP_012819841.1 NC_013408.1				100	97.7	27.3	97.5	26.4
YP_007195341.1 NC_019886.1					100	26.6	95.9	26.6
YP_003603582.1 NC_014115.1						100	27.4	70.1
YP_007195352.1 NC_019887.1							100	26.0
YP_003603439.1 NC_014110.1								100

4. Identification of NCBI contig QMOB01000129.1

The NCBI entry QMOB01000129.1, a 10757bp contig was assigned to a *Chloroflexi* bacterium metagenome (Dombrowski, et al., 2018). We submitted this entry to WASPS as a Fasta file and generated a synteny map using the default E-value of 10E-06, the 'primary hits only' option and BLAST. Very surprisingly, the six-frame translation generated 7 proteins displaying very high similarity (E-values ≤ 1.02E-12) to those encoded by 5 hyperthermophilic archaeal plasmids (Suppl. Fig.4). Since the *Chloroflexi* metagenomic samples originated from Guaymas Basin hydrothermal vents known to host Thermococcales archaea (Canganella, et al., 1998), we surmised that this particular contig in fact corresponded to a low level of archaeal DNA contamination in the metagenomics sample.



Supplemental Figure 4. *Synteny map for NCBI contig QMOB01000129.1*

5. Discussion

The plasticity of natural replicative plasmids contributes to the evolution of their host genome. Robust plasmid comparative genomics is therefore required to accurately assess the evolution of organisms. The conservation of gene order or synteny based on protein sequences has already proven successful to compare cellular genomes and infer evolutionary relationships. In this work, we have developed a novel database-backed natural plasmid synteny web service designed to overcome current limitations of current plasmid databases. The WASPS database is fully relational and carries all natural plasmids from the three domains of life. All plasmid-encoded proteins ranks in the database are pre-calculated according to a 'fully connected' clustering topology. The WASPS Webtool allows rigorous and straightforward analysis of user-submitted plasmid-related protein or DNA sequences. The highly significant and robust WASPS protein clustering allows the software to rapidly assign functions to submitted sequences and to infer their orthologous relationships. WASPS-generated synteny maps are almost identical to their manually computed and hand-drawn counterparts while requiring a fraction of the effort. WASPS' predictive capability allows users to easily identify mobile replicative mobile elements present in metagenomes as well as aid in the detection of potential DNA contaminations. The attractive and intuitive WASPS web interface incorporates the latest web standards and technologies. Among these, a navigable plasmid synteny map boasting striking iconic orthology symbols constitutes the flagship of this web service and should appeal to both wet and dry bench researchers. Important provisions have been adopted to keep the robust and lightweight WASPS database up to date by the means of an optimized and fully automated background task. The quality of the analyses provided by the WASPS Webtool are therefore destined to improve with time following the addition of new plasmid entries. Additional programs could be designed to take full advantage of the standalone WASPS database. The database model developed in this work could be further replicated to study viruses or target specifically integrative and conjugative elements.

6. References

Canganella, F.*, et al.* (1998) Thermococcus guaymasensis sp. nov. and Thermococcus aggregans sp. nov., two novel thermophilic archaea isolated from the Guaymas Basin hydrothermal vent site, *International journal of systematic bacteriology*, **48 Pt 4**, 1181-1185.

Dombrowski, N., Teske, A.P. and Baker, B.J. (2018) Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments, *Nature communications*, **9**, 4999.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST, *Bioinformatics*, **26**, 2460-2461.

Lerat, E., Daubin, V. and Moran, N.A. (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria, *PLoS biology*, **1**, E19.

Part 5. Towards a dissection of Int^{pTN3} site-specific and homologous recombinational activities

In the Article 1, we evidenced an unprecedented dual catalytic activity by the integrase Int^{pTN3}. It can catalyze both site-specific recombination and homologous recombination through a tyrosine recombinase mechanism. We now want to characterize the mechanism of the singular conservative homologous recombination activity.

No reaction conditions can discriminate the two activities

We wondered whether Int^{pTN3} homologous recombination activity was due to some reaction conditions. For instance, one could propose that it resulted from the high temperature of the reaction. This could not be true since Int$^{pT26\text{-}2}$ and Int^{SSV2} are active at the same or higher temperatures and do not catalyze homologous recombination (Article 3 and(Zhan et al., 2015)).

Divalent cations stimulated the activity of Topoisomerases IB (TopoIB) (Sissi and Palumbo, 2009) that share a common ancestry with tyrosine recombinases (Cheng et al., 1998). Divalent cations (1mM $MgCl_2$) were also included in Int^{pTN3} activity assays (Article 1). We wondered what their effect was. We therefore performed an integration assay with increasing concentrations of $MgCl_2$ or $MnCl_2$ and without any divalent cation (Figure 34). Int^{pTN3} catalyzed the formation of pUC18 dimers through homologous recombination and pJO322 dimers through site-specific recombination for all tested conditions. Notably, with no divalent cation, Int^{pTN3} could still catalyze the formation of pUC18 and pJO322 plasmid dimers (Figure 34.B). They are therefore not necessary for Int^{pTN3} dual activity. Dimer formation seemed to decrease at the higher 5mM $MgCl_2$ concentration, especially for pJO322 (Figure 34.B). Overall, $MgCl_2$ does not favor a particular activity of Int^{pTN3} but rather inhibits both activities at high concentrations. Similarly, $MnCl_2$ inhibited both pUC18 and pJO322 dimer formation at 7.5 mM (Figure 34.C). At 75 µM $MnCl_2$, pJO322 dimer formation was increased compared to the absence of $MnCl_2$ while pUC18 dimer formation remained stable. At low concentrations, Mn^{2+} slightly favors site-specific recombination over homologous recombination.

Overall, we didn't find any reaction conditions that clearly discriminated Int^{pTN3} site-specific and homologous recombination activities. We therefore hypothesized that the dual activity does not originate from a particular reaction condition but rather from the protein sequence itself. To explore this possibility, we decided to identify more sequences similar to Int^{pTN3}.

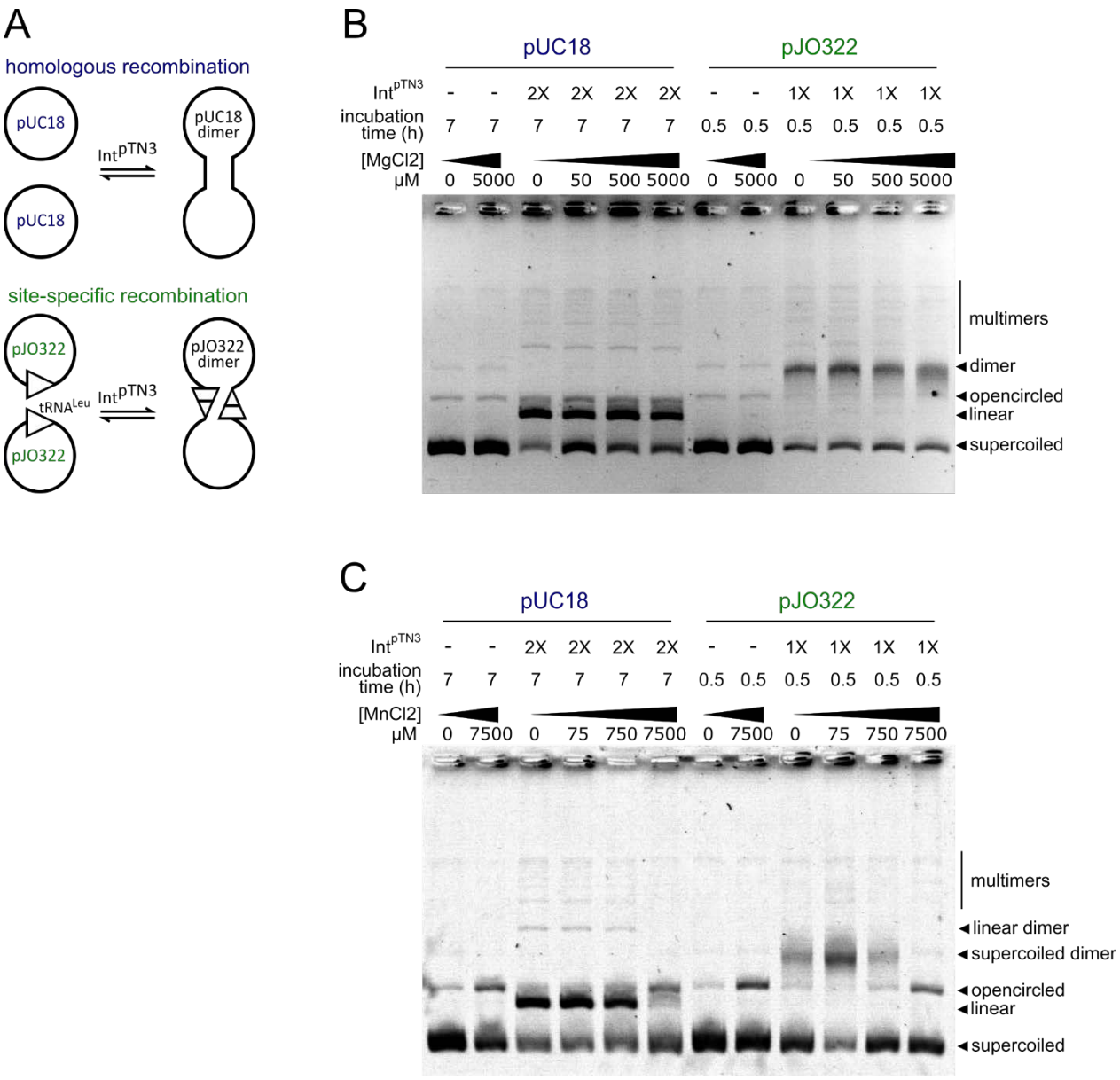


Figure 34. Divalent cation influence on Int^{pTN3} integration activity. A. Site-specific and homologous recombination activity are assayed by an integration assay the plasmid substrates pJO322 and pUC18 respectively. B and C. Increasing concentrations of $MgCl_2$ (B) and $MnCl_2$ (C) were used for the plasmid incubation with Int^{pTN3} at 65°C. The samples were treated with proteinase K and separated on a gel.

Table 6. Mobile genetic elements presenting an integrase of the Int^{pTN3} family

Element	Host	Integrated or episomal	Targeted tRNA gene	Element coordinates[1]	Plasmid / host access	Reference for the host
pTN3	T. nautili 30-1	Episomal and integrated	tRNAleu	/	NC_022527.1 / NZ_CP007264.1	(Gaudin et al., 2014; Gorlas et al., 2014)
TclCL1_IP1	T. cleftensis CL1	Integrated	tRNAleu	478738-492539	NC_018015.1	(Hensley et al., 2014)
TspJCM11816_IP3	T. sp. JCM11816 contig00004 and contig00005	Integrated	tRNAleu	ND	BBCU01000000	(Hoaki et al., 1994)
TKV4	T. kodakarensis KOD1	Integrated	tRNAleu	1182705-1201564	NC_006624.1	(Atomi et al., 2004)
TceDSM17994_IP2	T. celericrescens DSM 1799 Contig008	Integrated	tRNAleu	70088-87563	GCF_001484195.1	(Kuwabara et al., 2007)
TIRI33C_IP2	T. IRI33C	Integrated	tRNAleu	1474638-1454087	Lab collection	/
TIRI06C_IP2	T. IRI06C	Integrated	tRNAleu	144687-131977	Lab collection	/
T9-3_IP1	T. 9-3	Integrated	tRNAleu	312684-329157	Lab collection	/
T33-3_IP1	T. 33-3	Integrated	tRNAleu	198451-180585	Lab collection	/
TbaCH5_IP3	T. barophilus	Integrated	tRNAleu	2235665-2252849	NC_014804.1	(Oger et al., 2016)
pTF1	T. fumicolans	Episomal and integrated	tRNAser	/	Lab collection	(Godfroy et al., 1996)
TeuA501_IP1	T. eurythermalis A501	Integrated	tRNAser	1372825-1352621	NZ_CP008887.1	(Zhao et al., 2015)
TguDSM11113_IP2	T. guaymasensis DSM 11113	Integrated	tRNAser	1495271-1475772	NZ_CP007140.1	(Canganella et al., 1998)
TspEP1_IP1	T. sp. EP1 Contig14	Integrated	tRNAser	19626-13474	GCF_001317345.1	(Zhou et al., 2016)
TAMTc11_IP3	T. AMTc11	Integrated	tRNAser	1623206-1640955	Lab collection	(Gonnet et al., 2011)

1. Integrated elements are oriented from Int(C) to Int(N)

Integrases similar to Int^{pTN3} catalyze site-specific recombination in one of two distinct sites

Identification of integrases similar to Int^{pTN3}

We previously identified 5 pTN3-like integrated elements that all encoded an integrase similar to Int^{pTN3} (Article 1 (Cossu et al., 2017)). We searched for new similar integrases sequences by tblastn using all the known integrases as query. We found 1 new episomal integrase sequence and 8 new integrated integrase sequences that we reconstituted from the Int(N) and Int(C) parts as in Article 2, only in *Thermococcus* species (Table 6). All the integrase sequences are present on pTN3-like elements, except the one from TspEP1_IP1 (Figure 35). The search for new integrases therefore resulted in the discovery of new integrated elements.

The integrases have distinct site specificities

The 15 known integrases present their attachment site in tRNA^{Leu}CAA genes for 10 of them and tRNA^{Ser}CGA genes for 5 of them (Table 6). Both tRNAs have a supplementary loop (Figure 36.B-C). More precisely, the att site corresponds to the 5' half of the tRNA genes from the beginning of the tRNA gene or a few nucleotides upstream to the end of the anti-codon loop (Figure 36). The supplementary loop is not included in att site. The restriction to tRNA with supplementary loops is therefore surprising. The mean pairwise identity between all att sites is 89.0%. However, the att sites corresponding to tRNA^{Leu}CAA genes have a mean pairwise identity of 98.9% and those corresponding to tRNA^{Ser}CGA genes have a mean pairwise identity of 97.1%. Att sites can therefore be divided in two groups based on the tRNA gene they correspond to. The two groups have different sizes because of additional nucleotides present in the D-loop of tRNA^{Leu}CAA (Figure 36.B-C) .The mean pairwise identity is 78.4% between the two groups. This is similar to the 75% conservation level between all the att sites of Int$^{pT26\text{-}2}$-like integrases (Article 3). However, contrary to the Int$^{pT26\text{-}2}$-like integrases, there is no evident conserved core for the att site of Int^{pTN3} integrases.

The integrases can be divided in two groups: one for each specificity

To understand the evolutionary history of the two specificities, we decided to construct a phylogenetic tree of the integrases. Two problems arose. Firstly, the integrase from the element TbaCH5_IP3 is very divergent with a mean pairwise similarity of only 21.5% with the other integrases. This divergence might result from a selected diversification or from the accumulation of mutations after integration and the release of selection pressure. Either way, the divergence lead to a long-branch attraction phenomenon (data not shown) and we removed this integrase from our phylogenetic analysis. Secondly, integrases of the Int^{pTN3} family present additional sequences compared to their closest relatives (Article 1 (Cossu et al., 2017)). Their alignment with an outgroup is therefore difficult and highly dependent on the alignment method. Accordingly, we did not include any outgroup in the phylogenetic analysis and the tree is unrooted (Figure 37).

In the phylogenetic tree, integrases similar to Int^{pTN3} are clearly separated in two groups that each present one of the specificities (Figure 37). The specificity was probably switched once and then transmitted to the descendent integrases. We do not know which specificity is the ancestral one since the tree is unrooted. At the sequence level, each group of integrases is very similar with only a few varying amino-acids and a clear consensus sequence (Figure 38). The consensus sequences from each integrase group have a pairwise similarity of 72.1%. The two groups of integrases have very similar

protein sequences arguing for a recent specificity switch. The differences are predominantly located around the att site translation (Figure 38). It looks as if the att site was cut out and replaced by another sequence with very few other modifications (Figure 39-C).

The two specificities can efficiently be used for integration

We previously showed that purified Int^{pTN3} can efficiently catalyze site-specific recombination in $tRNA^{leu}$ genes (Article 1 (Cossu et al., 2017)). We wondered whether the other tRNA gene used by some integrases for integration is also functional. We therefore decided to purify an integrase with the $tRNA^{Ser}$ specificity. We chose the integrase from free plasmid pTF1 of *Thermococcus fumicolans* (Int^{pTF1}) (Table 6). We overproduced it in *E. coli* and purified it by affinity chromatography as explained in Material and Methods (page 291).

We assayed Int^{pTF1} site-specific activity with an integration assay. Site-specific recombination between two plasmids pCB624 carrying one att site copy results in plasmid dimer formation (Figure 40.A). After incubation with Int^{pTF1} at 65°C, we observed pCB624 dimers (Figure 40.B). Int^{pTF1} can efficiently catalyze site-specific integration *in vitro*.

Searching for the mechanism of specific site recognition

We demonstrated the capacity of one integrase of each group to catalyze site-specific recombination *in vitro* with its cognate attachment site. We should now determine whether the integrases could catalyze site-specific recombination with the other attachment site. If not, this set of very closely related integrases that use different sites for recombination would be a great model to study specificity determinism and specificity switch for suicidal integrases.

The amino-acids that are different between Int^{pTN3} and Int^{pTF1} sequences are probably involved in the recognition of the specific site (Figure 41.A). Some are exclusively present in all the integrases of one group (signature amino-acids). They tend to be grouped in three zones (Figure 41.A). These zones are candidate of choice for sequences that could recognize the specific site. Additionally, the zone around the att translation might also be involved in specific site recognition. Zhan et al. showed that residues 81-164 of $Int^{SSV2}$are involved in recognizing and binding the specific site (Zhan et al., 2015). For Int^{pTN3}, they correspond to the sequence between the att zone and zone 2.

To experimentally determine the sequence(s) involved in specific site recognition, we can consider testing the substrate binding and site-specific activities of Int^{pTN3} and Int^{pTF1} chimeras. First, we could test a chimera of the portion detected by Zhan et al. and another with all the detected zones (Figure 41.B). Then, if these chimeras present an altered specificity, we could test chimeras of individual zones.

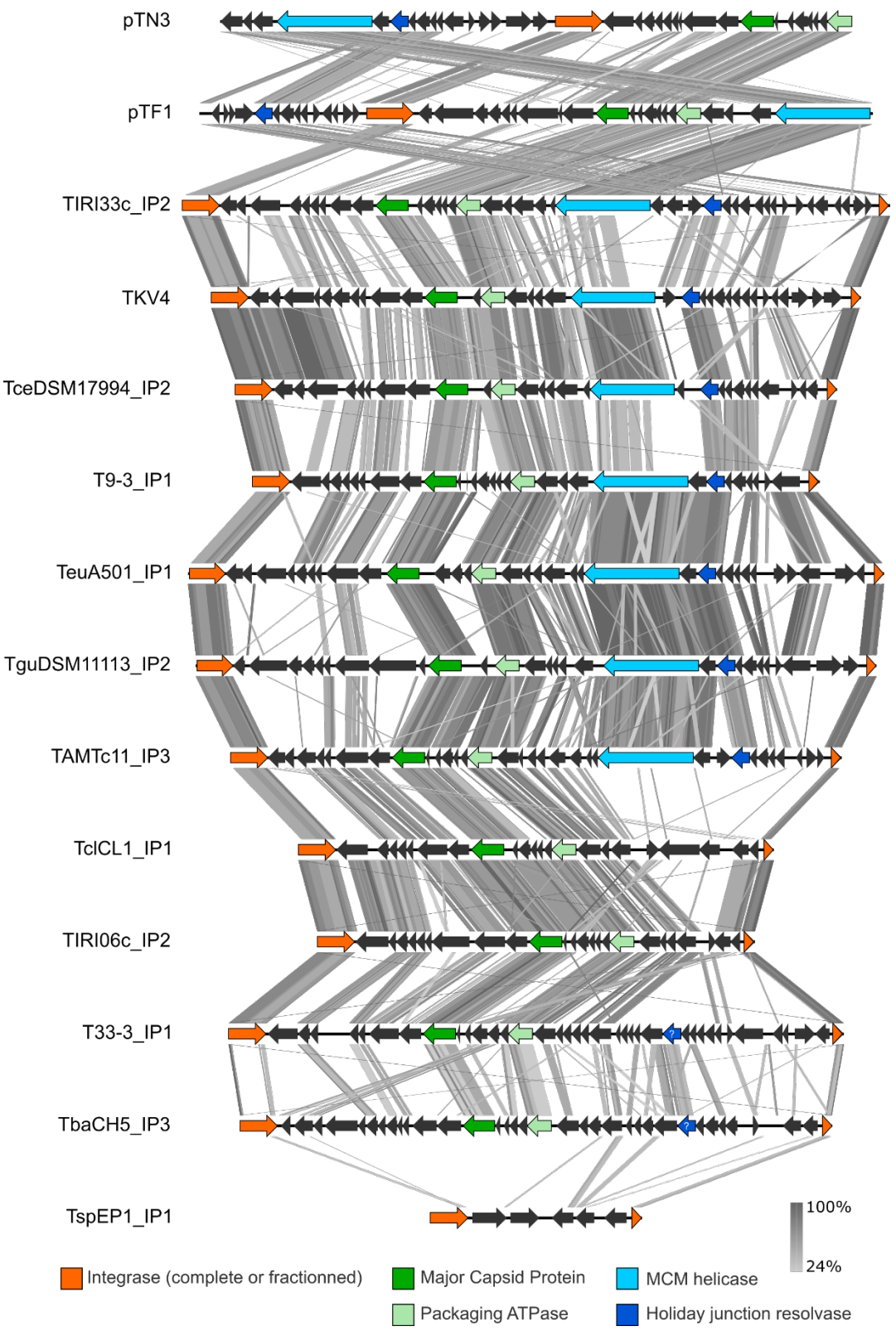


Figure 35. Synteny conservation between the different elements encoding an integrase of the Int^{pTN3} family. For each element, ORFs are represented by an arrow. Different colors indicate the known protein annotation. The grey scale represents pairwise similarity between the sequences calculated by tblastx. TspJCM11816_IP3 is excluded because it is present on two different contigs.

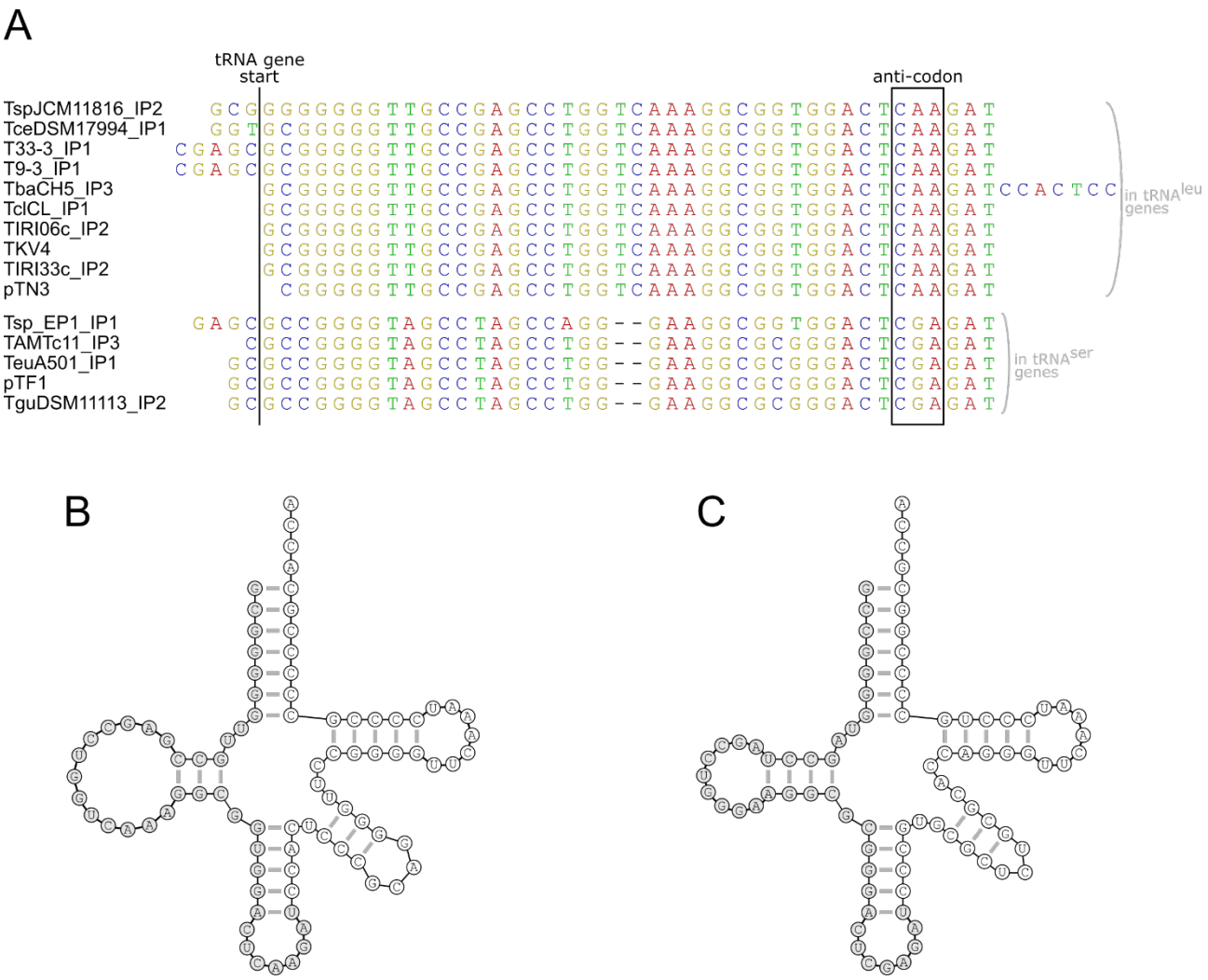


Figure 36. The Int^{pTN3} family of integrases presents two integration sites in tRNA^{Leu}CAA or tRNA^{Ser}CGA genes. A. Alignment of the att sites. B. Leaf-like structure of the tRNA^{Leu}CAA from *T. nautili*. C. Leaf-like structure of the tRNA^{Arg}CGA from *T. fumicolans*. Greyed nucleotides are present in the att site.

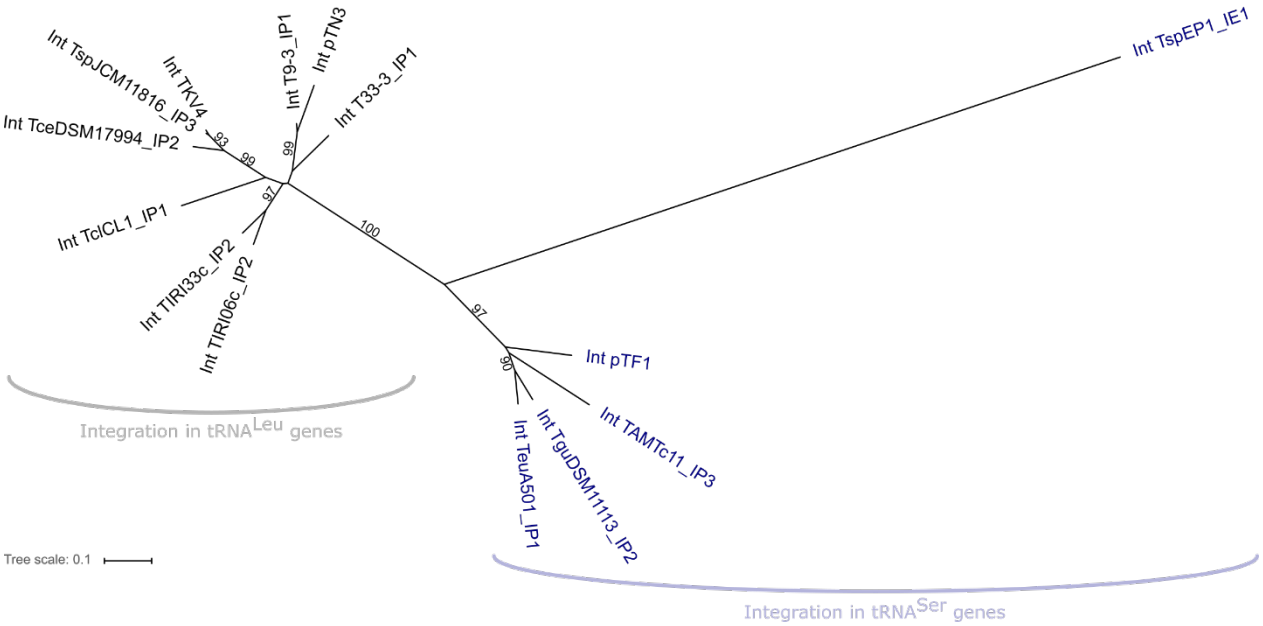
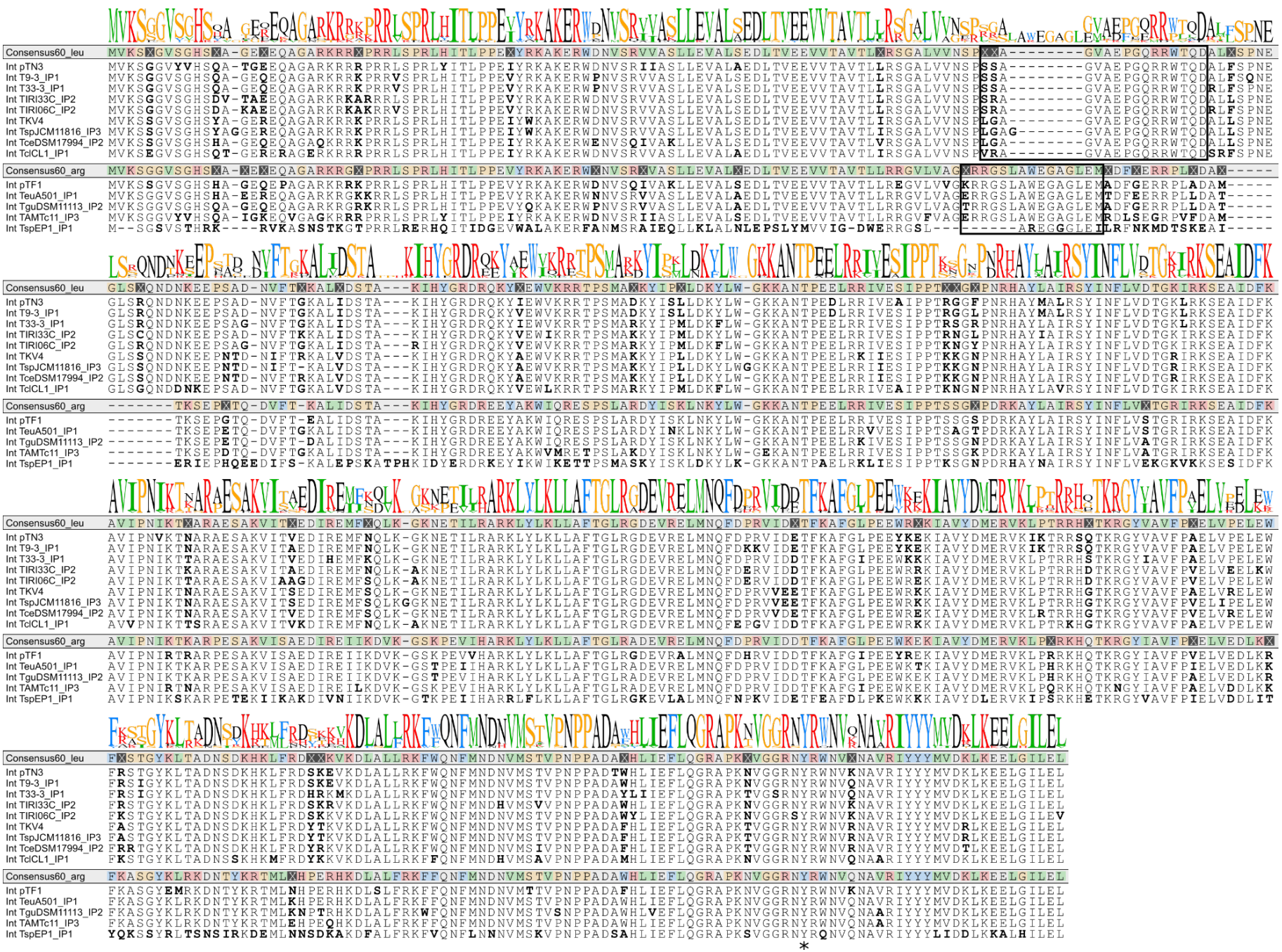


Figure 37. PhyML phylogenetic tree of the Int^{pTN3} family of integrases. The integrase of the element TbaCH5_IP3 is very divergent and attracted the integrase of the element TspEP1_IP1. It was therefore removed from the phylogenetic analysis. Sequences were aligned with MAFFT and trimmed with BMGE (442/503 remaining positions). Bootstrap values higher than 80% are indicated.

Next Page: **Figure 38. MAFFT sequence alignment of the integrases of Int^{pTN3} family** excluding the integrase of the element TbaCH5_IP3 which is very divergent. The sequence logo representing the conservation of all integrase sequences is indicated on top of the alignment. The 60% consensus sequences of all integrases targeting tRNA^{Leu} genes or of all integrases targeting tRNA^{Arg} genes are indicated. Positions varying from the consensus sequence are indicated in bold. The att site translation is boxed and the catalytic tyrosine is indicated by an asterisk.


Consensus60_leu
Int pTN3
Int T9-3_IP1
Int T33-3_IP1
Int TIRI33C_IP2
Int TIRI06C_IP2
Int TKV4
Int TspJCM11816_IP3
Int TceDSM17994_IP2
Int TclCL1_IP1
Consensus60_arg
Int pTF1
Int TeuA501_IP1
Int TguDSM11113_IP2
Int TAMTc11_IP3
Int TspEP1_IP1

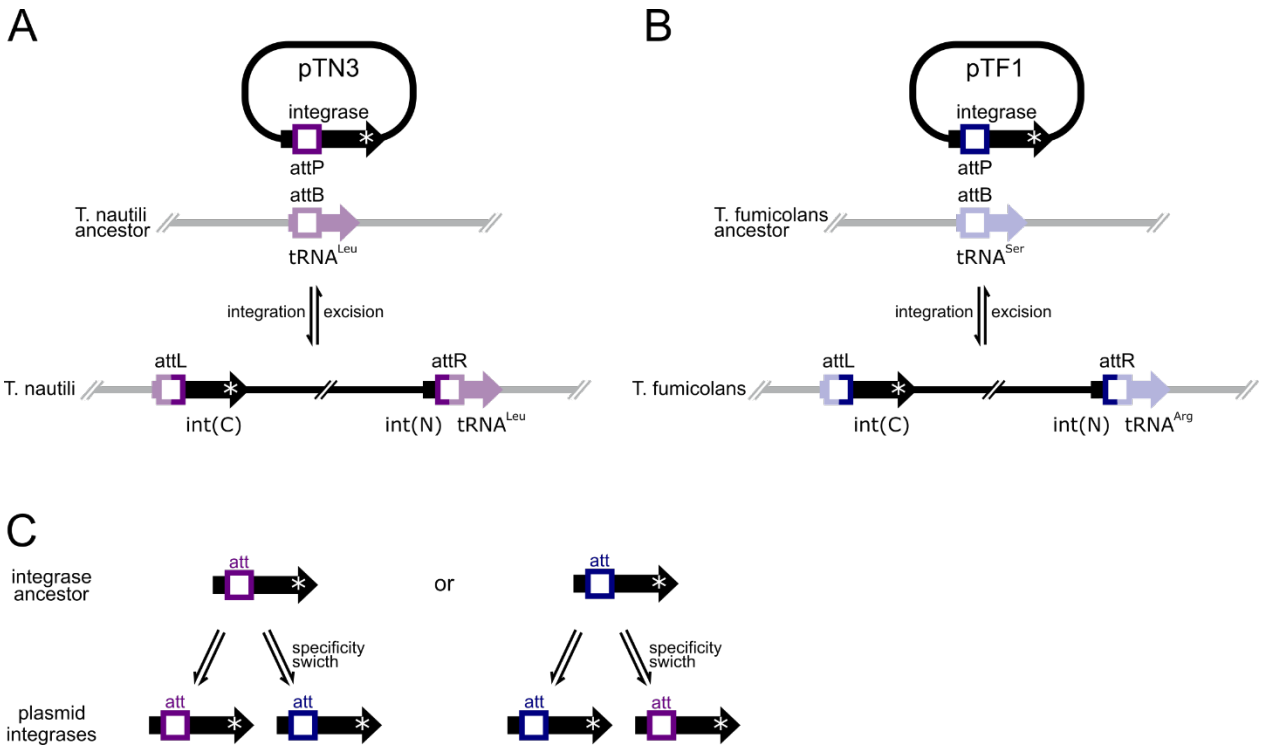


Figure 39. Plasmid integrations and integrase evolution models. A and B. Integration models for the plasmids pTN3 and pTF1, respectively. The two plasmids have two different integration specificities but the integrases are closely related. C. Evolution model for the specificity of Int^{pTN3} family of integrases from a common ancestor with one of the two specificities.

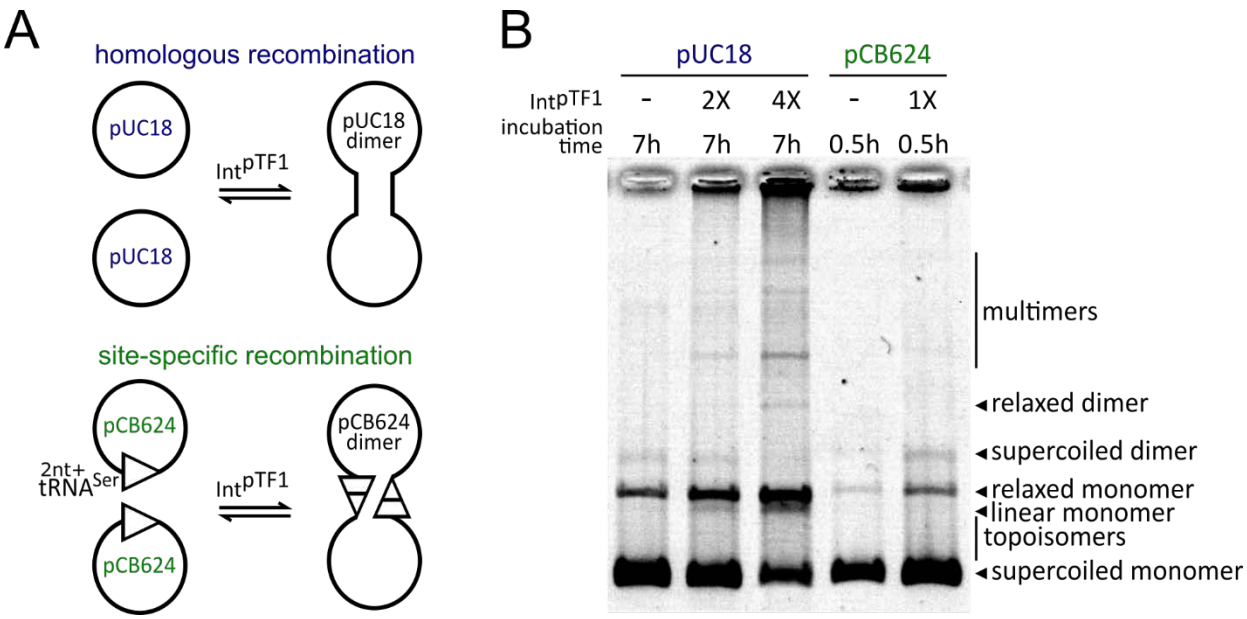


Figure 40. Int^{pTF1} can catalyze site-specific and homologous recombination *in vitro*. A. Site-specific and homologous recombination activity are assayed by an integration assay the plasmid substrates pCB624 and pUC18 respectively. B. The two plasmids were incubated with Int^{pTF1} at 65°C, treated with proteinase K and separated on a gel.

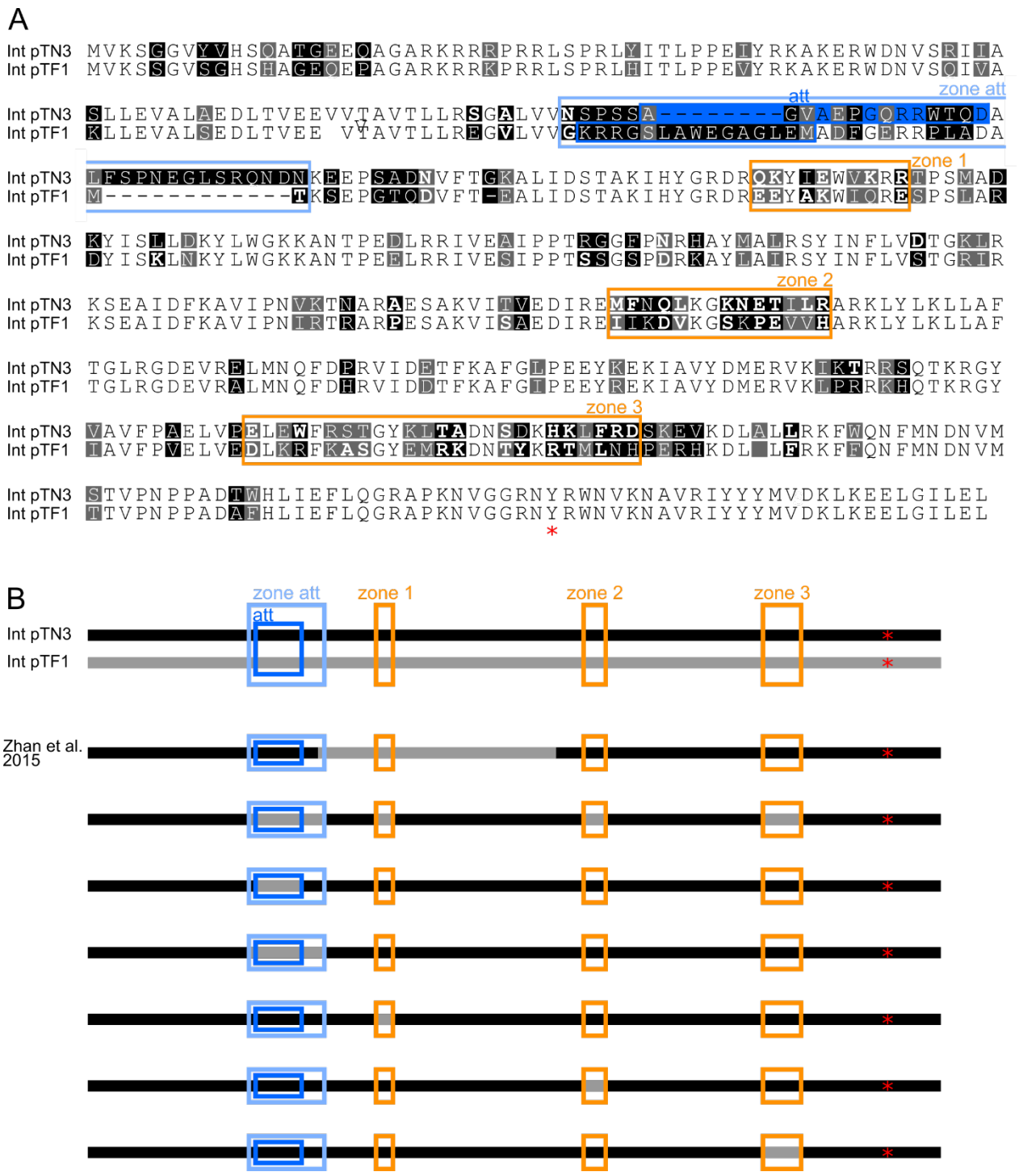


Figure 41. Integrases of plasmids pTN3 and pTF1 sequence alignment. A. The residues differing between the two integrases may be implicating in specific-site recognition and are indicated in black or in grey if they have similar properties. Bold amino-acid are signature of one integrase group. Zones with many signature amino-acid are boxed in orange. They are candidate of choice for the sequences responsible for specificity recognition. The att translation is highlighted in dark blue. The variable zone around the att translation is boxed in light blue. The catalytic tyrosine is indicated by a red asterisk. B. Int^{pTN3} and Int^{pTF1} chimers that could be tested for substrate binding and site-specific activities.

Searching for the sequence determinism of the homologous recombination activity

The homologous recombination activity is shared by all the members of the Int^{pTN3} family

As indicated above, Int^{pTN3} homologous recombination activity is not due to reaction conditions but is probably encoded by the integrase sequence. This raises the question whether only Int^{pTN3} can catalyze homologous recombination or whether this capacity is shared with the other similar integrases. We therefore decided to assay Int^{pTF1} capacity to catalyze homologous recombination *in vitro*. In an integration assay, homologous recombination between two nonspecific plasmids pUC18 leads to plasmid dimer formation (Figure 40.A page 245). After incubation with Int^{pTF1} at 65°C, we observed relaxed pUC18 dimer formation (Figure 40.B). Int^{pTF1} can efficiently catalyze homologous integration *in vitro* similarly to Int^{pTN3}. However, contrary to Int^{pTN3}, we observed very little formation of linear pUC18 (Figure 40.B). Int^{pTF1} catalyzes fewer double-stranded cuts than Int^{pTN3} in pUC18. In conclusion, both Int^{pTN3} and Int^{pTF1} can catalyze homologous recombination *in vitro*. We can reasonably assume that all integrases of the Int^{pTN3} family can catalyze homologous recombination.

Additional sequences of the Int^{pTN3} family

We previously showed that Int^{pTN3} presents additional sequences compared to the integrase Int^{SSV1} (Article 1). These additional sequences are clearly visible both in the primary sequence and in the predicted tertiary structure. The comparison of Int^{pTN3} and Int$^{pT26\text{-}2}$ evidence the same additional sequences. Since Int$^{pT26\text{-}2}$ cannot catalyze homologous recombination (Article 3), we can reasonably hypothesize that Int^{pTN3} additional sequences play a role in the homologous recombination activity. However, different alignment methods and different structure prediction programs place them differently. We therefore decided to solve Int^{pTN3} structure. We already obtained a diffracting crystal and we are analyzing the data (Figure 42). If the structure confirms the structure predictions, the first logical experiment would be to construct deletion mutant of Int^{pTN3} supplementary sequences and test whether the homologous recombination activity is singly eliminated. Once the structure is solved, we will also be able to better align primary sequences of Int^{pTN3} and Int$^{pT26\text{-}2}$ and detect individual residues that are conserved in all Int^{pTN3}-like sequences but not in Int$^{pT26\text{-}2}$-like sequences and are therefore potentially implicated in the homologous recombination activity.

In the meantime, we tried to improve the delimitation of Int^{pTN3} additional sequences. We aligned integrases of the Int^{pTN3} family with their closest relatives without any homologous recombination activity, i.e. integrases of the Int$^{pT26\text{-}2}$ family. Precisely, we aligned the 13 Int^{pTN3} family integrases available at the time with a selection of 13 integrases of the Int$^{pT26\text{-}2}$ family that are representative of their diversity. We compared the results of different alignment methods (MAFFT, PRALINE, Clustal ω or Muscle) as summarized in Figure 43. In Int^{pTN3} sequence, the residues included in the modeled supplementary loop corresponded to additional sequences or were at the limit of additional sequences compared to the outgroup sequences for each alignment method. Two Int^{pTN3} motifs that are absent from Int$^{pT26\ 2}$ were also detected: IPPT and PEVI/ETIL. Finally, a cluster of positively charged residues is present in the N-terminal fragment of Int^{pTN3} and Int$^{pT26\text{-}2}$ and is longer and more condensed for Int^{pTN3}. Overall, no obvious and well defined sequence was identified from the alignment that could be responsible for the homologous recombination activity. The structure resolution is really needed to confirm these predictions.

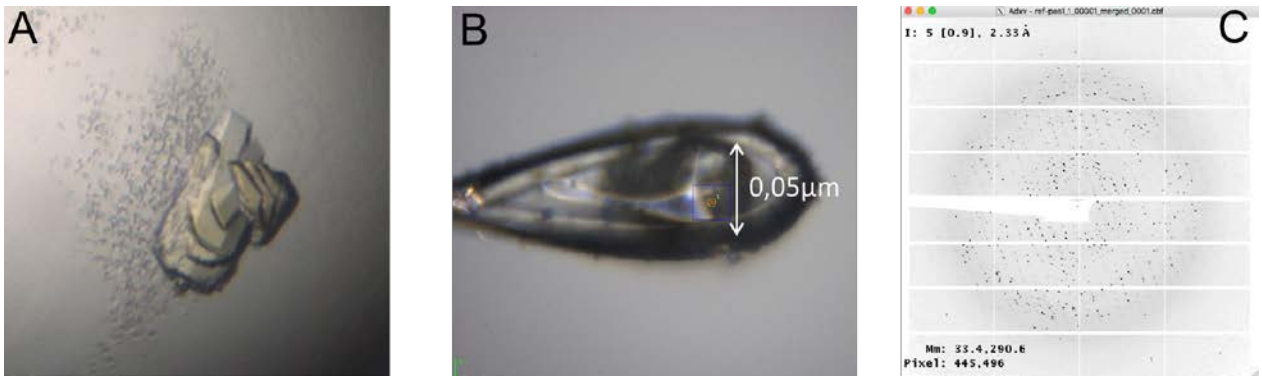
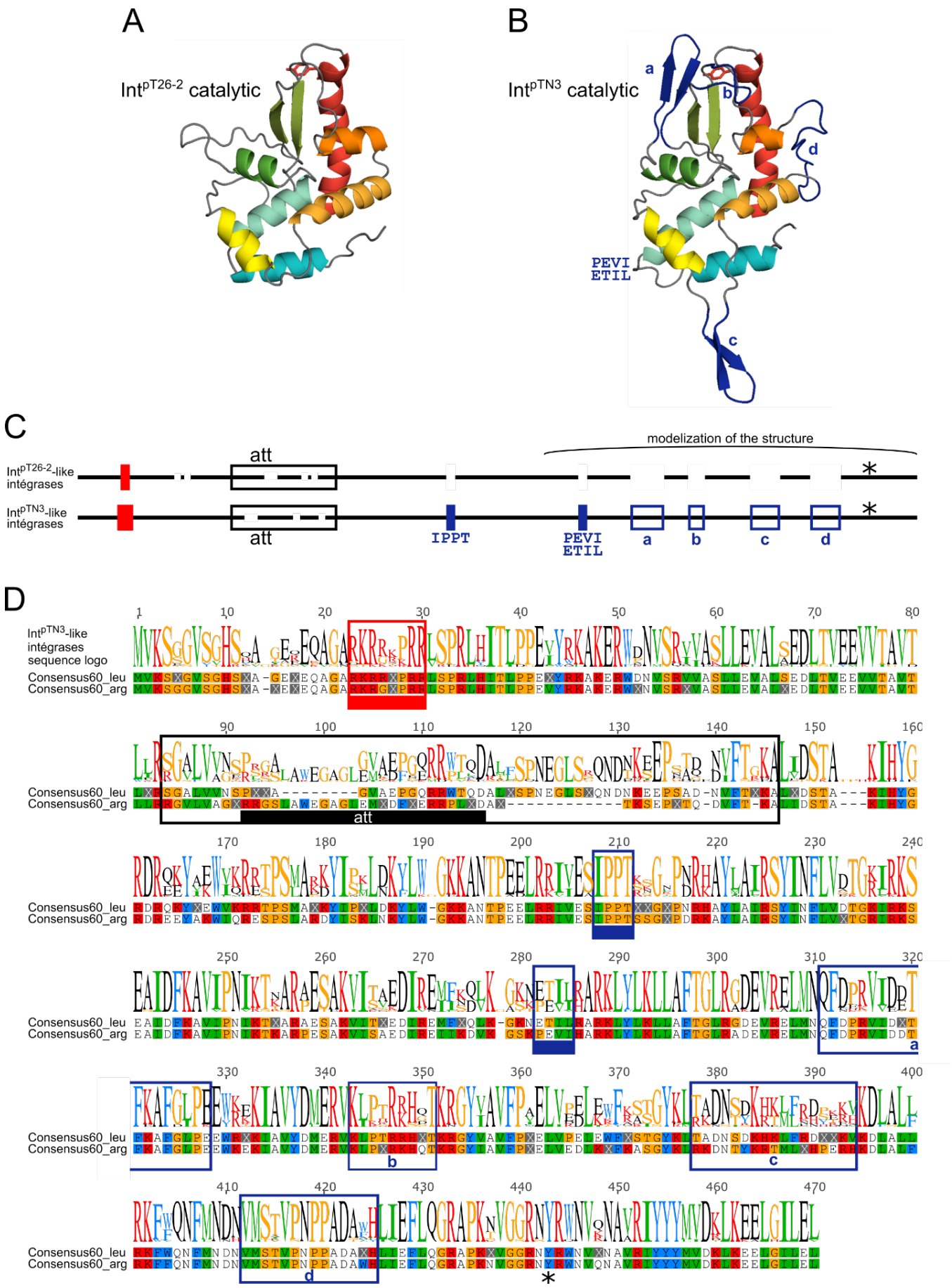


Figure 42. Int^{pTN3} crystal and its diffraction pattern. A. Image of Int^{pTN3} crystal. B. Portion of the Int^{pTN3} crystal mounted on Cryo Loop. C. The diffraction image collected in the PX1 SOLEIL beamline.

Next page: **Figure 43. The Int^{pTN3} family of integrases presents additional sequences compared to the Int^{pT26-2} family of integrases,** its closest known relative. The structures of the catalytic part of the integrases Int^{pT26-2} (A) and Int^{pTN3} (B) were modeled with Phyre2 based on the structure of the catalytic part of Int^{SSV1}. Int^{pTN3} additional structures are colored in dark blue and named from a to d. C. Schematic representation of the alignment of integrases from the Int^{pTN3} family and from the Int^{pT26-2} family by different methods (MAFFT, PRALINE, Clustal ω or Muscle). Sequences a to d corresponded to indels or to the borders of indels. Two additional sequences were detected that are specific to the Int^{pTN3} family of integrases and are indicated as full dark blue boxes with their corresponding amino-acids. The extended variable region around the att site is indicated by a black box. A cluster of positively charged amino-acids is present at the beginning of all sequences and is indicated as a red box. It longer for integrases of the Int^{pTN3} family. D. The above mentioned sequences are indicated on the 60% sequence consensus and sequence logo of the MAFFT sequence alignment of the Int^{pTN3} family of integrases.


A
Int^pT26-2 catalytic
B
Int^pTN3 catalytic
a
b
d
PEVI
ETIL
c
C
modelization of the structure
Int^pT26-2-like intégrases
Int^pTN3-like intégrases
att
att
IPPT
PEVI
ETIL
a
b
c
d
D
Int^pTN3-like intégrases sequence logo
Consensus60_leu
Consensus60_arg
att
a
b
c
d

Conclusions and perspectives

L'intégrase du plasmide pTN3 catalyse des inversions chromosomiques chez les *Thermococcus* par une activité de recombinaison homologue

Jusqu'à présent, aucun mécanisme n'avait été mis en évidence pour les nombreuses inversions observées dans les chromosomes de Thermococcales. Nous avons démontré que l'intégrase du plasmide pTN3 présente une activité catalytique classique de recombinaison site-spécifique doublée d'une activité inédite de recombinaison entre séquences répétées de plus de 100 paires de bases. Cette deuxième activité présente le mécanisme conservatif de recombinase à tyrosine mais utilise des séquences identiques non spécifiques comme la recombinaison homologue. Elle permet des recombinaisons entre séquences inversées répétées qui sont à l'origine des inversions observées dans le chromosome de *Thermococcus nautili* et probablement des autres espèces de *Thermococcus*. Nous pouvons maintenant envisager des applications biotechnologiques à l'activité de recombinasion homologue de l'intégrase du plasmide pTN3. Par exemple, l'intégrase pourrait utilisée pour modifier spécifiquement et précisément n'importe quelle séquence d'ADN (mutagenèse dirigiée).

La découverte de l'activité inédite de l'intégrase du plasmide pTN3 soulève deux nouveaux axes de recherche. Dans un premier axe, nous avons cherché à disséquer la double activité de l'intégrase. Un deuxième axe s'interroge sur les conséquences évolutives d'une telle activité pour le plasmide et pour l'hôte.

Dissection des deux activités catalytiques de l'intégrase du plasmide pTN3

Nous avons exploré deux pistes de recherche pour disséquer la double activité de l'intégrase du plasmide pTN3 qui devraient porter leurs fruits dans un futur proche. Premièrement, en caractérisant biochimiquement les intégrases des plasmides pTF1 et pT26-2 de Thermococcales, nous avons démontré que l'activité double est catalysée exclusivement par les intégrases de type pTN3. Cette approche a confirmé un rôle possible dans l'activité de recombinaison homologue des boucles spécifiques aux intégrases de type pTN3. Deuxièmement, nous avons entrepris la résolution de la structure de l'intégrase du plasmide pTN3 en collaboration avec l'équipe d'Hermann Van Tilbeurgh (I2BC). Nous avons réussi à obtenir des cristaux diffractants et devrions bientôt resoudre la structure de l'intégrase. Nous pourrions ensuite chercher à obtenir des co-cristaux l'intégrase avec différents substrats ADN pour déterminer la structure du complexe synaptique, à différents stades de réaction et pour les deux activités. Ceci éclairerait fortement les mécanismes moléculaires de la double activité de cette recombinase à tyrosine. La résolution de structures de l'intégrase permettrait aussi de confirmer (ou non) la présence de boucles spécifiques et de déterminer leur séquences. Nous pourrions alors identifier des séquences et des résidus importants à l'activité de recombinaison homologue et confirmer leur rôle par mutagénèse. Nous pourrions aussi greffer les résidus identifiés à l'intégrase $Int^{pT26\text{-}2}$ pour vérifier s'ils sont suffisants à conférer l'activité de recombinaison homologue.

Conséquences évolutives de l'activité de l'intégrase du plasmide pTN3

Du côté du plasmide

Le plasmide pTN3 encode une intégrase avec une activité de recombinaison inédite. Cette activité est-elle fortuite ? Est-elle directement impliquée dans le cycle de vie du plasmide ? Est-elle un cadeau du plasmide à l'hôte qui favoriserait sa rétention ?

La seconde activité des intégrases de type pTN3 pourrait être une conséquence fortuite du récent changement de spécificité observé dans cette famille. Il a été proposé qu'un changement de spécificité pourrait passer par une étape de relaxation de la spécificité (Dorgai et al., 1995; Voziyanov et al., 2003). La seconde activité des intégrases de la famille pTN3 pourrait alors correspondre à une activité de recombinaison site-spécifique relachée. Cela ne semble cependant pas être le cas pour au moins deux raisons. Premièrement, les deux activités de recombinaison site-spécifique et homologue ont des efficacités différentes qui les distinguent. Deuxièmement, l'activité de recombinaison homologue ne semble pas correspondre à une spécificité de site relâchée mais bien à une activité particulière de recombinaison entre deux séquences identiques sans résidus spécifiques particuliers. Dans tous les cas, la résolution de la structure de Int^{pTN3} suivie d'une analyse structure-fonction élucideront probablement cet aspect.

L'activité particulière des intégrases de type pTN3 pourrait apporter un avantage évolutif direct au plasmide. Cette activité de recombinaison homologue correspond peut-être à une forme d'évolvabilité programmée qui favorise la recombinaison entre éléments génétiques mobiles et donc l'évolution modulaire du plasmide pTN3. On s'attendrait alors à observer des intégrases de type pTN3 dans un groupe d'élément mobiles génétiques très modulaires. Ce n'est pas le cas puisque tous les plasmides portant une intégrase de type pTN3 forment un ensemble très homogène de plasmides de type pTN3, à une exception près.

Finalement, la seconde activité des intégrases de type Int^{pTN3} pourrait bénéficier indirectement au plasmide si elle est favorable à l'hôte. Elle augmenterait alors l'acceptation du plasmide par l'hôte. Nous allons aborder le bénéfice possible de l'intégrase pour l'hôte dans le paragraphe suivant.

Du côté des hôtes Thermococcales

Le chromosome des Thermococcales présente des substrats potentiels (multiples copies du chromosome et séquences répétées) pour l'activité de recombinaison homologue de l'intégrase. Cette activité peut donc impacter le chromosome, soit par une fonction biologique directe soit par l'intermédiaire des inversions. Les conséquences sont-elles favorables ou délétères pour l'hôte ?

Une activité de recombinaison homologue

L'activité de recombinaison homologue de l'intégrase entière pourrait être utilisée directement par l'hôte. Il est possible que l'intégrase fournisse une voie alternative à la recombinaison homologue dépendante de la recombinase RadA. Nous pourrions tester cette hypothèse en générant un mutant RadA dans une souche de *Thermococcus kodakarensis* exprimant l'intégrase Int^{pTN3}. RadA étant une protéine essentielle de *Thermococcus kodakarensis* (Fujikane et al., 2010), l'obtention d'un tel mutant indiquerait une complémentation fonctionnelle de la fonction biologie de la recombinaison homologue dépendant de RadA.

Des inversions dans le chromosome

Nous avons montré que l'activité de recombinaison homologue de l'intégrase Int^{pTN3} est responsable d'inversions dans le chromosome de l'hôte. Les inversions chromosomiques ont probablement des implications évolutives majeures puisqu'elles changent l'organisation millimétrée du chromosome. Ces changement peuvent auusi bien entrainer une chute de la fitness de l'hôte en désorganisant le chromosome ou au contraire l'augmenter en créant une nouvelle organisation plus favorable. Ils peuvent enfin avoir un effet neutre si les contraintes pesant sur l'oganisation du chromosome sont relachées.

L'organisation du chromosome archéen est principalement liée à des contraintes de réplication, et particulièrement à la progression de la réplication à partir d'une origine. Cependant, il a été démontré que certaines archées n'utilisent pas leur origine de réplication sous certaines conditions (Hawkins et al., 2013). Des résultats récents du laboratoire confirment que certaines souches de Thermococcales, dont *Thermococcus nautili* peuvent ne pas utiliser l'origine de réplication. La contrainte d'organisation du chromosome s'en trouve peut-être relâchée expliquant ainsi la fréquence d'accumulation d'inversions. En outre, il a été démontré au laboratoire que les inversions tendent à ne pas modifier exagérément l'organisation du chromosome en conservant la position de clusters de gènes importants à une distance constante de l'origine de réplication (Cossu et al., 2015). Les inversions pourraient donc avoir un effet relativement neutre pour l'hôte.

Au premier abord, les inversions observées chez *Thermococcus nautili* sont impressionnantes car elles ont été rapides et impactent environ un tiers du chromosome. Leur histoire est cependant plus complexe qu'il n'y parait. En effet, la comparaison des chromosomes des deux souches de *Thermococcus nautili* avec les chromosomes de souches proches montre que le chromosome « type » est en réalité celui de la souche 66G qui est le plus synténique avec les souches proches (Figure 44). Il semblerait donc que les inversions observées au laboratoire correspondent à un retour à l'état initial. L'intégrase catalyse donc des inversions temporaires qui réalisent des allers-retours dans le chromosome. Ces inversions n'ont peut-être pas le temps d'exprimer un potentiel effet évolutif.

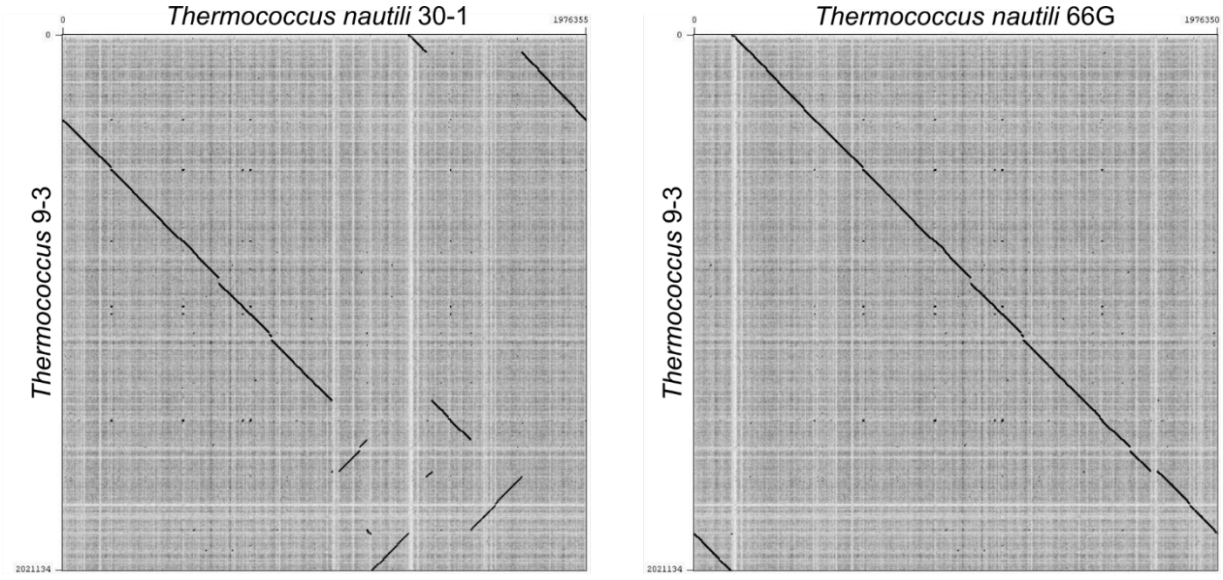


Figure 44. Comparaison par dotplot des chromosomes des deux souches de *Thermococcus nautili* avec *Thermococcus* 9-3.

Que les inversions soient neutres, favorables ou délétères, on peut se demander si elles sont recherchées, évitées ou alors subies par les chromosomes de Thermococcales. L'hôte possède deux leviers d'influence sur la fréquence des inversions. Le premier consiste à contrôler le nombre de copies de l'enzyme, par exemple en contrôlant le nombre de copies du plasmide. Le deuxième consiste à contrôler le nombre de copies du substrat, c'est-à-dire les répétitions de plus de 100 paires de bases. En cas de rôle favorable des inversions, le chromosome aurait avantage à conserver les répétitions. A l'inverse, en cas de rôle défavorable, le chromosome aurait avantage à les supprimer. Une analyse préliminaire du nombre de répétitions dans les chromosomes de Thermococcales montre qu'elles sont moins nombreuses dans les chromosomes de *Thermococcus* que dans ceux de *Pyrococcus*. Or, le plasmide pTN3 est seulement présent dans des espèces *Thermococcus*. Il semblerait donc que les répétitions soient éliminées des chromosomes rencontrant les intégrase de type pTN3, pointant vers un rôle délétère des inversions catalysées. Pour confirmer cette tendance, il faudrait maintenant étendre cette étude à tous les chromosomes disponibles au laboratoire en différenciant les souches ayant effectivement rencontré une intégrase de type pTN3 des autres.

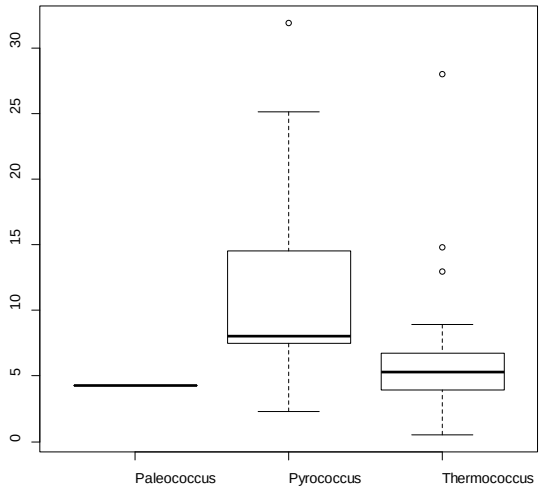


Figure 45. Nombre de répétition d'une taille supérieure à 100 pb par megabase dans les chromosomes de Thermococcales publiés avant janvier 2017. Les résultats sont indiqués sous forme de boite à moustache pour les espèces *Paleococcus*, *Pyrococcus* ou *Thermococcus* de gauche à droite.

Avantages évolutifs de l'activité suicidaire des intégrases d'archées

Des intégrases spécifiques des archées présentent une stratégie d'intégration singulière et assimilable à un suicide catalytique. Le gène codant pour l'intégrase est séparée en deux parties *int(N)* et *int(C)* lors de l'intégration du fait de la localisation intragénique du site de recombinaison attP (Figure 26 page 57). Cette fragmentation semble neutraliser l'activité de recombinaison codée par l'élément génétique mobile qui ne peut pas s'exciser. L'absence d'excision a effectivement été observée pour plusieurs éléments génétiques mobiles à intégrase de type suicide (She et al., 2001b; Wang et al., 2007). Nous avons confirmé l'absence d'excision spontanée de l'élément intégré TKV4 de *Thermococcus kodakarensis* (Article 1). La neutralisation de l'activité de recombinaison peut resulter de l'absence de promoteur pour le fragment *int(C)* ou de l'absence d'activité de l'intégrase fragmentée. Le fragment *int(C)* est probablement exprimé pour TKV4 puisqu'il complémente un mutant catalytique d'intégrase (Article 1). Cependant, l'élément TKV4 présente une intégration particulière où le fragment *int(C)* est sous le contrôle du promoteur du gène d'ARNt d'intégration. Le fragment *int(C)* n'est peut-être pas exprimé dans d'autres cas. Les polypeptides Int(C) de TKV4 et TPV1 ne semblent pas avoir d'activité autonome (Article 1, Part 2). Au contraire, la partie Int(C) de l'intégrase de SSV2 présente une activité catalytique *in vitro* (Zhan et al., 2015). Le mécanisme exact de l'inactivation post intégration de ces intégrases n'est pas élucidé.

La rétention évolutive de ces enzymes à l'activité catalytique suicidaire était jusqu'à présent restée énigmatique. Dans les mêmes conditions environnementales, le même background plasmidique peut recruter soit une intégrase suicidaire soit une intégrase classique. L'avantage des intégrases suicidaires n'est donc pas lié aux conditions environnementales ou à des fonctions plasmidiques particulières. A travers mes travaux de thèse, j'ai pu mettre en évidence des avantages évolutifs auparavant non identifiés apportés par la fragmentation suicidaire des intégrases (Article 3).

Un module d'intégration tout en un

Les systèmes d'intégration d'éléments génétiques mobiles incluent souvent des mécanismes de contrôle de la directionnalité de recombinaison qui sont complexes et coûteux en séquences codantes (voir par exemple le cas de l'intégrase du phage λ page 54). Ces mécanismes incluent notamment des protéines accessoires appelées facteurs de directionnalité de la recombinaison (RDF). La stratégie d'intégration suicidaire permet d'éviter d'avoir recours à de tels facteurs. Nous avons en effet montré que deux intégrases de type suicidaire sont actives pour toutes les directionnalités de recombinaison en absence de protéine accessoire (Part 1 et Article 3). Pour les intégrases de type suicide, la directionnalité de recombinaison n'est pas contrôlée lors de la réaction mais par la présence d'une intégrase catalytiquement active. Lorsque l'élément génétique mobile est libre, il présente une intégrase active et peut s'intégrer. Lorsqu'il est intégré, il ne présente pas d'intégrase active et ne peut pas s'exciser. En contrepartie de l'absence de RDF, l'élément génétique mobile n'a pas de contrôle sur son excision et est tributaire de l'arrivée d'une intégrase intacte. En plus du contrôle de directionnalité de réaction, la séquence codante de l'intégrase contient le site de recombinaison attP. Cette séquence codante est donc un système d'intégration tout en un : activité catalytique, séquence de recombinaison et contrôle de la directionnalité sont inclus. En conséquence, il est très facile de recruter ce système de recombinaison compact puisqu'il suffit d'acquérir une séquence relativement courte et continue.

Coopération salvatrice entre intégrases suicidaires

Les éléments génétiques mobiles intégrés par un système suicidaire ne sont pas définitivement piégés à l'intérieur du chromosome. Une intégrase suicide exogène et intacte peut en effet catalyser leur excision (Wang et al., 2007). Cette catalyse semble reposer sur une coopération entre les fragments Int(N) et Int(C) et l'intégrase exogène (Part 1). Ce sauvetage coopératif n'a pour l'instant été observé que pour des intégrases phylogénétiquement proches et nous ne savons pas s'il est possible entre intégrases suicidaires plus éloignées. Avec ce mécanisme d'excision, l'élément génétique mobile n'a aucun contrôle sur le timing d'excision. L'excision n'est pas déclenchée par un stress environnemental comme pour le phage λ, mais par l'arrivée d'un élément génétique mobile de type suicidaire. Cette stratégie de coopération entre intégrases de type suicidaire n'est rentable que s'il y a suffisamment d'intégrases de type suicidaire en circulation dans la population. La probabilité d'excision sera alors suffisamment élevée pour éviter le piégeage trop long de l'élément génétique mobile. Cette condition est bien respectée par les intégrases de type suicidaires qui sont très prévalentes chez les Thermococcales notamment (Article 2 et Article 3).

Un suicide coercitif

Les intégrases de type suicide ont été majoritairement identifiées pour des hôtes hyperthermophiles. Ils présentent des systèmes de défense contre les éléments génétiques mobiles particulièrement développés (Koonin et al., 2017), notamment les systèmes de type CRISPR-Cas. Fusco et al. ont proposé que la stratégie d'intégration suicidaire pourrait être un moyen de contraindre l'hôte à éteindre l'immunité CRISPR-Cas (Fusco et al., 2015). Après l'intégration d'un élément mobile suicidaire, l'hôte est confronté à deux solutions génétiques : conserver des spacers CRISPR actifs qui ciblent une portion du chromosome (auto-immunité) ou éteindre ces spacers. La deuxième solution est celle recherchée par l'élément génétique mobile.

La fragmentation comme mécanisme évolutif

L'analyse de la diversité de séquences des intégrases identifiées dans l'Article 3 nous a permis de proposer un mécanisme évolutif de ces intégrases lié à leur suicide. Le gène d'une intégrase intégrée est sous la forme de deux fragments *int(N)* et *int(C)* qui peuvent être mélangés par recombinaison homologue avec ceux d'une autre intégrase intégrée dans le même chromosome. Ceci aboutit à la formation d'une intégrase chimérique qui est potentiellement active. Nous avons observé ce mécanisme pour des intégrases proches phylogénétiquement mais nous ne savons pas s'il existe entre fragments d'intégrases différentes, de type Int^{pTN3} et $Int^{pT26\text{-}2}$ par exemple. Ce mécanisme permet de mixer et diffuser la diversité des intégrases suicidaires entre différents éléments génétiques mobiles. Il permet notamment l'échange de spécificité de recombinaison. Ce mécanisme nécessite alors une certaine souplesse de spécificité des intégrases que nous avons observé pour l'intégrase du plasmide pT26-2 (Article 3). Cette souplesse facilite notamment la probabilité d'activité pour l'intégrase nouvellement créée. Ce mécanisme requiert également la présence concomitante de deux intégrases suicide dans un chromosome. Il nécessite donc une forte présence des intégrases dans la population. Une limitation de ce mécanisme est qu'il permet seulement l'échange de diversité et non pas la création de diversité, notamment en terme de spécificité de recombinaison. Nous connaissons pour l'instant peu les mécanismes d'apparition d'une nouvelle spécificité de recombinaison.

Nous avons mis en évidence différents avantages à la stratégie de fragmentation des intégrases suicidaires. Du point de vue de l'élément génétique mobile, l'intégration n'aura pas les mêmes conséquences suivant qu'il utilise une intégrase classique ou une intégrase suicidaire. Notamment, les conditions et la probabilité d'excision seront différentes dans les deux cas. Le même plasmide peut d'ailleurs varier les stratégies d'intégrations. Nous ne savons pas quels paramètres favorisent le choix de l'une ou de l'autre stratégie. Les intégrases suicidaires présentent véritablement une histoire évolutive fascinante qui soulève encore de nombreuses questions portant sur leur origine, leur lien avec l'hyperthermophilie ou la reconnaissance de la spécificité de recombinaison.

Origine évolutive des intégrases de type suicide

Nous ne connaissons pas l'origine évolutive des intégrases suicidaires. Il est en effet difficile de les inclure dans une phylogénie avec les intégrases de classiques à cause de la divergence de leurs séquences. Il semble cependant que toutes les intégrases suicidaires aient une origine évolutive unique. De plus, l'apparition de ces intégrases interpelle. Si une seule intégrase suicidaire était apparue, elle aurait été inactivée dès son intégration et aurait disparu. En outre, les avantages évolutifs que nous avons mis en évidence pour la stratégie d'intégration suicidaire reposent pour beaucoup sur la présence d'une grande quantité d'intégrases dans la population, ce qui n'était pas le cas au début de leur histoire évolutive. Ces avantages ne permettent donc pas d'expliquer l'origine de ces intégrases.

Une hypothèse possible pour l'origine des intégrases de type suicide est la création d'un promoteur additionnel en amont d'une intégrase classique (Figure 46). Les intégrases classiques sont en effet souvent encodées à proximité immédiate du site d'intégration attP (Wang et al., 2018). Une étape évolutive intermédiaire avec deux promoteurs alternatifs aurait alors facilité l'évolution des particularités de l'intégration suicidaire dans un contexte de contraintes relâchées permis par la production et l'activité de l'intégrase initiale. Il a en effet été montré que l'évolution de nouvelles fonctions protéiques passe souvent par une étape de double fonctionnalité (Aharoni et al., 2005). Le promoteur initial aurait f progressivement disparu aboutissant à la création de la situation actuelle.

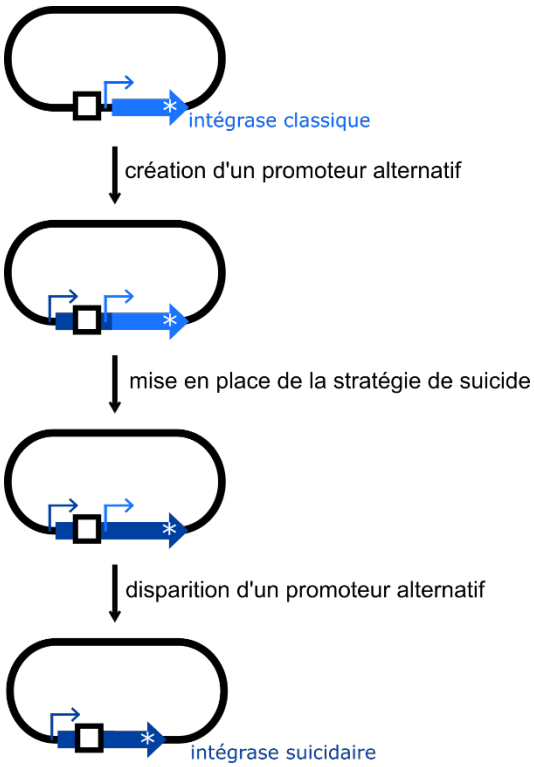


Figure 46. Une hypothèse pour l'origine des intégrases de type suicide à partir d'une intégrase classique

Intégrases suicidaires et hyperthermophilie

Toutes les intégrases de type suicide avaient jusqu'à présent été identifiées chez des archées hyperthermophiles. Nous avons confirmé cette tendance tout en identifiant pour la première fois des intégrases suicidaires mésophiles (Article 3). Nous ne savons pas si la présence des intégrases de type suicide dans des environnements hyperthermophiles est une coïncidence évolutive ou si elle reflète une réalité fonctionnelle. La caractérisation détaillée d'une intégrase classique hyperthermophile pourrait éclairer cette question. Nous avons ainsi commencé la caractérisation d'une intégrase de Methanococcales identifiée dans le même fond plasmidique que l'intégrase du plasmide pT26-2 (Part 2). La comparaison des deux intégrases pourrait permettre de comprendre pourquoi deux stratégies d'intégration distinctes sont utilisées par la même famille de plasmides hyperthermophiles. Une approche alternative serait de comparer des intégrases classiques et suicidaires présentes dans le même chromosome hôte. Une intégrase classique potentielle a été identifiée chez *Pyrococcus yayanosii* qui présente également une intégrase de type suicide (Li et al., 2016) (Article 2 Figure 47). La caractérisation de ces deux intégrases infectant le même hôte pourrait éclairer le lien entre hyperthermophilie et stratégie de suicide.

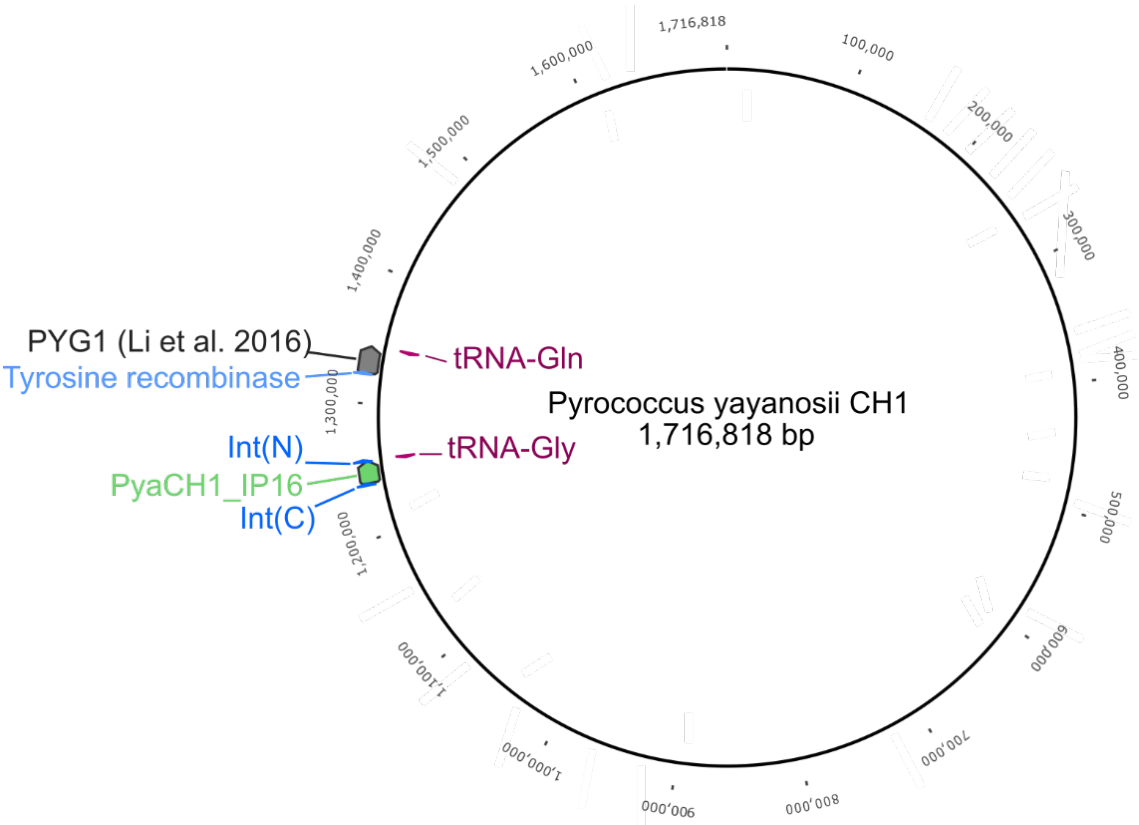


Figure 47. L'ilot génomique PYG1 et le plasmide intégré PyaCH1_IP16 sont intégrés au niveau de gènes codant pour des ARNt dans le chromosome de *Pyrococcus yayanosii* CH1. PYG1 présente une tyrosine recombinase classique et PyaCH1_IP16 code une intégrase de type suicide.

Reconnaissance de la spécificité par les intégrases suicidaires

Nous ne connaissons pas les mécanismes de reconnaissance de la spécificité pour les intégrases suicide. Les intégrases Int^{pTN3} and Int^{pTF1} sont très proches phylogénétiquement mais ne présentent pas la même spécificité d'intégration (Part 5). Nous ne savons pas si ces deux intégrases sont seulement capables de catalyser des recombinaisons avec leur propre spécificité de recombinaison ou si elles peuvent utiliser les deux spécificités de recombinaison. Si elles ne sont capables d'utiliser que leur propre site spécifique, alors elles pourraient être utilisées pour déterminer les mécanismes de reconnaissance spécifique de site en mettant en place une stratégie expérimentale similaire à celle utilisée pour les intégrases des phages λ et HK022 (Azaro and Landy, 2002). La caractérisation de chimères des deux intégrases et de mutants de l'intégrase du phage λ a permis d'identifier 5 résidus importants pour la reconnaissance spécifique dont certains limitent le spectre de spécificité et d'autres l'augmentent. Dans notre cas, la caractérisation de la spécificité de chimères de Int^{pTN3} and Int^{pTF1} comme celles proposées dans la Figure 41 (page 246) permettrait d'identifier les résidus potentiellement impliqués dans la reconnaissance de la spécificité. Cette étude permettrait également de mieux comprendre l'apparition de nouvelles spécificités de recombinaison chez les intégrases suicidaires.

Intégrases d'archées, revue de mes résultats et de la littérature

Article 5. Archaeal tyrosine integrases

Integrases are enzymes encoded by mobile genetic elements (MGE) that catalyze their integration into the host chromosome. They can be of serine or tyrosine type. Few serine integrases are encoded in archaea and none was characterized (Filée et al., 2007). On the contrary, many tyrosine integrases are encoded by archaea. They were reviewed twice in 2002 and 2004 (She et al., 2002, 2004). Since then, many new archaeal tyrosine integrase sequences were annotated (e.g. (Atanasova et al., 2018; Liu et al., 2015; Mochizuki et al., 2011) and several integrases were biochemically or structurally characterized (e.g. (Eilers et al., 2012; Wang et al., 2018; Zhan et al., 2012, 2015). Accounting for more than a decade of research, we propose an early version of a review of the literature on archaeal tyrosine integrases. Notably, we propose to rename the SSV-like integrases as suicide integrases. This renaming serves two purposes: (1) it clearly states the particularity of these integrases and (2) it clearly distinguishes the integrase whose gene is fragmented upon integration from the integrases whose sequence is similar to the SSV1 one.

Archaeal tyrosine integrases

The evolution of cellular genomes is shaped by mobile genetic elements (MGEs) as viruses, plasmids, pathogenicity islands or conjugative and mobilizable elements. They are capable of integrating their genes into their host genome or transferring them horizontally. MGE transfer rely on the concerted expression of their genes and of a number of cellular functions leading to the subversion of host physiology and the production of progeny. This process can be initiated shortly after cell entry, delayed at a later stage or even postponed indefinitely, depending on the MGE lifestyle. Some MGEs can only use one of these lifestyles and other are equipped with complex regulatory circuits allowing them to choose what fits best to their survival in varying environment conditions. In the phage λ paradigm, two lifestyles are defined as lysis and lysogeny and involve a bistable state of the phage DNA, respectively circular episome or prophage integrated into the host genome. The molecular interconversion between these two genomic states is catalyzed by a site-specific DNA recombinase or integrase. Site-specific recombinases are classified into two unrelated families, serine or tyrosine recombinases, referring to the catalytic amino-acid involved in the covalent link between the protein and the DNA substrate (Grindley et al., 2006). These two types of DNA recombinases are encountered in the three domains of life and have been characterized extensively in both Bacteria and Eukarya (Azaro and Landy, 2002; Duyne, 2015; Escudero et al., 2015; Jayaram et al., 2015; Landy, 2015; Meinke et al., 2016; Stark, 2015). Archaeal serine recombinases have been observed in transposons but were never fully investigated (Filée et al., 2007). On the other hand, several archaeal tyrosine recombinases have been analyzed and reviewed (She et al., 2002, 2004). She et al. defined two classes I and II of archaeal integrases (Figure 1). Class I corresponded to the SSV-like integrases found in Sulfolobales viruses and whose gene is neutralized by fragmentation upon integration. Class II corresponded to the pNOB8-like integrases that present a classical integration strategy. Recently, the study of archaeal tyrosine recombinases has produced a considerable amount of data which will be put in perspective in the present review. After an overview of archaeal tyrosine integrases diversity, we will present their catalytic activity and finally their biological functions. We will also consider archaeal tyrosine recombinases XerA, which present an activity similar to integrases.

Outline

Archaeal tyrosine recombinase diversity 3

Site-specific recombination from a biochemical point of view 3

A common catalytic mechanism for different reaction directionalities 3

Site-specific recombination activity of unescorted archaeal integrases 3

Insights into primary and tertiary sequence 4

DNA relaxation activity of archaeal integrases 5

Postmortem suicide integrase activity 5

The site specificity of recombination 5

Dimer resolution at dif sites 5

Integration in att sites 6

Integration in tRNA genes 6

Att site characteristics 7

Integration in other intragenic sequences and intergenic regions 8

Specificity switch 8

Integrase mobility between genomes and between mobile elements 9

Host specificity 9

Mobile element recruitment 9

Integrase primary function: MGE integration and excision 10

Integration is a major lifestyle for archaeal mobile elements 10

Evolutionary advantages to MGE integration: the fuselloviruses example 10

Suicidal integrase excision 12

Integration/excision temporality control in archaeal mobile elements 12

Integration/excision directionality control in archaeal mobile elements 13

Integrase related genome evolution 14

Mobile genetic elements modular evolution 14

Horizontal gene transfer 14

Chromosomal inversions 15

Future research directions 16

Tables and figures 16

References 19

Archaeal tyrosine recombinase diversity

We are currently running a bioinformatics analysis of the diversity of annotated archaeal tyrosine recombinases. We expect the following outcomes: sequence classification in various families, phylogenetic relationships between the families, identification of conserved residues.

Notably, we would like to determine whether the suicidal integrases from Sulfolobales (Goodman and Stedman, 2018; Pauly Matthew D. et al., 2019) and from Desulfurococcales (Mochizuki et al., 2011) form a coherent family (Table 1). We would also like to determine whether the suicidal integrases identified in the article 3 in Thermococcales, Archaeoglobales and Methanosarcinales correspond to a coherent family.

Site-specific recombination from a biochemical point of view

A common catalytic mechanism for different reaction directionalities

The working model of site-specific recombination catalyzed by tyrosine recombinase was proposed after the resolution of the co-crystal structure of the bacterial recombinase Cre with its Lox site (Guo et al., 1997). It can to all account be extended to archaeal tyrosine recombinases. The standard reaction requires a tetramer of recombinases and a pair of identical DNA sequences specific to the recombinases involved (Grindley et al., 2006). The identity constraint can be relaxed for one of the two sequences (Rajeev et al., 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. This stage is rate limiting for the integrase Int^{SSV1} (Serre et al., 2002). During the second stage, the recombinases catalyze a coordinated strand-exchange between the two DNA sequences through a covalent DNA-protein intermediate (Grindley et al., 2006). DNA cleavage and religation are mediated by a nucleophile tyrosine at specific phosphodiester bounds. Depending on the topology linkage of the two DNA sequences, the outcome of the recombination varies (Figure 1-A). In archaeal cells, site-specific recombination substrates are circular molecules. 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. Recombination between these two copies produces an excision and the two initial circular molecules are restored. Recombination between two sites carried in reverse orientation by a single circular molecule produces an inversion. 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 (Figure 1-B).

Site-specific recombination activity of unescorted archaeal integrases

Several archaeal tyrosine recombinases were proven to be active *in vitro* and *in vivo* through various activity assays (Table 1 to 3, Figure 3). The first archaeal integrase whose activity was tested is Int^{SSV1}. However, the activity observed *in vitro* by Muskhelishvili et al. (Muskhelishvili, 1993; Muskhelishvili et al., 1993) could not be reproduced except for the first step of the recombination reaction i.e. strand cleavage (Letzelter et al., 2004; Muskhelishvili et al., 1993; Serre et al., 2002; Zhan et al., 2015). Afterward, three integrases (Int^{pTN3}, Int^{SNJ2} and Int^{PYG1}) were shown to catalyze site-specific

recombination *in vivo* (Cossu et al., 2017; Li et al., 2016; Wang et al., 2018). It was not determined whether the integrase interacts with other cellular proteins. Additionally, the tyrosine recombinases PaXerA and TaXerA from *Pyrococcus abyssi* and *Thermoplasma acidophilus* respectively, usually resolve chromosome dimers in vivo but were shown to catalyze integration *in vitro* making them *bona fide* integrases (Cortez et al., 2010; Jo et al., 2017; Serre et al., 2013). The two Thermococcales integrases Int^{pTN3} and $Int^{pT26\text{-}2}$ from plasmids pTN3 and pT26-2 respectively, were also shown to catalyze site-specific recombination on circular substrates *in vitro* (Cossu et al., 2017) (This manuscript, articles 1 and 3). All tested arrangements of specific sites allowed recombination in the absence of any additional cofactor. This suggests that, contrary to most bacterial integrases (Landy, 2015), archaeal integrases do not require any recombination directionality factors (RDF) for efficient recombination *in vitro*. The integrases, Int^{SSV2}, PaXerA and TaXerA can also catalyze site-specific recombination on linear substrates *in vitro* (Cortez et al., 2010; Cossu et al., 2017; Jo et al., 2017; Serre et al., 2013; Zhan et al., 2015). Linear substrates are not their natural substrates but they are useful to characterize some aspects of the integrase activity as the strand cleavage site (Figure 3) (Serre et al., 2013). Overall, the activity of several integrases was characterized whose most remarkable aspect is the absence of necessary cofactor for catalysis.

Unessential cofactors were identified *in vivo* concurring in Int^{SNJ2} activity in *Natrinema sp.* J7-1 (Wang et al., 2018). The gene *orf1* coding for the integrase is transcribed in an operon with two other genes *orf2* and *orf3*. *orf2* and o*rf3* code for small proteins (111 aa and 140 aa respectively) containing a coiled-coil domain that could mediate protein-protein interactions or a MarR-like DNA binding domain, respectively. The presence of one or both proteins increased Int^{SNJ2} integration activity by 30% *in vivo* (Wang et al., 2018). For the inversion reaction, the recombination efficiency was increased 70 times in presence of one protein and 180 times in presence of the two proteins. They cooperatively activated Int^{SNJ2} recombination activity through an undetermined mechanism. Nevertheless, Int^{SNJ2} is active in their absence and the operons of many SNJ2-like integrases do not encode these cofactors (Wang et al., 2018).

Insights into primary and tertiary sequence

In the abovementioned tyrosine recombinase model, a nucleophilic tyrosine cleaves a phosphodiester bound and creates a covalent DNA-protein intermediate (Grindley et al., 2006). Int^{SSV1} and PaXerA were shown to form a covalent intermediate with the substrate DNA and the implication of the tyrosine Y314 was evidenced for Int^{SSV1} (Serre et al., 2002, 2013). This confirms that classical and suicide archaeal integrases are tyrosine recombinases and implement the mechanism described for previously characterized integrases. Several residues were identified in bacterial integrases in addition to the catalytic tyrosine that are conserved and necessary for catalysis (Esposito and Scocca, 1997; Grainge and Jayaram, 1999). Equivalent conserved residues were identified for archaeal integrases although differing from the bacterial consensus (She et al., 2004) (Figure 4) (*To update with the bioinformatics analysis*). Their mutation abolished Int^{SSV1} substrate cleavage activity confirming their importance for catalysis in archaea (Letzelter et al., 2004). All these catalytic residues are localized at the C-terminal end of the protein which is involved in DNA cleavage and ligation catalysis (Zhan et al., 2015) (Figure 4). The N-terminal extremity of Int^{SSV2} controls multimerization and the middle portion officiate in the specific DNA interaction (Zhan et al., 2015).

The structure resolution of Int^{SSV1}, PaXerA and TaXerA revealed that archaeal tyrosine recombinases present a catalytic fold similar to bacterial and eukaryotic integrases (Eilers et al., 2012; Jo et al., 2016; Serre et al., 2013). Moreover, archaeal PaXerA and TaXerA proteins display the canonical structure of tyrosine recombinases comprising two domains in a C-shape conformation around the DNA (Serre et al., 2013) (Figure 4). PaXerA and TaXerA active sites also assemble *in cis* meaning that the same integrase monomer supplies the entire active site, similarly to the majority of bacterial integrases (Jayaram et al., 2015; Jo et al., 2016; Serre et al., 2013). On the contrary, Int^{SSV1} and Int^{pTN3} catalyze DNA cleavage *in trans* where one integrase monomer activates the sessile phosphodiester bond while the adjacent monomer supplies the catalytic tyrosine (Cossu et al., 2017; Eilers et al., 2012; Letzelter et al., 2004). For PaXerA, the last helix was shown to be crucial for the assembly of the protein complex and the completion of the recombination reaction (Serre et al., 2013).

DNA relaxation activity of archaeal integrases

Eukaryotic and bacterial tyrosine recombinases are related to topoisomerases IB as highlighted by the conservation of a catalytic core (Cheng et al., 1998). This relationship explains why tyrosine recombinases can often catalyze DNA relaxation (Abremski et al., 1986; Landy, 2015). Int^{pT26-2}, Int^{SSV1} and TaXerA also presented non-specific DNA relaxation activity (Jo et al., 2017; Letzelter et al., 2004) (this manuscript, Article 3). This activity underlines the relationship between archaeal tyrosine recombinases and topoisomerases IB.

Postmortem suicide integrase activity

After integration, the suicide integrase gene is split in two with a *int(N)* gene encoding the N-terminal part of the integrase Int(N) and a *int(C)* gene encoding the C-terminal part of the integrase Int(C) including the catalytic residues (Figures 1 and 4). It is intriguing whether Int(C) can catalyze recombination and excise the integrated MGE alone or in association with Int(N). The Int(N) and Int(C) moieties of Int^{SSV2}do not interact in solution in absence of DNA suggesting that they do not cooperate to assemble an entire functional enzyme (Zhan et al., 2015). On its own, Int(C) cannot form multimers since the N-terminal part of the integrase is responsible for multimerization (Int^{SSV1} dimerization and Int^{SSV2} tetramerization) (Zhan et al., 2012, 2015). Int(C)^{SSV1} and Int(C)^{SSV2} were shown to be to be sufficient for *in vitro* activity (Zhan et al., 2012, 2015). On the contrary, in *Thermococcus kodakarensis*, Int(C)^{TKV4} is not sufficient for *in vivo* activity since the element TKV4 is not excised in its presence.(Cossu et al., 2017). From these two conflicting results, it remains unclear whether Int(C) can catalyze recombination *in vivo*.

The site specificity of recombination

Dimer resolution at dif sites

Site-specific recombinases, among which tyrosine recombinases, catalyze recombination between two identical and precisely defined sequences called sites. The sites of Xer recombinases are named dif and are present in a single copy on circular chromosomes of archaea and bacteria (Castillo et al., 2017; Cossu et al., 2015). If a chromosome dimer is formed, the dif sequence is then present in two copies and the chromosome dimer is resolved through site-specific recombination between the two dif sites (Figure 6.B). 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., 2015; Jo et al., 2017) (Figure 5). The extremities of the spacer correspond to the position of the tyrosine catalyzed cleavage (Castillo et al., 2017; Serre et al., 2013). The precise sequence is variable except for some conserved positions (Cortez et al., 2010). It seems that both the stem-loop structure and the sequence are important for the functionality of dif sites. Interestingly, PaXer can bind its dif substrate with a high affinity and stem-looped structured 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 et al., 2017). Similarly, TaXerA and PaXerA can bind dif inverted repeats (Jo et al., 2017; Serre et al., 2013). The activity of TaXerA was assayed on a series of dif sites variant 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). Some dif positions are involved in sequence recognition and others in strand-transfer reaction.

Integration in att sites

Integrases catalyze the recombination between an attachment site attP on the episomal MGE and attB on the chromosome (Landy, 2015) (Figure 6.A). As a result, two attachment sites attL and attR are present at the extremities of the integrated MGE. In the canonical model, the four attachment sites are strictly identical and correspond to the integrase specific site. For exemple, the Sulfolobales SSV2 virus was only found integrated in its cognate attB site and not in any other slightly divergent sites (Contursi et al., 2006). The recombination site specificity is more relaxed for some other archaeal integrases. For example, in absence of its cognate attB site, the virus SSV1 could integrate in a sequence differing in two nucleotides from att (Contursi et al., 2006; Schleper et al., 1992; She et al., 2001a).

Integration in tRNA genes

For bacterial integrases, many attB sites lie within tRNA genes (Williams, 2002). This trend is also true for archaeal integrases (Krupovic et al., 2019; She et al., 2002; Wang et al., 2018) (This manuscript, Articles 1 to 3). 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). We inventoried the tRNA genes used as integration site in Archaea (Figure 7) and only 7 out of the 44 tRNA genes were not used as attB site. Those tRNA genes are either deleterious when used as integration site or they are targeted by still undetected MGE. Overall, it seems that almost any tRNA gene can be use as integration site. Some tRNA genes as the $tRNA^{Glu}TTC$, $tRNA^{Arg}TCT$ or $tRNA^{Val}CAC$ are used more frequently than others (Figure 7). It would be interesting to analyze the cause of this frequent use.

Looking in more detail, different segments of the tRNA gene can be used for integration (Williams, 2002) (Figure 5.B). The attB site mostly correspond to the 3' part of the tRNA gene with various 5' limits (Serre et al., 2002) (This manuscript, Articles 2 and 3). The att site can overlap the anti-codon and T stem-loops (e.g. virus SSV2) (Contursi et al., 2006) or can be as short as the amino-acid attachment site (e.g. virus SNJ2) (Liu et al., 2015). The pTN3-like integrases are unique in archaea in that their attB site corresponds to the 5' half of the tRNA gene (Cossu et al., 2017). In all cases, whether the attB site

lies within to the 5' or 3' part, the DNA sequence corresponding to the tRNA is reconstituted after integration. With an att site in the 5' end of the tRNA gene, after integration, the tRNA gene is under the control of the integrase promotor (Cossu et al., 2017; Krupovič and Bamford, 2008). In a single occasion, the integration event changed the tRNA gene. The non-specific integration of the virus SSVK1 into a $tRNA^{Glu}$ gene sporadically generated a $tRNA^{Asp}$ gene (Wiedenheft et al., 2004).

Several advantages for targeting tRNA genes were proposed. Firstly, tRNA genes are very stable through time (Williams, 2002). It is possible that ancestral tyrosine recombinases targeted other sequences but that they disappeared when the target sequence changed. Also, tRNA genes are quite conserved between species and targeting them allow to integrate in multiple genomes. This was observed for the SSV2 virus which can integrate in both *S. islandicus* and *S. solfataricus* genomes (Contursi et al., 2006). Secondly, tRNA genes are a multigene family that offers a multitude of potential target sites with only limited nucleotide changes (Winckler et al., 2005). For example, two different $tRNA^{Glu}$ genes are targeted by the closely related integrases from virus SSV4 and plasmid pXZ1 and present only a single nucleotide mutation (Peng, 2008). Thirdly, tRNA sequences might be easier to recognize in genomes. It is possible that the some particular structure are present in tRNA genes that facilitate their recognition by the integrase (Williams, 2002). Alternatively, similarly to rDNA operons (Nalabothula et al., 2013), the highly expressed tRNA genes might also lack histone coverage and be more accessible to recombinases. Finally, it was suggested that tRNA genes are common integration site because they were the ancestral target of tyrosine recombinases (Campbell, 1992).

Att site characteristics

The attachment sites are canonically defined as the common sequence between the MGE and the host chromosome or as the direct repeat at the extremities of the integrated element (Figure 6). They usually extend aver 40 to 50 bp for suicide integrases (Cossu et al., 2015; She et al., 2002), 40 to 50 bp for Sulfolobales pNOB8 integrases (Erauso et al., 2006; She et al., 2002) and around 50 to 60 bp for some Methanococcales integrases (ref Badel). They can be as short as 8 bp in a Thaumarchaeota integrated element (Krupovic et al., 2019), 11 bp in a Methanobacteriales integrated virus (Krupovič et al., 2010) or 13 bp in a Halovivax integrated virus (Liu et al., 2015) or they can be longer than 100 nt for some Thermococcales elements (102 bp for PkuNCB100_IP1 or 243 bp for TIRI33c_IE1) (This manuscript, Articles 2 and 3). For these two Thermococcales MGE, the long att site encompassed the shorter att site of the closely related integrases. For some unknown reasons, the att site is enlarged in those elements.

The "repeated sequence" definition of the att site does not assure its functionality for recombination activity and several studies endeavored to determine the role of the different positions for efficient recombination in archaea. Those studies postulated that a minimal necessary and sufficient recombination site exists. For Int^{SSV1} and Int^{SSV2}, two sequences were suggested to be sufficient for recombination *in vitro* (Figure 5.A): the *stricto sensu* att site or the inverted-repeats separated by an overlap region (core-type sites defined for the lambda recombination) (Muskhelishvili et al., 1993; Serre et al., 2002; Zhan et al., 2015). The two sequences do not completely overlap and the sequence at the intersection was not tested alone for recombination. The necessary positions are therefore not known. For Int^{SSV1}, the cleavage and religation take place at the 5' boundaries of the overlap region (anti-codon loop) as for classical bacterial tyrosine recombinases (Grindley et al., 2006;

Serre et al., 2002). The observation of an *in vivo* aspecific integration suggested that the strand cleavage position might be variable (Wiedenheft et al., 2004).

For Int^{pTN3} and $Int^{pT26\text{-}2}$, the identical DNA stretch shared by the attL and attR sites is not sufficient for recombination *in vitro* (Cossu et al., 2017) (This manuscript, Articles 1 and 3). Additional nucleotides are necessary in the incomplete anti-codon arm (Figure 5). This suggest that the cleavage site could be at the extremities of the anti-codon loop as reported for Int^{SSV1}. But it could also be located at the extremities of the D or T loop for Int^{pTN3} or $Int^{pT26\text{-}2}$ respectively. The cleavage site at the extremities of a tRNA loop can also be considered for Int^{PYG1} but not for Int^{SNJ2} whose att site is very short (Figure 5.B). It would be interesting to determine whether the 14nt long and stem-loop free att site from Int^{SNJ2} is sufficient for recombination. Finally, For $Int^{pT26\text{-}2}$ 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. $Int^{pT26\text{-}2}$ effective recombination site is not centered in the att site but located towards the 5'end, similarly to Int^{SSV1} (Serre et al., 2002), Int^{SSV2} (Zhan et al., 2015) and numerous bacterial integrases (Campbell, 1992).

Integration in other intragenic sequences and intergenic regions

Several archaeal attB sites were reported within intergenic sequences or in protein coding genes (Krupovic et al., 2019; Krupovič et al., 2010; Luo et al., 2001). For example, in *Methanothermobacter wolfeii,* the att sites of the prophage ΨM100 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 (Brüssow and Desiere, 2001) and presents an AT-rich att site in an intergenic region similarly to phage λ (Campbell, 1992). Att sites are also found in coding regions (Krupovič et al., 2010). When the function of the targeted gene is known, it can be as diverse as a gene coding fir 3-hydroxy-3-methylglutaryl-coenzyme A reductase for the virus BJ1 of *Halorubrum sp.* (Krupovič et al., 2010), or heavy metal cation efflux system for the provirus Msmi-Pro1 of *Methanobrevibacter smithii* (Krupovič et al., 2010) or 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 for recombination. Especially, the presence of stem-loop structures is crucial for the activity of tyrosine recombinases analyzed to date and we wonder whether those regions present such a structure.

Specificity switch

Closely related integrases do not always target the same att site. For example, the classical integrases identified in pT26-2 related plasmids from Methanococcales can target $tRNA^{Ser}TGA$, $tRNA^{Ser}GCT$ or $tRNA^{Leu}TAA$ (This manuscript, Article 2). The suicide integrases identified in pT26-2 related plasmids from Thermococcales can target 14 different tRNA genes (This manuscript, Article 3). In both cases, the most probable evolutionary scenario includes an ancestral integrase with one specificity followed by specificity diversification in the descendant lineages. One limitation was identified in specificity switch between tRNA genes: closely related integrases target only a subpopulation of tRNA genes harboring or missing a supplementary loop (This manuscript, Article 2).

Little is known about specificity innovation. Does it happen gradually with an intermediate state of double specificity? Is it a sudden change triggered by an aspecific recombination? Two partners control integrase specificity: the integrase and the att site. Suicidal integrases harbor the translation of the att site in their protein sequence (Figure 4.A) 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 it would compromise protein integrity, but it was on the contrary shown that the att site translation is quite variable in closely related sequences without obvious deleterious effect (This manuscript, Article 3). Notably, length variations are compensated around the att site avoiding any frameshifts in the C-terminal region.

A mechanism was proposed to explain specificity switches in suicidal integrases. Two integrases integrated in the same chromosome mix their int(N) and int(C) fragments through homologous recombination creating two chimeric integrases. These new integrases can potentially efficiently recombine one of the two original att sites.

Integrase mobility between genomes and between mobile elements

Host specificity

Tyrosine integrases were detected in many archaeal phyla. However, closely related integrases are mostly always found in a narrower range of organisms (Table 1). For example, two distinct integrase families are present in pT26-2 plasmids from Methanococcales and Thermococcales (This manuscript, Article 2). Related integrases are not restricted to a local area but are found all around the globe where their host is present. For example, similar integrases from pleolipovirus were identified in chromosomes from 10 different genera from Halobacteiaceae from all around the globe (Liu et al., 2015).

Mobile element recruitment

In the archaeal domain, integrases are carried by a wide variety of MGEs: conjugative plasmids (She et al., 2002) or cryptic plasmids (Erdmann et al., 2017) (This manuscript, Article 2), viruses from several viral families (e.g. Myoviridae (Klein et al., 2002; Tang et al., 2002), Pleolipoviridae (Atanasova et al., 2018a), Fuselloviridae (Goodman and Stedman, 2018)) and unidentified MGE (Li et al., 2016). Integrases from the same family can be recruited by several mobile elements. For example, 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 MGE (This manuscript, Article 3). In *Sulfolobus solfataricus*, a plasmid and a virus were isolated that both encoded a very similar integrase (86% nucleotide similarity, 94% amino acid similarity) (Peng, 2008). Additionally, some MGE families include integrative and non-integrative members as the Thermococcales pAMT11 family (This manuscript, Article 3) or the haloarchaeal pleolipoviruses (Liu et al., 2015; Roine et al., 2010). Strikingly, the sequences of the two pleolipoviruses HHPV3 and HHPV4 are very syntenic and similar except for a HHPV4 specific 3kb region including an integrase gene (Atanasova et al., 2018b). For suicidal integrases, the att site is included in the integrase gene resulting in a compaction that could favor exchange between mobile elements (This Manuscript, Article 3). Finally, even when the integration module is conserved in a plasmid family, its evolutionary history can be complex. For example, all the conjugative Sulfolobales plasmids exhibit conserved conjugation and integration modules but 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 MGE. Not all archaeal integrases are frequently exchanged between various mobile element types.

Integrases can be recruited by MGE, and conversely MGE can also lose integrase genes. One plasmid pING1 was identified for which no attP site could be determined (Erauso et al., 2006). The plasmid encodes an integrase exhibiting all the conserved residues of its family (pNOB8-like integrases). It is possible that in that case, the attP site loss would lead to the integrase gene degeneration and/or loss.

Integrase primary function: MGE integration and excision

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 (Figure 6.A). Such integrase-encoding MGEs are present in archaeal genomes (Gaudin et al., 2014; Krupovic et al., 2019; Soler et al., 2010). In Thermococcales, it was shown that more than 30% of the published genomes contain an integrated element encoding an integrase of the pT26-2 family (This manuscript, Article 3). The proportion of genomes presenting any integrase-encoding MGE is most probably higher than that. In the phylum Thaumarchaea, integrated MGEs were systematically detected (Krupovic et al., 2019) and 20 over 21 analyzed genomes presented some. In this systematic search, no integrase could be detected in some integrated MGE. This presumably results from the integrase gene loss after integration similarly to what was observed for integrated plasmids of Methanococcales (This manuscript, Article 2). Several related or unrelated MGEs can be integrated in the same chromosome at different loci (Pauly Matthew D. et al., 2019) (This manuscript, Article 3) or integrated in tandem at the same locus (Krupovic et al., 2019; Krupovič et al., 2010). Overall, integrated MGEs are widely present in archaeal genomes.

Evolutionary advantages to MGE integration: the fuselloviruses example

The canonical integration model developed from the phage lambda analysis propose that the integrative state helps the MGE to survive through certain environmental conditions. When stressful conditions are encountered by the cell, the MGE excises, enters the lytic cycle and is released into the environment (Gandon, 2016; Paul, 2008). Depending on the environmental conditions, the MGE chooses to reproduce vertically (integration) or horizontally (infection). The same lifestyle was observed for *Acidianus convivator* bicaudavirus ATV (Prangishvili et al., 2006). It adopts a lysogenic lifestyle and integrates into the host chromosome under optimal growth temperature conditions. Inversely, the virus adopts a lytic lifestyle resulting in host cell lysis under suboptimal growth temperature conditions. For the archaeal fuselloviruses, which are the most studied archaeal MGE with an integrase and exemplified by the model virus SSV1, the integration implications are different on several aspects (Prangishvili et al., 2001). (1) 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 the lambda virus, but cell are not massively lysed after viral production and return to the lysogenic sate (Martin et al., 1984). The virus TPV1 replication is also induced by a UV-treatment without any massive cellular lysis (Gorlas et al., 2012). (2) During SSV1 integrative stage, a few circular copies of the virus genome remain in the cell (Pauly Matthew D. et al., 2019; Yeats et al., 1982). Similarly, a high copy number of TPV1 circular DNA is present in its host cells (Gorlas et al., 2012) (3) During SSV1 integrative stage, the majority of the viral ORFs are expressed, including the integrase gene and the structural proteins (Fröls et al., 2007). It is not known whether the transcription template correspond 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., 2015a) but it does not result in provirus silencing like for the phage lambda. Contrastingly, SSV2 integrase is not basally expressed (Fusco et al., 2015a). (4) For the virus SSV1, evidence point towards 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 lambda (Fusco et al., 2015b). Overall, it seems that the integrase of lysogenic fuselloviruses is not involved in the regulation of virus replication and virion production.

Nevertheless, almost all fuselloviruses encode a suicidel tyrosine integrase (Goodman and Stedman, 2018; Gorlas et al., 2012) suggesting a probable evolutionary importance for the virus survival. SSV1 viruses lacking the integrases genes were indeed proved to be outcompeted by wild-type viruses (Clore and Stedman, 2007). However the mutant viruses were infectious and stably maintained in *Sulfolobus* and no clear benefit was identified for the integrase activity. Several advantages were proposed for integration in general. (1) Increasing long-term maintenance of the MGE (Echols, 1972). (2) Staying in the cell as a solution when chances of finding a new suitable host are low (Levin et al., 1977) (3) The integrated MGE can provide functions that are beneficiary for the host and therefore increase MGE retention probability. For example, the integrated element PYG1 was shown to increase its host resistance to temperature (Li et al., 2016). (4) Integrating the host chromosome might force the cell into accepting the MGE and shutting down its defense systems (as the CRISPR-Cas system), especially with a suicide integrase as observed in a SSV2 infection (Fusco et al., 2015a). After the infection, SSV2 episomal viral DNA was cured in all likelihood by the CRISPR-Cas system. As a result, the intact episomal-encoded integrase was lost, excision was abolished and the provirus was effectively trapped in the chromosome. A deletion in the CRIPSR spacer responsible for SSV2 recognition was subsequently observed, suppressing the now self-targeting spacers and allowing the host cell to survive in presence of the trapped provirus (Fusco et al., 2015a).

Additionally, several plasmids were identified that are present both in the integrated and episomal states in cells (Basta et al., 2009; Cossu et al., 2017; Gaudin et al., 2014). Similarly to SSV2 infection, the pTN3 plasmid was initially present both in integrated and episomal states in a population of *Thermococcus nautili.* The circular state was almost lost after several subcultures (Cossu et al., 2017) probably as a result of plasmid targeting by spacers from the host CRISPR-Cas9 defense system (Oberto et al., 2014). The integrated form can then act as a safekeeping copy of the disappearing plasmid. The pAF1 plasmid was evidenced to be stably maintained simultaneously as integrated and episomal state (Basta et al., 2009). When the host was co-infected with the virus AFV1, the plasmid episomal form disappeared rapidly but the integrated form remained. The integrated form here again can act as a safekeeping copy. For these two plasmids, the integration did not correspond to a chosen and controlled lifestyle of hideout like for the phage lambda but to a safekeeping strategy in case normal plasmid replication was impaired by a host defense system or another MGE.

Suicidal integrase excision

For suicidal integrases, integrated MGE might be "kamikazes" that ensure the inactivation of the host CRISPR-system. Nevertheless, it is difficult to consider integrase evolution in the absence of excision. Excision was shown to happen in the presence of the episomal MGE coding for the complete integrase (Fusco et al., 2015a). Can excision also happen in the absence of any episomal form when only fragments of integrase are present?

Firstly, the partitioned integrase could catalyze excision. Evidences are so far contradictory for the activity of the fragmented integrase. Int(C)^{SSV2} can catalyze excision but Int(C)^{pTN3} cannot. To effectively catalyze excision *in vivo*, partitioned integrase should be produced. The *int(N)* fragment is under the control of the integrase promotor and the protein can be translated from the start codon. The fragment *int(C)* is in some cases included in a transcription unit as for int(C)^{pTN3} that is under the control of the promotor of a tRNA gene (Cossu et al., 2017). An in frame start codon is often present near the beginning of *int(C)* suggesting that the catalytic part of the integrase can be produced (personal observation). However, in *Sulfolobus solfataricus* P2, only the int(N) moiety is transcribed for pXQ1 and XQ2 integrated elements (Jäger et al., 2014; She et al., 2001b). No expression was detected for Int(N) moiety or the complete integrase gene from integrated plasmid pSSVi in *Sulfolobus solfataricus*. In *Thermococcus kodakarensis*, the complementation of a tyrosine integrase variant by the endogenous Int(C) suggests that Int(C) is expressed (This manuscript, Article 1 (Cossu et al., 2017). Observations concerning the expression of the Int(N) and Int(C) moiety are so far contradictory and it is unclear whether the fragmented integrase could catalyze excision.

Secondly, an integrase from exogenous circular element could catalyze the integrase excision. This was observed several times. In the strain *Sulfolobus solfataricus* P2(pSSVi), the integrated plasmid pSSVi was excised after the introduction of the virus SSV2 (Ren et al., 2013). The chronology of event was dissected in details and pointed towards the catalysis of pSSVi excision by the SSV2 integrase. In *Thermococcus kodakarensis*, trans-complementation was evidenced between an exogenous catalytic integrase mutant and the endogenous fragmented integrase that lead to the integrated element excision (Cossu et al., 2017). This suggest that the Int(C) moiety might play a role in the excision of a mobile element by an exogenous integrase. In both cases, fragmented integrase and exogenous integrases were closely related. This excision catalysis by another MGE depend on the widespread of closely related integrases in the population and in various MGE which is was observed for the pT26-2 family of integrases (This manuscript, Article 3).

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: (1) reaction temporality control and (2) reaction directionality control (integration or excision). It is interesting 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 present 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.

Some archaeal halophilic tailed viruses belong to the Caudovirales, which also include tailed bacteriophages (Krupovic et al., 2018; Senčilo et al., 2013). 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). The 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 lambda cI regulator protein (Iro et al., 2007). The integrase gene involvement was not determined in the regulatory network. 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 (Atanasova et al., 2018a; Chen et al., 2014; Liu et al., 2015) suggesting that this mechanisms of lysis-lysogeny regulation is widely shared among halophilic viruses.

The first level of integration/excision temporal regulation is the regulation of the 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-shapes viruses, transcription temporality was investigated by several studies. In SSV1, the integrase is under the control of an early promotor that allow a rapid expression after UV-induction (Fröls et al., 2007) and the F55 repressor downregulates expression of the integrase operon in 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 more than 7 hours after infection (Ren et al., 2013). The mechanisms of integrase expression regulation might differ between the various SSV viruses. 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).

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 lambda directionality regulation (Landy, 2015). However, the integrase Int^{SNJ2} activity 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. However, in naturally occurring conditions, suicidal integrases are partitioned after integration. Excision would then require the activity of the split integrase which 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 related genome evolution

Mobile genetic elements modular evolution

MGE evolution proceed mainly through module exchange (Basta et al., 2009; Hendrix et al., 2000; Iranzo et al., 2016; Krupovič et al., 2010). For example, the integrated element PYG1 from *Pyrococcus yayanosii* presents a module similar to an integrated element of *Thermococcus barophilus* MP and another module similar to the plasmid pTBMP1 (Li et al., 2016). A mechanism of module exchange was proposed that involves tandem MGE integration (Redder et al., 2009). In that configuration, several MGE are integrated in the same chromosomal locus resulting in an array of consecutive MGEs. Recombination between the integrated MGEs can then lead to the creation of a new module arrangements. Alternatively, illegitimate excision can lead to the MGE excision embarking a portion of the neighbor MGE.

Several halophilic viruses were identified that encode tyrosine recombinases that seem implicated in viral DNA rearrangements (Rössler et al., 2004; Senčilo et al., 2013). One of the DNA rearrangements is involved in the generation of protein variants presenting various cell surface adhesion specificities (Klein et al., 2012)

Horizontal gene transfer

Horizontal gene transfer (HGT) is the transmission of a DNA sequence from an individual to another independently of reproduction. HGT is recognized as a driving force of archaeal evolution (Wagner et al., 2017). Several successive steps are required for an effective HGT: (1) DNA is transferred into the cell through transformation, membrane vesicle, viral infection, conjugation, cell fusion or other specialized cellular apparatus (Wagner et al., 2017). (2) The DNA is incorporated into the host chromosome through homologous recombination if the DNA is similar to the host genome or through site-specific integration if the DNA contains an integration module. (3) In case of MGE-catalyzed integration, the DNA should be immobilized in the host chromosome to be effectively characterized as HGT. This can happen through the mutation or loss of the integrase gene. For example, a non-sense mutation is present inside the integrase coding-sequence of the *Sulfolobus* integrated element pST4 (She et al., 2004). Similarly, mutations into the attL and attR sites would result in the MGE fixation into the host chromosome. Integrated plasmids that lack detectable att sites were recently identified that might correspond to fixed elements (This manuscript, Article 2). Additionally, She et al. suggested that integration in secondary attachment sites, i.e. att sites with a mismatch, could prevent efficient MGE excision and lead to the fixation of MGE genes in the host chromosome (She et al., 2004)

Potentially all functions can be the subject of MGE-mediated HGT. The most studied was replication. The archaea *Sulfolobus islandicus* and *Haloferax volcanii* possess several active chromosomal origins of replication, some of which were acquired from integrated MGE (Hawkins et al., 2013; Robinson and Bell, 2007; 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). MGE were also proposed to be implicated in the HGT of introns in tRNA genes (Sugahara et al., 2012). Sugahara et al. argued that recombination between an intron-containing attP and an intron free attB in a tRNA gene could be a mechanism of intron acquisition in tRNA. The MGE-carried attP site serves as a tRNA intron vector between tRNAs and between cells.

Chromosomal inversions

Among MGE, Transposable Elements (TE) are known to be frequently involved in generating inversions in the host chromosome (Darmon and Leach, 2014; Eickbush and Furano, 2002; Redder and Garrett, 2006; Vandecraen et al., 2017; Weckselblatt and Rudd, 2015; Zivanovic et al., 2002). Inversions proceed through homologous recombination between two paralogous integrated elements. Other integrated MGE are also involved in similar processes. For example, in *Thermococcus kodakarensis*, a large-scale inversion was identified that occurred between the integrated elements TKV2 and TKV3 (This manuscript, Article 2). Tyrosine recombinases are fundamental in that process because they catalyze the integrations that bring MGE paralogous copies into the chromosome.

Another unique mechanism was identified in archaea that lead to chromosomal inversions. The tyrosine-integrase from Thermococcus plasmid pTN3 "catalyzes low sequence specificity recombination reactions with the same outcome as homologous recombination events on DNA segments as short as 104bp" (Cossu et al., 2017). This homologous recombination activity resulted in four large-scale chromosomal inversions over the span of 60 generations in *Thermococcus nautili*. (Cossu et al., 2017). It is also probably the cause of many of the large chromosomal inversions observed in Thermococcales (Zivanovic et al., 2002).

Archaea tyrosine recombinases are thus involved in chromosomal inversion either indirectly through the integration of multiple, recombinable, MGE copies or directly through a homologous recombination activity. 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.

Future research directions

• The detailed characterization of minimal recombination sites and cleavage sites of more archaeal integrase would enlighten the role of the secondary structure for catalysis.

• The elucidation of the ambiguous expression and activity of suicide integrase fragments would help interpreting suicidal activity implications.

• Archaeal virus lysogeny was never investigated in details despite the obvious differences with the canonical lysogeny from phage lambda. Such investigation would allow to better understand the variability of lysogeny in natural communities. The exploration of archaeal MGE integration and excision dynamics would similarly be interesting.

• No global evolutionary analysis of all tyrosine recombinases from archaeal, bacterial and eukaryotes was ever published. It would though determine integrase origin and whether they are exchanged between the three domains. It would also dtermine whether integration is preferentially implemented in certain environmental or genetic conditions.

Tables and figures

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)
SSV1 integrase	Suicide	Sulfolobales (Desulfurococcales?)	yes	yes	yes	(Serre et al., 2002; Zhan et al., 2015)
pNOB8 integrase	Classical	Sulfolobales	no	no	no	(She et al., 2002)
PYG1 integrase	Classical	Thermococcales	yes	no	no	(Li et al., 2016)
pTN3 integrase	Suicide	Thermococcales	yes	yes	no	This manuscript, Article 1 i.e. (Cossu et al., 2017)
pT26-2 integrase	Suicide	Thermococcales, Archaeoglobales (Methanosarcinales?)	yes	yes	no	This manuscript, Article 3
SNJ2 integrase	Classical	Halobacteriales	yes	no	no	(Wang et al., 2018)

Table 2. Archaeal integrases characterized in vitro

Integrase name	Integrase type	Encoding genetic element	Host	Site-specific *in vitro* activity	*in vitro* assay conditions					Reference
					Temperature	Enzyme/ substrate molar ratio	Salts	pH	Additional reagent	
Int^{SSV1}	Suicide	virus	*Saccharolobus shibatae*	strand cleavage, DNA relaxation	65°C	40/1	125 mM NaCl	7.5	50 µg/mL BSA	(Letzelter et al., 2004; Serre et al., 2002)
Int^{SSV2}	Suicide	virus	*Sulfolobus islandicus*	recombination on linear substrates	65°C	1000/1	150 mM NaCl	6.7	50 µg/mL BSA	(Zhan et al., 2015)
Int^{pTN3}	Suicide	plasmid	*Thermococcus nautili*	integration, excision, inversion	65°C	10/1	300 mM KCl 1mM MgCl2	8	/	This manuscript, Article 1
Int^{pT26-2}	Suicide	plasmid	*Thermococcus sp. 26-2*	integration, excision, inversion, recombination on linear substrates	75°C	17/1	300 mM KCl	8	/	This manuscript, Article 3
PaXerA	Classical	chromosome	*Pyrococcus abyssi*	recombination on linear substrates, excision, intégration	65°C	133/1	50 mM NaCl	7,5	50 µg/mL BSA	(Cortez et al., 2010)
TaXerA	Classical	chromosome	*Thermoplasma acidophilum*	recombination on linear substrates, excision	63°C	NI	50 mM NaCl	7.5	1 mM EDTA	(Jo et al., 2017)

Table 3. Archaeal integrases characterized in vivo

Integrase name	Integrase type	Encoding genetic element	Host	Site-specific *in vivo* activity
Int^{pTN3}	Suicide	plasmid	*Thermococcus nautili*	excision
Int^{PYG1}	Classical	genomic island	*Pyrococcus yayanosii*	excision, integration
Int^{SNJ2}	Classical	virus	*Natrinema sp.* J7-1	inversion, integration

Figure 1. The two classes of archaeal integrases as defined by She et al. (She et al., 2002). The black and grey squares correspond to the specific recombination sites.

Figure 2. The different outcomes of site-specific recombination. The black rectangles correspond to the specific recombination sites. A. With circular DNA molecules as substrates, the recombination outcome can be integration, excision or inversion, depending on the relative position of the two specific sites. B. Recombination between two linear DNA molecules results in two chimeric linear DNA molecules.

Figure 3. Various activity assay were implemented to detect archaeal integrases activities *in vitro* and *in vivo*. A. Different substrates harboring the specific site (black box) were incubated with the integrases *in vitro* and the products were monitored (this manuscript, Articles 1 and 3) (Cortez et al., 2010; Jo et al., 2017; Zhan et al., 2015). **B.** A half-site strand transfer assay was first implemented *in vitro* by Serre et al (Serre et al., 2013). It allows the verification of the strand cleavage site (in blue). 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 entire specific site reconstruction only if the 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 the 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 PCR with different pairs of 4 primers (black and grey arrows). (Cossu et al., 2017; Li et al., 2016; Wang et al., 2018)

Figure 4. Archaeal integrases sequence domains and conserved residues. **A.** Suicidal integrase can present three configurations: the Int(N) protein contains the N-terminal portion of the intact Int protein while the Int(C) protein contains the C-terminal portion. **B.** The conserved catalytic residues are indicated for all characterized archaeal tyrosine integrases from Table 1. Some domains of particular interest are indicated. Functional domains were dissected for Int^{SSV2} (Zhan et al., 2015). Int^{pTN3} present 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 (Jo et al., 2016; Serre et al., 2013).

Figure 5. 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 in lambda (Landy, 2015). The blue box correspond to the att site. The black sequences are necessary and sufficient for recombination *in vitro* (This manuscript, Articles 1 and 3). The black arrows indicate the cleavage site when experimentally determined (Jo et al., 2017; Serre et al., 2002, 2013) **B.** Att sites often correspond to tRNA sequences. The leaf-like structure of the targeted tRNA are indicated with att site nucleotides circled in blue.

Figure 6. The roles of tyrosine recombinases in archaea. A. Integrases catalyze site-specific recombination between att sequences resulting in MGE integration or excision from the host chromosome. B. XerA recombinases catalyze site-specific recombination between dif sites resulting in chromosome dimer resolution.

Figure 7. Almost all tRNA genes are targeted by archaeal tyrosine recombinases. All tRNA anti-codon combinations are listed along with their corresponding amino-acids. Anti-codons that are not found in archaeal tRNA are indicated in light grey. An array of symbols indicates the utilization of a tRNA gene with the specified anti-codon as attB site for Thermococcales (Cossu et al., 2017; Li et al., 2016) (This manuscript, Articles 1 to 3), Halobacteriales (Krupovič et al., 2010; Liu et al., 2015), Archaeoglobales (This manuscript, Article 3), Methanococcales (Krupovič et al., 2010) (This manuscript, Article 2), Methanosarcinales (This manuscript, Article 3), Sulfolobales (Peng, 2008; Redder et al., 2009; She et al., 2001b, 2002; Wang et al., 2007) and Thaumarchaeota (Krupovic et al., 2019). Anti-codons corresponding to untargeted tRNA genes are highlighted in bold green.

References

Abremski, K., Wierzbicki, A., Frommer, B., and Hoess, R.H. (1986). Bacteriophage P1 Cre-loxP site-specific recombination. Site-specific DNA topoisomerase activity of the Cre recombination protein. J. Biol. Chem. *261*, 391–396.

Atanasova, N.S., Demina, T.A., Krishnam Rajan Shanthi, S.N.V., Oksanen, H.M., and Bamford, D.H. (2018a). Extremely halophilic pleomorphic archaeal virus HRPV9 extends the diversity of pleolipoviruses with integrases. Research in Microbiology.

Atanasova, N.S., Heiniö, C.H., Demina, T.A., Bamford, D.H., and Oksanen, H.M. (2018b). The Unexplored Diversity of Pleolipoviruses: The Surprising Case of Two Viruses with Identical Major Structural Modules. Genes *9*, 131.

Azaro, M.A., and Landy, A. (2002). λ Integrase and the λ Int Family. Mobile DNA II 118–148.

Basta, T., Smyth, J., Forterre, P., Prangishvili, D., and Peng, X. (2009). Novel archaeal plasmid pAH1 and its interactions with the lipothrixvirus AFV1. Molecular Microbiology *71*, 23–34.

Brüssow, H., and Desiere, F. (2001). Comparative phage genomics and the evolution of Siphoviridae: insights from dairy phages. Molecular Microbiology *39*, 213–223.

Campbell, A.M. (1992). Chromosomal insertion sites for phages and plasmids. J. Bacteriol. *174*, 7495–7499.

Castillo, F., Benmohamed, A., and Szatmari, G. (2017). Xer Site Specific Recombination: Double and Single Recombinase Systems. Front. Microbiol. *8*.

Chen, S., Wang, C., Xu, J.-P., and Yang, Z.L. (2014). Molecular characterization of pHRDV1, a new virus-like mobile genetic element closely related to pleomorphic viruses in haloarchaea. Extremophiles *18*, 195–206.

Cheng, C., Kussie, P., Pavletich, N., and Shuman, S. (1998). Conservation of Structure and Mechanism between Eukaryotic Topoisomerase I and Site-Specific Recombinases. Cell *92*, 841–850.

Clore, A.J., and Stedman, K.M. (2007). The SSV1 viral integrase is not essential. Virology *361*, 103–111.

Contursi, P., Jensen, S., Aucelli, T., Rossi, M., Bartolucci, S., and She, Q. (2006). Characterization of the Sulfolobus host–SSV2 virus interaction. Extremophiles *10*, 615–627.

Cortez, D., Quevillon-Cheruel, S., Gribaldo, S., Desnoues, N., Sezonov, G., Forterre, P., and Serre, M.-C. (2010). Evidence for a Xer/ dif System for Chromosome Resolution in Archaea. PLOS Genet *6*, e1001166.

Cossu, M., Da Cunha, V., Toffano-Nioche, C., Forterre, P., and Oberto, J. (2015). Comparative genomics reveals conserved positioning of essential genomic clusters in highly rearranged Thermococcales chromosomes. Biochimie *118*, 313–321.

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

Darmon, E., and Leach, D.R.F. (2014). Bacterial Genome Instability. Microbiol. Mol. Biol. Rev. *78*, 1–39.

Duyne, G.D.V. (2015). Cre Recombinase. Microbiology Spectrum *3*.

Echols, H. (1972). Developmental Pathways for the Temperate Phage: Lysis Vs Lysogeny. Annual Review of Genetics *6*, 157–190.

Eickbush, T.H., and Furano, A.V. (2002). Fruit flies and humans respond differently to retrotransposons. Current Opinion in Genetics & Development *12*, 669–674.

Eilers, B.J., Young, M.J., and Lawrence, C.M. (2012). The Structure of an Archaeal Viral Integrase Reveals an Evolutionarily Conserved Catalytic Core yet Supports a Mechanism of DNA Cleavage in trans. J. Virol. *86*, 8309–8313.

Erauso, G., Stedman, K.M., van de Werken, H.J.G., Zillig, W., and van der Oost, J. (2006). Two novel conjugative plasmids from a single strain of Sulfolobus. Microbiology *152*, 1951–1968.

Erdmann, S., Tschitschko, B., Zhong, L., Raftery, M.J., and Cavicchioli, R. (2017). A plasmid from an Antarctic haloarchaeon uses specialized membrane vesicles to disseminate and infect plasmid-free cells. Nature Microbiology *2*, 1446.

Escudero, J.A., Loot, C., Nivina, A., and Mazel, D. (2015). The Integron: Adaptation On Demand. 139–161.

Esposito, D., and Scocca, J.J. (1997). The integrase family of tyrosine recombinases: evolution of a conserved active site domain. Nucleic Acids Res *25*, 3605–3614.

Filée, J., Siguier, P., and Chandler, M. (2007). Insertion Sequence Diversity in Archaea. Microbiol. Mol. Biol. Rev. *71*, 121–157.

Fröls, S., Gordon, P.M.K., Panlilio, M.A., Schleper, C., and Sensen, C.W. (2007). Elucidating the transcription cycle of the UV-inducible hyperthermophilic archaeal virus SSV1 by DNA microarrays. Virology *365*, 48–59.

Fusco, S., She, Q., Bartolucci, S., and Contursi, P. (2013). Tlys, 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. Journal of Virology *87*, 5926–5936.

Fusco, S., Liguori, R., Limauro, D., Bartolucci, S., She, Q., and Contursi, P. (2015a). Transcriptome analysis of Sulfolobus solfataricus infected with two related fuselloviruses reveals novel insights into the regulation of CRISPR-Cas system. Biochimie *118*, 322–332.

Fusco, S., She, Q., Fiorentino, G., Bartolucci, S., and Contursi, P. (2015b). Unravelling the Role of the F55 Regulator in the Transition from Lysogeny to UV Induction of Sulfolobus Spindle-Shaped Virus 1. Journal of Virology *89*, 6453–6461.

Gandon, S. (2016). Why Be Temperate: Lessons from Bacteriophage λ. Trends in Microbiology *24*, 356–365.

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

Goodman, D.A., and Stedman, K.M. (2018). Comparative genetic and genomic analysis of the novel fusellovirus Sulfolobus spindle-shaped virus 10. Virus Evol *4*.

Gorlas, A., Koonin, E.V., Bienvenu, N., Prieur, D., and Geslin, C. (2012). TPV1, the first virus isolated from the hyperthermophilic genus Thermococcus. Environmental Microbiology *14*, 503–516.

Grainge, I., and Jayaram, M. (1999). The integrase family of recombinases: organization and function of the active site. Molecular Microbiology *33*, 449–456.

Grindley, N.D.F., Whiteson, K.L., and Rice, P.A. (2006). Mechanisms of Site-Specific Recombination. Annual Review of Biochemistry *75*, 567–605.

Guo, F., Gopaul, D.N., and Duyne, G.D.V. (1997). Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. Nature *389*, 40.

Hawkins, M., Malla, S., Blythe, M.J., Nieduszynski, C.A., and Allers, T. (2013). Accelerated growth in the absence of DNA replication origins. Nature *503*, 544–547.

Hendrix, R.W., Lawrence, J.G., Hatfull, G.F., and Casjens, S. (2000). The origins and ongoing evolution of viruses. Trends in Microbiology *8*, 504–508.

Iranzo, J., Koonin, E.V., Prangishvili, D., and Krupovic, M. (2016). Bipartite Network Analysis of the Archaeal Virosphere: Evolutionary Connections between Viruses and Capsidless Mobile Elements. J. Virol. *90*, 11043–11055.

Iro, M., Klein, R., Gálos, B., Baranyi, U., Rössler, N., and Witte, A. (2007). The lysogenic region of virus ϕCh1: identification of a repressor-operator system and determination of its activity in halophilic Archaea. Extremophiles *11*, 383–396.

Jäger, D., Förstner, K.U., Sharma, C.M., Santangelo, T.J., and Reeve, J.N. (2014). Primary transcriptome map of the hyperthermophilic archaeon Thermococcus kodakarensis. BMC Genomics *15*, 684.

Jayaram, M., Ma, C.-H., Kachroo, A.H., Rowley, P.A., Guga, P., Fan, H.-F., and Voziyanov, Y. (2015). An Overview of Tyrosine Site-specific Recombination: From an Flp Perspective. Microbiol Spectr *3*.

Jo, C.H., Kim, J., Han, A., Park, S.Y., Hwang, K.Y., and Nam, K.H. (2016). Crystal structure of Thermoplasma acidophilum XerA recombinase shows large C-shape clamp conformation and cis-cleavage mode for nucleophilic tyrosine. FEBS Letters *590*, 848–856.

Jo, M., Murayama, Y., Tsutsui, Y., and Iwasaki, H. (2017). In vitro site-specific recombination mediated by the tyrosine recombinase XerA of Thermoplasma acidophilum. Genes Cells.

Ken, R., and Hackett, N.R. (1991). Halobacterium halobium strains lysogenic for phage phi H contain a protein resembling coliphage repressors. Journal of Bacteriology *173*, 955–960.

Klein, R., Baranyi, U., Rössler, N., Greineder, B., Scholz, H., and Witte, A. (2002). Natrialba magadii virus ϕCh1: first complete nucleotide sequence and functional organization of a virus infecting a haloalkaliphilic archaeon. Molecular Microbiology *45*, 851–863.

Klein, R., Rössler, N., Iro, M., Scholz, H., and Witte, A. (2012). Haloarchaeal myovirus ϕCh1 harbours a phase variation system for the production of protein variants with distinct cell surface adhesion specificities. Molecular Microbiology *83*, 137–150.

Krupovič, M., and Bamford, D.H. (2008). Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota. Virology *375*, 292–300.

Krupovič, M., Forterre, P., and Bamford, D.H. (2010). Comparative Analysis of the Mosaic Genomes of Tailed Archaeal Viruses and Proviruses Suggests Common Themes for Virion Architecture and Assembly with Tailed Viruses of Bacteria. Journal of Molecular Biology *397*, 144–160.

Krupovic, M., Cvirkaite-Krupovic, V., Iranzo, J., Prangishvili, D., and Koonin, E.V. (2018). Viruses of archaea: Structural, functional, environmental and evolutionary genomics. Virus Research *244*, 181–193.

Krupovic, M., Makarova, K.S., Wolf, Y.I., Medvedeva, S., Prangishvili, D., Forterre, P., and Koonin, E.V. (2019). Integrated mobile genetic elements in Thaumarchaeota. Environmental Microbiology *0*.

Landy, A. (2015). The λ Integrase Site-specific Recombination Pathway. Microbiology Spectrum *3*.

Letzelter, C., Duguet, M., and Serre, M.-C. (2004). Mutational Analysis of the Archaeal Tyrosine Recombinase SSV1 Integrase Suggests a Mechanism of DNA Cleavage in trans. J. Biol. Chem. *279*, 28936–28944.

Levin, B.R., Stewart, F.M., and Chao, L. (1977). Resource-Limited Growth, Competition, and Predation: A Model and Experimental Studies with Bacteria and Bacteriophage. The American Naturalist *111*, 3–24.

Li, Z., Li, X., Xiao, X., and Xu, J. (2016). An Integrative Genomic Island Affects the Adaptations of the Piezophilic Hyperthermophilic Archaeon Pyrococcus yayanosii to High Temperature and High Hydrostatic Pressure. Front. Microbiol. *7*.

Liu, D., and Huang, L. (2002). Induction of the Sulfolobus shibatae virus SSV1 DNA replication by mitomycin C. Chin. Sci. Bull. *47*, 923–927.

Liu, Y., Wang, J., Liu, Y., Wang, Y., Zhang, Z., Oksanen, H.M., Bamford, D.H., and Chen, X. (2015). Identification and characterization of SNJ2, the first temperate pleolipovirus integrating into the genome of the SNJ1-lysogenic archaeal strain. Molecular Microbiology *98*, 1002–1020.

Luo, Y., Pfister, P., Leisinger, T., and Wasserfallen, A. (2001). The Genome of Archaeal Prophage ΨM100 Encodes the Lytic Enzyme Responsible for Autolysis ofMethanothermobacter wolfeii. Journal of Bacteriology *183*, 5788–5792.

Martin, A., Yeats, S., Janekovic, D., Reiter, W.-D., Aicher, W., and Zillig, W. (1984). SAV 1, a temperate u.v.-inducible DNA virus-like particle from the archaebacterium Sulfolobus acidocaldarius isolate B12. The EMBO Journal *3*, 2165–2168.

Meinke, G., Bohm, A., Hauber, J., Pisabarro, M.T., and Buchholz, F. (2016). Cre Recombinase and Other Tyrosine Recombinases. Chem. Rev.

Mochizuki, T., Sako, Y., and Prangishvili, D. (2011). Provirus Induction in Hyperthermophilic Archaea: Characterization of Aeropyrum pernix Spindle-Shaped Virus 1 and Aeropyrum pernix Ovoid Virus 1. Journal of Bacteriology *193*, 5412–5419.

Muskhelishvili, G. (1993). The Archaeal SSV Integrase Promotes Intermolecular Excisive Recombination in Vitro. Systematic and Applied Microbiology *16*, 605–608.

Muskhelishvili, G., Palm, P., and Zillig, W. (1993). SSV1-encoded site-specific recombination system in Sulfolobus shibatae. Mol. Gen. Genet. *237*, 334–342.

Nalabothula, N., Xi, L., Bhattacharyya, S., Widom, J., Wang, J.-P., Reeve, J.N., Santangelo, T.J., and Fondufe-Mittendorf, Y.N. (2013). Archaeal nucleosome positioning in vivo and in vitro is directed by primary sequence motifs. BMC Genomics *14*, 391.

Oberto, J., Gaudin, M., Cossu, M., Gorlas, A., Slesarev, A., Marguet, E., and Forterre, P. (2014). Genome Sequence of a Hyperthermophilic Archaeon, Thermococcus nautili 30-1, That Produces Viral Vesicles. Genome Announc. *2*, e00243-14.

Oppenheim, A.B., Kobiler, O., Stavans, J., Court, D.L., and Adhya, S. (2005). Switches in Bacteriophage Lambda Development. Annu. Rev. Genet. *39*, 409–429.

Paul, J.H. (2008). Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? The ISME Journal *2*, 579–589.

Pauly Matthew D., Bautista Maria A., Black Jesse A., and Whitaker Rachel J. (2019). Diversified local CRISPR-Cas immunity to viruses of Sulfolobus islandicus. Philosophical Transactions of the Royal Society B: Biological Sciences *374*, 20180093.

Peng, X. (2008). 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 *154*, 383–391.

Prangishvili, D., Stedman, K., and Zillig, W. (2001). Viruses of the extremely thermophilic archaeon Sulfolobus. Trends in Microbiology *9*, 39–43.

Prangishvili, D., Vestergaard, G., Häring, M., Aramayo, R., Basta, T., Rachel, R., and Garrett, R.A. (2006). Structural and Genomic Properties of the Hyperthermophilic Archaeal Virus ATV with an Extracellular Stage of the Reproductive Cycle. Journal of Molecular Biology *359*, 1203–1216.

Rajeev, L., Malanowska, K., and Gardner, J.F. (2009). Challenging a Paradigm: the Role of DNA Homology in Tyrosine Recombinase Reactions. Microbiol Mol Biol Rev *73*, 300–309.

Raymann, K., Forterre, P., Brochier-Armanet, C., and Gribaldo, S. (2014). Global Phylogenomic Analysis Disentangles the Complex Evolutionary History of DNA Replication in Archaea. Genome Biol Evol *6*, 192–212.

Redder, P., and Garrett, R.A. (2006). Mutations and Rearrangements in the Genome of Sulfolobus solfataricus P2. J. Bacteriol. *188*, 4198–4206.

Redder, P., Peng, X., Brügger, K., Shah, S.A., Roesch, F., Greve, B., She, Q., Schleper, C., Forterre, P., Garrett, R.A., et al. (2009). Four newly isolated fuselloviruses from extreme geothermal environments reveal unusual morphologies and a possible interviral recombination mechanism. Environmental Microbiology *11*, 2849–2862.

Ren, Y., She, Q., and Huang, L. (2013). Transcriptomic analysis of the SSV2 infection of Sulfolobus solfataricus with and without the integrative plasmid pSSVi. Virology *441*, 126–134.

Robinson, N.P., and Bell, S.D. (2007). Extrachromosomal element capture and the evolution of multiple replication origins in archaeal chromosomes. PNAS *104*, 5806–5811.

Roine, E., Kukkaro, P., Paulin, L., Laurinavičius, S., Domanska, A., Somerharju, P., and Bamford, D.H. (2010). New, Closely Related Haloarchaeal Viral Elements with Different Nucleic Acid Types. Journal of Virology *84*, 3682–3689.

Rössler, N., Klein, R., Scholz, H., and Witte, A. (2004). Inversion within the haloalkaliphilic virus phi Ch1 DNA results in differential expression of structural proteins. Mol. Microbiol. *52*, 413–426.

Samson, R.Y., Xu, Y., Gadelha, C., Stone, T.A., Faqiri, J.N., Li, D., Qin, N., Pu, F., Liang, Y.X., She, Q., et al. (2013). Specificity and Function of Archaeal DNA Replication Initiator Proteins. Cell Reports *3*, 485–496.

Schleper, C., Kubo, K., and Zillig, W. (1992). 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 *89*, 7645–7649.

Selb, R., Derntl, C., Klein, R., Alte, B., Hofbauer, C., Kaufmann, M., Beraha, J., Schöner, L., and Witte, A. (2017). The Viral Gene ORF79 Encodes a Repressor Regulating Induction of the Lytic Life Cycle in the Haloalkaliphilic Virus φCh1. Journal of Virology *91*, e00206-17.

Senčilo, A., Jacobs-Sera, D., Russell, D.A., Ko, C.-C., Bowman, C.A., Atanasova, N.S., Österlund, E., Oksanen, H.M., Bamford, D.H., Hatfull, G.F., et al. (2013). Snapshot of haloarchaeal tailed virus genomes. RNA Biology *10*, 803–816.

Serre, M.-C., Letzelter, C., Garel, J.-R., and Duguet, M. (2002). Cleavage Properties of an Archaeal Site-specific Recombinase, the SSV1 Integrase. J. Biol. Chem. *277*, 16758–16767.

Serre, M.-C., El Arnaout, T., Brooks, M.A., Durand, D., Lisboa, J., Lazar, N., Raynal, B., van Tilbeurgh, H., and Quevillon-Cheruel, S. (2013). The Carboxy-Terminal αN Helix of the Archaeal XerA Tyrosine Recombinase Is a Molecular Switch to Control Site-Specific Recombination. PLoS One *8*.

She, Q., Singh, R.K., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M.J., Chan-Weiher, C.C.-Y., Clausen, I.G., Curtis, B.A., Moors, A.D., et al. (2001a). The complete genome of the crenarchaeon Sulfolobus solfataricus P2. PNAS *98*, 7835–7840.

She, Q., Peng, X., Zillig, W., and Garrett, R.A. (2001b). Genome evolution: Gene capture in archaeal chromosomes. Nature *409*, 478–478.

She, Q., Brügger, K., and Chen, L. (2002). Archaeal integrative genetic elements and their impact on genome evolution. Research in Microbiology *153*, 325–332.

She, Q., Shen, B., and Chen, L. (2004). Archaeal integrases and mechanisms of gene capture. Biochemical Society Transactions *32*, 222–226.

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

Stark, W.M. (2015). The Serine Recombinases. 73–89.

Stolt, P., and Zillig, W. (1992). In vivo studies on the effects of immunity genes on early lytic transcription in the Halobacterium salinarium phage φ H. Molec. Gen. Genet. *235*, 197–204.

Sugahara, J., Fujishima, K., Nunoura, T., Takaki, Y., Takami, H., Takai, K., Tomita, M., and Kanai, A. (2012). Genomic Heterogeneity in a Natural Archaeal Population Suggests a Model of tRNA Gene Disruption. PLOS ONE *7*, e32504.

Tang, S.-L., Nuttall, S., Ngui, K., Fisher, C., Lopez, P., and Dyall-Smith, M. (2002). HF2: a double-stranded DNA tailed haloarchaeal virus with a mosaic genome. Molecular Microbiology *44*, 283–296.

Vandecraen, J., Chandler, M., Aertsen, A., and Houdt, R.V. (2017). The impact of insertion sequences on bacterial genome plasticity and adaptability. Critical Reviews in Microbiology *43*, 709–730.

Wagner, A., Whitaker, R.J., Krause, D.J., Heilers, J.-H., van Wolferen, M., van der Does, C., and Albers, S.-V. (2017). Mechanisms of gene flow in archaea. Nat Rev Micro *15*, 492–501.

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

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

Weckselblatt, B., and Rudd, M.K. (2015). Human Structural Variation: Mechanisms of Chromosome Rearrangements. Trends in Genetics *31*, 587–599.

Wiedenheft, B., Stedman, K., Roberto, F., Willits, D., Gleske, A.-K., Zoeller, L., Snyder, J., Douglas, T., and Young, M. (2004). Comparative Genomic Analysis of Hyperthermophilic Archaeal Fuselloviridae Viruses. J Virol *78*, 1954–1961.

Williams, K.P. (2002). Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. Nucleic Acids Res *30*, 866–875.

Winckler, T., Szafranski, K., and Glöckner, G. (2005). Transfer RNA gene-targeted integration: an adaptation of retrotransposable elements to survive in the compact Dictyostelium discoideum genome. Cytogenetic And Genome Research *110*, 288–298.

Witte, A., Baranyi, U., Klein, R., Sulzner, M., Luo, C., Wanner, G., Kru¨ger, D.H., and Lubitz, W. (1997). Characterization of Natronobacterium magadii phage ΦCh1, a unique archaeal phage containing DNA and RNA. Molecular Microbiology *23*, 603–616.

Yeats, S., McWilliam, P., and Zillig, W. (1982). A plasmid in the archaebacterium Sulfolobus acidocaldarius. The EMBO Journal *1*, 1035–1038.

Zhan, Z., Ouyang, S., Liang, W., Zhang, Z., Liu, Z.-J., and Huang, L. (2012). Structural and functional characterization of the C-terminal catalytic domain of SSV1 integrase. Acta Cryst. D *68*, 659–670.

Zhan, Z., Zhou, J., and Huang, L. (2015). Site-Specific Recombination by SSV2 Integrase: Substrate Requirement and Domain Functions. J. Virol. *89*, 10934–10944.

Zivanovic, Y., Lopez, P., Philippe, H., and Forterre, P. (2002). Pyrococcus genome comparison evidences chromosome shuffling-driven evolution. Nucleic Acids Res *30*, 1902–1910.

Figure 1

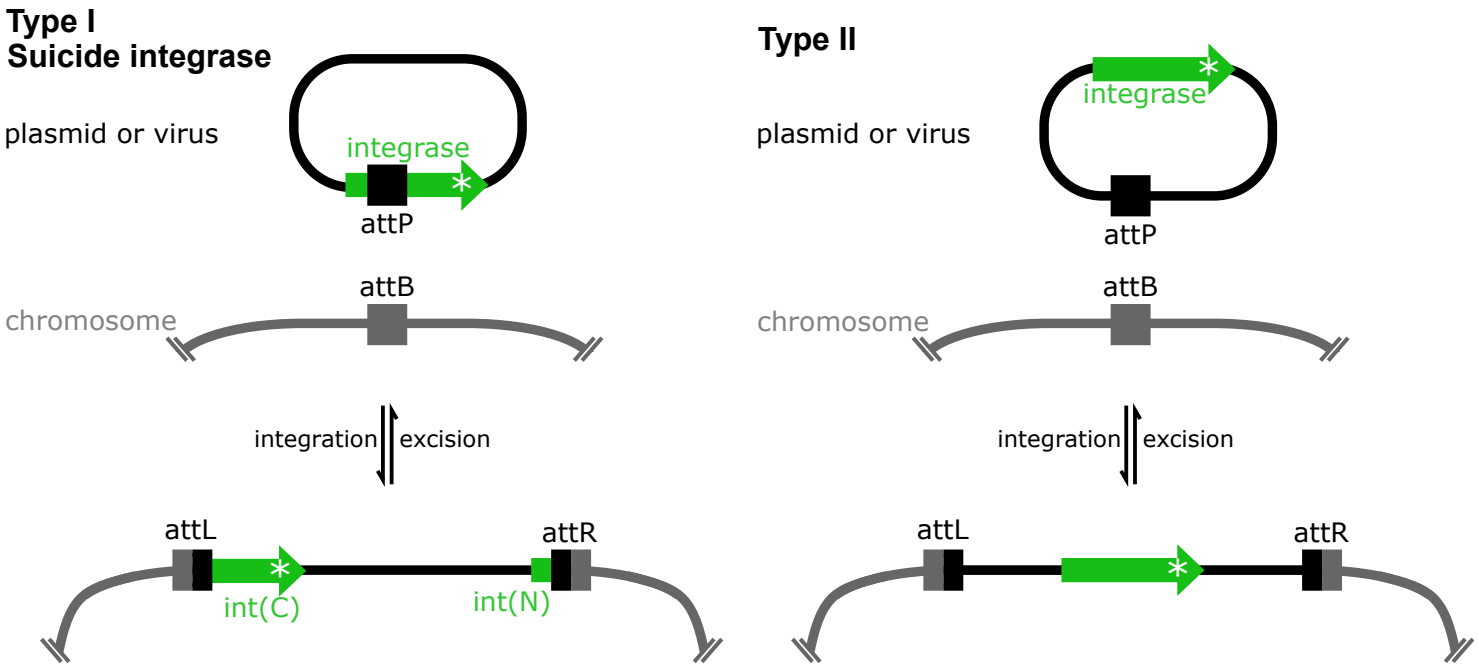

Type I
Suicide integrase
plasmid or virus
integrase
attP
attB
chromosome
integration
excision
attL
attR
int(C)
int(N)
Type II
plasmid or virus
integrase
attP
attB
chromosome
integration
excision
attL
attR

Figure 2

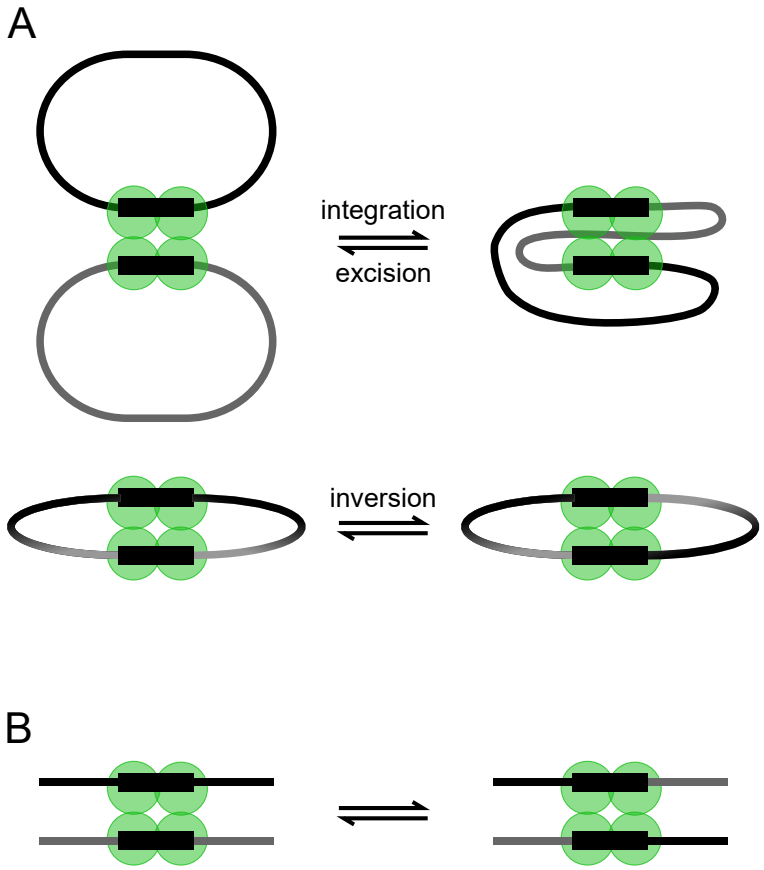


Figure 3

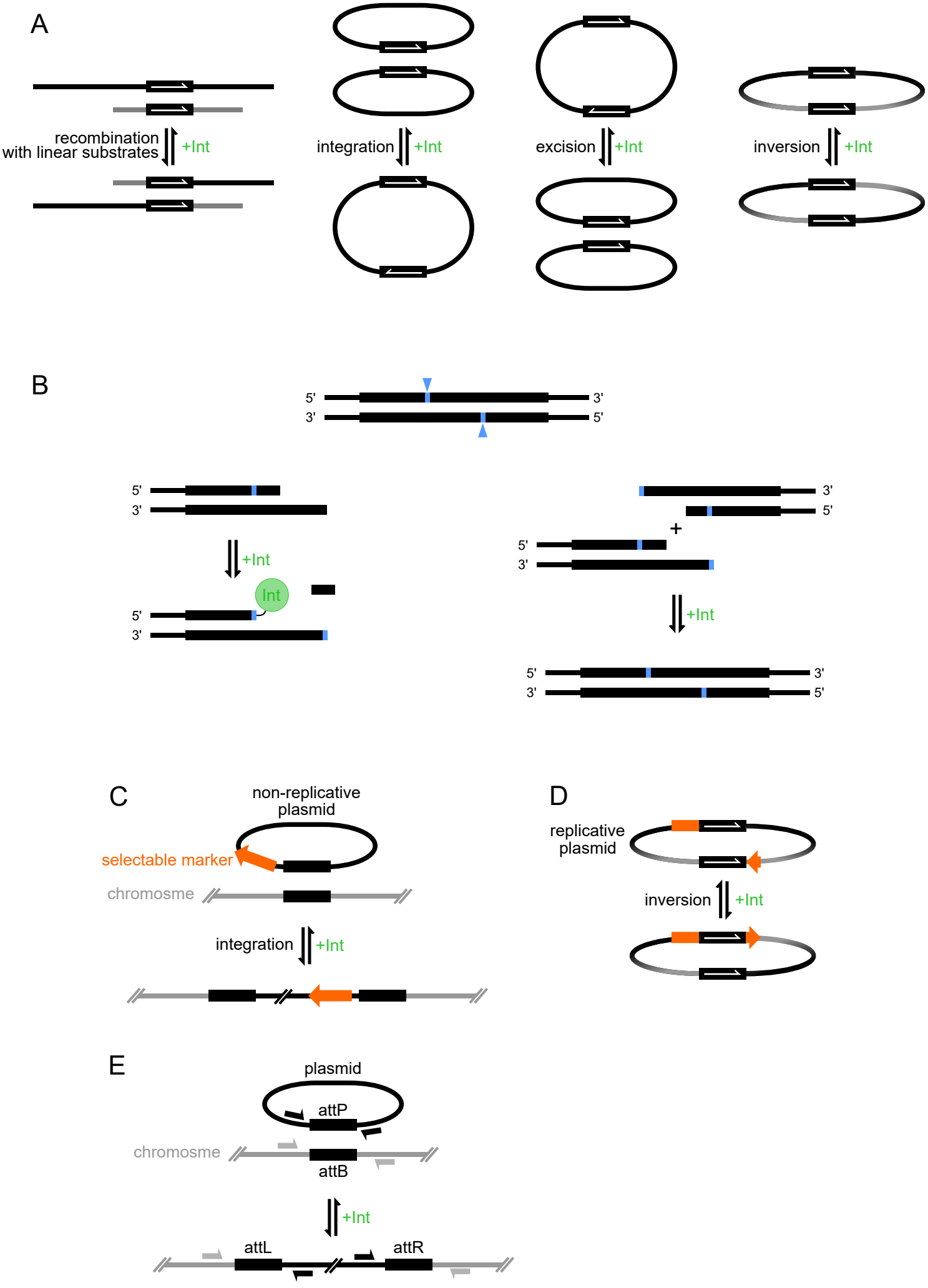

A
recombination with linear substrates
+Int
integration
+Int
excision
+Int
inversion
+Int
B
5'
3'
3'
5'
5'
3'
+Int
Int
5'
3'
3'
5'
+
5'
3'
+Int
5'
3'
3'
5'
C
non-replicative plasmid
selectable marker
chromosme
integration
+Int
D
replicative plasmid
inversion
+Int
E
plasmid
attP
chromosme
attB
+Int
attL
attR

Figure 4

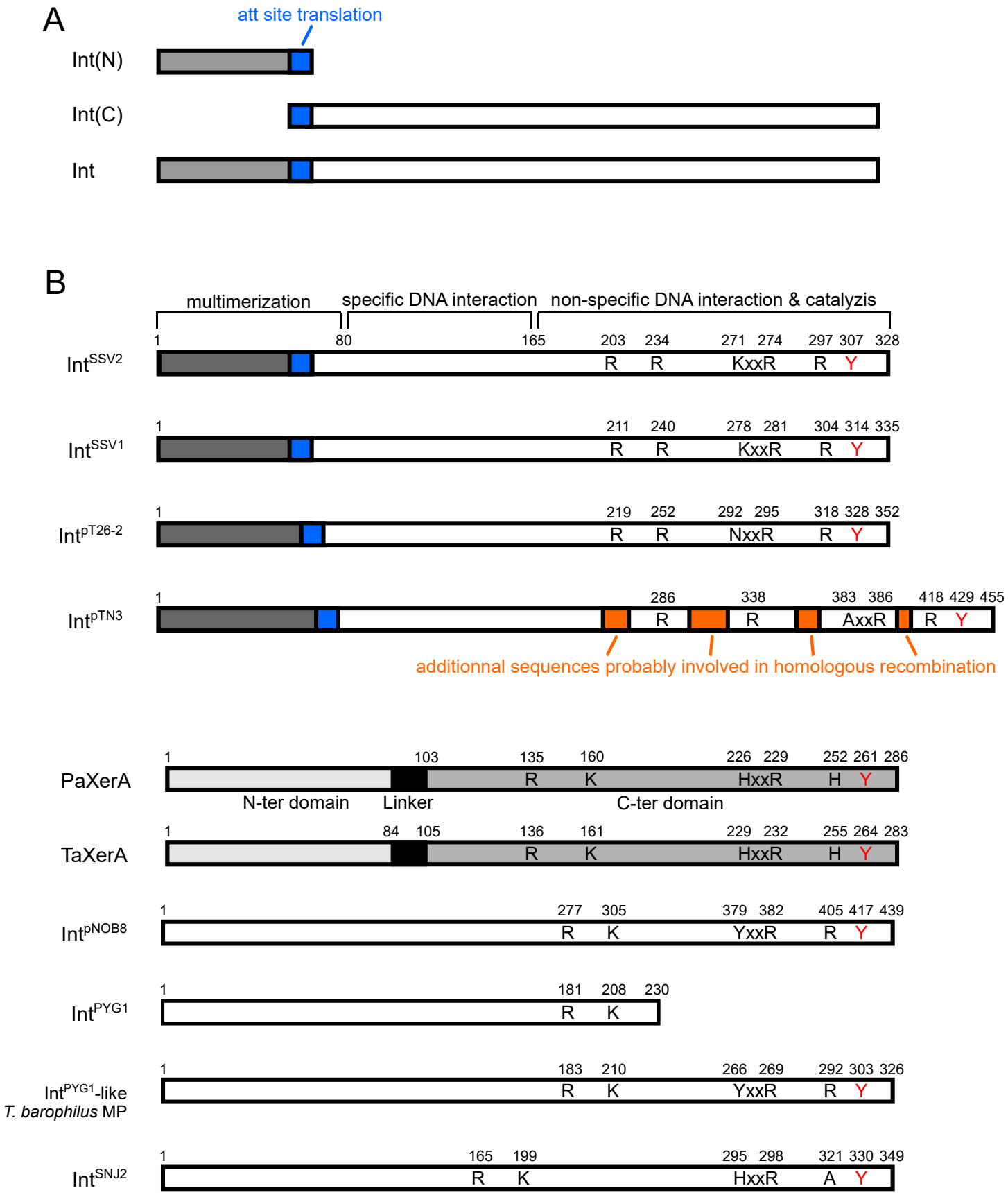

A
att site translation
Int(N)
Int(C)
Int
B
multimerization
specific DNA interaction
non-specific DNA interaction & catalyzis
Int^SSV2
1 80 165 203 234 271 274 297 307 328
R R KxxR R Y
Int^SSV1
1 211 240 278 281 304 314 335
R R KxxR R Y
Int^pT26-2
1 219 252 292 295 318 328 352
R R NxxR R Y
Int^pTN3
1 286 338 383 386 418 429 455
R R AxxR R Y
additionnal sequences probably involved in homologous recombination
PaXerA
1 103 135 160 226 229 252 261 286
R K HxxR H Y
N-ter domain
Linker
C-ter domain
TaXerA
1 84 105 136 161 229 232 255 264 283
R K HxxR H Y
Int^pNOB8
1 277 305 379 382 405 417 439
R K YxxR R Y
Int^PYG1
1 181 208 230
R K
Int^PYG1-like
T. barophilus MP
1 183 210 266 269 292 303 326
R K YxxR R Y
Int^SNJ2
1 165 199 295 298 321 330 349
R K HxxR A Y

Figure 5

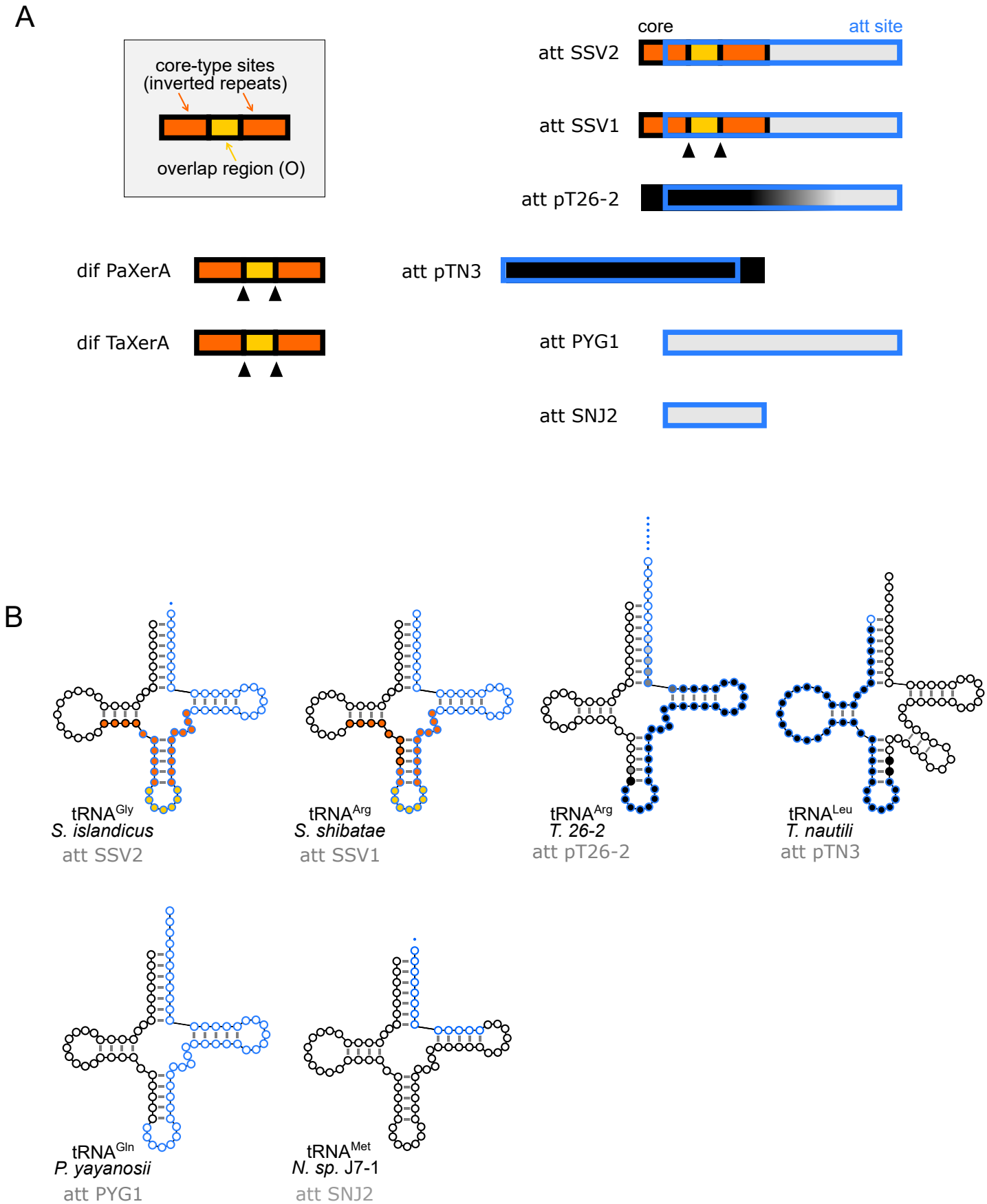

A
core-type sites
(inverted repeats)
overlap region (O)
core
att site
att SSV2
att SSV1
att pT26-2
dif PaXerA
att pTN3
dif TaXerA
att PYG1
att SNJ2
B
tRNA^Gly
S. islandicus
att SSV2
tRNA^Arg
S. shibatae
att SSV1
tRNA^Arg
T. 26-2
att pT26-2
tRNA^Leu
T. nautili
att pTN3
tRNA^Gln
P. yayanosii
att PYG1
tRNA^Met
N. sp. J7-1
att SNJ2

Figure 6

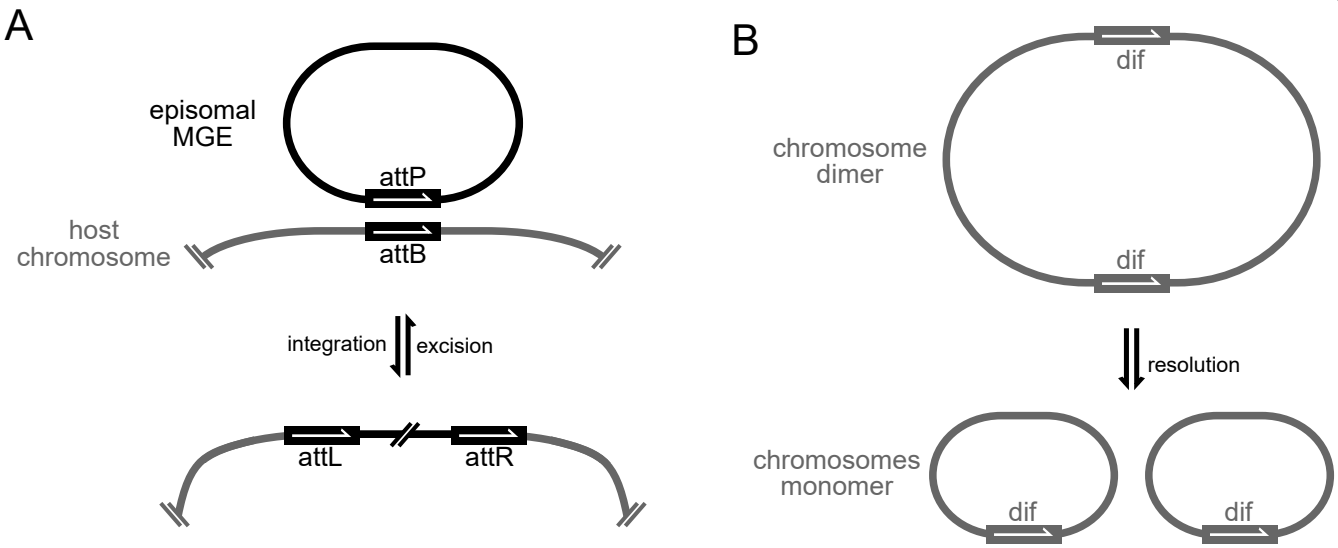

A
episomal
MGE
attP
host
chromosome
attB
integration
excision
attL
attR
B
dif
chromosome
dimer
dif
resolution
chromosomes
monomer
dif
dif

Figure 7

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 ▨

Supplementary material and methods

For the experiments presented in the articles, the material and methods are included in the article. Here, we present supplementary material and methods for the additional experiments.

In vitro recombination activity assays

Expression vector construction

The procedure indicated in Article 3 was used for the construction of the expression vectors pAN610 and pAN612 for the integrases identified in *Methanocaldococcus fervens* (NCBI protein ID WP_015792118.1) and *Methanocaldococcus sp.* FS406-22 (NCBI protein ID ADC69779.1) respectively. The PCR template was total DNA extracts graciously supplemented by William B Whitman (University of Georgia) for *Methanocaldococcus fervens* AG96 and by Tom Lie (Leigh Lab, University of Washington) for *Methanocaldococcus sp.* FS406-22. The primers IntMfer_FOR and _REV and IntM40622_FOR and _REV were used respectively. The same procedure was used for the construction of the expression vectors pCB557 and pCB577 for Int^{TPV1} (Genbank AEY69059.1) and its catalytic mutant Int^{TPV1}Y291F respectively, with the primers IntTPV1_FOR and _REV and 557-Y291F_FOR and _REV respectively. For the wild type integrase cloning, the PCR template was virus DNA extract graciously supplemented by Aurore Gorlas (I2BC, Orsay). The same procedure was also used for the construction of the expression vector pCB622 for Int^{pTF1} with the primers IntpTF1_FOR and _REV and with *Thermococcus fumicolans* genomic DNA graciously supplemented by Phil Oger (INSA, Lyon) as PCR template.

Recombinant protein production and purification

Int^{TPV1} and its catalytic mutant Int^{TPV1}Y291F were produced and purified as indicated in the Article 3 with several modifications. Induction time was 2.5 hours and soluble protein extracts were treated 15 min at 55°C. The integrases identified in *Methanocaldococcus fervens* and *Methanocaldococcus sp.* FS406-22 Int^{pTF1} were produced and purified as indicated in the Article 3.

For the purification of Int^{pTF1}, plasmids pCB622 was introduced in *Escherichia coli* Rosetta BL21 (DE3). Cells were grown in LB medium to OD 0.5 and recombinant protein production was induced with 250 µM IPTG. After 2.5 hour, cells were harvested and resuspended in the purification buffer (1 M KCl, 40 mM Tris HCl pH=8, 10 % glycerol and 5 mM B-mercaptoethanol) supplemented with a protease inhibitor cocktail (cOmplete ULTRA Tablets, EDTA-free, Roche). Cell were lyzed by a pressure shock with a one shot cell disruptor (Constant Systems Ltd) and centrifuged for 30 min at 18000 g at 4°C. The soluble fraction was recovered and warmed at 65°C for 10 minutes. Int^{pTF1} was then purified using Strep-Tactin® Spin Column (IBA) accordingly to the manufacturer instructions. Proteins were eluted three consecutive times with 150 µL purification buffer supplemented with 2.5 mM d-biotin. D-biotin was subsequently removed by buffer exchange with a Vivaspin® Centrifugal Concentrators (Sartorius). The purified proteins contained the N-ter strep-tag and their concentration was determined by spectrophotometry.

Integrase substrate construction and production

To construct the plasmid pCB628, pCB512 and pCB624, the method presented in the Article 3 for plasmid pCB568 was implemented. For pCB628, the annealed oligonucleotides were tRNAser406-22+1_A and tRNAser406-22+1_B. For pCB512, the annealed oligonucleotides were tRNAglyTPV_A and tRNAglyTPV_B. For pCB624, the annealed oligonucleotides were 2+tRNAserFUMI_A and 2+tRNAserFUMI_B.

Plasmid pCB594 was obtained by Gibson assembly of the following three fragments: (1) pCB512 digested by NdeI, (2) a PCR product amplified from pUC4K with the primers KanR-inv1BIS and KanR-inv2 and corresponding to the KmR gene and (3) a PCR product amplified from pCB512 with the primers tRNAgly-inv1 and 2 (Table 8) and corresponding to $tRNA^{Gly}$ from *Thermococcus sp.* TPV. The assembled product was cloned into *Escherichia coli* strain XL1-Blue. All plasmids were verified by sequencing and extracted using the kit NucleoSpin Plasmid (Macherey Nagel).

In vitro integrase enzymatic assay

For integration assay, a plasmid carrying one attachment site was commonly incubated with the adequate integrase, treated with proteinase K and ran on a gel at low voltage (50 V) for 3 hours. The gel was stained with Ethidium Bromide after migration. For inversion and excision assays, a plasmid carrying two attachment sites was commonly incubation with the adequate integrase. The plasmid was then purified with the kit NucleoSpin Gel and PCR clean-up (Macherey Nagel), digested with the adequate restriction enzymes (Fast Digest, ThermoScientific) and separated by a classical gel electrophoresis.

In details, for the integration assay of the integrase identified in *Methanocaldococcus sp.* 406-22, 500 µg pCB628 DNA and 300 ng integrase were incubated at 65°C in 300 mM KCl, 25 mM Tris HCl pH=8, 0.25 % glycerol and 125 µM B-mercaptoethanol for a total volume of 20 µL. The reaction mixture was treated with proteinase K and run on a gel at low voltage (50 V) for 3 hours. The gel was stained with ethidium bromide after migration.

For Int^{TPV1} inversion assay, 500 µg pCB594 DNA was incubated with 300 ng Int^{TPV1} or Int^{TPV1}Y291F at 55°C for 1h or 4h. The reaction buffer was 200 mM KCl, 4 mM Tris HCl pH=8, 1% glycerol and 500 µM B-mercaptoethanol for a total volume of 20 µL. Reaction products were purified with the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel), digested with the restriction enzymes XhoI and ScaI (Fast-digest, ThermoFisher) and separated by gel electrophoresis.

For Int^{pTF1} integration assay, 500 µg pUC18 or pCB624 were incubated with 170 ng (1X), 340 ng (2X) or 720 ng (4X) Int^{pTF1} at 65°. The incubation buffer was 300 mM KCl, 1.9 to 8 mM Tris HCl pH=8, 0.5% to 2% glycerol and 250 to 1000 µM B-mercaptoethanol for a total volume of 20 µL. The reaction mixture was treated with proteinase K and run on a gel at low voltage (50 V) for 3 hours. The gel was stained with Ethidium Bromide after migration.

Crystallography

To purify Int^{pTN3}, plasmid pJO344 (Article 1 (Cossu et al., 2017)) was introduced in *Escherichia coli* BL21 (DE3). Cells were grown in 4 L of LB medium, induced at DO=0.5 with 1 mM IPTG and harvested after 2.5 h. Cell pellets were resuspended in the purification buffer (1M KCl, 40 mM Tris pH=8, 5 mM B-mercaptoethanol, 5% glycerol) supplemented with a protease inhibitor cocktail (cOmplete ULTRA Tablets, EDTA-free, Roche). Cell were lyzed by a pressure shock with a one shot cell disruptor (Constant Systems Ltd). The soluble fraction was recovered, warmed at 65°C for 10 minutes, centrifuged at 5000 g for 15 min and filtered. The solution was then loaded on a StrepTrap HP 5mL column (GE Healthcare). STREP-tagged Int^{pTN3} was eluted with the purification buffer supplemented with 2.5 mM d-desthiobiotin. Fractions containing Int^{pTN3} were concentrated with a Vivaspin® Centrifugal Concentrators (Sartorius). Protein solution was then diluted with unsalted purification buffer to obtain a 500 mM KCl concentration and loaded on a HiTrap™ Heparin HP 1 mL column (GE Healthcare). Proteins were eluted in a KCl gradient from 0.5 M to 1.5M. Fractions containing Int^{pTN3} ranged from 0.75 M KCl to 1M KCl with a maximal elution at 0.88 M KCl. They were concentrated with a Vivaspin® Centrifugal Concentrators (Sartorius) and loaded on a HiLoad 16/600 75 prep grade column (GE Healthcare). Fractions containing Int^{pTN3} were concentrated to 30 mg/mL with a Vivaspin® Centrifugal Concentrators (Sartorius). Purified Int^{pTN3} was then used for crystallogenesis in various conditions and one crystal was obtained that diffracted at 2.1Å.

Several test were performed to lower the KCl concentration either during the exclusion chromatography or during the final concentration, but they all resulted in protein precipitation.

Table 7. Plasmids[1] constructed during the thesis

Name	Backbone	Insertion	Resistance
pAN610	pET-26b**(+)**	Wild type gene coding for the integrase identified in *Methanocaldococcus fervens*	KmR
pAN612	pET-26b**(+)**	Wild type gene coding for the integrase identified in *Methanocaldococcus sp.* FS406-22	KmR
pCB557	pET-26b**(+)**	Wild type gene coding for Int^{TPV1}	KmR
pCB577	pET-26b**(+)**	Y291F mutant of the gene coding for Int^{TPV1}	KmR
pCB622	pET-26b**(+)**	Wild type gene coding for Int^{pTF1}	KmR
pCB628	pUC18	tRNA^{Ser}TGA gene and 1nt downstream from *Methanocaldococcus sp.* FS406-22	AmpR
pCB512	pUC18	tRNA^{Gly}CCC gene from *Thermococcus sp. TPV*	AmpR
pCB594	pCB512	Reverse repetition of the pCB512 insert	AmpR, KmR
pCB624	pUC18	tRNA^{Ser}CGA gene and 2nt upstream from *Thermococcus fumicolans*	AmpR

[1]Plasmids used in the experiments presented in the articles are excluded from this table.

Table 8. Oligonucleotides[1] used during the thesis

Name	Sequence
IntTPV1_FOR	TTAAGAAGGAGATATACATATGTGGAGCCACCCGCAGTTCGAAAAAAACGAGATGGGCATCAAC
IntTPV1_REV	CAGTGGTGGTGGTGGTGGTGCTCGAGTGCGGCCGCTCATTCCCCATCCTCCAGC
557-Y291F_FOR	CCAGCTGGACAAAATACGCATCG
557-Y291F_REV	AAAAACTCGCCGACGAGG
IntMfer_FOR	GGTGCTCGAGTGCGGCCGCATTACTGGATTCTCAACAAC
IntMfer-REV	CTTTAAGAAGGAGATATACAATGTGGAGCCACCCGCAGTTCGAAAAAAGGGAGAAAGAGCTTAAAC
IntM40622_FOR	GGTGCTCGAGTGCGGCCGCATCACAACAACTTTTTTATCGTCTTCAAC
IntM40622_REV	CTTTAAGAAGGAGATATACAATGTGGAGCCACCCGCAGTTCGAAAAAAAGGGGAACGGGCTTAATTTTAAAAATC
IntpTF1_FOR	CTTTAAGAAGGAGATATACATATGTGGAGCCACCCGCAGTTCGAAAAAGTAAAATCGTCTGGTGTGTC
IntpTF1_REV	CAGTGGTGGTGGTGGTGGTGCTCGAGTGCGGCCGCTAAAGCTCCAAAATCCC
tRNAser406-22+1_A	CTAGGGATCCGCAGGGGTAGCCAAGCCAGGACTACGGCGCTGGACTTGAGATCCAGTGGGGCTTTGCCCCGCCTGGGTTCAAATCCCAGCCCCTGCGCCATGAATTCGGTC
tRNAser406-22+1_B	GACCGAATTCATGGCGCAGGGGCTGGGATTTGAACCCAGGCGGGGCAAAGCCCCACTGGATCTCAAGTCCAGCGCCGTAGTCCTGGCTTGGCTACCCCTGCGGATCCCTAG
tRNAglyTPV_A	CGGGATCCGCGGTGGTAGTCTAGCCTGGTCCAGGACACCGGCCTCCCAAGCCGGTGACCCGGGTTCAAATCCCGGCCACCGCACCAGAATTCGGTC
tRNAglyTPV_B	GACCGAATTCTGGTGCGGTGGCCGGGATTTGAACCCGGGTCACCGGCTTGGGAGGCCGGTGTCCTGGACCAGGCTAGACTACCACCGCGGATCCCG
2+tRNAserFUMI_A	CTAGGGATCCGCGCCGGGGTAGCCTAGCCTGGGAAGGCGCGGGACTCGAGATCCCGTGCGCTCTGCGCACCAGGGTTCAAATCCCTGCCCCGGCGCCAGAATTCGGTC
2+tRNAserFUMI_B	GACCGAATTCTGGCGCCGGGGCAGGGATTTGAACCCTGGTGCGCAGAGCGCACGGGATCTCGAGTCCCGCGCCTTCCCAGGCTAGGCTACCCCGGCGCGGATCCCTAG
KanR-inv1BIS	CTAGACTACCACCGCCGCTGAGGTCTGCCTCGT
KanR-inv2	GTGCGGTATTTCACACCGCAAAAGCCACGTTGTGTCTCAAAATC
tRNAgly-inv1	ATTGTACTGAGAGTGCACCATGGTGCGGTGGCCGGGAT
tRNAgly-inv1	AGGCAGACCTCAGCGGCGGTGGTAGTCTAGCCTGGTC

[1]Oligonucleotides used in the experiments presented in the articles are excluded from this table.

References

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

Adli, M. (2018). The CRISPR tool kit for genome editing and beyond. Nat. Commun. *9*, 1911.

Aharoni, A., Gaidukov, L., Khersonsky, O., Gould, S.M., Roodveldt, C., and Tawfik, D.S. (2005). The "evolvability" of promiscuous protein functions. Nat. Genet. *37*, 73.

Albers, S.-V., Forterre, P., Prangishvili, D., and Schleper, C. (2013). The legacy of Carl Woese and Wolfram Zillig: from phylogeny to landmark discoveries. Nat. Rev. Microbiol. *11*, 713–719.

Anderson, I., Rodriguez, J., Susanti, D., Porat, I., Reich, C., Ulrich, L.E., Elkins, J.G., Mavromatis, K., Lykidis, A., Kim, E., et al. (2008). Genome Sequence of Thermofilum pendens Reveals an Exceptional Loss of Biosynthetic Pathways without Genome Reduction. J. Bacteriol. *190*, 2957–2965.

Anderson, R.E., Sogin, M.L., and Baross, J.A. (2014). Evolutionary Strategies of Viruses, Bacteria and Archaea in Hydrothermal Vent Ecosystems Revealed through Metagenomics. PLOS ONE *9*, e109696.

Andersson, A.F., Pelve, E.A., Lindeberg, S., Lundgren, M., Nilsson, P., and Bernander, R. (2010). Replication-biased genome organisation in the crenarchaeon Sulfolobus. BMC Genomics *11*, 454.

Andrake, M.D., and Skalka, A.M. (2015). RETROVIRAL INTEGRASE: THEN AND NOW. Annu. Rev. Virol. *2*, 241.

Arndt, D., Marcu, A., Liang, Y., and Wishart, D.S. PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes. Brief. Bioinform.

Arnold, H.P., She, Q., Phan, H., Stedman, K., Prangishvili, D., Holz, I., Kristjansson, J.K., Garrett, R., and Zillig, W. (1999). The genetic element pSSVx of the extremely thermophilic crenarchaeon Sulfolobus is a hybrid between a plasmid and a virus. Mol. Microbiol. *34*, 217–226.

Atanasova, N.S., Demina, T.A., Krishnam Rajan Shanthi, S.N.V., Oksanen, H.M., and Bamford, D.H. (2018). Extremely halophilic pleomorphic archaeal virus HRPV9 extends the diversity of pleolipoviruses with integrases. Res. Microbiol.

Atomi, H., Fukui, T., Kanai, T., Morikawa, M., and Imanaka, T. (2004). Description of Thermococcus kodakaraensis sp. nov., a well studied hyperthermophilic archaeon previously reported as Pyrococcus sp. KOD1. Archaea *1*, 263–267.

Atomi, H., Imanaka, T., and Fukui, T. (2012). Overview of the genetic tools in the Archaea. Evol. Genomic Microbiol. *3*, 337.

Ausiannikava, D., and Allers, T. (2017). Diversity of DNA Replication in the Archaea. Genes *8*, 56.

Azaro, M.A., and Landy, A. (2002). λ Integrase and the λ Int Family. Mob. DNA II 118–148.

Bamford, D.H., Ravantti, J.J., Rönnholm, G., Laurinavičius, S., Kukkaro, P., Dyall-Smith, M., Somerharju, P., Kalkkinen, N., and Bamford, J.K.H. (2005). Constituents of SH1, a Novel Lipid-Containing Virus Infecting the Halophilic Euryarchaeon Haloarcula hispanica. J. Virol. *79*, 9097–9107.

Bannert, N., and Kurth, R. (2006). The Evolutionary Dynamics of Human Endogenous Retroviral Families. Annu. Rev. Genomics Hum. Genet. *7*, 149–173.

Barksdale, L., and Arden, S.B. (1974). Persisting Bacteriophage Infections, Lysogeny, and Phage Conversions. Annu. Rev. Microbiol. *28*, 265–300.

Barre, F.-X., and Midonet, C. (2015). Xer Site-Specific Recombination: Promoting Vertical and Horizontal Transmission of Genetic Information. In Mobile DNA III, A.M. Lambowitz, M. Gellert, M. Chandler, N.L. Craig, S.B. Sandmeyer, and P.A. Rice, eds. (American Society of Microbiology), pp. 163–182.

Baxter, J.C., and Funnell, B.E. (2015). Plasmid Partition Mechanisms. Plasmids Biol. Impact Biotechnol. Discov. 135–155.

Béguin, P., Charpin, N., Koonin, E.V., Forterre, P., and Krupovic, M. (2016). Casposon integration shows strong target site preference and recapitulates protospacer integration by CRISPR-Cas systems. Nucleic Acids Res. *44*, 10367–10376.

Bell, S.D., and Jackson, S.P. (1998). Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features. Trends Microbiol. *6*, 222–228.

Bentley, S.D., and Parkhill, J. (2004). Comparative Genomic Structure of Prokaryotes. Annu. Rev. Genet. *38*, 771–791.

Benzer, S. (1955). FINE STRUCTURE OF A GENETIC REGION IN BACTERIOPHAGE. Proc. Natl. Acad. Sci. U. S. A. *41*, 344–354.

Bertelli, C., Laird, M.R., Williams, K.P., Lau, B.Y., Hoad, G., Winsor, G.L., and Brinkman, F.S. (2017). IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. Nucleic Acids Res. *45*, W30–W35.

Bimboim, H.C., and Doly, J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res. *7*, 1513–1523.

Bize, A., Karlsson, E.A., Ekefjärd, K., Quax, T.E.F., Pina, M., Prevost, M.-C., Forterre, P., Tenaillon, O., Bernander, R., and Prangishvili, D. (2009). A unique virus release mechanism in the Archaea. Proc. Natl. Acad. Sci. *106*, 11306–11311.

Bobay, L.-M., Rocha, E.P.C., and Touchon, M. (2013). The Adaptation of Temperate Bacteriophages to Their Host Genomes. Mol. Biol. Evol. *30*, 737–751.

Bobay, L.-M., Touchon, M., and Rocha, E.P.C. (2014). Pervasive domestication of defective prophages by bacteria. Proc. Natl. Acad. Sci. U. S. A. *111*, 12127–12132.

Botstein, D. (2006). A THEORY OF MODULAR EVOLUTION FOR BACTERIOPHAGES*. Ann. N. Y. Acad. Sci. *354*, 484–491.

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., et al. (2018). Ten things you should know about transposable elements. Genome Biol. *19*, 199.

Bridger, S.L., Lancaster, W.A., Poole, F.L., Schut, G.J., and Adams, M.W.W. (2012). Genome Sequencing of a Genetically Tractable Pyrococcus furiosus Strain Reveals a Highly Dynamic Genome. J. Bacteriol. *194*, 4097–4106.

Brileya, K., and Reysenbach, A.-L. (2014). The Class Archaeoglobi. In The Prokaryotes: Other Major Lineages of Bacteria and The Archaea, E. Rosenberg, E.F. DeLong, S. Lory, E. Stackebrandt, and F. Thompson, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 15–23.

Brochier-Armanet, C., Forterre, P., and Gribaldo, S. (2011). Phylogeny and evolution of the Archaea: one hundred genomes later. Curr. Opin. Microbiol. *14*, 274–281.

Brockhurst, M.A., Chapman, T., King, K.C., Mank, J.E., Paterson, S., and Hurst, G.D.D. (2014). Running with the Red Queen: the role of biotic conflicts in evolution. Proc. R. Soc. B Biol. Sci. *281*.

Brügger, K., Redder, P., She, Q., Confalonieri, F., Zivanovic, Y., and Garrett, R.A. (2002). Mobile elements in archaeal genomes. FEMS Microbiol. Lett. *206*, 131–141.

Cabezón, E., Ripoll-Rozada, J., Peña, A., de la Cruz, F., and Arechaga, I. (2015). Towards an integrated model of bacterial conjugation. FEMS Microbiol. Rev. *39*, 81–95.

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

Campbell, A.M. (1963). Episomes. In Advances in Genetics, E.W. Caspari, and J.M. Thoday, eds. (Academic Press), pp. 101–145.

Canganella, F., Jones, W.J., Gambarcota, A., and Antranikian, G. (1998). Thermococcus guaymasensis sp. nov. and Thermococcus aggregans sp. nov., two novel thermophilic archaea isolated from the Guaymas Basin hydrothermal vent site. Int. J. Syst. Evol. Microbiol. *48*, 1181–1185.

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

Casjens, S.R., and Hendrix, R.W. (2015). Bacteriophage lambda: Early pioneer and still relevant. Virology *479–480*, 310–330.

Castillo, F., Benmohamed, A., and Szatmari, G. (2017). Xer Site Specific Recombination: Double and Single Recombinase Systems. Front. Microbiol. *8*.

Catchpole, R., Gorlas, A., Oberto, J., and Forterre, P. (2018). A series of new E. coli Thermococcus shuttle vectors compatible with previously existing vectors. Extremophiles 1–8.

Cavicchioli, R. (2011). Archaea — timeline of the third domain. Nat. Rev. Microbiol. *9*, 51–61.

Chaconas, G., and Harshey, R.M. (2002). Transposition of Phage Mu DNA. Mob. DNA II 384–402.

Chandler, M., de la Cruz, F., Dyda, F., Hickman, A.B., Moncalian, G., and Ton-Hoang, B. (2013). Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. Nat. Rev. Microbiol. *11*, 525–538.

Chen, C.W. (2007). Streptomyces Linear Plasmids: Replicationand Telomeres. In Microbial Linear Plasmids, F. Meinhardt, and R. Klassen, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 33–61.

Chen, Y., and Rice, P.A. (2003). New Insight into Site-Specific Recombination from Flp Recombinase-DNA Structures. Annu. Rev. Biophys. Biomol. Struct. *32*, 135–159.

Cheng, C., Kussie, P., Pavletich, N., and Shuman, S. (1998). Conservation of Structure and Mechanism between Eukaryotic Topoisomerase I and Site-Specific Recombinases. Cell *92*, 841–850.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. Nat. Rev. Genet. *18*, 71–86.

Contursi, P., Jensen, S., Aucelli, T., Rossi, M., Bartolucci, S., and She, Q. (2006). Characterization of the Sulfolobus host–SSV2 virus interaction. Extremophiles *10*, 615–627.

Cortez, D., Forterre, P., and Gribaldo, S. (2009). A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. Genome Biol. *10*, R65.

Cortez, D., Quevillon-Cheruel, S., Gribaldo, S., Desnoues, N., Sezonov, G., Forterre, P., and Serre, M.-C. (2010). Evidence for a Xer/ dif System for Chromosome Resolution in Archaea. PLOS Genet *6*, e1001166.

Cossu, M., Da Cunha, V., Toffano-Nioche, C., Forterre, P., and Oberto, J. (2015). Comparative genomics reveals conserved positioning of essential genomic clusters in highly rearranged Thermococcales chromosomes. Biochimie *118*, 313–321.

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

Craigie, R., and Bushman, F.D. (2012). HIV DNA Integration. Cold Spring Harb. Perspect. Med. *2*.

Crick, F.H.C., Barnett, L., Brenner, S., and Watts-Tobin, R.J. (1961). General Nature of the Genetic Code for Proteins. Nature *192*, 1227.

Cunha, V.D., Gaia, M., Gadelle, D., Nasir, A., and Forterre, P. (2017). Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. PLOS Genet. *13*, e1006810.

Curcio, M.J., and Derbyshire, K.M. (2003). The outs and ins of transposition: from Mu to Kangaroo. Nat. Rev. Mol. Cell Biol. *4*, 865–877.

Cury, J., Touchon, M., and Rocha, E.P.C. (2017). Integrative and conjugative elements and their hosts: composition, distribution and organization. Nucleic Acids Res. *45*, 8943–8956.

Dalmasso, C., Oger, P., Selva, G., Courtine, D., L'Haridon, S., Garlaschelli, A., Roussel, E., Miyazaki, J., Reveillaud, J., Jebbar, M., et al. (2016). Thermococcus piezophilus sp. nov., a novel hyperthermophilic and piezophilic archaeon with a broad pressure range for growth, isolated from a deepest hydrothermal vent at the Mid-Cayman Rise. Syst. Appl. Microbiol. *39*, 440–444.

Darmon, E., and Leach, D.R.F. (2014). Bacterial Genome Instability. Microbiol. Mol. Biol. Rev. *78*, 1–39.

Delwart, E.L. (2007). Viral metagenomics. Rev. Med. Virol. *17*, 115–131.

Dorgai, L., Yagil, E., and Weisberg, R.A. (1995). Identifying Determinants of Recombination Specificity: Construction and Characterization of Mutant Bacteriophage Integrases. J. Mol. Biol. *252*, 178–188.

Dubin, M.J., Mittelsten Scheid, O., and Becker, C. (2018). Transposons: a blessing curse. Curr. Opin. Plant Biol. *42*, 23–29.

Duyne, G.D.V. (2008). Chapter 12:Site-specific Recombinases. In Protein-Nucleic Acid Interactions, pp. 303–332.

Duyne, G.D.V. (2015). Cre Recombinase. Microbiol. Spectr. *3*.

Dy, R.L., Richter, C., Salmond, G.P.C., and Fineran, P.C. (2014). Remarkable Mechanisms in Microbes to Resist Phage Infections. Annu. Rev. Virol. *1*, 307–331.

Echols, H. (1972). Developmental Pathways for the Temperate Phage: Lysis Vs Lysogeny. Annu. Rev. Genet. *6*, 157–190.

Eilers, B.J., Young, M.J., and Lawrence, C.M. (2012). The Structure of an Archaeal Viral Integrase Reveals an Evolutionarily Conserved Catalytic Core yet Supports a Mechanism of DNA Cleavage in trans. J. Virol. *86*, 8309–8313.

Elbarbary, R.A., Lucas, B.A., and Maquat, L.E. (2016). Retrotransposons as regulators of gene expression. Science *351*, aac7247.

Elkins, J.G., Podar, M., Graham, D.E., Makarova, K.S., Wolf, Y., Randau, L., Hedlund, B.P., Brochier-Armanet, C., Kunin, V., Anderson, I., et al. (2008). A korarchaeal genome reveals insights into the evolution of the Archaea. Proc. Natl. Acad. Sci. *105*, 8102–8107.

Erauso, G., Stedman, K.M., van de Werken, H.J.G., Zillig, W., and van der Oost, J. (2006). Two novel conjugative plasmids from a single strain of Sulfolobus. Microbiology *152*, 1951–1968.

Erdmann, S., Tschitschko, B., Zhong, L., Raftery, M.J., and Cavicchioli, R. (2017). A plasmid from an Antarctic haloarchaeon uses specialized membrane vesicles to disseminate and infect plasmid-free cells. Nat. Microbiol. *2*, 1446.

Erwin, J.A., Marchetto, M.C., and Gage, F.H. (2014). Mobile DNA elements in the generation of diversity and complexity in the brain. Nat. Rev. Neurosci. *15*, 497–506.

Escudero, J.A., Loot, C., Nivina, A., and Mazel, D. (2015). The Integron: Adaptation On Demand. 139–161.

Escudero, J.A., Loot, C., Parissi, V., Nivina, A., Bouchier, C., and Mazel, D. (2016). Unmasking the ancestral activity of integron integrases reveals a smooth evolutionary transition during functional innovation. Nat. Commun. *7*, 10937.

Esposito, D., and Scocca, J.J. (1997). The integrase family of tyrosine recombinases: evolution of a conserved active site domain. Nucleic Acids Res. *25*, 3605–3614.

Farkas, J., Stirrett, K., Lipscomb, G.L., Nixon, W., Scott, R.A., Adams, M.W.W., and Westpheling, J. (2012). Recombinogenic Properties of Pyrococcus furiosus Strain COM1 Enable Rapid Selection of Targeted Mutants. Appl Env. Microbiol *78*, 4669–4676.

Feschotte, C., and Pritham, E.J. (2007). DNA Transposons and the Evolution of Eukaryotic Genomes. Annu. Rev. Genet. *41*, 331–368.

Fiala, G., and Stetter, K.O. (1986). Pyrococcus furiosus sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C. Arch. Microbiol. *145*, 56–61.

Filée, J., Siguier, P., and Chandler, M. (2007). Insertion Sequence Diversity in Archaea. Microbiol. Mol. Biol. Rev. *71*, 121–157.

Flores, G.E., and Reysenbach, A.-L. (2011). Hydrothermal Environments, Marine. In Encyclopedia of Geobiology, J. Reitner, and V. Thiel, eds. (Dordrecht: Springer Netherlands), pp. 456–467.

Forterre, P., Krupovic, M., Raymann, K., and Soler, N. (2014). Plasmids from Euryarchaeota. Microbiol. Spectr. *2*.

Forterre, P., Cunha, V.D., and Catchpole, R. (2017). Plasmid vesicles mimicking virions. Nat. Microbiol. *2*, 1340.

Fröls, S., Gordon, P.M.K., Panlilio, M.A., Schleper, C., and Sensen, C.W. (2007). Elucidating the transcription cycle of the UV-inducible hyperthermophilic archaeal virus SSV1 by DNA microarrays. Virology *365*, 48–59.

Frost, L.S., Leplae, R., Summers, A.O., and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. Nat. Rev. Microbiol. *3*, 722–732.

Fujikane, R., Ishino, S., Ishino, Y., and Forterre, P. (2010). Genetic analysis of DNA repair in the hyperthermophilic archaeon, Thermococcus kodakaraensis. Genes Genet. Syst. *85*, 243–257.

Fukui, T., Atomi, H., Kanai, T., Matsumi, R., Fujiwara, S., and Imanaka, T. (2005). Complete genome sequence of the hyperthermophilic archaeon Thermococcus kodakaraensis KOD1 and comparison with Pyrococcus genomes. Genome Res. *15*, 352–363.

Fusco, S., Liguori, R., Limauro, D., Bartolucci, S., She, Q., and Contursi, P. (2015). Transcriptome analysis of Sulfolobus solfataricus infected with two related fuselloviruses reveals novel insights into the regulation of CRISPR-Cas system. Biochimie *118*, 322–332.

Gandon, S. (2016). Why Be Temperate: Lessons from Bacteriophage λ. Trends Microbiol. *24*, 356–365.

Garcillán-Barcia, M.P., and de la Cruz, F. (2008). Why is entry exclusion an essential feature of conjugative plasmids? Plasmid *60*, 1–18.

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

Geslin, C., Romancer, M.L., Erauso, G., Gaillard, M., Perrot, G., and Prieur, D. (2003). PAV1, the First Virus-Like Particle Isolated from a Hyperthermophilic Euryarchaeote, "Pyrococcus abyssi." J. Bacteriol. *185*, 3888–3894.

Geslin, C., Gaillard, M., Flament, D., Rouault, K., Le Romancer, M., Prieur, D., and Erauso, G. (2007). Analysis of the First Genome of a Hyperthermophilic Marine Virus-Like Particle, PAV1, Isolated from Pyrococcus abyssi. J. Bacteriol. *189*, 4510–4519.

Ghosh, K., Guo, F., and Duyne, G.D.V. (2007). Synapsis of loxP Sites by Cre Recombinase. J. Biol. Chem. *282*, 24004–24016.

Giedraitienė, A., Vitkauskienė, A., Naginienė, R., and Pavilonis, A. (2011). Antibiotic Resistance Mechanisms of Clinically Important Bacteria. Medicina (Mex.) *47*, 19.

Gifford, W.D., Pfaff, S.L., and Macfarlan, T.S. (2013). Transposable elements as genetic regulatory substrates in early development. Trends Cell Biol. *23*, 218–226.

Godfroy, A., Meunier, J.-R., Guezennec, J., Lesongeur, F., Raguénès, G., Rimbault, A., and Barbier, G. (1996). Thermococcus fumicolans sp. nov., a New Hyperthermophilic Archaeon Isolated from a Deep-Sea Hydrothermal Vent in the North Fiji Basin. Int. J. Syst. Evol. Microbiol. *46*, 1113–1119.

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

González, J.M., Sheckells, D., Viebahn, M., Krupatkina, D., Borges, K.M., and Robb, F.T. (1999). Thermococcus waiotapuensis sp. nov., an extremely thermophilic archaeon isolated from a freshwater hot spring. Arch. Microbiol. *172*, 95–101.

Goodman, D.A., and Stedman, K.M. (2018). Comparative genetic and genomic analysis of the novel fusellovirus Sulfolobus spindle-shaped virus 10. Virus Evol. *4*.

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

Gorlas, A., Alain, K., Bienvenu, N., and Geslin, C. (2013). Thermococcusprieurii sp. nov., a hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent. Int. J. Syst. Evol. Microbiol. *63*, 2920–2926.

Gorlas, A., Croce, O., Oberto, J., Gauliard, E., Forterre, P., and Marguet, E. (2014). Thermococcusnautili sp. nov., a hyperthermophilic archaeon isolated from a hydrothermal deep-sea vent. Int. J. Syst. Evol. Microbiol. *64*, 1802–1810.

Grabundzija, I., Messing, S.A., Thomas, J., Cosby, R.L., Bilic, I., Miskey, C., Gogol-Döring, A., Kapitonov, V., Diem, T., Dalda, A., et al. (2016). A *Helitron* transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. Nat. Commun. *7*, 10716.

Grainge, I., and Jayaram, M. (1999). The integrase family of recombinases: organization and function of the active site. Mol. Microbiol. *33*, 449–456.

Grindley, N.D.F., Whiteson, K.L., and Rice, P.A. (2006). Mechanisms of Site-Specific Recombination. Annu. Rev. Biochem. *75*, 567–605.

Guan, Y., Ngugi, D.K., Blom, J., Ali, S., Ferry, J.G., and Stingl, U. (2014). Draft Genome Sequence of an Obligately Methylotrophic Methanogen, Methanococcoides methylutens, Isolated from Marine Sediment. Genome Announc. *2*.

Guédon, G., Libante, V., Coluzzi, C., Payot, S., and Leblond-Bourget, N. (2017). The Obscure World of Integrative and Mobilizable Elements, Highly Widespread Elements that Pirate Bacterial Conjugative Systems. Genes *8*, 337.

Guo, F., Gopaul, D.N., and Duyne, G.D.V. (1997). Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. Nature *389*, 40.

Guy, L., and Ettema, T.J.G. (2011). The archaeal 'TACK' superphylum and the origin of eukaryotes. Trends Microbiol. *19*, 580–587.

Hain, J., Reiter, W.-D., Hüdepohl, U., and Zillig, W. (1992). Elements of an archaeal promoter defined by mutational analysis. Nucleic Acids Res. *20*, 5423–5428.

Hakimi, J.M., and Scocca, J.J. (1996). Purification and characterization of the integrase from the Haemophilus influenzae bacteriophage HP1; identification of a four-stranded intermediate and the order of strand exchange. Mol. Microbiol. *21*, 147–158.

Harms, A., Brodersen, D.E., Mitarai, N., and Gerdes, K. (2018). Toxins, Targets, and Triggers: An Overview of Toxin-Antitoxin Biology. Mol. Cell *70*, 768–784.

Harrison, E., and Brockhurst, M.A. (2012). Plasmid-mediated horizontal gene transfer is a coevolutionary process. Trends Microbiol. *20*, 262–267.

Hawkins, M., Malla, S., Blythe, M.J., Nieduszynski, C.A., and Allers, T. (2013). Accelerated growth in the absence of DNA replication origins. Nature *503*, 544–547.

Held, N.L., and Whitaker, R.J. (2009). Viral biogeography revealed by signatures in Sulfolobus islandicus genomes. Environ. Microbiol. *11*, 457–466.

Henderson, I.R., Owen, P., and Nataro, J.P. (1999). Molecular switches — the ON and OFF of bacterial phase variation. Mol. Microbiol. *33*, 919–932.

Hendrix, R.W. (2002). Bacteriophages: Evolution of the Majority. Theor. Popul. Biol. *61*, 471–480.

Hendrix, R.W., Smith, M.C.M., Burns, R.N., Ford, M.E., and Hatfull, G.F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. Proc. Natl. Acad. Sci. *96*, 2192–2197.

Hensley, S.A., Jung, J.-H., Park, C.-S., and Holden, J.F. (2014). Thermococcus paralvinellae sp. nov. and Thermococcus cleftensis sp. nov. of hyperthermophilic heterotrophs from deep-sea hydrothermal vents. Int. J. Syst. Evol. Microbiol. *64*, 3655–3659.

Heyer, W.-D., Ehmsen, K.T., and Liu, J. (2010). Regulation of Homologous Recombination in Eukaryotes. Annu. Rev. Genet. *44*, 113–139.

Hickman, A.B., and Dyda, F. (2015a). The casposon-encoded Cas1 protein from Aciduliprofundum boonei is a DNA integrase that generates target site duplications. Nucleic Acids Res. *43*, 10576–10587.

Hickman, A.B., and Dyda, F. (2015b). Mechanisms of DNA Transposition. Microbiol. Spectr. *3*.

Hickman, A.B., Chandler, M., and Dyda, F. (2010). Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. Crit. Rev. Biochem. Mol. Biol. *45*, 50–69.

Hileman, T.H., and Santangelo, T.J. (2012). Genetics Techniques for Thermococcus kodakarensis. Front. Microbiol. *3*.

Hille, F., Richter, H., Wong, S.P., Bratovič, M., Ressel, S., and Charpentier, E. (2018). The Biology of CRISPR-Cas: Backward and Forward. Cell *172*, 1239–1259.

Hinnebusch, J., and Tilly, K. (1993). Linear plasmids and chromosomes in bacteria. Mol. Microbiol. *10*, 917–922.

Hoaki, T., Nishijima, M., Kato, M., Adachi, K., Mizobuchi, S., Hanzawa, N., and Maruyama, T. (1994). Growth requirements of hyperthermophilic sulfur-dependent heterotrophic archaea isolated from a shallow submarine geothermal system with reference to their essential amino acids. Appl Env. Microbiol *60*, 2898–2904.

Horvath, P., and Barrangou, R. (2010). CRISPR/Cas, the Immune System of Bacteria and Archaea. Science *327*, 167–170.

Howard-Varona, C., Hargreaves, K.R., Abedon, S.T., and Sullivan, M.B. (2017). Lysogeny in nature: mechanisms, impact and ecology of temperate phages. ISME J. *11*, 1511–1520.

Hülter, N., Ilhan, J., Wein, T., Kadibalban, A.S., Hammerschmidt, K., and Dagan, T. (2017). An evolutionary perspective on plasmid lifestyle modes. Curr. Opin. Microbiol. *38*, 74–80.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860.

Iranzo, J., Koonin, E.V., Prangishvili, D., and Krupovic, M. (2016). Bipartite Network Analysis of the Archaeal Virosphere: Evolutionary Connections between Viruses and Capsidless Mobile Elements. J. Virol. *90*, 11043–11055.

Jamet, A., Touchon, M., Ribeiro-Gonçalves, B., Carriço, J.A., Charbit, A., Nassif, X., Ramirez, M., and Rocha, E.P.C. (2017). A widespread family of polymorphic toxins encoded by temperate phages. BMC Biol. *15*, 75.

Jayaram, M., Ma, C.-H., Kachroo, A.H., Rowley, P.A., Guga, P., Fan, H.-F., and Voziyanov, Y. (2015). An Overview of Tyrosine Site-specific Recombination: From an Flp Perspective. Microbiol. Spectr. *3*.

Jo, C.H., Kim, J., Han, A., Park, S.Y., Hwang, K.Y., and Nam, K.H. (2016). Crystal structure of Thermoplasma acidophilum XerA recombinase shows large C-shape clamp conformation and cis-cleavage mode for nucleophilic tyrosine. FEBS Lett. *590*, 848–856.

Jo, M., Murayama, Y., Tsutsui, Y., and Iwasaki, H. (2017). In vitro site-specific recombination mediated by the tyrosine recombinase XerA of Thermoplasma acidophilum. Genes Cells Devoted Mol. Cell. Mech.

Juhas, M., Meer, V.D., Roelof, J., Gaillard, M., Harding, R.M., Hood, D.W., and Crook, D.W. (2009). Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol. Rev. *33*, 376–393.

Kellner, S., Spang, A., Offre, P., Szöllősi, G.J., Petitjean, C., and Williams, T.A. (2018). Genome size evolution in the Archaea. Emerg. Top. Life Sci. ETLS20180021.

Khan, S.A. (2005). Plasmid rolling-circle replication: highlights of two decades of research. Plasmid *53*, 126–136.

Kim, M.-S., Choi, A.R., Lee, S.H., Jung, H.-C., Bae, S.S., Yang, T.-J., Jeon, J.H., Lim, J.K., Youn, H., Kim, T.W., et al. (2015). A Novel CO-Responsive Transcriptional Regulator and Enhanced H2 Production by an Engineered Thermococcus onnurineus NA1 Strain. Appl Env. Microbiol *81*, 1708–1714.

King, A.M., Lefkowitz, E., Adams, M.J., and Carstens, E.B. (2011). Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses (Elsevier).

Kobryn, K. (2007). The Linear Hairpin Replicons of Borrelia burgdorferi. In Microbial Linear Plasmids, F. Meinhardt, and R. Klassen, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 117–140.

Koonin, E.V., and Wolf, Y.I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. *36*, 6688–6719.

Koonin, E.V., Makarova, K.S., and Wolf, Y.I. (2017). Evolutionary Genomics of Defense Systems in Archaea and Bacteria. Annu. Rev. Microbiol. *71*, 233–261.

Koonin Eugene V. (2016). Viruses and mobile elements as drivers of evolutionary transitions. Philos. Trans. R. Soc. B Biol. Sci. *371*, 20150442.

Krejci, L., Altmannova, V., Spirek, M., and Zhao, X. (2012). Homologous recombination and its regulation. Nucleic Acids Res. *40*, 5795–5818.

Kroll, J., Klinter, S., Schneider, C., Voß, I., and Steinbüchel, A. (2010). Plasmid addiction systems: perspectives and applications in biotechnology. Microb. Biotechnol. *3*, 634–657.

Krupovič, M., and Bamford, D.H. (2008). Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota. Virology *375*, 292–300.

Krupovic, M., Gonnet, M., Hania, W.B., Forterre, P., and Erauso, G. (2013). Insights into Dynamics of Mobile Genetic Elements in Hyperthermophilic Environments from Five New Thermococcus Plasmids. PLoS ONE *8*, e49044.

Krupovic, M., Makarova, K.S., Forterre, P., Prangishvili, D., and Koonin, E.V. (2014a). Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. BMC Biol. *12*, 36.

Krupovic, M., Quemin, E.R.J., Bamford, D.H., Forterre, P., and Prangishvili, D. (2014b). Unification of the Globally Distributed Spindle-Shaped Viruses of the Archaea. J. Virol. *88*, 2354–2358.

Krupovic, M., Béguin, P., and Koonin, E.V. (2017). Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. Curr. Opin. Microbiol. *38*, 36–43.

Krupovic, M., Cvirkaite-Krupovic, V., Iranzo, J., Prangishvili, D., and Koonin, E.V. (2018). Viruses of archaea: Structural, functional, environmental and evolutionary genomics. Virus Res. *244*, 181–193.

Krupovic, M., Makarova, K.S., Wolf, Y.I., Medvedeva, S., Prangishvili, D., Forterre, P., and Koonin, E.V. (2019). Integrated mobile genetic elements in Thaumarchaeota. Environ. Microbiol. *0*.

Kuwabara, T., Minaba, M., Ogi, N., and Kamekura, M. (2007). Thermococcus celericrescens sp. nov., a fast-growing and cell-fusing hyperthermophilic archaeon from a deep-sea hydrothermal vent. Int. J. Syst. Evol. Microbiol. *57*, 437–443.

Labrie, S.J., Samson, J.E., and Moineau, S. (2010). Bacteriophage resistance mechanisms. Nat. Rev. Microbiol. *8*, 317–327.

Lambie, S.C., Kelly, W.J., Leahy, S.C., Li, D., Reilly, K., McAllister, T.A., Valle, E.R., Attwood, G.T., and Altermann, E. (2015). The complete genome sequence of the rumen methanogen Methanosarcina barkeri CM1. Stand. Genomic Sci. *10*, 57.

Landy, A. (2015). The λ Integrase Site-specific Recombination Pathway. Microbiol. Spectr. *3*.

Langille, M.G.I., Hsiao, W.W.L., and Brinkman, F.S.L. (2010). Detecting genomic islands using bioinformatics approaches. Nat. Rev. Microbiol. *8*, 373–382.

Lecompte, O., Ripp, R., Puzos-Barbe, V., Duprat, S., Heilig, R., Dietrich, J., Thierry, J.-C., and Poch, O. (2001). Genome Evolution at the Genus Level: Comparison of Three Complete Genomes of Hyperthermophilic Archaea. Genome Res. *11*, 981–993.

Leigh, J.A., Albers, S.-V., Atomi, H., and Allers, T. (2011). Model organisms for genetics in the domain Archaea: methanogens, halophiles, Thermococcales and Sulfolobales. FEMS Microbiol. Rev. *35*, 577–608.

Lepage, E., Marguet, E., Geslin, C., Matte-Tailliez, O., Zillig, W., Forterre, P., and Tailliez, P. (2004). Molecular Diversity of New Thermococcales Isolates from a Single Area of Hydrothermal Deep-Sea Vents as Revealed by Randomly Amplified Polymorphic DNA Fingerprinting and 16S rRNA Gene Sequence Analysis. Appl. Environ. Microbiol. *70*, 1277–1286.

Letzelter, C., Duguet, M., and Serre, M.-C. (2004). Mutational Analysis of the Archaeal Tyrosine Recombinase SSV1 Integrase Suggests a Mechanism of DNA Cleavage in trans. J. Biol. Chem. *279*, 28936–28944.

Li, Z., Li, X., Xiao, X., and Xu, J. (2016). An Integrative Genomic Island Affects the Adaptations of the Piezophilic Hyperthermophilic Archaeon Pyrococcus yayanosii to High Temperature and High Hydrostatic Pressure. Front. Microbiol. *7*.

Lilly, J., and Camps, M. (2015). Mechanisms of Theta Plasmid Replication. Microbiol. Spectr. *3*.

Lindås, A.-C., and Bernander, R. (2013). The cell cycle of archaea. Nat. Rev. Microbiol. *11*, 627–638.

Liu, D., and Huang, L. (2002). Induction of the Sulfolobus shibatae virus SSV1 DNA replication by mitomycin C. Chin. Sci. Bull. *47*, 923–927.

Liu, Y., Wang, J., Liu, Y., Wang, Y., Zhang, Z., Oksanen, H.M., Bamford, D.H., and Chen, X. (2015). Identification and characterization of SNJ2, the first temperate pleolipovirus integrating into the genome of the SNJ1-lysogenic archaeal strain. Mol. Microbiol. *98*, 1002–1020.

Lopez, P., Philippe, H., Myllykallio, H., and Forterre, P. (1999). Identification of putative chromosomal origins of replication in Archaea. Mol. Microbiol. *32*, 883–886.

López-García, P., Forterre, P., Oost, J. van der, and Erauso, G. (2000). Plasmid pGS5 from the Hyperthermophilic ArchaeonArchaeoglobus profundus Is Negatively Supercoiled. J. Bacteriol. *182*, 4998–5000.

Lossouarn, J., Dupont, S., Gorlas, A., Mercier, C., Bienvenu, N., Marguet, E., Forterre, P., and Geslin, C. (2015). An abyssal mobilome: viruses, plasmids and vesicles from deep-sea hydrothermal vents. Res. Microbiol. *166*, 742–752.

Luo, Y., Leisinger, T., and Wasserfallen, A. (2001). Comparative Sequence Analysis of Plasmids pME2001 and pME2200 of Methanothermobacter marburgensis Strains Marburg and ZH3. Plasmid *45*, 18–30.

Luria, S.E., and Delbrück, M. (1943). Mutations of Bacteria from Virus Sensitivity to Virus Resistance. Genetics *28*, 491–511.

Lwoff, A. (1953). LYSOGENY1. Bacteriol. Rev. *17*, 269–337.

Makarova, K.S., Yutin, N., Bell, S.D., and Koonin, E.V. (2010). Evolution of diverse cell division and vesicle formation systems in Archaea. Nat. Rev. Microbiol. *8*, 731–741.

Marraffini, L.A. (2015). CRISPR-Cas immunity in prokaryotes. Nature *526*, 55–61.

Marraffini, L.A., and Sontheimer, E.J. (2010). CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat. Rev. Genet. *11*, 181–190.

Martin, A., Yeats, S., Janekovic, D., Reiter, W.-D., Aicher, W., and Zillig, W. (1984). SAV 1, a temperate u.v.-inducible DNA virus-like particle from the archaebacterium Sulfolobus acidocaldarius isolate B12. EMBO J. *3*, 2165–2168.

Medema, M.H., Trefzer, A., Kovalchuk, A., van den Berg, M., Müller, U., Heijne, W., Wu, L., Alam, M.T., Ronning, C.M., Nierman, W.C., et al. (2010). The Sequence of a 1.8-Mb Bacterial Linear Plasmid Reveals a Rich Evolutionary Reservoir of Secondary Metabolic Pathways. Genome Biol. Evol. *2*, 212–224.

Mochizuki, T., Yoshida, T., Tanaka, R., Forterre, P., Sako, Y., and Prangishvili, D. (2010). Diversity of viruses of the hyperthermophilic archaeal genus Aeropyrum, and isolation of the Aeropyrum pernix bacilliform virus 1, APBV1, the first representative of the family Clavaviridae. Virology *402*, 347–354.

Mochizuki, T., Sako, Y., and Prangishvili, D. (2011). Provirus Induction in Hyperthermophilic Archaea: Characterization of Aeropyrum pernix Spindle-Shaped Virus 1 and Aeropyrum pernix Ovoid Virus 1. J. Bacteriol. *193*, 5412–5419.

Mojica, F.J.M., Díez-Villaseñor, C., Soria, E., and Juez, G. (2000). Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. Mol. Microbiol. *36*, 244–246.

Montaño, S.P., and Rice, P.A. (2011). Moving DNA around: DNA transposition and retroviral integration. Curr. Opin. Struct. Biol. *21*, 370–378.

Munson-McGee, J.H., Snyder, J.C., and Young, M.J. (2018). Archaeal Viruses from High-Temperature Environments. Genes *9*, 128.

Muskhelishvili, G., Palm, P., and Zillig, W. (1993). SSV1-encoded site-specific recombination system in Sulfolobus shibatae. Mol. Gen. Genet. MGG *237*, 334–342.

Nesmelova, I.V., and Hackett, P.B. (2010). DDE transposases: Structural similarity and diversity. Adv. Drug Deliv. Rev. *62*, 1187–1195.

Ng, W.L., Kothakota, S., and DasSarma, S. (1991). Structure of the gas vesicle plasmid in Halobacterium halobium: inversion isomers, inverted repeats, and insertion sequences. J. Bacteriol. *173*, 1958–1964.

Nunes-Düby, S.E., Kwon, H.J., Tirumalai, R.S., Ellenberger, T., and Landy, A. (1998). Similarities and differences among 105 members of the Int family of site-specific recombinases. Nucleic Acids Res. *26*, 391–406.

Oberto, J., Sloan, S.B., and Weisberg, R.A. (1994). A segment of the phage HK022 chromosome is a mosaic of other lambdoid chromosomes. Nucleic Acids Res. *22*, 354–356.

Offre, P., Spang, A., and Schleper, C. (2013). Archaea in Biogeochemical Cycles. Annu. Rev. Microbiol. *67*, 437–457.

Ofir, G., and Sorek, R. (2018). Contemporary Phage Biology: From Classic Models to New Insights. Cell *172*, 1260–1270.

Oger, P., Sokolova, T.G., Kozhevnikova, D.A., Taranov, E.A., Vannier, P., Lee, H.S., Kwon, K.K., Kang, S.G., Lee, J.-H., Bonch-Osmolovskaya, E.A., et al. (2016). Complete Genome Sequence of the Hyperthermophilic and Piezophilic Archaeon Thermococcus barophilus Ch5, Capable of Growth at the Expense of Hydrogenogenesis from Carbon Monoxide and Formate. Genome Announc. *4*, e01534-15.

Oppenheim, A.B., Kobiler, O., Stavans, J., Court, D.L., and Adhya, S. (2005). Switches in Bacteriophage Lambda Development. Annu. Rev. Genet. *39*, 409–429.

Oren, A. (2014a). The Family Methanocaldococcaceae. In The Prokaryotes: Other Major Lineages of Bacteria and The Archaea, E. Rosenberg, E.F. DeLong, S. Lory, E. Stackebrandt, and F. Thompson, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 201–208.

Oren, A. (2014b). The Family Methanococcaceae. In The Prokaryotes: Other Major Lineages of Bacteria and The Archaea, E. Rosenberg, E.F. DeLong, S. Lory, E. Stackebrandt, and F. Thompson, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 215–224.

Oren, A. (2014c). The Family Methanosarcinaceae. In The Prokaryotes: Other Major Lineages of Bacteria and The Archaea, E. Rosenberg, E.F. DeLong, S. Lory, E. Stackebrandt, and F. Thompson, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 259–281.

Partridge, S.R., Kwong, S.M., Firth, N., and Jensen, S.O. (2018). Mobile Genetic Elements Associated with Antimicrobial Resistance. Clin. Microbiol. Rev. *31*, e00088-17.

Pauly Matthew D., Bautista Maria A., Black Jesse A., and Whitaker Rachel J. (2019). Diversified local CRISPR-Cas immunity to viruses of Sulfolobus islandicus. Philos. Trans. R. Soc. B Biol. Sci. *374*, 20180093.

Peng, X. (2008). 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 *154*, 383–391.

Peng, X., Holz, I., Zillig, W., Garrett, R.A., and She, Q. (2000). Evolution of the family of pRN plasmids and their integrase-mediated insertion into the chromosome of the crenarchaeon Sulfolobus solfataricus11Edited by J. Karn. J. Mol. Biol. *303*, 449–454.

Peters, J.E., and Craig, N.L. (2001). Tn7: smarter than we thought. Nat. Rev. Mol. Cell Biol. *2*, 806–814.

Pfister, P., Wasserfallen, A., Stettler, R., and Leisinger, T. (1998). Molecular analysis of Methanobacterium phage ΨM2. Mol. Microbiol. *30*, 233–244.

Pietilä, M.K., Atanasova, N.S., Manole, V., Liljeroos, L., Butcher, S.J., Oksanen, H.M., and Bamford, D.H. (2012). Virion Architecture Unifies Globally Distributed Pleolipoviruses Infecting Halophilic Archaea. J. Virol. *86*, 5067–5079.

Pinto, U.M., Pappas, K.M., and Winans, S.C. (2012). The ABCs of plasmid replication and segregation. Nat. Rev. Microbiol. *10*, 755–765.

Poulter, R.T.M., and Butler, M.I. (2015). Tyrosine Recombinase Retrotransposons and Transposons. 1271–1291.

Prangishvili, D., Bamford, D.H., Forterre, P., Iranzo, J., Koonin, E.V., and Krupovic, M. (2017). The enigmatic archaeal virosphere. Nat. Rev. Microbiol. *15*, nrmicro.2017.125.

Prieur, D. (1997). Microbiology of deep-sea hydrothermal vents. Trends Biotechnol. *15*, 242–244.

Quemin, E.R.J., Lucas, S., Daum, B., Quax, T.E.F., Kühlbrandt, W., Forterre, P., Albers, S.-V., Prangishvili, D., and Krupovic, M. (2013). First Insights into the Entry Process of Hyperthermophilic Archaeal Viruses. J. Virol. *87*, 13379–13385.

Rajeev, L., Malanowska, K., and Gardner, J.F. (2009). Challenging a Paradigm: the Role of DNA Homology in Tyrosine Recombinase Reactions. Microbiol. Mol. Biol. Rev. MMBR *73*, 300–309.

Rankin, D.J., Rocha, E.P.C., and Brown, S.P. (2011). What traits are carried on mobile genetic elements, and why? Heredity *106*, 1–10.

Raoult, D., and Forterre, P. (2008). Redefining viruses: lessons from Mimivirus. Nat. Rev. Microbiol. *6*, 315–319.

Ravin, N.V. (2011). N15: The linear phage–plasmid. Plasmid *65*, 102–109.

Raymann, K., Brochier-Armanet, C., and Gribaldo, S. (2015). The two-domain tree of life is linked to a new root for the Archaea. Proc. Natl. Acad. Sci. U. S. A. *112*, 6670–6675.

Reiter, W.-D., Palm, P., Yeats, S., and Zillig, W. (1987). Gene expression in archaebacteria: Physical mapping of constitutive and UV-inducible transcripts from the Sulfolobus virus-like particle SSV1. Mol. Gen. Genet. MGG *209*, 270–275.

Reiter, W.-D., Palm, P., and Yeats, S. (1989). Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. Nucleic Acids Res. *17*, 1907–1914.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. Nature *499*, 431–437.

Roberts, A.P., Chandler, M., Courvalin, P., Guédon, G., Mullany, P., Pembroke, T., Rood, J.I., Jeffery Smith, C., Summers, A.O., Tsuda, M., et al. (2008). Revised nomenclature for transposable genetic elements. Plasmid *60*, 167–173.

Rocha, E.P.C. (2008). The Organization of the Bacterial Genome. Annu. Rev. Genet. *42*, 211–233.

Rohwer, F., and Barott, K. (2013). Viral information. Biol. Philos. *28*, 283–297.

Rosario, K., and Breitbart, M. (2011). Exploring the viral world through metagenomics. Curr. Opin. Virol. *1*, 289–297.

Sadowski, P.D. (1995). The Flp Recombinase of th 2-µm Plasmid of Saccharomyces cerevisiae. Prog. Nucleic Acid Res. Mol. Biol. *51*, 53–91.

Salanoubat, M., Genin, S., Artiguenave, F., Gouzy, J., Mangenot, S., Arlat, M., Billault, A., Brottier, P., Camus, J.C., Cattolico, L., et al. (2002). Genome sequence of the plant pathogen Ralstonia solanacearum. Nature *415*, 497.

Salmond, G.P.C., and Fineran, P.C. (2015). A century of the phage: past, present and future. Nat. Rev. Microbiol. *13*, 777–786.

San Filippo, J., Sung, P., and Klein, H. (2008). Mechanism of Eukaryotic Homologous Recombination. Annu. Rev. Biochem. *77*, 229–257.

Schleper, C., Kubo, K., and Zillig, W. (1992). 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. *89*, 7645–7649.

Schleper, C., Jurgens, G., and Jonuscheit, M. (2005). Genomic studies of uncultivated archaea. Nat. Rev. Microbiol. *3*, 479–488.

Schut, G.J., Lipscomb, G.L., Han, Y., Notey, J.S., Kelly, R.M., and Adams, M.M.W. (2014). The Order Thermococcales and the Family Thermococcaceae. In The Prokaryotes, E. Rosenberg, E.F. DeLong, S. Lory, E. Stackebrandt, and F. Thompson, eds. (Springer Berlin Heidelberg), pp. 363–383.

Selb, R., Derntl, C., Klein, R., Alte, B., Hofbauer, C., Kaufmann, M., Beraha, J., Schöner, L., and Witte, A. (2017). The Viral Gene ORF79 Encodes a Repressor Regulating Induction of the Lytic Life Cycle in the Haloalkaliphilic Virus φCh1. J. Virol. *91*, e00206-17.

Senčilo, A., Jacobs-Sera, D., Russell, D.A., Ko, C.-C., Bowman, C.A., Atanasova, N.S., Österlund, E., Oksanen, H.M., Bamford, D.H., Hatfull, G.F., et al. (2013). Snapshot of haloarchaeal tailed virus genomes. RNA Biol. *10*, 803–816.

Serre, M.-C., Letzelter, C., Garel, J.-R., and Duguet, M. (2002). Cleavage Properties of an Archaeal Site-specific Recombinase, the SSV1 Integrase. J. Biol. Chem. *277*, 16758–16767.

Serre, M.-C., El Arnaout, T., Brooks, M.A., Durand, D., Lisboa, J., Lazar, N., Raynal, B., van Tilbeurgh, H., and Quevillon-Cheruel, S. (2013). The Carboxy-Terminal αN Helix of the Archaeal XerA Tyrosine Recombinase Is a Molecular Switch to Control Site-Specific Recombination. PLoS ONE *8*.

Shapiro, J. (1983). Mobile Genetic Elements (Elsevier).

She, Q., Phan, H., Garrett, R.A., Albers, S.-V., Stedman, K.M., and Zillig, W. (1998). Genetic profile of pNOB8 from Sulfolobus: the first conjugative plasmid from an archaeon. Extremophiles *2*, 417–425.

She, Q., Singh, R.K., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M.J., Chan-Weiher, C.C.-Y., Clausen, I.G., Curtis, B.A., Moors, A.D., et al. (2001a). The complete genome of the crenarchaeon Sulfolobus solfataricus P2. Proc. Natl. Acad. Sci. *98*, 7835–7840.

She, Q., Peng, X., Zillig, W., and Garrett, R.A. (2001b). Genome evolution: Gene capture in archaeal chromosomes. Nature *409*, 478–478.

She, Q., Brügger, K., and Chen, L. (2002). Archaeal integrative genetic elements and their impact on genome evolution. Res. Microbiol. *153*, 325–332.

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

Simon, R.D. (1978). Halobacterium strain 5 contains a plasmid which is correlated with the presence of gas vacuoles. Nature *273*, 314.

Sissi, C., and Palumbo, M. (2009). Effects of magnesium and related divalent metal ions in topoisomerase structure and function. Nucleic Acids Res. *37*, 702–711.

Skala, A.M. (2014). Retroviral DNA Transposition: Themes and Variations. Microbiol. Spectr. *2*.

Snyder, J.C., Bolduc, B., and Young, M.J. (2015). 40 Years of archaeal virology: Expanding viral diversity. Virology *479–480*, 369–378.

del Solar, G., Giraldo, R., Ruiz-Echevarría, M.J., Espinosa, M., and Díaz-Orejas, R. (1998). Replication and Control of Circular Bacterial Plasmids. Microbiol. Mol. Biol. Rev. *62*, 434–464.

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

Spaans, S.K., Oost, J. van der, and Kengen, S.W.M. (2015). The chromosome copy number of the hyperthermophilic archaeon Thermococcus kodakarensis KOD1. Extremophiles *19*, 741–750.

Spang, A., Caceres, E.F., and Ettema, T.J.G. (2017). Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. Science *357*, eaaf3883.

Stark, W.M. (2015). The Serine Recombinases. 73–89.

Stedman, K.M., She, Q., Phan, H., Arnold, H.P., Holz, I., Garrett, R.A., and Zillig, W. (2003). Relationships between fuselloviruses infecting the extremely thermophilic archaeon Sulfolobus: SSV1 and SSV2. Res. Microbiol. *154*, 295–302.

Stewart, F.M., and Levin, B.R. (1984). The population biology of bacterial viruses: Why be temperate. Theor. Popul. Biol. *26*, 93–117.

Stieglmeier, M., Klingl, A., Alves, R.J.E., Rittmann, S.K.-M.R., Melcher, M., Leisch, N., and Schleper, C. (2014). Nitrososphaera viennensis gen. nov., sp. nov., an aerobic and mesophilic, ammonia-oxidizing archaeon from soil and a member of the archaeal phylum Thaumarchaeota. Int. J. Syst. Evol. Microbiol. *64*, 2738–2752.

Straub, C.T., Counts, J.A., Nguyen, D.M.N., Wu, C.-H., Zeldes, B.M., Crosby, J.R., Conway, J.M., Otten, J.K., Lipscomb, G.L., Schut, G.J., et al. (2018). Biotechnology of extremely thermophilic archaea. FEMS Microbiol. Rev. *42*, 543–578.

Sun, C., Zhou, M., Li, Y., and Xiang, H. (2006). Molecular Characterization of the Minimal Replicon and the Unidirectional Theta Replication of pSCM201 in Extremely Halophilic Archaea. J. Bacteriol. *188*, 8136–8144.

Sung, P., and Klein, H. (2006). Mechanism of homologous recombination: mediators and helicases take on regulatory functions. Nat. Rev. Mol. Cell Biol. *7*, 739–750.

Swarts, D.C., Hegge, J.W., Hinojo, I., Shiimori, M., Ellis, M.A., Dumrongkulraksa, J., Terns, R.M., Terns, M.P., and van der Oost, J. (2015). Argonaute of the archaeon Pyrococcus furiosus is a DNA-guided nuclease that targets cognate DNA. Nucleic Acids Res. *43*, 5120–5129.

Thiel, A., Michoud, G., Moalic, Y., Flament, D., and Jebbar, M. (2014). Genetic Manipulations of the Hyperthermophilic Piezophilic Archaeon Thermococcus barophilus. Appl Env. Microbiol *80*, 2299–2306.

Torsvik, T., and Dundas, I.D. (1974). Bacteriophage of Halobacterium salinarium. Nature *248*, 680.

Touchon, M., and Rocha, E.P.C. (2016). Coevolution of the Organization and Structure of Prokaryotic Genomes. Cold Spring Harb. Perspect. Biol. *8*, a018168.

Touchon, M., Bernheim, A., and Rocha, E.P. (2016). Genetic and life-history traits associated with the distribution of prophages in bacteria. ISME J. *10*, 2744–2754.

Tumbula, D.L., Bowen, T.L., and Whitman, W.B. (1997). Characterization of pURB500 from the archaeon Methanococcus maripaludis and construction of a shuttle vector. J. Bacteriol. *179*, 2976–2986.

Valen, L.V. (1974). Two modes of evolution. Nature *252*, 298.

Van Duyne, G.D. (2001). A Structural View of Cre-loxP Site-Specific Recombination. Annu. Rev. Biophys. Biomol. Struct. *30*, 87–104.

Van Duyne, G.D. (2005). Lambda Integrase: Armed for Recombination. Curr. Biol. *15*, R658–R660.

Villanueva, L., Damsté, J.S.S., and Schouten, S. (2014). A re-evaluation of the archaeal membrane lipid biosynthetic pathway. Nat. Rev. Microbiol. *12*, 438–448.

Viret, J.F., Bravo, A., and Alonso, J.C. (1991). Recombination-dependent concatemeric plasmid replication. Microbiol. Mol. Biol. Rev. *55*, 675–683.

Voziyanov, Y., Konieczka, J.H., Francis Stewart, A., and Jayaram, M. (2003). Stepwise Manipulation of DNA Specificity in Flp Recombinase: Progressively Adapting Flp to Individual and Combinatorial Mutations in its Target Site. J. Mol. Biol. *326*, 65–76.

Wagner, A., Whitaker, R.J., Krause, D.J., Heilers, J.-H., van Wolferen, M., van der Does, C., and Albers, S.-V. (2017). Mechanisms of gene flow in archaea. Nat. Rev. Microbiol. *15*, 492–501.

Wang, H., Peng, N., Shah, S.A., Huang, L., and She, Q. (2015). Archaeal extrachromosomal genetic elements. Microbiol. Mol. Biol. Rev. MMBR *79*, 117–152.

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

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

Watson, J.D. (2014). Molecular Biology of the Gene (Pearson/CSH Press).

Weitz, J.S., Hartman, H., and Levin, S.A. (2005). Coevolutionary arms races between bacteria and bacteriophage. Proc. Natl. Acad. Sci. *102*, 9535–9540.

White, M.F. (2011). Homologous recombination in the archaea: the means justify the ends. Biochem. Soc. Trans. *39*, 15–19.

White, J.R., Escobar-Paramo, P., Mongodin, E.F., Nelson, K.E., and DiRuggiero, J. (2008). Extensive Genome Rearrangements and Multiple Horizontal Gene Transfers in a Population of Pyrococcus Isolates from Vulcano Island, Italy. Appl. Environ. Microbiol. *74*, 6447–6451.

Wiedenheft, B., Stedman, K., Roberto, F., Willits, D., Gleske, A.-K., Zoeller, L., Snyder, J., Douglas, T., and Young, M. (2004). Comparative Genomic Analysis of Hyperthermophilic Archaeal Fuselloviridae Viruses. J. Virol. *78*, 1954–1961.

Williams, K.P. (2002). Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. Nucleic Acids Res. *30*, 866–875.

Winckler, T., Szafranski, K., and Glöckner, G. (2005). Transfer RNA gene-targeted integration: an adaptation of retrotransposable elements to survive in the compact Dictyostelium discoideum genome. Cytogenet. Genome Res. *110*, 288–298.

Witte, A., Baranyi, U., Klein, R., Sulzner, M., Luo, C., Wanner, G., Kru¨ger, D.H., and Lubitz, W. (1997). Characterization of Natronobacterium magadii phage ΦCh1, a unique archaeal phage containing DNA and RNA. Mol. Microbiol. *23*, 603–616.

Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U. S. A. *74*, 5088–5090.

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. *87*, 4576–4579.

Wolkowicz, R., and Schaechter, M. (2008). What makes a virus a virus? Nat. Rev. Microbiol. *6*, 643.

Wood, A.G., Whitman, W.B., and Konisky, J. (1989). Isolation and characterization of an archaebacterial viruslike particle from Methanococcus voltae A3. J. Bacteriol. *171*, 93–98.

Wozniak, R.A.F., and Waldor, M.K. (2010). Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. Nat. Rev. Microbiol. *8*, 552–563.

Yamashiro, K., Yokobori, S., Oshima, T., and Yamagishi, A. (2006). Structural analysis of the plasmid pTA1 isolated from the thermoacidophilic archaeon Thermoplasma acidophilum. Extremophiles *10*, 327.

Yang, G., Nagel, D.H., Feschotte, C., Hancock, C.N., and Wessler, S.R. (2009). Tuned for Transposition: Molecular Determinants Underlying the Hyperactivity of a Stowaway MITE. Science *325*, 1391–1394.

Yeats, S., McWilliam, P., and Zillig, W. (1982). A plasmid in the archaebacterium Sulfolobus acidocaldarius. EMBO J. *1*, 1035–1038.

Zaremba-Niedzwiedzka, K., Caceres, E.F., Saw, J.H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K.W., Anantharaman, K., Starnawski, P., Kjeldsen, K.U., et al. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature *541*, 353–358.

Zhan, Z., Ouyang, S., Liang, W., Zhang, Z., Liu, Z.-J., and Huang, L. (2012). Structural and functional characterization of the C-terminal catalytic domain of SSV1 integrase. Acta Crystallogr. Sect. D *68*, 659–670.

Zhan, Z., Zhou, J., and Huang, L. (2015). Site-Specific Recombination by SSV2 Integrase: Substrate Requirement and Domain Functions. J. Virol. *89*, 10934–10944.

Zhang, Z., Liu, Y., Wang, S., Yang, D., Cheng, Y., Hu, J., Chen, J., Mei, Y., Shen, P., Bamford, D.H., et al. (2012). Temperate membrane-containing halophilic archaeal virus SNJ1 has a circular dsDNA genome identical to that of plasmid pHH205. Virology *434*, 233–241.

Zhao, W., Zeng, X., and Xiao, X. (2015). Thermococcus eurythermalis sp. nov., a conditional piezophilic, hyperthermophilic archaeon with a wide temperature range for growth, isolated from an oil-immersed chimney in the Guaymas Basin. Int. J. Syst. Evol. Microbiol. *65*, 30–35.

Zhou, M., Liu, Q., Xie, Y., Dong, B., and Chen, X. (2016). Draft genome sequence of Thermococcus sp. EP1, a novel hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent on the East Pacific Rise. Mar. Genomics *26*, 9–11.

Zivanovic, Y., Lopez, P., Philippe, H., and Forterre, P. (2002). Pyrococcus genome comparison evidences chromosome shuffling-driven evolution. Nucleic Acids Res. *30*, 1902–1910.

Zivanovic, Y., Armengaud, J., Lagorce, A., Leplat, C., Guérin, P., Dutertre, M., Anthouard, V., Forterre, P., Wincker, P., and Confalonieri, F. (2009). Genome analysis and genome-wide proteomics of Thermococcus gammatolerans, the most radioresistant organism known amongst the Archaea. Genome Biol. *10*, R70.

Résumé long grand public - Thèse de doctorat - Catherine Badel

Eléments génétiques mobiles et évolution génomique chez les Archées Thermococcales

Tous les êtres vivants possèdent une notice de fonctionnement appelée information génétique sous forme d'ADN linéaire ou circulaire (chromosome). Des portions de cet ADN (séquences codantes aussi appelées gènes) permettent la production de protéines. Les protéines sont les ouvrières de la cellule qui assurent son fonctionnement. Certaines protéines d'un même organisme ou d'organismes différents sont similaires ; elles forment une famille protéique.

Les êtres vivants peuvent être classés dans trois groupes appelés domaines sur la base de leur histoire évolutive. Si l'ensemble des êtres vivants correspondait à une cousinade, les domaines seraient des fratries. Un premier domaine correspond aux Eucaryotes dont les cellules contiennent un noyau. Il comprend notamment les animaux, plantes ou champignons. Un deuxième domaine correspond aux Bactéries, micro-organismes unicellulaires, pathogènes ou non. Le troisième domaine a été découvert il y a 40 ans et correspond aux Archées. Comme les bactéries, les archées sont des micro-organismes unicellulaires dont le chromosome est libre à l'intérieur de la cellule. En revanche, les archées ont un fonctionnement cellulaire proche des eucaryotes, avec notamment des protéines similaires. Les archées sont présentes dans tous les milieux terrestres. On en trouve communément dans les lacs, dans les sols ou sur notre peau. Certaines archées habitent dans des environnements extrêmes comme les sédiments de la mer morte ou les eaux très acides des lacs de Yellowstone. Les archées Thermococcales habitent dans les parois de cheminées hydrothermales à proximité des dorsales océaniques, à plusieurs centaines de mètres de profondeur. Elles y rencontrent des températures supérieures à 80°C, une forte pression et une concentration élevée en métaux lourds.

Des parasites génétiques, appelés éléments génétiques mobiles, utilisent la machinerie d'un hôte cellulaire pour se multiplier et se disperser. Ils peuvent infecter les trois domaines du vivant. Les éléments génétiques mobiles sont soit simplement composés d'un ADN (plasmide) ou contiennent des composants supplémentaires comme pour les virus où une capside sert notamment à protéger et transmettre l'ADN. Les éléments génétiques mobiles emploient différentes stratégies de parasitisme. La plus simple est le parasitisme strict. L'élément génétique mobile est alors un fardeau pour l'hôte qui s'emploie à l'éliminer. Une deuxième stratégie consiste à assurer sa survie en fournissant des fonctions utiles à l'hôte. Enfin, une troisième stratégie de leurre utilise l'intégration de l'ADN de l'élément génétique mobile dans un chromosome de l'hôte. L'élément génétique mobile devient temporairement un morceau de chromosome, jusqu'à son excision où il redevient libre. L'intégration et l'excision sont réalisées par différentes familles de protéines codées par les éléments génétiques mobiles, dont des recombinases à tyrosines. Elles permettent une intégration dite site-spécifique, c'est à dire au niveau d'une séquence déterminée du chromosome.

Les archées sont remarquables par leur grande diversité d'éléments génétiques mobiles. Certains de ces éléments utilisent une famille de recombinases à tyrosine étonnante : le gène de la recombinase est fragmenté lors de l'intégration. La recombinase ne peut alors plus être produite et l'élément génétique mobile ne peut pas s'exciser de manière autonome. Il est bloqué dans le chromosome. L'intégration représente dans ce cas une forme de suicide dont l'utilité était jusqu'à présent incomprise. Au cours de ma thèse, j'ai identifié 62 recombinases à tyrosine suicidaires appartenant à une même famille. Elles sont présentes dans de nombreux chromosomes d'archées, dont les Thermococcales, indiquant leur succès à catalyser des intégrations. Elles sont d'ailleurs utilisées par différents types d'éléments génétiques mobiles. J'ai pu reproduire l'activité d'intégration d'une de ces recombinases suicidaire : la recombinase du plasmide pT26-2 de Thermococcales. Pour cela, j'ai incubé la recombinase purifiée avec son ADN cible spécifique et j'ai mesuré la quantité d'intégrations produites. La recombinase du plasmide pT26-2 est active à des températures proches de l'ébullition de l'eau, représentant un avantage dans les environnements très chauds des Thermococcales. Cette étude d'activité *in vitro* ainsi que la reconstruction de l'histoire évolutive des 62 recombinases a permis de proposer des utilités à l'activité d'intégration suicide. Notamment, le suicide pourrait correspondre à un mécanisme évolutif permettant aux recombinases d'échanger les fragments de gènes créés lors de l'intégration.

Les chromosomes subissent couramment de grandes modifications comme l'incorporation, la perte ou le changement de position de longs morceaux ; on parle alors d'évolution génomique. Notamment, il peut se produire une inversion où un morceau de chromosome change d'orientation. Il est en quelque sorte coupé puis recollé dans l'autre sens. Ces inversions sont très fréquentes chez les archées Thermococcales mais on n'en connaissait jusqu'à présent pas le mécanisme. J'ai démontré que la recombinase à tyrosine du plasmide pTN3 de l'hôte *Thermococcus nautili* est responsable de certaines des inversions. Elle utilise pour cela un mécanisme inédit d'inversion entre séquences non spécifiques. Pour disséquer ce mécanisme, j'ai comparé l'activité de la recombinase à tyrosine du plasmide pTN3 avec deux autres recombinases proches. Deux des intégrases présentent l'activité inédite d'inversion. Les portions fortement similaires entre ces deux recombinases participent probablement à cette activité. La poursuite de la comparaison des trois recombinases et la détermination de la structure tridimensionnelle de la recombinase à tyrosine du plasmide pTN3 nous informeront plus en détail sur les mécanismes mis en jeu.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Eléments génétiques mobiles et évolution génomique chez les archées Thermococcales

Mots clés : intégrase, recombinaison à site spécifique, recombinaison homologue, réarrangement chromosomique, archées hyperthermophiles, élément génétique mobile

Résumé : Les réarrangements permettent une évolution rapide du génome par l'acquisition de séquences codantes exogènes, la perte de fonctions non-essentielles ou la création de nouvelles organisations génomiques. Différents mécanismes de réarrangements impliquant des éléments génétiques mobiles (EGM) ont été identifiés chez les archées, les bactéries et les eucaryotes. En revanche, on ignore l'origine des nombreuses inversions génomiques détectées pour les espèces du genre archéen *Thermococcus*. Mes travaux de thèse visent à améliorer la compréhension de l'évolution génomique chez les Thermococcales à travers l'étude de deux familles d'EGM : les familles de plasmides pTN3 et pT26-2. Plus précisément, je me suis intéressée aux recombinases à tyrosine (ou intégrases) que ces plasmides encodent et qui permettent leur intégration dans le chromosome de l'hôte. J'ai montré que l'intégrase plasmidique Int^{pTN3} est responsable d'inversions dans le chromosome de son hôte *Thermococcus nautili* grâce à une activité catalytique inédite de recombinaison homologue. J'ai par la suite caractérisé deux autres intégrases de Thermococcales reliés phylogénétiquement à Int^{pTN3} dont seulement une présente une activité de recombinaison homologue. La comparaison de leurs séquences primaires et la résolution de la structure de Int^{pTN3} vont maintenant éclairer les déterminants génétiques responsables de la spécificité de site et de l'activité de recombinaison homologue.
Les trois intégrases appartiennent à une classe de recombinases spécifique des archées qui catalyse une intégration suicidaire. Lors de l'intégration, le gène de l'intégrase est fragmenté et probablement désactivé. L'EGM intégré se retrouve piégé dans le chromosome. Les avantages évolutifs d'une telle activité suicidaire restent pour l'instant mystérieux. J'ai identifié 62 intégrases hyperthermophiles suicidaires et reconstruit leur histoire évolutive. Ces intégrases sont très prévalentes et recrutées par différents EGM. De plus, j'ai montré que l'une de ces intégrases présente *in vitro* une activité de recombinaison site-spécifique à des températures proches de l'ébullition de l'eau, représentant un avantage dans les environnements hyperthermophiles.

Title : Mobile genetic elements and genome evolution in the archaea Thermococcales

Keywords : integrase, site-specific recombination, homologous recombination, chromosomal rearrangement, hyperthermophilic archaea, mobile genetic element

Abstract : Genomes rapidly evolve through rearrangements that can generate new genome organizations or lead to the acquisition of foreign coding sequences or the loss of non-essential functions. Several mechanisms of rearrangement were uncovered for Archaea, Bacteria and Eukaryotes that involve mobile genetic elements (MGE). Species from the archaeal genera *Thermococcus* present numerous genomic inversions but none of the previously known inversion drivers. To better understand the genomic evolution of Thermococcales, I investigated two of their MGE families: the pTN3 and pT26-2 plasmid families. Specifically, I focused on the tyrosine recombinases (or integrase) that these plasmids encode and that catalyze their site-specific integration in the host chromosome. I demonstrated that the plasmidic integrase Int^{pTN3} is responsible for chromosomal inversions in the host *Thermococcus nautili* through an unprecedented homologous recombination catalytic activity. I also characterized two other related *Thermococcus* integrases and only one catalyzes homologous recombination. The structure resolution of Int^{pTN3} and primary sequence comparisons will now provide clues about the genetic determinants of site specificity and of the homologous recombination activity.
The three integrases all belong to an archaeal-specific class of integrases that catalyzes a suicidal integration. The integrase gene is partitioned and presumably inactivated upon integration. The integrated MGE is then trapped into the chromosome. The evolutionary benefits of this suicide activity are puzzling. I identified 62 related suicidal hyperthermophilic integrases and reconstructed their evolutionary history. They are highly prevalent and recruited by diverse MGE. I also showed that one of these integrases can catalyze *in vitro* site-specific recombination at near-boiling water temperature, representing an advantage in hyperthermophilic environments.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France