Co-option de systèmes moléculaires complexes de la membrane bactérienne et archéenne

Rémi Denise



SORBONNE UNIVERSITÉ

École doctorale Complexité du Vivant

Institut Pasteur
Génomique Évolutive des Microbes

Co-option de systèmes moléculaires complexes de la membrane bactérienne et archéenne

Par Rémi Denise

Thèse de doctorat en Biologie

Pour obtenir le grade de Docteur de Sorbonne Université

Sous la direction de Eduardo PC Rocha et Sophie Abby

Soutenue publiquement le 04 octobre 2019, devant un jury composé de :

Vincent DAUBIN	Rapporteur	CNRS, Université Lyon 1, Lyon
Romé VOULHOUX	Rapporteur	CNRS, Aix-Marseille Université, Marseille
Alexandra CALTEAU	Examinatrice	CEA, Genoscope, Evry
Laura EME	Examinatrice	Université Paris-Sud, Orsay
Ingrid LAFONTAINE	Examinatrice	IBPC, Sorbonne Université, Paris
Sophie ABBY	Co-encadrante de thèse	CNRS, Université Grenoble Alpes, Grenoble
Eduardo ROCHA	Directeur de thèse	CNRS, Institut Pasteur, Paris

À Claude et Louis Raymond.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABBRÉVIATIONS

μm Micro mètre

AA Acide aminé

Aap Archaeal adhesive pilus

Archaeal-T4P Pilus de type IV chez les archées

Bas Bindosome

Com Competence pilus

ComM Com des monodermes

Cryo-EM Cryogenic electron microscopy

Epd EppA-dependant

HGT Transfert horizontal de gènes

HMM Hidden Markov model

IM Integral membrane

indel Insertion-Délétion

kb Kilo base

MSH Manose-sensitive hemaglutinin

S-layer Couche-S

SDA Secretin-dynamic–associated

T2SS Type II secretion system

T3SS Type III protein secretion system

T4aP Type IVa pilus

T4bP Type IVb pilus

T4P Type IV pilus

T4SS Type IV protein secretion system

T5SS Type V protein secretion system

T6SS Type VI protein secretion system

Tad Tight adherence

TFF Type IV filament

TFF-SF Super-famille des filaments de type IV

Ups UV-inducible pilus in *Sulfolobus*

REMERCIEMENTS

Ce n'est qu'en de rares occasions que l'on peut se retrouver devant une page blanche, libre d'écrire des mots de gratitudes aux personnes que l'on apprécie. Même si ceux qui me connaisse savent que je n'aime pas écrire et que je prends un temps monstrueux à écrire. Je vais cependant faire un effort pour remercier toutes ces personnes qui m'ont accompagnée tout au long de ces trois belles années de thèse, aussi bien professionnellement que personnellement.

Tout d'abord, Je tiens à remercier Eduardo, mon directeur de thèse et Sophie, ma co-encadrante de thèse. Toute cette thèse ne serait pas grand-chose sans votre encadrement, votre disponibilité, votre savoir et votre confiance. Ce fut un réel plaisir de travailler avec vous et j'ai beaucoup appris à votre contact. Eduardo, j'ai vraiment apprécié toutes les relectures, corrections et tout le temps que tu as pris pour toujours t'assurer que ma thèse avançait et que je ne m'enfermais pas dans une boucle infinie d'améliorations minimes. Enfin, merci pour ton calme et ta bienveillance qui m'ont toujours apaisé quand le stress venait à monter. Sophie, merci d'avoir été disponible, souvent dans la minute, même si tu n'étais pas physiquement là (au début étant à Vienne et maintenant à Grenoble). J'ai beaucoup apprécié nos rendez-vous par Skype qui me permettait à chaque fois de remettre de l'ordre dans les analyses que je venais de faire. J'ai également beaucoup appris grâce à toi à phylogénie et d'avoir pris le temps de m'expliquer (à de multiples reprises) comment bien inférer et analyser ces arbres. Merci aussi à tous les deux d'avoir cru en moi dès le départ, quand je ne savais ce qu'était un système de sécrétion.

Je tiens à remercier les membres du jury, qui seront les premiers à lire ces mots. Mesdames Alexandra Calteau, Laura Eme et Ingrid Lafontaine, je vous remercie de votre temps et votre contribution à l'évaluation de mon travail. Messieurs Vincent Daubin et Romé Voulhoux, merci de votre travail minutieux et de votre contribution à ce manuscrit, mais surtout de m'avoir considérée digne de soutenir ma thèse et partager mon travail.

Cette aventure a également été possible grâce aux différents membres de l'unité de Génomique Évolutive des Microbes que j'ai côtoyé au cours de ces 3 années.

Merci à Marie pour les multiples conversations qu'on a pu partager aussi bien sur la science que sur tout autre chose. Merci à Jorge et Jean pour ne jamais m'avoir éjecté de leur bureau quand je venais à l'improviste leur poser des questions dont je trouvais souvent la réponse par moi-même. Merci à Amandine P. pour son aide sur des questions de programmation. Merci à Matthieu pour les moments détente sur la terrasse. Merci à Aude et Camille D. pour les appels à l'aide qu'elles m'ont demandés ce qui m'a permis d'en apprendre plus sur des sujets que je ne connaissais pas. Merci à Fanny pour m'aider dans les choix cornéliens que j'avais sur le design de mes figures. Merci à Brigitte pour toute sa disponibilité. Et merci à tous les autres que je n'ai pas mentionné pour les nombreuses discussions et toute l'aide que vous m'avez apportée.

Je tiens aussi à remercier Hilde de Reuse sans qui je n'aurais jamais postulé dans ce labo et qui m'a conseillé indirectement d'aller faire mon stage de M2 chez Eduardo.

Un grand merci aux membres de mon comité de suivi de thèse : mesdames Olivera Francetic, Simonetta Gribaldo, Ingrid Lafontaine et Allison Williams. Merci pour les discussions intéressantes et de l'intérêt que vous avez apporté sur mes travaux de recherches. Cela m'a beaucoup aidé à me remotiver sur l'intérêt de mes recherches et au fait que ça pouvait intéresser des gens.

Merci nombreuses personnes qui étaient juste en face de mon bureau pendant ma première année de thèse, à savoir le Hub de bio-informatique. Merci de m'avoir à la fois aidé à me détendre et en même temps avoir des discussions souvent utiles pour certaines analyses durant ma thèse, merci Bertrand, Vincent, Alexis, Christophe, Thomas C., Julien et les autres personnes que je ne mentionne pas.

Merci à Martin, Steven et Min pour nos déjeuner le jeudi (plus ou moins hebdomadaire) qui ont été une source à la fois de lâcher prise et de défouloir.

Merci à Marie pour les soirées « je veux bien sortir mais demain il faut que je fasse une manip' tôt donc pas longtemps » qui n'ont jamais été fini à l'heure que prévu. Merci aussi à toutes les personnes que j'ai connues par STAPA et le social time : Borja, Emeline, Mathieu, Justine, Brenna, Alexandre, Alexis, Alexis, Florian, Alicia et tous les autres que je ne mentionne pas.

Merci à Swann et Méthilde d'habiter juste à côté de l'Institut Pasteur pour me sortir rapidement de mon tourment quand je finissais tard le soir et que je cherchais désespérément quelqu'un avec qui partagé mon infortune.

Merci à toutes les personnes que j'ai connues par YRLS et Doc&Co, que j'ai plaisir à revoir en dehors de ces deux occasions et qui m'ont permis de rencontrer encore plus d'autres personnes chouettes.

Merci à mes amis de longue date qui m'ont permis de sortir de tout ce monde de la recherche et qui m'a permis d'accéder à des bonnes bouffées d'air frais quand j'en avait le plus besoin.

Merci à ma famille de toujours m'avoir soutenu durant mes trois ans de thèse. Merci tous particulièrement à ma mère, Agnès, et ma tante, Elsa, pour avoir relue cette thèse et permis de réduire drastiquement le nombre de faute d'orthographe et d'avoir un regard extérieur sur mes phrases qui n'avaient pas beaucoup de sens.

Pour finir, merci à mes grands-parents, Claude et Louis Raymond, qui ont disparu durant ma thèse et qui aurait été ravie et très content de partager ce moment avec moi. C'est pour cela que je leur dédie cette thèse.

PREAMBULE

Le but de cette thèse est d'étudier l'évolution d'une famille de systèmes macromoléculaires présente à la fois dans la membrane des bactéries et dans celle des archées, la super-famille des filaments de type IV.

Les membres de cette super-famille sont connus pour être impliqués dans l'adhésion, la sécrétion de protéines, la motilité par contraction, la motilité flagellaire (chez les archées, sans rapport avec le flagelle bactérien), l'absorption d'ADN, la conduction électrique. Cette super-famille est aussi largement répandue parmi les différent phyla bactériens et archéens. Cependant, l'histoire évolutive de cette famille de systèmes reste peu claire et il n'existe pas d'outil précis qui permet de détecter ces systèmes dans les génomes bactériens et archéens.

Dans une première partie, je vais introduire des concepts clés sur les cellules bactériennes et archéennes pour permettre de comprendre l'environnement globale dans lesquels les systèmes se situent. Je vais ensuite introduire les mécanismes évolutifs qui se sont succédé au cours de l'évolution de cette super-famille. Les membres de cette super-famille se sont répandu parmi les bactéries et les archées par transferts horizontaux et certains membres participent à ces mécanismes, je vais donc détailler un peu plus en détail ces mécanismes. Enfin j'introduirai chaque membre de la super-famille de type IV pour montrer la diversité de systèmes et de fonctions qui existe au sein de cette famille.

Dans une deuxième partie, je vais m'intéresser aux mécanismes de co-option et « bricolage » moléculaire, qui ont permis, par exemple, dans le cas de la super-famille de type IV de permettre la diversification ses membres autour d'un ensemble de briques communes. Cette partie ne se limitera pas seulement au cas de la super-famille de type IV, mais exposera comment ces mécanismes évolutifs ont permis chez les bactéries et archées l'apparition d'une majorité de différents systèmes de sécrétions.

Enfin dans une troisième et dernière partie, je vous présenterai la contribution que j'ai apporté à l'étude de la super-famille des filaments de type IV. Cette étude la super-famille a permis de clarifier l'histoire évolutive de cette dernière. Elle a également permis de mettre en lumière des proches parentés entre des système ar-

chéens et bactériens. Et également a montré grâce à des approches de génomique comparative et d'analyses phylogénétiques comment une poignée de composants clés ont donné lieu à une profusion d'appareils impliqués dans d'importants processus cellulaires bactériens et archéens (motilité, adhésion, sécrétion de protéines et absorption d'ADN).

Ce travail a eu pour but d'avoir une meilleure compréhension de ces systèmes, mais également de fournir un outil pour permettre leur détection. Il offre également un éclairage sur l'évolution des systèmes membranaires et augmente, à son échelle, la compréhension globale de l'évolution bactérienne et archéenne.

Première partie

Introduction

1 Brève introduction sur les bactéries et les archées

1.1 Définition Bactérie et Archée

Les bactéries, les archées et les eucaryotes composent les trois domaines du vivant. Les bactéries et les archées, contrairement aux eucaryotes, ont leur matériel génétique libre dans le cytoplasme et non pas stocké dans le noyau. De ce fait, ces organismes sont appelés des procaryotes. Les procaryotes sont des organismes unicellulaires que l'on peut trouver dans la majorité des environnements sur Terre. Dans les deux paragraphes suivants, je reviendrai brièvement sur ce qu'est une bactérie et ce qu'est une archée ainsi que sur ce qui les différencie. Par la suite, je détaillerai plus particulièrement deux points : 1) la composition de la membrane de ces organismes et en quoi le transport à travers cette membrane est important ; 2) les moyens utilisés par ces organismes pour se mouvoir dans l'environnement et quelles stratégies ils utilisent pour se fixer à leur nouvel environnement.

Bactéries En 1656, Antonie Van Leeuwenhoek a observé, sous un microscope de sa propre conception, pour la première fois, une bactérie [1]. Mais ce n'est qu'en 1828 que le mot « bacterium » fut introduit par Christian Gottfried Ehrenberg [2]. Les bactéries se trouvent dans la majorité des environnements sur Terre. On retrouve dans le corps humain une quantité presque similaire de cellules bactériennes que de cellules humaines (1,3 cellule bactérienne pour une cellule humaine) [3]. Par exemple, on trouve au niveau du microbiote intestinal 10^{13} à 10^{14} bactéries correspondant à plus de 500 espèces différentes [4].

Historiquement, on a divisé les bactéries en deux grands types de bactéries mis en évidence par le protocole du bactériologiste danois Hans Christian Gram en 1884. Ce protocole permet d'observer les propriétés de la paroi bactérienne. Les bactéries, qui répondent positivement à la coloration de Gram (gram +), sont généralement dotées d'une simple paroi avec une grande quantité de peptidoglycane en surface, et celles qui répondent négativement (gram -) sont généralement composées de moins de peptidoglycane mais pourvues d'une membrane externe supplémentaire, je reviendrais dans la section suivante sur les différent types de membranes. Les bactéries peuvent se trouver sous différentes formes et différentes tailles. Elles ont typiquement une taille comprise entre 0,5-5 μm et des formes qui peuvent être une sphère (coque), une ellipse (bacille), on peut même en trouver qui ont des formes de filaments...

Archées En 1977, Woese et Fox et al. ont défini les archées comme un groupe procaryote différent de celui des bactéries en se basant sur des études des séquences génétiques de l'ARN ribosomal 16S [5, 6]. Les analyses phylogénétiques ont ainsi conduit à la proposition d'une nouvelle division de la vie en trois domaines : les bactéries, les archées et les eucaryotes. Certains des organismes que nous connais-

sons aujourd'hui sous le nom d'archées étaient déjà connus et étudiés. Le groupe des archébactéries comprenait les organismes méthanogènes, halophiles et extrémophiles, qui faisaient déjà l'objet d'études [7]. Elles ont été trouvées dans des environnements extrêmes comme les lacs salés et sources d'eau chaude acide, de températures très élevées... Cependant, les progrès récents du séquençage environnemental et des analyses phylogénétiques et métagénomiques ont permis de découvrir que les archées sont omniprésentes. Plus spécifiquement, elles ont été identifiées dans l'océan, les sols, des sédiments, des lacs, des environnements souterrains, en association avec des bactéries et des eucaryotes et dans différentes parties du corps humain, y compris la cavité buccale, l'intestin et le vagin [8].

Plusieurs caractéristiques distinguent l'enveloppe cellulaire des archées de celle des bactéries, notamment la composition lipidique de la membrane cytoplasmique, qui est principalement composée d'éther-lipides, contrairement à la liaison ester que l'on trouve dans les glycérophospholipides plus courants chez les bactéries. De plus, une couche superficielle (le « S-Layer » ou couche-S) est présente dans la partie la plus externe de l'enveloppe cellulaire de la majorité des archées. La couche-S est composée de glycoprotéines et forme une structure poreuse paracristalline. La couche-S des archées représente la surface en contact avec l'environnement extracellulaire, détermine la forme des cellules et sert de barrière protectrice et sélective [9, 10].

1.2 Les types de membranes et le besoin de transport

Toute cellule, peu importe son origine (archée, bactérie ou eucaryote), est délimitée par une membrane lipidique. Cette membrane permet à la cellule d'avoir une forme définie et la protège de l'environnement en servant de barrière sélective qui permet le transport de molécules [11]. Comme expliqué dans la section précédante, on peut ainsi classifier les organismes en se basant sur le nombre de membranes cellulaires que possède la cellule : ainsi les organismes avec une seule membrane sont appelés monodermes (fig. 1b) et ceux avec deux membranes sont appelés didermes (fig. 1a) [12]. Chez les bactéries, la membrane interne, la paroi et la membrane externe sont les composants principaux de l'enveloppe bactérienne.

Chez les didermes, la membrane externe est la couche la plus externe de l'enveloppe cellulaire. Cette membrane contient des phospholipides au niveau de sa couche interne et des glycolipides au niveau de sa couche externe. Ces glycolipides vont avoir tendances à se lier entre eux créant ainsi une barrière pour les molécules hydrophobes. On retrouve également dans cette membrane des protéines telles que des récepteurs et des porines. Parmi ces protéines membranaires, on retrouve deux types de protéines : les protéines transmembranaires, qui vont permettre la diffusion passive des petites molécules hydrophiles, et les lipoprotéines, qui vont servir de liaisons entre la membrane externe et le peptidoglycane [14].

En dessous de la membrane externe, on retrouve la paroi cellulaire qui est présente dans toutes les bactéries [15]. Dans les cellules didermes, elle se compose d'une couche fine de peptidoglycane, et est aussi appelée périplasme servant de

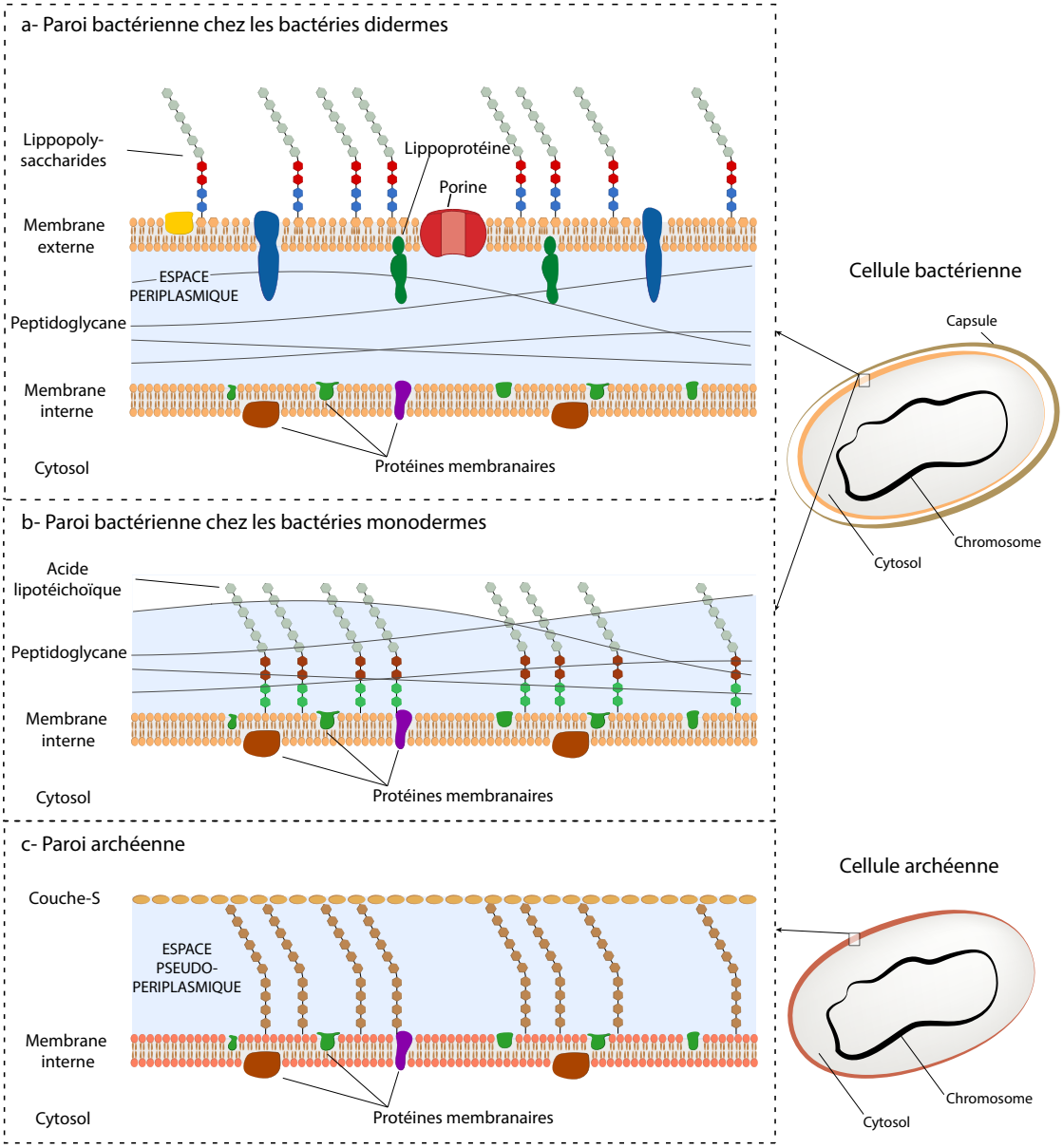


Figure 1 – Paroi des cellules bactériennes et archéennes. (a) Paroi des bactéries didermes. (b) Paroi des bactéries monodermes. (c) Paroi des cellules archéennes. Figure inspirée de [13].

compartiment supplémentaire et de barrière pour le passage d'ions et molécules [11]. Chez les monodermes, elle est présente sous forme d'une couche épaisse de peptidoglycane. Le peptidoglycane sert d'exosquelette à la cellule [16]. C'est lui qui détermine la forme de la cellule.

La membrane interne, ou cytoplasmique, est la couche la plus interne de l'enveloppe cellulaire. Elle contient des protéines qui ont un rôle, par exemple, dans la production d'énergie, dans la sécrétion ou insertion de protéine ainsi que d'internalisation de nutriments. On retrouve dans cette membrane des glycérophospholipides, des phosphatidyléthanolamines et du phosphatidylglycérol [17]. La mem-

brane interne sert de barrière sélective, elle est imperméable aux ions et aux protons, ce qui permet le stockage d'énergie sous forme de gradient de protons et d'ions. Cependant, elle possède des protéines qui permettent le transport de molécules, la translocation de protéines et le transport d'électrons.

La présence de ces structures dans l'enveloppe cellulaire permet une compartimentation et donc un espace qui abrite des machineries transportant des molécules à travers cette enveloppe. Il est à noter que certaines de ces machineries sont nécessaires à l'assemblage de structures permettant la motilité (pili, flagelles) et d'autres à la sécrétion ou absorption de macromolécules (systèmes de sécrétion, transporteurs, pompes à efflux...) [18].

Chez certaines bactéries, il existe des couches protectrices à l'extérieur. Par exemple, la couche-S, qui est une couche monomoléculaire composée d'une répétition de protéines identiques ou encore la capsule, composée généralement de polysaccharides.

1.3 Motilité et adhésion

Pour coloniser un environnement, les bactéries et archées peuvent avoir besoin de deux choses : 1) de pouvoir se déplacer pour aller explorer leur environnement et ainsi trouver un terrain favorable à leur développement ou pour trouver des nutriments (sans forcément aller ensuite s'y installer) 2) de pouvoir se fixer à un endroit propice à leur développement. Je vais donc développer ici des moyens connus chez ces organismes pour se mouvoir (motilité) ainsi que des moyens à leur disposition pour se fixer aux surfaces (adhérences).

1.3.1 La motilité

Motilité par rotation. Ce type de motilité est présent chez les bactéries et les archées et est dû à la rotation d'une structure filamenteuse appelée flagelle. Ce type de motilité permet à la cellule de nager dans les environnements liquides (fig. 2a). Il est à noter que même si le flagelle présent chez les bactéries et chez les archées sont des structures qui produisent le même type de motilité, ils ne sont pas issus de la même famille de système moléculaire et ne sont pas homologues. Ainsi le flagelle bactérien fait partie de la même super-famille que le système de sécrétion de type III (T3SS) [19], tandis que le flagelle archéen (aussi appelé archaellum) est un membre de la super-famille des filaments de type IV, comme le pilus de type IV (T4P) et le système de sécrétion de type II (T2SS) [20]. Dans ce type de motilité, le mouvement est contrôlé par la vitesse à laquelle le flagelle tourne sur lui-même [21]. Par exemple, chez *Halobacterium salinarium* l'archaellum à fréquence de rotation moyenne de 23 $\pm$ 5 Hz pour une vitesse moyenne de nage de 3,3 $\pm$ 0,9 μm par seconde [22]. Les flagelles bactériens peuvent avoir une vitesse allant de 2 μm par seconde (p.ex. chez les *Beggiatoa*) à 200 μm (p. ex. chez les *Vibrio*) [23].

Motilité par contraction. En plus de la motilité passant par l'action du flagelle, les bactéries possèdent un autre type de motilité appelé motilité par contraction (twitching motility, Fig. 2b). Ce type de motilité utilise un filament appelé le pilus de type IV (T4P) et passe par des cycles d'extension et rétraction de ce pilus pour produire un mouvement en deux dimensions [24–28]. L'extension du pilus est induite par la polymérisation de la fibre et va permettre d'adhérer à la surface, la rétraction est induite par la dépolymérisation de la fibre et permet la libération et la propulsion de la cellule. Ce mouvement est souvent mis en action par la coopération de plusieurs pili [29–31]. Il est intéressant de savoir qu'un seul pilus peut générer une force de 100 pN et un amas de pili peut atteindre une force de 1-2 nN pour une vitesse de 1-2 μm par seconde [29, 32, 33]. Une récente étude chez *Pseudomonas aeruginosa* [34] décrit les différentes étapes de ce mouvement. En premier lieu, on va avoir une exploration de l'espace par le pilus, au contact de la surface celui-ci va adhérer et commencer à rétracter le pilus. Enfin la cellule bactérienne va avancer grâce à la rétraction du pilus pour arriver au point d'attachement.

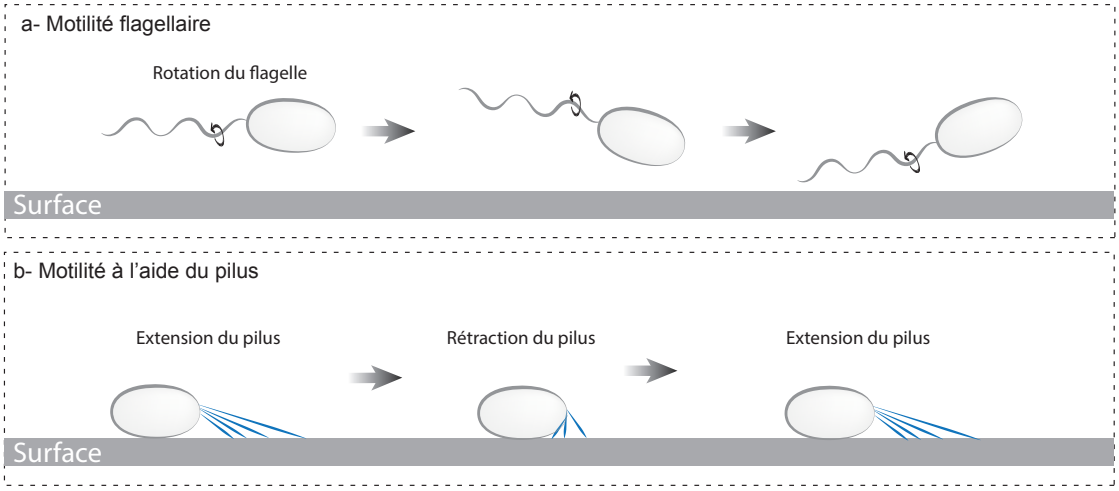


Figure 2 – Mécanismes de motilité des cellules bactériennes et archéennes. (a) Motilité utilisant le flagelle comme moteur de déplacement. (b) Motilité utilisant le pilus de type IV comme moteur de déplacement.

1.3.2 Adhésion

Pour pouvoir coloniser un environnement, il est important que les cellules bactériennes (ou archéennes) se fixent à une surface pour ensuite se reproduire et former des colonies. Ainsi, la faculté de la cellule à atteindre la surface et à se fixer à un point donné est soumise à plusieurs contraintes et met en jeu différents facteurs. Par exemple, la force ionique et le pH sont des facteurs majeurs qui peuvent altérer la capacité d'une bactérie à atteindre et à adhérer à la surface. Il est à noter que l'accessibilité aux nutriments peut aussi nuire ou favoriser l'adhérence des bactéries [35, 36]. Par exemple, pour *Agrobacterium tumefaciens*, les limitations en fer et en manganèse altèrent indépendamment l'adhésion cellulaire [37], alors que

la limitation en phosphate déclenche la fixation cellulaire [38]. En plus des facteurs de l'environnement, on peut retrouver à la surface des cellules des machineries qui vont lui permettre d'adhérer plus facilement aux différentes surfaces. Ainsi, par exemple, les flagelles et les pili, qui en plus d'avoir un rôle dans la motilité, vont dans le cas des flagelles permettre à la cellule d'adhérer de façon réversible à la surface et pour les pili jouer un rôle dans l'adhésion des bactéries didermes [39].

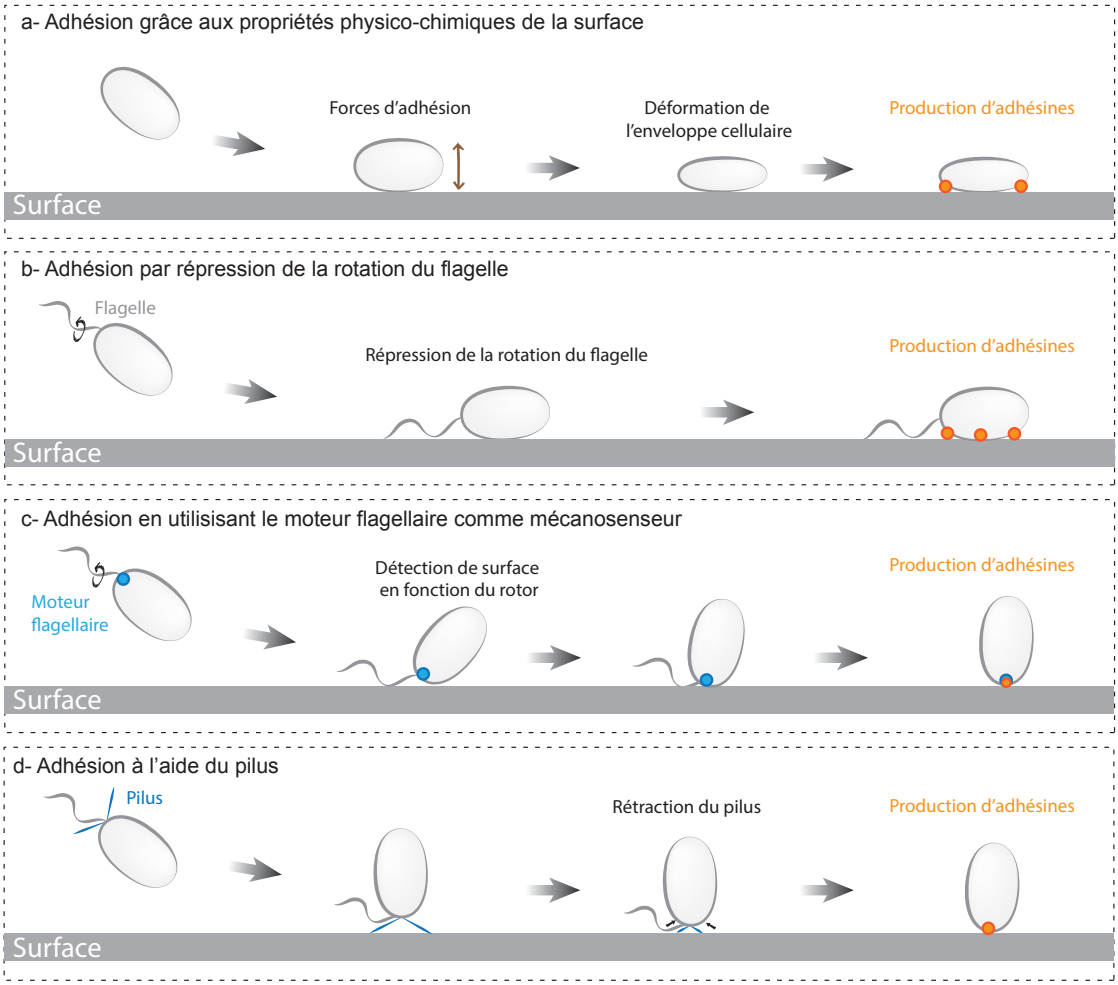


Figure 3 – Mécanismes majeurs d'adhésion aux surfaces. (a) Après un contact avec la surface, la cellule est sujette aux forces d'adhésion venant de la surface (flèche marron), lesquelles vont entraîner la déformation de la cellule. (b) Le contact initial avec la surface permet de diminuer la rotation du flagelle, qui va entraîner un signal de production de gènes liés à l'adhésion. (c) Un contact physique entre la cellule et la surface entraîne un changement de conformation du moteur qui va permettre de stimuler la production d'adhésine. (d) L'intéraction entre le pilus et la surface entraîne la rétraction du pilus qui va permettre la production d'adhésine. Figure adaptée de [40].

Les premières forces physico-chimiques que les bactéries subissent à l'approche d'une surface solide sont : les interactions de Van der Waals, qui sont généralement attractives, les interactions électrostatiques (modulées par le pH et la force ionique de l'environnement) et enfin les interactions hydrophobes et acido-basiques, qui varient d'attractives à répulsives selon le milieu et les bactéries [41]. L'enveloppe cellulaire bactérienne est souvent chargée négativement, ainsi les surfaces

sont chargées positivement ou qui sont neutres sont plus facilement colonisées que celles étant chargées négativement [42]. La même remarque est également valable en ce qui concerne l'hydrophobicité de la surface sur l'adhésion bactérienne : les bactéries dont la surface cellulaire est plus hydrophobe colonisent de préférence les matériaux hydrophobes et vice versa [42]. L'hydrophobicité des bactéries varie selon les espèces bactériennes et est influencée par le milieu de croissance, l'âge des bactéries et la structure de leur surface [42].

En plus des interactions physico-chimiques, on retrouve également à la surface de l'enveloppe cellulaire bactérienne diverses protéines, lipides et polysaccharides ainsi que des structures filamenteuses et non filamenteuses [39]. De nombreuses cellules bactériennes abritent diverses machineries extracellulaires composées de protéines qui ont un rôle direct ou indirect dans l'adhésion. Par exemple, en plus de propulser les bactéries, le flagelle joue un rôle important dans l'adhésion en assurant un contact physique avec la surface mais aussi en fonctionnant comme adhésif [43, 44]. Tandis que les différents pili vont être impliqués dans l'adhésion initiale aux surfaces [39]. Il existe ainsi différents types de pili qui sont impliqués dans les premières étapes de l'adhésion. On peut citer, par exemple, le pilus de type I (T1P) [45], le pilus à adhérence étroite (« tight adherence pilus », Tad) [46, 47] et le pilus de type IVa (T4aP) [48].

Après avoir atteint la surface, les cellules bactériennes, pour renforcer la fixation à la surface, vont adapter certaines interactions entre la cellule et la surface. Ceci se produit grâce au repositionnement du corps cellulaire et des structures de surface, mais également par la production de molécules d'adhésine [40]. Une fois l'adhésion irréversible des cellules obtenue, les cellules peuvent commencer la colonisation de la surface et l'établissement d'un biofilm [40].

2 Mécanismes d'évolution moléculaires

Au sein des bactéries et archées, une multitude de machineries ont évoluées pour répondre aux besoins primordiaux des cellules. Je vais donc dans la cette partie m'intéresser à leurs mécanismes d'évolution. Il existe chez les êtres vivants des mécanismes qui leur permettent d'évoluer, de s'adapter à un changement de leur environnement. Dans cette section nous allons présenter les moyens qu'ont les êtres vivants pour évoluer et les forces évolutives auxquelles ils sont soumis.

2.1 Mutation, sélection et le devenir des gènes

Au cours du temps, des variations peuvent survenir dans le matériel génétique d'un individu, c'est ce qu'on appelle une mutation. Ainsi à chaque génération au niveau d'une population donnée, un individu peut présenter dans son génome un état différent de celui de ses parents, créant un polymorphisme. Par la suite, le devenir de cette mutation reste incertain. Si le mutant initial ne se reproduit pas, la mutation disparaît de la population avec lui. Si la mutation ne disparaît pas

directement, les mutants ont éventuellement un poids sélectif. La fréquence de la mutation dans la population va changer en fonction de cela et de la taille efficace de la population. Si la fréquence diminue dans la population la mutation peut disparaître de celle-ci. *A contrario*, si la fréquence de cette mutation ne disparaît pas et envahit la population, on parle de fixation.

Lors de la comparaison de génomes, on rencontre souvent des changements ponctuels des quatre nucléotides A, T, G ou C, mais aussi des insertions/délétions, des duplications de gènes et des inversions de gènes... Parmi les mutations ponctuelles, il existe différents types qui n'ont pas le même effet en ce qui concerne les processus biologiques qui vont se produire après (transcription, traduction). On peut ainsi avoir des mutations dites synonymes, qui vont modifier la base nucléotidique mais qui ne changeront pas l'acide aminé au niveau de la séquence protéique qui correspond, ou des mutations non synonymes qui vont, elles, modifier la séquence protéique associée et donc altérer la fonction et ou la structure de la protéine (de façon positive ou négative). La modification la séquence peut aussi provoqué une arrivée d'un codon stop prématuré qui peut provoqué, suivant sa position dans la séquence, une protéine plus petite et non fonctionnelle qui sera dégradée.

Associée aux mutations qui permettent l'évolution des populations, il existe une force qui s'applique aux mutations et qui joue un rôle dans sa fixation, c'est la sélection naturelle. Cette force va avoir tendance à ramener la fréquence d'une mutation défavorable vers 0 et d'une mutation favorable vers 1. La probabilité de fixation va être plus importante dans le cas d'une mutation favorable que dans le cas d'une mutation défavorable, et ne va pas avoir de grand effet quand il s'agit d'une mutation neutre. Cependant certaines mutations restent en fréquences intermédiaires, soit parce que l'hétérozygote a plus d'avantage (eucaryotes), soit parce qu'il y a de la sélection qui dépend de la fréquence (p. ex. certains traits ne sont adaptatifs que quand ils sont rares). Certaines mutations peuvent être également fixées par hasard, par exemple, parce qu'elles sont liées génétiquement à une mutation bénéfique (phénomène de « hitchhiking ») [49].

De récentes critiques soutiennent que le biais de mutation et le biais de développement peuvent expliquer l'origine des adaptations indépendamment, ou en plus, de la sélection naturelle [50, 51]. Les mutations ne sont certainement pas réparties au hasard dans le génome, et les différentes régions du génome présentent des taux de mutation différents, soit en raison de contraintes structurelles [52], soit en raison de la sélection pour différents taux de mutation optimaux [53]. Les taux de substitutions peuvent également évoluer dû à des changements dans l'efficacité de la sélection naturelle avec un changement dans la taille efficace de la population [54] ou par une sélection de deuxième ordre lors de l'adaptation à de nouveaux environnements [55]. Enfin, les taux de substitutions sont directement influencés par des facteurs environnementaux externes tels que le rayonnement UV et la température [56, 57].

Les gènes sont aussi soumis à d'autres processus biologiques en plus des mutations ponctuelles et de la sélection. Un processus, appelé la duplication, conduit à

la présence au sein d'un même génome d'une séquence en deux copies, initialement unique. Les deux copies, initialement identiques, peuvent par la suite diverger au cours de l'évolution. L'existence de duplications nombreuses et de longueurs variées est très visible dans les génomes. De ce fait, la plupart des génomes contiennent des familles de gènes homologues. Ces familles multigéniques, qui peuvent comprendre de deux jusqu'à plusieurs centaines de gènes, sont le résultat de processus de duplication de gènes à l'intérieur des génomes, mais aussi de transferts de gènes (que je détaillerai dans ma prochaine sous-section). Les gènes homologues issus de duplication sont appelées paralogues, pour les différencier des orthologues qui sont eux issus d'un évènement de spéciation. Il a été souligné le rôle important des duplications durant l'évolution des êtres vivants, en particulier pour l'acquisition de nouvelles fonctions [58], cependant cette information est à modérer chez les bactéries où la grande majorité des expansions des familles de protéines sont dues à des transferts de protéines [59].

La fusion de gènes, quant à elle, est un mécanisme qui aboutit à la concaténation de deux gènes à côté dans le génome pour en faire un seul. Elle se traduit, d'un point de vue protéique, par une protéines unique possédant les domaines de deux protéines initiale. *A contrario*, la fission constitue le phénomène inverse de la fusion et conduit à la division d'un gène en deux autres gènes. Il est important de noter, que s'il existe la forme fissionnée d'un gène et la forme fusionnée dans deux génomes différents, il est difficile de faire l'hypothèse d'une fusion plutôt que d'une fission. Cependant il a été montré, que les fusions sont en moyenne quatre fois plus fréquentes que les fissions [60]. On peut également rajouter que si une fission se produit sur une protéine qui possède deux domaines identiques, la fission de celle-ci peut mener à deux protéiques qui possèdent une fonction similaire. Les gènes ainsi nouvellement crées sont encodés côte à côte dans le génome, ce qui peut faire penser à une duplication de gènes. Il est donc important de prendre en compte l'histoire générale d'une protéine pour lui attribuer un évènement de duplication ou fission de gène.

2.2 Transfert horizontal de gènes (HGT)

Les bactéries et les archées croissent par division cellulaire. Chaque nouvelle cellule est le clone de la cellule mère. Par cette méthode, la transmission de gènes (entre la cellule mère et la cellule fille) est dit vertical. Il a été montré que les bactéries et les archées sont capables d'acquérir des gènes venant d'autres espèces de bactéries et/ou d'archées et/ou d'eucaryotes par plusieurs moyens dont les plus étudiés sont : la conjugaison [61], la transduction [62] et la transformation [63,64]. Ces mécanismes sont responsables des transferts horizontaux de gènes (HGT) ou transferts latéraux de gènes (LGT) pour les opposer à la transmission verticale de gènes. Les HGTs sont connus pour être un mécanisme permettant de transférer des gènes de résistance aux antibiotiques entre différentes espèces [65]. Il a également été découvert que les transferts horizontaux de gènes n'existent pas seulement au sein des bactéries et des archées mais on aussi entre tous les domaines du vivant.

C'est par exemple le cas de *Agrobacterium tumefaciens*, qui induit la production de tumeur chez la plante en transférant un plasmide dans la plante pour faciliter l'infection bactérienne [66]. On peut aussi citer le cas du transfert d'une synthétase tyrosyl-ARNt d'une haloarchée chez l'ancêtre commun des opisthoconthes qui constitue un caractère dérivé partagé de ce groupe, et donc suggère qu'ils ont un ancêtre commun à l'exception des autres espèces [67]. Ces HGTs peuvent donc se produire entre différents domaines du vivant et ont été découverts assez tôt au court de l'histoire de la biologie moléculaire [63, 68]. Ces mécanismes de transferts de gènes sont importants pour l'évolution des espèces, plus nombreux chez bactéries et archées mais également montré comme cruciaux chez les eucaryotes, permettent l'obtention de nouveaux gènes et fonctions. Je vais détailler ces différents mécanisme de transferts dans la suite de cette section.

2.2.1 Conjugaison

Le mécanisme de conjugaison est un mécanisme qui implique un contact direct entre deux cellules voisines. La machinerie de conjugaison est encodée dans le génome sur des éléments génétiques mobiles, qui vont être intégrés dans le génome de la bactérie. La conjugaison bactérienne comprend deux mécanisme distincts, impliquant le transfert soit d'ADN double brin soit d'ADN simple brin. Un troi-

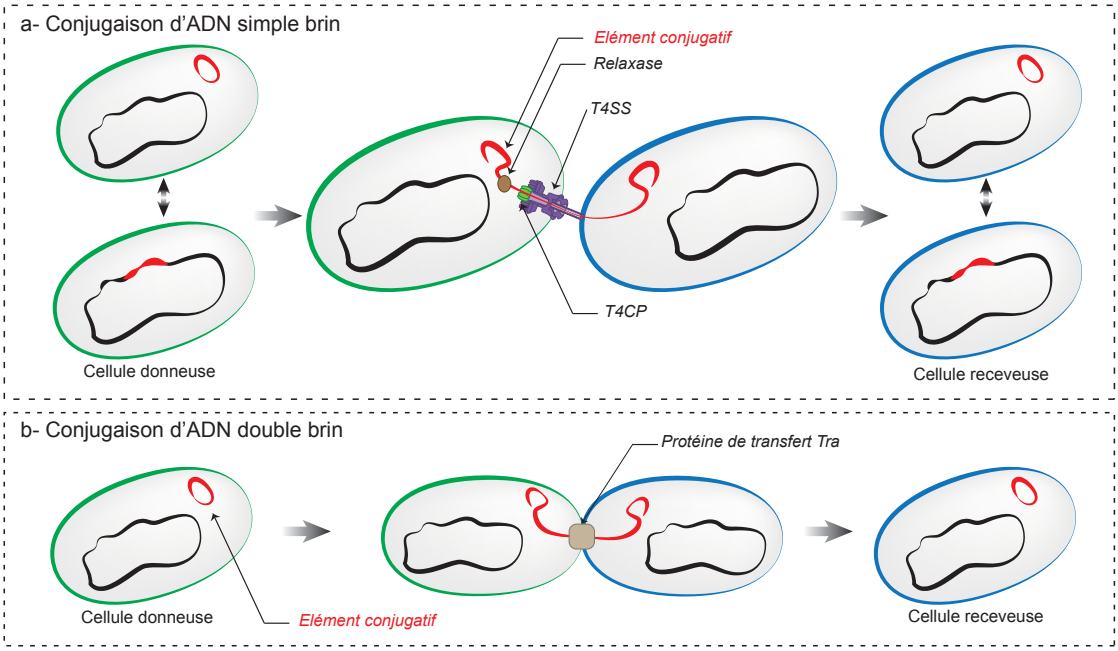


Figure 4 – Mécanismes majeurs de conjugaison. (a) Conjugaison utilisant un ADN simple brin. L'élément conjugatif (rouge), intégré ou non, va être coupé par la relaxase sur un de ses deux brins dans la cellule donneuse (vert) et un des deux brins va s'attacher à la relaxase. Il va ensuite être sécrété au niveau de la cellule receveuse (bleu) en passant par le T4SS et s'intégrer ou non dans le génome de la cellule receveuse. (b) Conjugaison double brin. L'élément conjugatif (rouge), intégré ou non, va être transferé de la cellule donneuse (vert) à la cellule receveuse (bleu) par l'intermédiaire de la protéine de transfert Tra et sera intégré ou non dans le génome de la cellule receveuse. Figure adaptée de [69].

sième mécanisme plus rare, mal compris, spécifique aux *Mycobacterium spp.* est appelé transfert conjugatif distributif [70]. Ce dernier mécanisme transfère des fragments d'ADN de taille variable (50 bp jusqu'à 200 kbp) qui vont être intégrés à différentes positions dans le chromosome [70].

La conjugaison utilisant l'ADN simple brin a été trouvée dans la majorité des bactéries impliquant des éléments conjugatifs [71]. Cette nanomachinerie est composée : du système de sécrétion de type IV (T4SS), d'une protéine de couplage de type IV (T4CP) et d'une relaxase. La relaxase va produire une coupure simple-brin et se lier de façon covalente à ce brin. Ensuite, en utilisant la T4CP, le complexe relaxase-ADN va être sécrété par le T4SS dans l'autre cellule (fig. 4a).

La conjugaison utilisant un double brin d'ADN est, elle, spécifique aux *Actinobacteria*. Cette conjugaison utilise une seule protéine [72] et n'implique pas le T4SS et ne transfère de l'ADN qu'à l'intérieur du mycélium formé par plusieurs actinobactéries (fig. 4b).

La conjugaison utilisant un seul brin d'ADN est un mécanisme de transfert horizontal qui permet de transférer sur une très grande distance phylogénétique, par exemple, entre bactéries et plantes ainsi que bactéries et levures [66].

2.2.2 Transduction

La transduction a été le dernier mécanisme de transfert horizontal découvert [62]. Ce type de mécanisme met en oeuvre des virus bactériens appelés phages (ou bactériophages) ou des virus d'archées. Il existe différents types de phages qui impliquent donc différents types de transduction. Au sein des phages, on va trouver des phages dits virulents, qui se répliquent intensivement dans la cellule et lysent la cellule (cycle lytique) (fig. 5A) [73], et dits tempérés, qui peuvent soit agir comme des phages virulents, soit rester sous forme latente en s'intégrant au chromosome bactérien ou en restant sous une forme plasmidique dans la cellule (cycle lysogénique) (fig. 5a) [73]. Quand le phage est en cycle lysogénique, il est appelé prophage. La transduction est un processus de transfert de gènes d'une bactérie à l'autre par l'intermédiaire d'un phage.

Ce mécanisme peut se trouver sous plusieurs formes et dépend du cycle de vie du phage et du mécanisme d'encapsidation [69]. On a tout d'abord la transduction dite spécialisée (fig. 5b) qui se produit lorsqu'un phage tempéré passe en phase lytique et récupère, au moment de l'excision, un ou plusieurs gènes de son voisinage qui seront transférés au prochain hôte de ce phage [73]. Une autre forme de ce mécanisme est la transduction dite généralisée (fig. 5c), qui se produit durant la phase lytique quand un phage encapsule de l'ADN de son hôte et le transfère dans un autre hôte. Cet évènement est relativement rare, il a été estimé pour le phage P22 qu'à peu près 2% des virions contiennent de l'ADN du chromosome donneur [74]. Cependant, une étude récente [75] a montré que ce taux était plus élevé au sein de communautés naturelles.

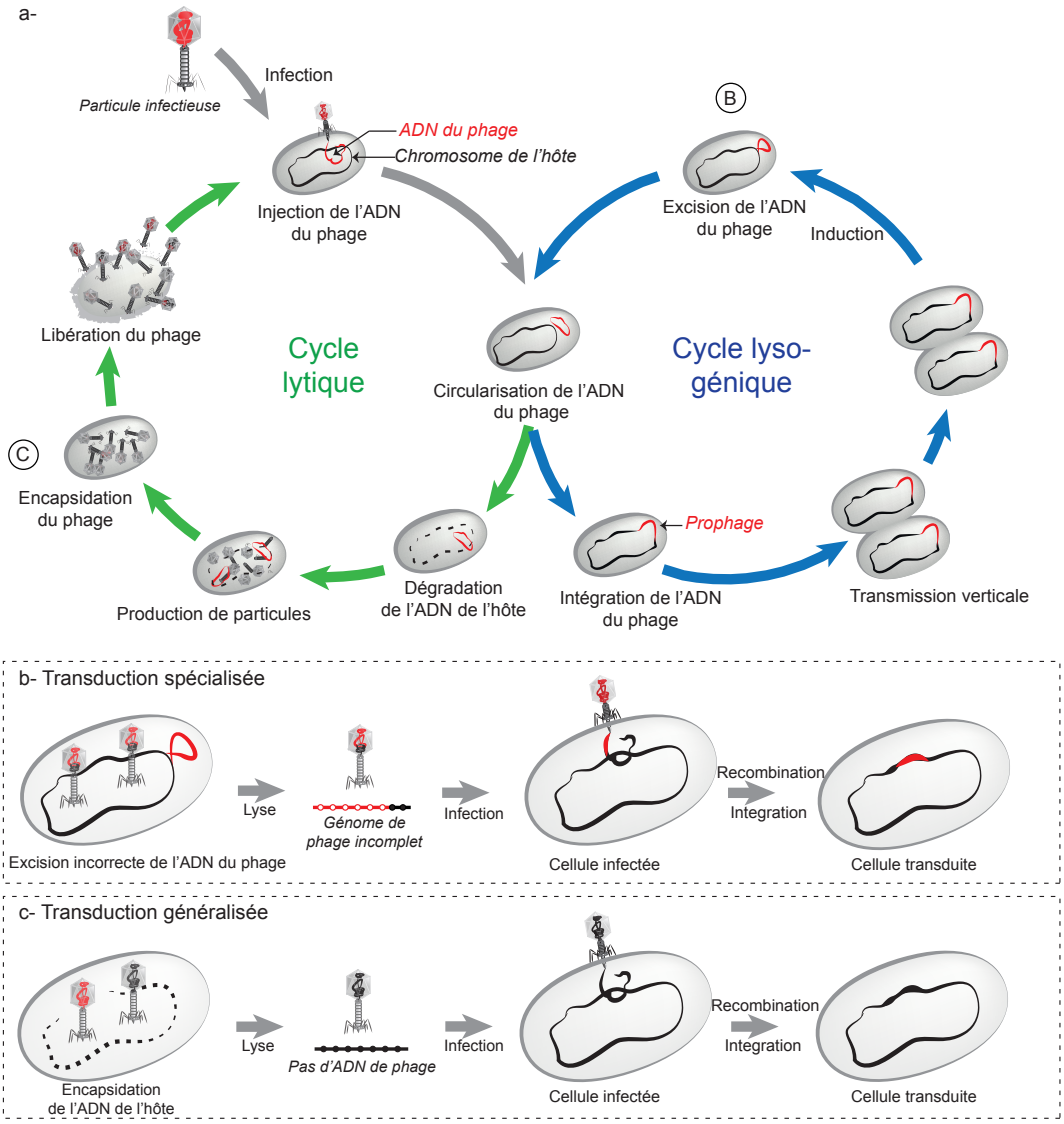


Figure 5 – Mécanismes majeurs d'HGT médiés par les phages. (a) L'infection par un phage peut mener soit à un cycle lytique (en vert) si le phage est virulent ou tempéré, soit à un cycle lysogénique si le phage est tempéré. (b) La transduction spécialisée est due à un défaut d'excision d'un prophage, créant un génome de phage avec de l'ADN de phage (en rouge) et de l'ADN de l'hôte (en noir). (c) La transduction généralisée peut transférer un fragment aléatoire de l'ADN de l'hôte dans des cellules sensibles à ce phage. Cela se produit quand un phage encapsule l'ADN de l'hôte (en noir) à la place de l'ADN du phage (en rouge). Figure adaptée de [69].

2.2.3 Compétence

La compétence, ou transformation naturelle, est le premier mécanisme de transfert horizontal de gènes à avoir été découvert [63]. Ce mécanisme est caractérisé par la faculté à récupérer de l'ADN exogène dans l'environnement. Les gènes codant les systèmes permettant la compétence semblent être présents dans la majorités des espèces bactériennes, seulement environ 80 espèces ont été montrées comme transformables [76]. Les gènes permettant la compétence sont, pour la majorité des bactéries, apparentés à la super-famille de nanomachines comprenant le pilus

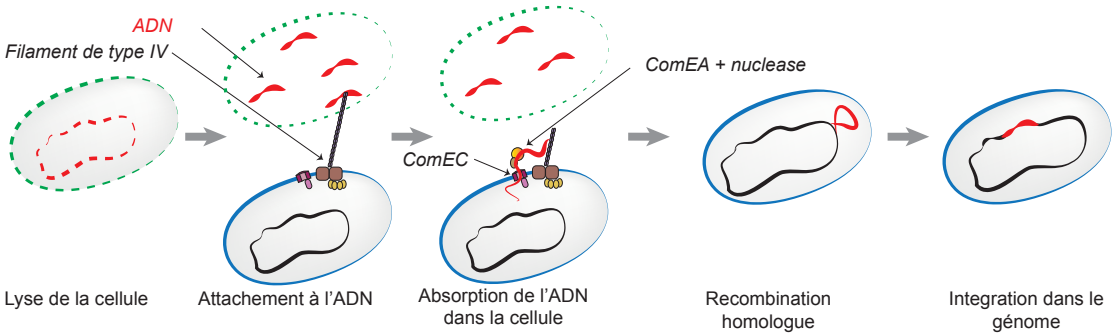


Figure 6 – Mécanisme majeur de compétence. Après lyse de la cellule, l'ADN de la cellule (en rouge) se retrouve libre dans l'environnement. Une cellule compétente (en bleu) va alors pouvoir attacher son pilus de compétence a un fragment d'ADN. La cellule va ensuite rétracter son pilus dans le périplasme pour que le récepteur à l'ADN (ComEA) s'attache à celui-ci, l'ADN sera ensuite clivé par la nucléase et un des deux brins d'ADN va passer dans le cytoplasme de la cellule grâce au pore ComEC. Une fois l'ADN simple brin dans la cellule, on aura une recombinaison homologue qui va permettre à l'ADN simple brin d'être intégré dans le génome de la cellule (en noir). Figure adaptée de [69].

de type IV (T4P) et le système de sécrétion de type II (T2SS) [76], à l'exception des *Helicobacter* et des *Campylobacter* qui utilisent un système de sécrétion de type IV (T4SS) [77].

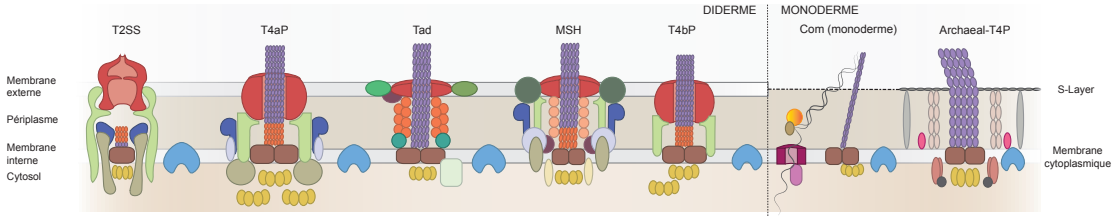
La compétence est divisée en quatre étapes (fig. 6) : 1) Libération de l'ADN dans l'environnement par lyse d'une cellule bactérienne ou par sécrétion de l'ADN par la bactérie [78]. 2) Absorption de l'ADN grâce au pilus de compétence [79]. L'absorption de l'ADN fait référence au transport de l'ADN à travers l'enveloppe bactérienne. 3) La translocation de l'ADN à travers la membrane cytoplasmique [80]. Généralement, ce processus actif commence par la liaison de l'ADN au pilus ; l'absorption se fera probablement par rétraction de pilus de type IV pour tirer l'ADN dans le périplasme en passant par les pores de la sécrétine dans le cas des bactéries didermes [81]. Une fois dans le périplasme, un seul brin de l'ADN est transféré à travers la membrane cytoplasmique tandis que l'autre brin est dégradé. Cette dernière étape menant à une transformation naturelle nécessite un mécanisme d'absorption de l'ADN qui est composé d'un récepteur membranaire de liaison à l'ADN (ComEA), d'une protéine formant un canal à la membrane cytoplasmique (ComEC) et d'une ATPase (ComFA) [80]. Pour finir, l'ADN simple brin entre dans la cellule et est éventuellement intégré au génome de la bactérie par recombinaison homologue.

3 La super-famille des filaments de type IV (TFF-SF)

Durant ma thèse, je me suis intéressé à la super-famille des systèmes bactériens et archéens, qui comprend : le système de sécrétion de protéines de type II (T2SS), le pilus de type IVa (T4aP), le pilus de type IVb (T4bP), le « mannose-sensitive hemagglutinin » pilus (MSH), le pilus à adhésion étroite (Tad), l'appareil

de compétence (Com) et le pilus de type IV chez les archées (Archaeal-T4P). Ces systèmes ont des composants homologues centraux, parfois en plusieurs copies, et présentent des similitudes en termes d'architecture macromoléculaire [20, 82, 83]. L'organisation de ces systèmes et de leurs fonctions sera détaillée système par système dans cette section.

Les systèmes de la super-famille des filaments de type IV (TFF-SF) sont impliqués dans des fonctions typiquement associées aux pili extracellulaires chez les procaryotes, y compris la fixation cellulaire et la formation de biofilms, et sont exploités par les phages pour l'infection cellulaire [84–86]. Les T4aP, T2SS, T4bP et Tad sont également des facteurs de virulence importants des bactéries pathogènes [87–92]. Néanmoins, et malgré leur homologie, les différentes familles de la TFF-SF ont développé des fonctions biochimiques spécifiques. Les T4aP et T4bP permettent aux bactéries de se déplacer (grâce à la motilité par contraction, un mouvement causé par des cycles répétés d'extension-rétraction du pilus) [25, 93]. Certains TFFs sont impliqués dans l'adhérence efficace aux surfaces abiotiques facilitant la formation de biofilms [94]. Le T2SS sécrète les protéines du périplasme à travers la membrane externe [95]. Certains T4aP, Com et Archaeal-T4Ps facilitent l'absorption de l'ADN de l'espace extracellulaire dans la cellule [76, 96]. Dans les bactéries, ces systèmes sont de loin les appendices les plus fréquemment impliqués dans la transformation naturelle. Les Archaeal-T4Ps comprennent le flagelle archéen impliqué dans la motilité par rotation de l'appendice (Archaellum), le pilus impliqué dans l'absorption du sucre (bindosome ou Bas), le pilus Ups impliqué dans l'établissement de contacts cellule-cellule pour permettre la réparation de l'ADN, et plusieurs pili aux fonctions encore mal définies [97–99].



Function	Protéine SDA	Sécretine	ATPase		IM platforme	Piline majeure	Piline mineure	Protéine d'assemblage						Prepiline peptidase	Pilotine	Fonctions diverses								
T2SS	GspC	GspD	GspE		GspF	GspG	GspH-I-J-K	GspL	GspM					GspO	GspS									
T4P	PilP	PilQ	PilB-T-U		PilC	PilA	PilE-V-W-X FimT	PilM	PilO	PilN				PilD										
Tad		RcpA	TadA	TadZ	TadB-C	Flp	TadE-F					TadG	RcpB	TadV	TadD	RcpC								
Com monoderme			ComGA		ComGB	ComGC								ComC		ComEA	EndA	ComEC	ComFA					
Archaeal-T4P			FlaI	FlaH	FlaJ	FlaB								PibD/FlaK						FlaF	FlaG	FlaX		
MSH	MshK	MshL	MshE	MshM	MshG	MshA	MshB-C-D-O	MshI2	MshJ	MshI	MshQ				MshN									MshF-P
T4bP	PilP	PilQ	PilB-T			PilA	PilE			PilN				PilD										

Figure 7 – Schéma de la super-famille des filaments de type IV. Les composants homologues sont représentés avec la même couleur. La table en légende indique le code couleur et le nom des composants dans les différents systèmes. SDA est l'initiale de « secretin-dynamic associated » et IM de « internal membrane ».

3.1 T4aP : Type IVa pilus

Les pili de type IV (T4P) sont des structures filamenteuses minces, longues, flexibles et assemblées à la surface de nombreuses espèces bactériennes. Les T4P

sont impliqués dans de nombreuses fonctions telles que la motilité cellulaire, l'adhérence aux surfaces, la prédation, la formation de biofilms, l'absorption de l'ADN et la sécrétion de ses protéines extracellulaires [28,32,100,101]. L'adhérence peut être assurée par les propriétés adhésives de la piline majeure ou par certaines pilines mineures, comme PilV dans *Neisseria meningitidis* et *Neisseria gonorrhoeae* [83,102] ; ou par des protéines associées au pilus, comme l'adhésine PilY1 de *Pseudomonas aeruginosa* [103].

Le phénomène d'aggrégation/agglutination des bactéries voisines, favorisé par les T4P, permet la formation des microcolonies par une interaction pilus-pilus produite à l'aide des sous-unités de pilines mineures, telle que PilX chez *Neisseria meningitidis* [104]. La formation de microcolonies groupe les cellules pour concentrer les toxines sécrétées (p. ex. *Vibrio cholerae*), cette agglomération est également utile comme protection contre la réponse immunitaire de l'hôte. La formation de microcolonies par les T4P est connue pour participer à la formation du biofilm, où l'attachement efficace à une surface et l'aggrégation cellulaire sont essentiels [105–107].

La pathogénicité liée à la motilité par contraction a été caractérisée chez les espèces de *Neisseria*, en particulier chez *Neisseria gonorrhaeae*, qui est capable d'activer certaines des cascades de signalisation de l'hôte. La motilité par contraction se produit dans les surfaces semi-solides comme l'épithélium muqueux et in vitro sur les surfaces humides à viscosité modérée [48]. Outre la motilité, les contractions sont également nécessaires pour la colonisation de l'hôte, et la rétraction est requise chez certaines espèces pour l'absorption de l'ADN pendant la compétence [108].

Les T4Ps se composent de différents groupes phylogénétiquement distincts et sont considérés comme des facteurs clés de virulence de nombreux agents pathogènes de végétaux, d'animaux et d'humains tels que *Francisella tularensis*, *Neisseria gonorrhaeae*, *Oichelabacter nodosus* [101,109,110]. Ainsi, on retrouve sous le terme de T4P plusieurs groupes phylogénétiquement distincts : les T4aP, les T4bP et les Tad (ou T4cP). Sous cette sous-section je vais m'intéresser exclusivement au T4aP.

Grâce aux dernières données de cryo-EM, on sait que le T4aP a une architecture globale qui est divisée en quatre parties principales [100] (fig. 8), dont les gènes sont généralement localisés en plusieurs groupes dans différents loci sur le génome [112] :

— *Le complexe de la membrane externe* (absent chez les bactéries monodermes) : comprenant la sécrétine, PilQ, une protéine associée à la sécrétine PilP et dans quelques cas une lipoprotéine, la pilotine PilF. La sécrétine est un multimère de sous-unités PilQ qui forme un canal dans la membrane externe pour la sortie du pilus. La lipoprotéine PilF est impliquée dans l'oligomérisation de PilQ et aide à une localisation correcte de PilQ dans la membrane externe [113]. Par ailleurs, PilP semble être nécessaire dans la croissance des fibres, cependant cette protéine n'est pas présente chez les monodermes et donc son rôle ne semble pas avéré [114, 115]. PilP peut également être un constituant spécifique aux didermes servant d'ancre macromoléculaire néces-

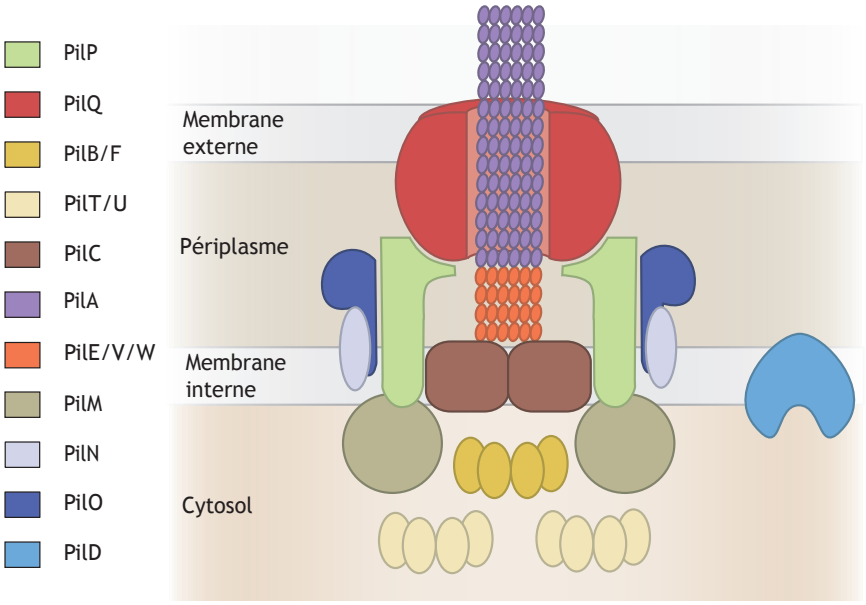


Figure 8 – Schéma détaillé du pilus de type IVa. Figure inspirée de [100, 111].

saire pour promouvoir un assemblage efficace du pilus [100].

— *La plateforme d'assemblage* : localisée dans la membrane cytoplasmique de la bactérie. Elle comprend les protéines PilC, PilM, PilN et PilO. PilC est une protéine polytopique qui agit comme une plateforme durant la formation du pilus [116, 117]. Elle est essentielle pour la polymérisation du pilus et la motilité par contraction [118]. De plus, les domaines cytoplasmiques de PilC semblent interagir avec les ATPases PilB et PilT [48,117]. Les protéines PilM, PilN et PilO forment un complexe associé à la membrane cytoplasmique par l'interaction de PilM et PilN [119]. PilM, PilN et PilO sont structurés en forme de « cage » dans laquelle PilC est située au centre [100]. Il a été proposé que PilC joue un rôle dans le mouvement de rotation du pilus pendant l'assemblage [100, 117], comme chez le T2SS [120]. Le complexe PilM-PilN-PilO joue un rôle de maintien de la machinerie pour aligner la plateforme d'assemblage avec la sécrétine, par interaction avec PilP, pour assurer la sortie du pilus au moment de l'assemblage [100, 111, 119, 121].

— *Le complexe moteur* : formé soit par l'ATPase d'extension, PilB, ou par l'ATPase de rétraction, PilT et/ou PilU [122]. PilB, PilT et PilU ont une localisation cytoplasmique ; PilB est impliquée dans la polymérisation du pilus tandis que PilT est impliquée dans la rétraction du pilus par dépolymérisation [123]. Des analyses structurales [119] révèlent une rotation dans le sens horaire de la cavité centrale de PilB qui peut se traduire par la rotation de PilC et du pilus. La force générée par un seul filament rétractable (100 pN) [124] permet à la cellule de déplacer 10 000 fois sa propre masse [114], faisant du T4aP le moteur linéaire le plus puissant rapporté à ce jour. La

tension exercée sur la fibre peut affecter de façon réversible la morphologie du pilus, comme dans le T4aP de *Neisseria gonorrhoeae* [30]. Les données structurelles suggèrent que la rétraction de l'ATPase PilT fonctionne par rotation inverse par rapport à l'ATPase PilB pour favoriser le démontage de la fibre [125].

— *Le filament* : fibre de largeur comprise entre 60 et 80 acides aminés (AA) et longue de plusieurs μm. Elle est formée par la piline majeure PilA et les pilines mineures FimT, FimU, PilV, PilW, PilX, PilE. L'adhésine PilY1 est nécessaire à la formation des filaments et semble faire partie du même complexe, cependant sa localisation n'est pas claire [126]. Le filament (ou pilus) est composé de milliers d'exemplaires de la sous-unité principale (la piline majeure) et de plusieurs pilines mineures en faible abondance. Les pilines sont de petites protéines d'un poids moléculaire inférieur à 20 kDa ; elles sont synthétisées sous forme de préprotéines, clivées par la prépiline peptidase PilD [127]. Ce processus est essentiel pour la poursuite de l'assemblage du pilus. Les pilines ont une structure « de sucette » (tête globulaire suivie d'une queue). Les têtes globulaires fournissent au T4aP leurs propriétés fonctionnelles, et les interactions du domaine de la tête déterminent la spécificité de l'assemblage et contribuent à la stabilité des fibres [128]. Les glucides de surface associés au pilus peuvent interférer avec la reconnaissance des anticorps anti-T4aP en imitant les antigènes de l'hôte. Des études du T4aP chez *Geobacter sulfurreducens*, composé d'une petite piline majeure (66 AA), ont montré des propriétés remarquables de ce pilus, comme la capacité d'agir comme des nanofils électriquement conducteurs [129, 130]. L'adhésine PilY1, associée au pilus chez *Pseudomonas aeruginosa*, se lie à l'intégrine jouant un rôle dans l'adhérence. De plus, il a récemment été démontré que PilY1 fonctionne comme un mécanosenseur induisant la virulence lors la fixation à l'hôte [103, 126, 131]. Chez *Neisseria gonorrhaeae*, PilY1 est située à l'extrémité de la fibre et participe à la reconnaissance d'un récepteur glycoprotéique de cellules humaines. Les pilines mineures, forment un complexe d'initiation qui amorce l'assemblage du pilus [126]. Certaines d'entre elles (PilV, PilW et PilX) interagissent et participent à la présentation en surface de l'adhésine PilY1 [132, 133]. Les pilines mineures sont nécessaires à une piliation efficace et sont utiles dans des fonctions diverses comme l'aggrégation ou la motilité [134–136]. Les pilines mineures ne sont pas localisées aux extrémités des fibres, mais semblent plutôt être distribuées dans tout le pilus [135].

3.2 T4bP : Type IVb pilus

Au sein des systèmes dénommés T4P, on retrouve un groupe phylogénétiquement distinct, le groupe des T4bP [137]. Les T4bP ont des pilines plus grandes (180-208 AA) que celles des T4aP (150-160 AA) [92]. Les gènes des T4bPs encodent un plus petit ensemble de protéines codantes (10-12), qui sont toujours majoritairement regroupées dans un seul opéron et peuvent se situer dans des plasmides ou des

éléments mobiles [138]. T4bP sont communément trouvés dans des pathogènes (p. ex. le pilus formant des faisceaux (Bfp) d'*Escherichia coli* entéropathogène [139] et le pilus corégulé par des toxines (Tcp) de *Vibrio cholerae* [140]. Les T4bP vont avoir une architecture similaire à celle des T4aP (fig. 9), ils possèdent un filament composé d'une sous-unité de piline majeure, PilA ainsi qu'une ou plusieurs pilines mineure, PilE ; une prépiline peptidase PilD ; un complexe moteur avec une ATPase, PilB, dans quelques cas une deuxième ATPase, PilT ; une plateforme d'assemblage composée d'une protéine membranaire intégrale cytoplasmique, PilC, qui est liée à deux protéines qui forment un complexe cyclique autour de PilC, PilM et PilN, mais qui n'ont pas de similarité de séquences avec leur homologues chez T4aP ; et un complexe de la membrane extérieure composé d'une sécrétine, PilQ et une protéine associée à la sécrétine, PilP [141].

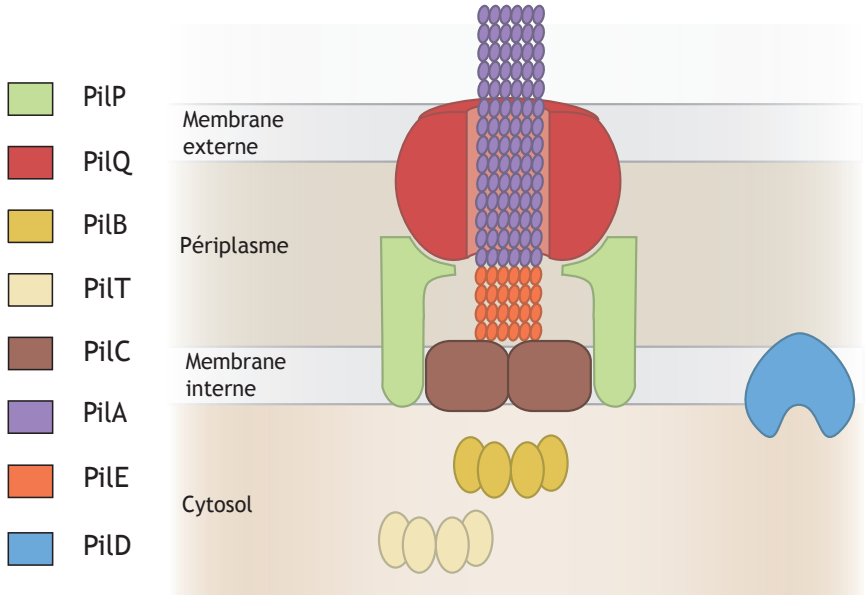


Figure 9 – Schéma du détaillé du pilus de type IVb. Figure inspirée de [141].

Parmi les T4bP décrit il existe :

Lng. Le pilus long (« longus pilus », Lng) est présent chez *Echerichia coli* entérotoxigène. Ce pilus est exclusivement exprimé à la surface de la bactérie si la température de culture est à 37 °C et à un pH de 7,5. Cette expression du pilus Lng suggère qu'il doit être présent dans les différents segments de l'appareil gastro-intestinal en réponse au changement interne de pH [142], mais on en sait encore peu sur ce pilus.

Bfp. Le pilus formant des faisceaux (« bundle forming pilus », Bfp) est trouvé chez *Echerichia coli* entéropathogène (EPEC). Les gènes du pilus Bfp sont en-

codés sur un plasmide appelé EAF (EPEC adherence factor) [143]. Le pilus Bfp possède des sous-unités qui lui permettent de reconnaître les récepteurs N-acétylgalactosamine présents dans les cellules humaines intestinales lors d'une déficience de microvillosités [143,144]. Il a une fonction d'adhérence qui peut être médiée par les propriétés adhésives de sa piline majeure [144].

Tcp. Le pilus corrégulé par des toxines (toxin-coregulated pilus, Tcp), présent chez *Vibrio cholerae*, ne possède pas d'ATPase de rétraction (homologue de PilT) et est donc considéré comme incapable d'effectuer la rétraction du pilus et des fonctions associées à la rétraction [145,146]. Le Tcp favorise l'aggrégation des bactéries voisines pour former des microcolonies. Cela peut être dû à la piline majeure et est utile lors de colonisation de l'hôte [146]. Le Tcp est un récepteur connu du bactériophage CTX qui porte les gènes qui encode une toxine cholérique [141]. Le Tcp est capable de sécréter des protéines solubles, notamment le facteur de colonisation TcpF [147]. Son architecture a été révélé récemment par cryotomographie électronique [141].

R64. Le pilus R64 est un pilus fin dont les gènes ont été découverts sur le plasmide conjugatif R64 chez *Salmonella enterica* [138]. Il est requis pour l'appariement des cellules en milieu liquide et pourrait jouer un rôle lors de la conjugaison [148], cependant peu de choses sont connues sur ce pilus.

Cof. Le pilus à facteur de colonisation (colonization factor, Cof) a été découvert chez les *Echerichia coli* entérotoxigéniques (ETEC). Le domaine C-terminal du pilus adopte un repliement spécifique qui suggère que la piline majeure du pilus Cof permet l'attachement et la colonisation de l'épithélium intestinal de l'hôte [149]. De plus, il a été montré que ce pilus était capable de sécréter une protéine, CofJ, à l'extrémité de son pilus qui permet l'ancrage lors de l'adhésion et qui augmente la pathogenèse [149].

3.3 Tad (T4cP) : « tight adherence pilus »

Le pilus à adhérence étroite (« tight adherence pilus », Tad) a d'abord identifié chez les *Aggregatibacter actinomycetemcomitans*. Il représente un sous-groupe distinct chez les T4P et sont aussi appelés T4cP [46, 150, 151]. Ce pilus est présent chez plusieurs bactéries didermes et monodermes [152] et est connu pour participer à l'adhérence sur différentes surfaces, la formation de biofilms extrêmement tenaces en milieux liquides, la pathogénèse et la transformation naturelle [28, 83, 150, 151, 153–156]. Il est également connu comme un récepteur au bactériophage CbK [157]. Cette machinerie est généralement composée de douze gènes encodés en un seul cluster.

Le Tad a, comme les autres systèmes homologues de la TFF-SF, une architecture globale divisée en quatre parties principales [154] (fig. 10) :

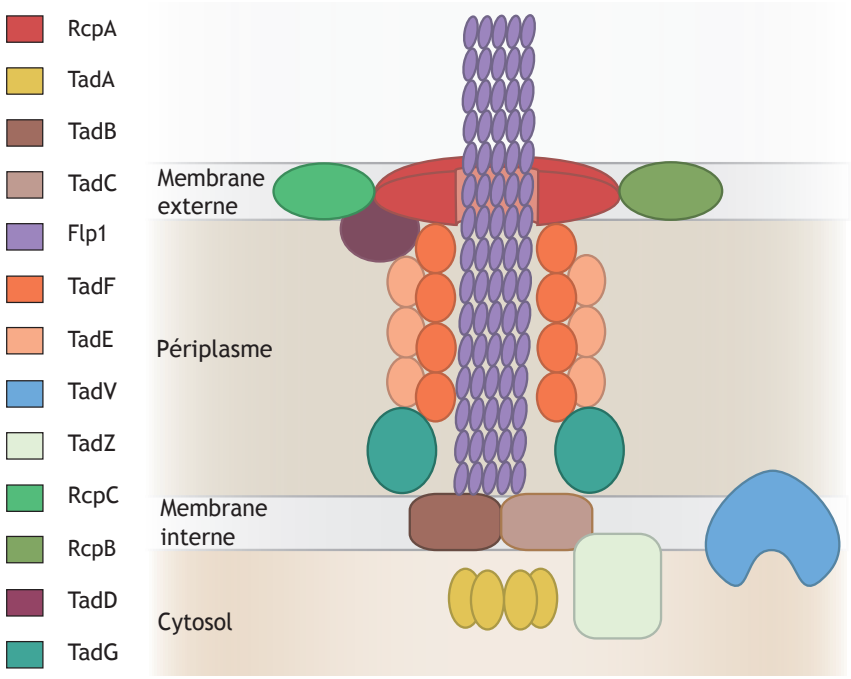


Figure 10 – Schéma du détaillé du pilus à adhérence étroite (Tad, T4cP). Figure inspirée de [154].

— *Le filament* : fibre composée de sous-unités de piline majeure, Flp (fimbrial low molecular weight protein) [158]. Cette piline est plus courte (50-70 AA) que celle présente chez les T4aP et T4bP [137]. La piline majeure, joue un rôle dans la formation de biofilm et l'adhérence car une mutation de cette protéine altère ces fonctions [159]. Les pilines mineures TadE et TadF semblent former une structure oligomérique dans le périplasme, afin de permettre l'assemblage du pilus ainsi qu'assurer un contact entre les deux membranes chez les bactéries didermes [154]. Une protéine, TadG, présente dans l'opéron semblerait avoir une fonction d'ancrage du pilus au niveau de la membrane cytoplasmique [154]. Les pilines sont synthétisées avant d'être assemblées dans le pilus en pré-pilines, qui sont clivées par la prépiline peptidase TadV et ensuite incorporées dans le pilus.

— *Le complexe d'assemblage* : complexe composé de deux protéines homologues transmembranaires , TadB et TadC, elles forment la plateforme sur laquelle le pilus s'assemble. Associée à cette plateforme, on retrouve une protéine cytoplasmique, TadZ, qui semble appartenir à la super-famille des ATPase MinD/ParA mais pour laquelle nous n'avons que peu d'information sur sa fonction dans la machinerie Tad et qui n'a aucun homologue connu parmi les protéines des autres machineries de la TFF-SF [154].

— *Le complexe moteur* : TadA, ATPase située dans le cytoplasme, elle est nécessaire à la formation du pilus en apportant l'énergie nécessaire à l'assemblage du pilus, facilité par le complexe d'assemblage [160]. Fait intéressant, cette ATPase semble avoir un lien phylogénétique avec l'ATPase non ubiquitaire

du système de sécrétion de type IV, VirB11 [154].

— *Le complexe de la membrane externe* : présent exclusivement chez les didermes, ce complexe est composé d'un pore (la sécrétine RcpA) qui permet à la fibre de sortir à travers la membrane externe des bactéries didermes [154, 161]. De la protéine RcpB, localisée dans la membrane externe, qui semble requise pour la stabilité et l'assemblage de la sécrétine en association avec une autre protéine localisée dans la membrane externe, TadD [154, 158]. Cependant nécessité de la protéine RcpB ne semble pas totalement prouvée mais semble essentielle à la formation du pilus [162]. Enfin, une dernière protéine, RcpC, également localisée dans la membrane externe, semble être impliquée dans la modification post-traductionnelle de la piline [154].

3.4 T2SS : Type II secretion system

Le système de sécrétion de type II (T2SS) est utilisé chez plusieurs espèces bactériennes pathogènes et non pathogènes, il permet de s'adapter aux conditions environnementales (en sécrétant des enzymes de dégradation) [163]. Le T2SS est également connu comme l'une des branches terminales de la voie sécrétoire générale (« general secretory pathway ») des bactéries didermes [164–166]. Cette machinerie est composée de douze à quinze protéines différentes, appelées Gsp (general secretion pathway), qui, en général, sont encodées en un seul cluster de trois opérons [95]. L'opéron gsp est généralement composée de treize gènes nommés C à O.

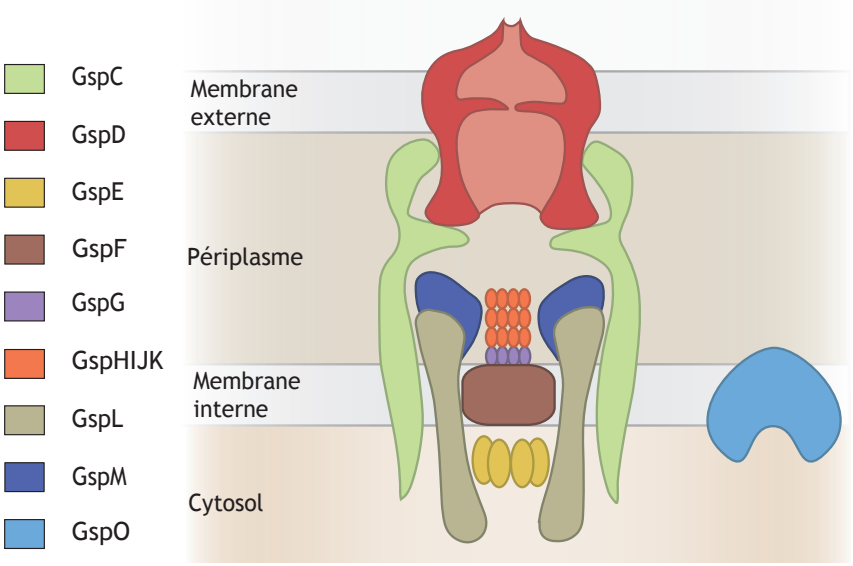


Figure 11 – Schéma du détaillé de l'appareil de sécrétion de type II. Figure inspirée de [163].

Grâce aux dernières données de cryo-EM, on sait que le T2SS à une architecture globale qui est divisée en quatre parties principales [167] (fig. 11) :

— *La plateforme d'assemblage* : complexe composé de quatre protéines membranaires intégrales qui se situe dans la membrane interne : GspC, GspF, GspL et GspM [168]. Cette plateforme permet l'extraction membranaire des sous-unités de pseudopiline et leur incorporation dans la fibre, entraînant la croissance de pseudopilus. Les protéines GspC, GspL et GspM sont disposées de manière cylindrique avec leurs domaines périplasmiques formant le bord extérieur du T2SS dans le périplasme, tandis que GspF est située au centre de cet anneau formé par les trois autres composants [95, 169]. Parmi les protéines de la plateforme, GspC permet la connexion de la plateforme d'assemblage avec la sécrétine, GspD [170, 171]. GspL et GspM sont des protéines bitopiques de structures similaires [172, 173]. GspL permet le lien avec GspE [171], qui est renforcé par la GspF, qui se lie aux deux protéines et les stabilise [174]. GspF, contrairement aux autres protéines, est une protéine polytopique intégrale de la membrane nécessaire à la sécrétion de protéines et la construction du pseudo-pilus [171, 175, 176].

— *Le pseudo-pilus* : fibre hélicoïdale assemblée dans le périplasme au niveau de la plateforme d'assemblage [120, 177]. Ce sous-complexe est principalement formé par la pseudopiline majeure, GspG, qui est la sous-unité principale, mais aussi par les pseudopilines mineures GspH, GspI, GspJ et GspK, qui forment un complexe tertiaire impliqué dans l'initiation de la fibre et se retrouvent donc à l'extrémité du pseudo-pilus [178–180]. Les pseudopilines sont de petites protéines, variant de 12 à 20 kDa, situées dans la membrane interne avant assemblage [181, 182]. Ces protéines sont synthétisées sous forme de pré-pseudopilines et sont coupées du côté N-terminal par la prépiline peptidase dédiée, GspO, pour obtenir des pseudopilines qui sont incorporées dans le pseudo-pilus par la suite [101]. Le pseudo-pilus est un filament qui croît à partir de la base, en ajoutant des sous-unités au niveau de la membrane interne.

— *Le complexe moteur* : GspE, ATPase située sur la face cytoplasmique de la membrane interne, associée à la plateforme d'assemblage. Elle fournit l'énergie nécessaire à la polymérisation de la fibre, par l'hydrolyse de l'ATP en ADP, permettant la formation du pseudo-pilus mais également la sécrétion de substrat [183–185]. L'ATPase interagit avec GspF, entraînant la rotation de GspF et l'assemblage du pseudopilus [120]. GspE est liée au T2SS par son association à GspL [168, 171, 186, 187].

— *La sécrétine* : pore présent dans la membrane externe qui est formé par l'assemblage des sous-unités de sécrétine en dodécamère [188, 189]. Elle permet la sécrétion des protéines à l'extérieur de la bactérie. Elle est formée par oligomérisation de GspD et son insertion dans la membrane [190]. Dans certains cas, l'insertion dans la membrane externe de GspD est facilitée par une lipoprotéine, la pilotine GspS, cependant la présence de GspS n'est pas nécessaire [191, 192]. GspD est un complexe dodécamérique fermé, cependant il existe un petit pore constitutivement ouvert qui permet le passage de pe-

tites molécules (<600 Da) [193]. Il y a donc un réarrangement du canal qui se produit lors de la sécrétion des exoprotéines ou du pseudopilus de 6 nm de large.

3.5 MSH : mannose-sensitive hemagglutinin pilus

Le « mannose-sensitive hemagglutinin » pilus (MSH pilus) est un pilus qui tient son nom du fait qu'en présence de D-mannose il ne permet plus d'agglutiner les globules rouges (activité d'hémaglutination) [194]. Ce pilus peut cependant avoir cette activité en absence de D-mannose, et est un facteur de colonisation chez *Vibrio cholerae* [195].

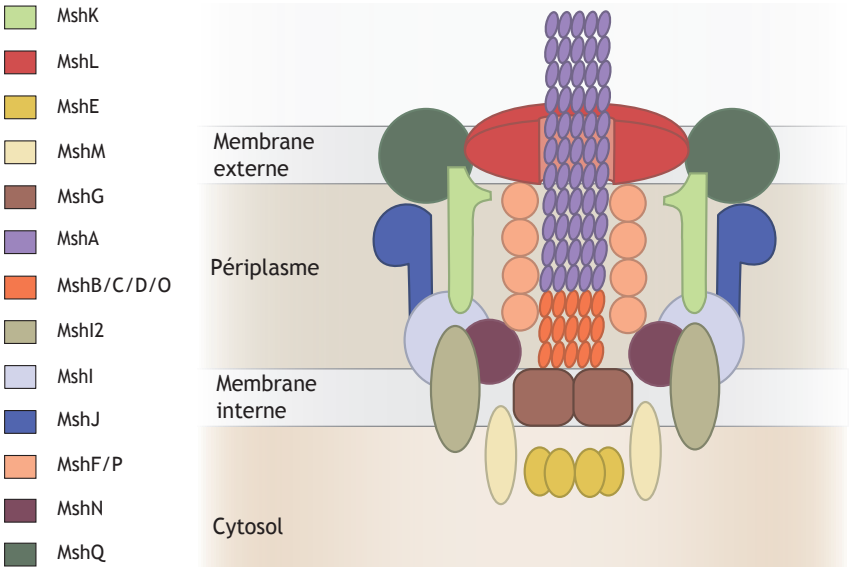


Figure 12 – Schéma du détaillé du pilus « mannose-sensitive hemagglutinin ». Figure inspirée de [86].

Ce pilus est composé de dix-sept sous-unités [86] (fig. 12). Parmi ses sous-unités, on retrouve l'ATPase (MshE) permettant la synthèse du pilus depuis la protéine membranaire intégrale (MshG) jusqu'au milieu extérieur en passant par la sécrétine (MshL). On trouve également une ATPase secondaire (MshM) dont la fonction n'est pas très bien connue. Le pilus est composé majoritairement d'une piline majeure (MshA) et de pilines mineures (MshB, MshC, MshD, MshO). C'est grâce à la piline majeure MshA que le pilus est capable d'avoir la fonction hémagglutinante [195]. Enfin, le système possède des protéines de diverses autres fonctions associées au maintien du pilus dans le périplasme (MshF, MshH, MshI, MshJ, MshK, MshN, MshP, MshQ) [86].

La présence de pili sur les cellules bactériennes est souvent associée à la capacité de coloniser les surfaces. Une étude portant sur des souches mutantes $\Delta mshA$ a démontré le rôle du MSH dans la colonisation et la formation de biofilms sur les surfaces abiotiques et les surfaces biotiques, mais n'a pas de rôle dans la vi-

rulence [196]. Le pilus MSH est un récepteur connu de bactériophages (p. ex. le choleraphage 493) et permet donc aux phages d'infecter la bactérie [84].

3.6 Le pilus de compétence

La transformation naturelle est un mécanisme qui implique le transport actif de l'ADN exogène à travers l'enveloppe bactérienne dans le cytoplasme. Pour se transformer, les bactéries développent un état appelé compétence, caractérisé par l'expression de gènes de compétence (Com) et, chez la majorité des espèces, l'expression de gènes des machineries appartenant à la TFF-SF sont également impliqués [81]. La transformation naturelle a été étudiée chez plusieurs bactéries modèles. Les modèles les mieux étudiés parmi les espèces monodermes sont *Bacillus subtilis* et *Streptococcus pneumoniae*, tandis que les représentants des bactéries didermes sont *Haemophilus influenzae* et *Neisseria gonorrhoeae* [197]. Cependant ici, nous n'allons pas nous intéresser à la compétence chez *Helicobacter pylori*, car cette bactérie utilise un système apparenté au système de sécrétion de type IV (T4SS) [198, 199].

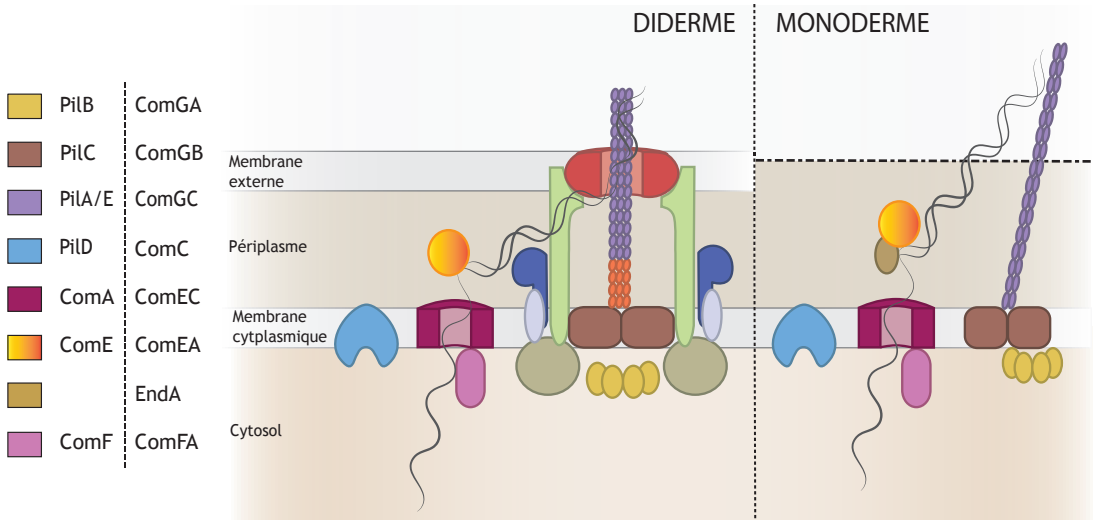


Figure 13 – Schéma du détaillé du pilus de compétence chez les didermes et les monodermes. La légende de cette figure ne montre que les protéines communes aux deux pili, pour plus de légende sur le pilus de compétence des didermes voir figure 8. Figure inspirée de [76].

Toutes les bactéries monodermes naturellement transformables possèdent dans leur chromosome un opéron comG, codant pour une ATPase (ComGA), une protéine membranaire intégrale (ComGB), une piline majeure (ComGC) et des pilines mineures (ComGD, ComGE, ComGF, ComGG), y compris un gène codant pour une prépiline peptidase (ComC) [200]. Les protéines de l'opéron comG sont impliquées dans l'assemblage de ce qui a été proposé comme étant un pilus de compétence. Ces gènes sont essentiels à la transformation chez *Bacillus subtilis* et *Streptococcus pneumoniae* [80, 201] (fig. 13).

Le pilus de compétence n'avait pas été visualisé en action jusqu'à récemment [79, 202]. C'est chez *Streptococcus pneumoniae*, que pour la première fois, on a visualisé le pilus de compétence d'une bactérie monoderme par fluorescence et microscopie électronique. Cette étude a montré que le pilus de compétence est une fibre longue, mince et flexible composée de la piline majeure ComGC, qui se lie directement à l'ADN extracellulaire.

Les modèles actuels proposent que la fonction du pilus de compétence est de lier l'ADN extracellulaire et de permettre sa translocation à travers l'enveloppe cellulaire [203]. Le pilus de compétence permettrait à l'ADN exogène d'accéder au complexe d'absorption de l'ADN de la membrane cytoplasmique composé du récepteur de l'ADN double brin ComEA [76, 204], de la nucléase EndA, qui dénature l'ADN double brin en ADN simple brin et le transmet au pore ComEC au niveau de la membrane cytoplasmique, qui permet le passage de l'ADN simple brin dans le cytoplasme [205]. ComEC est aidé par l'ATPase ComFA au moment du transport de l'ADN simple brin à travers la membrane [206].

Dans les bactéries didermes, la compétence nécéssite majoritairement le T4aP, possèdant une piline mineure qui semble avoir la faculté d'attacher l'ADN [80, 207] (fig. 13). L'ADN est ensuite transporté à travers la double membrane par le canal de la sécrétine (PilQ) et livré à la protéine périplasmique de liaison à l'ADN (ComE) [208, 209]. L'ATPase de rétraction du T4aP (PilT) pourrait jouer un rôle actif en rétractant le pilus et en tirant le complexe ComP-ADN lié dans le périplasme [207]. Une fois que l'ADN double brin atteint le recepteur à l'ADN, ComE, qui relie l'ADN au canal cytoplasmique formé par le ComA, un seul brin d'ADN entre dans le cytoplasme [210]. Chez *Vibrio cholerae*, les gènes de la compétence sont induits par la présence de chitine dans l'environnement, ils comprennent des gènes codant pour un T4aP et des composants de son mécanisme de biogénèse [211–213]. Ce T4aP a récemment été visualisé par microscopie d'immunofluorescence (IF) et son abondance moyenne est d'un pilus par cellule [79, 214].

3.7 Les différents pili chez les archées

Les archées interagissent avec l'environnement par l'intermédiaire de structures extracellulaires qui interviennent dans la colonisation des surfaces. La variété des filaments de surface que possèdent les archées permet l'adhésion aux surfaces biotiques et abiotiques, l'échange d'ADN, les interactions intercellulaires, la formation de biofilms et la motilité. Les machineries de surfaces peuvent être classés en fonction de la composition protéique en plusieurs groupes : comme les filaments de surface de type IV (archaellum, pili adhésifs, bindosome, pilus Epd et pilus inductible par les UVs) ou encore d'autres machineries de surface qui ne sont pas de type IV (p. ex. le système Ced, le filament « hamus », les cannulae) [215].

3.7.1 L'Archaellum

Longtemps connu comme le flagelle archéen ou l'archaellum n'a aucune relation (structurelle ou évolutive) avec le flagelle bactérien [82]. C'est un système macromoléculaire homologue au pilus de type IV [216], contrairement au flagelle bactérien (qui a des homologies avec le système de sécrétion de type III, T3SS [19]) l'archaellum est un filament de type IV constitué de sous-unités (les pilines ou aussi nommé archaellines) qui sont associées grâce à une ATPase cytoplasmique homologue à l'ATPase PilB [217].

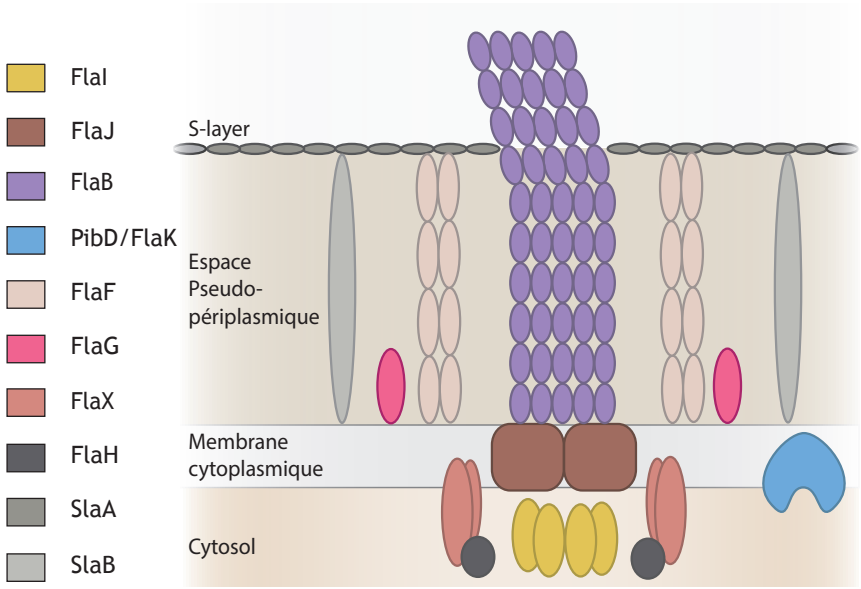


Figure 14 – Schéma du détaillé de l'archaellum. Figure inspirée de [218, 219].

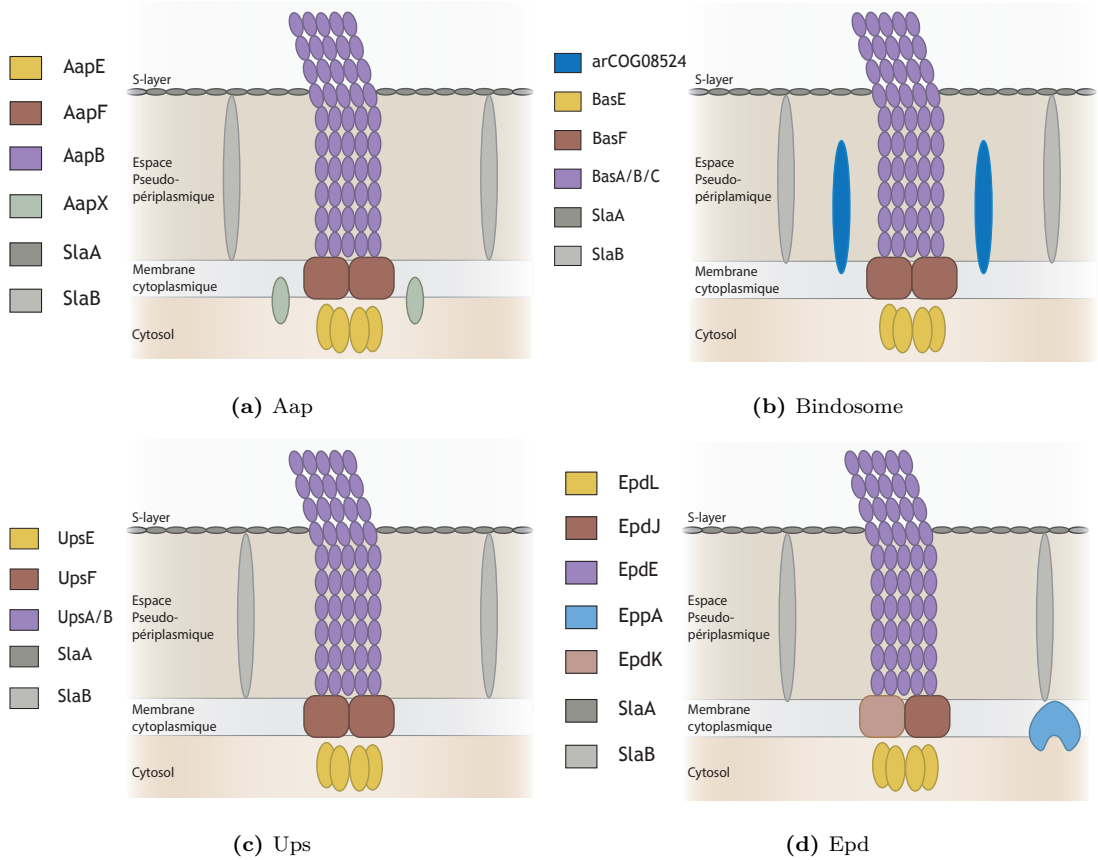
L'archaellum permet à la cellule de nager grâce à une force de rotation similaire à celle du flagelle bactérien, à la différence que l'énergie utilisée pour la rotation de l'archaellum est d'origine ATPasique. De plus, l'archaellum joue un rôle dans diverses activités comme la formation de biofilms, l'interaction cellule-cellule et l'adhésion [220–222].

L'archaellum est composé de sept à quinze protéines organisées en trois parties principales (fig. 14) : le corps basal, le crochet et le filament [223]. Dans l'un des systèmes les mieux étudiés chez *Sulfolbus acidocaldarius*, le filament est formé par des sous-unités FlaB qui sont assemblées de la même façon que tous les systèmes de la TFF-SF. En comparaison avec le flagelle bactérien, l'archaellum ne possède pas de lumière centrale et est plus fin (10-14 nm contre 18-24 nm), cependant son épaisseur est plus grande que celle des T4aP (6-9 nm) [223]. Le nombre d'archaellines varie selon les espèces : chez *Sulfolobus solfataricus*, l'archaellum ne possède qu'un seul gène de l'archaelline, tandis que pour l'archaellum de *Halobacterium salinarum* il y a cinq archaellines codées dans le génome [224]. Toutes les pilines qui composent ce filament sont clivées par la prépiline peptidase FlaK (aussi appelée

PibD) et assemblées en filament par l'ATPase FlaI [225].

Bien qu'il appartienne à la TFF-SF, on en sait peu sur l'architecture globale de l'archaellum. Comme les autres membres de la TFF-SF, il possède une ATPase hexamèrique cytoplasmique, FlaI, impliquée dans l'assemblage des pilines dans le filament mais aussi dans la rotation de l'archaellum permettant la mobilité de la cellule [226, 227]. L'ATPase est associée à une protéine membranaire cytoplasmique, FlaJ, ainsi qu'à une ATPase, FlaH, appartenant à la super-famille des ATPases RecA [215]. L'ATPase FlaH ne semble pas hydrolyser l'ATP mais en a besoin pour se lier à FlaI [215]. FlaI et FlaH se lient également à FlaX pour former un complexe moteur stable [215]. Deux autres protéines, FlaF et FlaG, sont également utiles pour la rotation et l'assemblage de l'archaellum [228, 229]. Ces protéines membranaires se situent principalement dans le pseudo-périplasme de la membrane et FlaF se lie également au « S-Layer » [219, 228]. Il a été proposé que FlaF forme un canal entre la membrane plasmique et le « S-Layer » qui ancre la machinerie à l'enveloppe de l'archée et permet à l'archaellum en croissance de traverser l'espace pseudo-périplasmique [219].

3.7.2 Les autres pili : Aap, Bindosome, Ups et Epd



(a) Aap **(b)** Bindosome

(c) Ups **(d)** Epd

Figure 15 – Schéma du détaillé des autres TFFs chez les archées. Figures inspirées de [99]

Aap : Archaeal adhesive pilus. Le pilus Aap (fig. 15a) est le filament le plus abondant observé chez *Sulfolobus acidocaldarius* en conditions de croissance en milieu riche. Ce pilus sert à l'adhésion mais il a aussi un rôle dans la formation de biofilms en facilitant les structures en forme de tours [230]. Il a été montré que l'Aap et l'Ups pilus (*voir ci-dessous*) sont nécessaires à la structure du biofilm chez *Sulfolobus acidocaldarius* [231]. Il est encodé dans le génome en un locus de cinq gènes codant pour les protéines de la piline mineure (AapA), de la piline majeure (AapB), de l'ATPase (AapE), de la protéine membranaire intégrale et d'une putative iron-sulfur oxidoreductase (AapX) [231].

Bindosome. Le bindosome (aussi appelé Bas pilus, fig. 15b) est un pilus présent chez *Sulfolobus solfataricus*, il intervient dans l'expression des protéines de fixation aux sucres dans l'enveloppe cellulaire. Le pilus est composé de l'ATPase d'assemblage (BasE), la protéine membranaire intégrale (BasF) et trois pilines (BasA, BasB et BasC) [232]. La délétion de l'opéron du bindosome suggère que le pilus est impliqué dans la localisation correcte des protéines permettant la fixation, et donc l'absorption de sucres.

Ups : UV-inducible pilus of *Sulfolobus*. Le pilus Ups (fig. 15c) est exprimé en réponse à l'exposition aux UV chez les *Sulfolobolus*. Il favorise l'aggrégation des cellules, ce qui facilite l'échange d'ADN chromosomique entre les archées. Cet ADN est ensuite utilisé lors de la recombinaison homologue au moment de la réparation de l'ADN. C'est un des seuls systèmes de pilus connu associé à l'échange d'ADN chez les archées.

Les gènes du pilus Ups sont organisés dans un système d'opéron et encodent une ATPase (UpsE), une protéine membranaire intégrale (UpsF), deux pilines (appelées UpsA et UpsB) et un gène qui code pour une protéine de fonction inconnue (UpsX). Les pilus Ups sont des structures rigides de l0 nm de diamètre et de longueur variable [233].

Epd : EppA-dependent pilus. Le pilus Epd (fig. 15d) n'a été décrit que chez *Methanococcus maripaludis* [234,235]. Ce pilus se compose de douze gènes répartis en quatre loci. Le premier locus a été décrit par Szabo et al. [234] et contient les gènes codant pour : la prépiline peptidase (EppA, euryarchaeal type IV prepilin peptidase A), des pilines mineures (EpdA, EpdB et EpdC), ainsi que des gènes de fonctions inconnues mais nécessaires à la pilliation (EpdF, EpdG, EpdH et EpdI). Les deux suivants ont été décrits par Nair et al. [235] et sont répartis en deux gènes isolés dans le génome, la piline majeure (EpdE) et une piline mineure (EpdD), et un dernier locus qui contient les gènes de deux protéines membranaire intégrales (EpdJ et EpdK) et une ATPase (EpdL). La fonction de ce pilus reste malheureusement inconnue à l'heure actuelle.

3.8 Tableau récapitulatif des systèmes de la super-famille

Fonctions	T4aP	T4bP	Tad (T4cP)	T2SS	MSH	Com	Archaellum	Aap	Bindosome	Epd	Ups
Formation de biofilms	•		•		•		•	•			•
Motilité	•						•				
Adhérence	•	•	•		•		•	•			
Pathogénicité	•	•	•	•	•						
Récepteur de bactériophages		•	•		•						
Echange d'ADN	•		•			•					•
Absorption de substrat									•		
Interaction cellules-cellules	•	•	•				•				•

Table 1 – Tableau récapitulatif des systèmes de la super-famille des filaments de type IV.

Deuxième partie

Co-option et bricolage moléculaire

CHAPITRE 1

CO-OPTION ET BRICOLAGE MOLÉCULAIRE

1.1 Article 1 : Playing molecular building sets : the evolution of protein secretion systems and related cellular appendages

Ce chapitre introduit la notion de co-option et de « bricolage » moléculaire. Ces notions sont détaillées dans cette revue qui sera publié dans Trends in Microbiology. Je vais dans un premier temps énoncer les points importants détaillés dans la revue qui seront suivis de la revue elle même.

Les familles de systèmes, y compris les systèmes de sécrétion de protéines, participent à la virulence, au mutualisme, à la compétition, à la motilité, à l'adhésion et à de nombreux autres processus d'importance capitale chez les procaryotes. Par exemple, la superfamille des filaments de type IV, qui comprend le système de sécrétion de type II (T2SS), a évolué vers des systèmes spécialisés dans l'aggrégation, la transformation naturelle et plusieurs types de motilité [236] qui révèlent encore de nouvelles fonctions [46]. Des histoires évolutives complexes ont également conduit à l'innovation fonctionnelle d'autres systèmes associés aux membranes. Le système de sécrétion de type VI (T6SS), éventuellement dérivés de phages [237,238], et l'inhibition dépendante du contact par le système de sécrétion de type V (T5SS), qui fait partie d'un grand clan de porines bactériennes, ont des rôles clés dans les interactions sociales antagonistes entre bactéries [239, 240]. Le système de sécrétion de type IV (T4SS) sécrète des protéines, mais est également responsable de la conjugaison de l'ADN [241, 242]. Le système de sécrétion de type III (T3SS) est un facteur clé de virulence et provient d'une structure homologue présente dans le flagelle bactérien [19]. Ces systèmes fournissent des exemples uniques de la façon dont un ensemble commun de composants centraux peut se diversifier en une multitude de fonctions moléculaires aux rôles écologiques

variés. Au-delà de l'analyse de l'évolution des systèmes de sécrétion, nous souhaitons attirer l'attention des microbiologistes sur les contributions de ces études pour comprendre l'innovation fonctionnelle.

Ce manuscrit est divisé en sections mettant en évidence les différents mécanismes qui agissent sur l'évolution et la diversification de ces systèmes : 1) comment la promiscuité protéique peut favoriser l'innovation fonctionnelle, 2) comment le recrutement de nouveaux composants se fait par cooptation de gènes impliqués dans d'autres processus, 3) comment la spécialisation des systèmes conduit à l'innovation fonctionnelle, 4) comment cela conduit à de grandes familles de nanomachineries qui se sont diversifiées par rapport à un ensemble commun de composants, et enfin 5) comment cela peut conduire à de nouvelles fonctions radicales dans les communautés microbiennes suite à des mutations et au transfert génétique horizontal. La conclusion de l'article soulignera les tendances communes aux processus évolutifs affectant les systèmes de sécrétion de protéines et la façon dont ils peuvent être exploités pour identifier de nouveaux systèmes.

Playing with molecular building sets: the evolution of protein secretion systems

Rémi Denise[1,2], Sophie S Abby[3,*], Eduardo P C Rocha[1,*]

[1] Microbial Evolutionary Genomics, Institut Pasteur, CNRS, UMR3525, Paris, 75015, France.

[2] Sorbonne Université, Collège doctoral, F-75005 Paris, France

[3] Univ. Grenoble Alpes, CNRS, Grenoble INP, TIMC-IMAG, 38000 Grenoble, France

[*] corresponding authors

Abstract

The emergence of complex molecular systems in natural history remains an intriguing evolutionary process. Protein secretion systems of Bacteria are complex membrane-associated nanomachines important for many biotic and abiotic interactions. They constitute remarkable examples of systems that evolved by the combination of a multitude of molecular evolution mechanisms from the co-option of parts of other membrane-associated nanomachines. Here, we review what is known on the evolution of protein secretion systems. We highlight how these evolutionary processes depend on the complexity of the systems, the functional differences between extant systems and the original co-opted system, and the mechanisms of effector recognition. Understanding these processes provides clues on how to find novel systems in poorly sampled bacterial lineages.

Keywords: Molecular evolution; horizontal gene transfer; exaptation; secretion; functional innovation.

Glossary

Co-option/Exaptation: use of an existing biological function for a novel adaptive purpose.

Diversifying selection: when natural selection leads to the rapid diversification of a gene; frequent in proteins involved in pathogenesis.

Effector protein: a secreted protein that has an effect on another cell, *e.g.* a virulence factor that subverts the function of an eukaryotic cell.

Functional innovation: change in a function to provide an adaptive response.

Monoderms/Diderms: Bacteria lacking/having an outer membrane. Diderms tend to be gram negative and monoderms gram positive, but several exceptions are known.

Neo-functionalization: a process where a gene acquires a new function.

Protein secretion systems: systems allowing the transfer of proteins across the outer membrane of diderm Bacteria. Systems with homologous components may exist in monoderms, and are sometimes also called protein secretion systems.

Sub-functionalization: a process where a gene with several functions specializes in a subset of them.

Tinkering: recruitment of a component into a new biological system (pathway, protein complex, regulon).

Main text

The mechanisms and the plausibility of the evolution of complex functions have been one of the most intriguing and contentious points in evolutionary biology [1]. Microbes have an increasingly important role in studies aiming to unravel the evolutionary mechanisms of functional innovation because they can be easily manipulated, evolved, and analyzed in the laboratory. In nature, their large effective population sizes and their ability to exchange genes horizontally render selection processes efficient [2, 3]. Furthermore, Prokaryotes emerged in the planet billions of years before Eukaryotes, and were responsible for many structural, biochemical and genetic innovations.

The creation of novel genes from random sequences seems rare in most phyla. Instead, novel functions arose by the modification of previous genes and genetic structures (tinkering) [4], by mutations, deletions or accretions of genetic material and by recombination. The resulting variants — in terms of biochemical reactions, cell localization, genetic regulation — were then purged or amplified by natural selection. Horizontal gene transfer spreads these novel adaptive functions when there are ecological opportunities, as currently observed in the evolution of antibiotic-resistant Bacteria [5, 6]. The transfer of genes into a genome also provides genetic variants that may often be under relaxing selection and may become the substrate of further functional innovation by mutation and recombination. For example, genomes contain many mobile genetic elements that can evolve novel functions for the host advantage [7]. In this process, a gene or system that originally evolved to respond to a given adaptive need was co-opted (or exapted) to provide a function that tackles a different need [8]. Comparative genomics, phylogenetics and experimental studies showed that the tinkering of existing cellular machineries was at the origin of most, if not all, extant protein secretion systems of Bacteria.

Protein secretion systems and related nanomachines

Prokaryotes use secreted proteins (effectors and auxiliary proteins) to protect themselves, manipulate their environment, and interact with other individuals [9, 10]. Protein secretion systems are defined as machines that transfer proteins across the outer membrane in diderm

Bacteria [11]. They are called “TXSS” for type “X” secretion system, where X is a number from 1 to 9 (so far, for broad reviews see [12, 13]). Some systems transfer proteins either directly from the cytoplasm (T1SS, T3SS, T4SS, T6SS, T7SS) or from the periplasm (T2SS, T5SS, T8SS, T9SS). Protein secretion systems have different numbers of components for assembling and transporting proteins across the outer membrane, across the cytoplasmic membrane (when relevant), and a system to recognize effectors. Some systems further deliver proteins into other cells using specialized pili (T3SS, T4SS, T6SS).

Evolutionary associations underlie the numerous structural and sequence similarities between components of secretion systems and other cellular nanomachines (Figure 1). For example, the T2SS is related to the super-family of type IV filaments (TFF), the T3SS is homologous to the secretion system of the bacterial flagellum, the T4SS is closely related to the conjugative pilus, and some T6SS components shares striking similarities with bacteriophage proteins. The study of the evolutionary processes leading to protein secretion systems provides clues on their structural biology, assembly, genetics, and distribution across taxa (and vice-versa). It may also provide means of finding novel systems and ways to manipulate secretion [14]. In the following sections, we exemplify evolutionary processes that were at the origin of the best-studied protein secretion systems.

Minor tweaks for novel functions

Many biological systems have multiple functions that can foster processes of specialization. An interesting example is provided by the AAA+ ATPases of the T4SS, a system that originated for DNA exchange by conjugation (Figure 2b). In the latter process, a nucleoprotein complex composed of a single-stranded DNA molecule and the relaxase is transferred into another cell by the mating pair formation (MPF) apparatus (reviewed in [15-17]). The MPF contains a T4SS whose phylogeny can be divided into eight major clades with some different accessory components [18]. Such analyses indicate that the T4SS was originally involved in conjugation in diderm Bacteria and only later was transferred to monoderms (Bacteria and Archaea). Two AAA+ ATPases of all T4SS —VirB4 and the coupling protein (T4CP) — arose from an event of gene duplication preceding the emergence of modern T4SS. The T4CP specialized in mediating the interaction between the relaxase and the T4SS, whereas VirB4 became tightly involved in

the function and/or assembly of T4SS. Other homologous ATPases are found in some subtypes of T4SS, *e.g.* the VirB11 ATPase that was recruited from a TFF [19]. Alternative mechanisms of conjugation that transport double stranded DNA, such as those associated with TraB and TdtA, also have homologs of VirB4/T4CP [20, 21].

Conjugation is a protein secretion mechanism because the relaxase is recognized by the T4CP and then secreted by the T4SS with the covalently linked DNA [22]. This has facilitated its co-option into a machine specialized in protein secretion. Pathogenic bacteria use T4SS to secrete virulence factors into Eukaryotic cells [15] and a T4SS of *Xanthomonas citri* was recently shown to secrete toxins into other bacteria [23]. Phylogenetic analyses show many independent co-options of T4SS into specialized protein secretion systems, but known secretion systems are only found in two of the eight T4SS clades (named T and I) [18]. The reasons why specialized protein secretion is so far known to be restricted to these types of T4SS are unknown and could relate to specific characteristics of their T4SS allowing them to interact with eukaryotic cells, as suggested by the observation that broad host range conjugative systems tend to be of type T [24]. Two of the eight T4SS clades are present exclusively in monoderms and these are devoid of systems able to deliver effectors directly into other cells. The functional promiscuity of the T4SS, *i.e.* its ability to intrinsically secrete proteins, implicates that the evolution of a protein secretion system from the adequate conjugation pilus can take place in a very small number of steps: acquisition of a T4CP-interacting domain by the effector (or horizontal transfer of the effector) and loss of the relaxase. T4SS with intermediate properties have been observed, including secretion of the relaxase without ssDNA by conjugative systems [25], and DNA conjugation and protein secretion of virulence factors by the same system [26]. In contrast, the evolution of a T4SS of Campylobacterales into a competence pilus for DNA uptake discovered in *Helicobacter pylori* required many more changes and seems to have taken place only once [27]. The specialization of the T4SS into a protein secretion system may facilitate the subsequent expansion of its ability to secrete multiple effectors (see Box 1). For example, the T4SS of *Legionella* spp. can deliver numerous different effectors into eukaryotic cells because its T4CP binds adaptors that recruit distinct subsets of effectors [28, 29].

Specialization stimulates subsequent innovation

As the conjugation pilus is inherently able to secrete DNA-associated proteins, the bacterial flagellum includes a secretion system — the flagellar T3SS (F-T3SS) — to export the components of the flagellar filament during its biosynthesis. Flagella are complex machines containing the F-T3SS, the motor, the hook and the filament (Figure 2a). While the key function of the flagellum is to allow motility, it can provide many other functions including adhesion, biofilm formation, and interaction with immune systems [30]. The secretion of toxins and other proteins by flagella was reported in several bacteria [31-33], showing that protein secretion to the extracellular milieu has evolved multiple times in the F-T3SS. The flagellar basal bodies devoid of most extracellular structures observed at the cell surface of *Buchnera* spp. [34] could function as secretion systems, since the extracellular components of the flagellum are not required for this function. This reduction of the flagellar structure is reminiscent of the mechanism used by some bacteria to cut the costs on cell motility under nutrient depletion by ejecting the flagellum hook and filament [35]. Hence, simple gene losses could have driven the specialization of the F-T3SS into systems able to secrete proteins in the environment, but unable to provide cell motility.

One particular case of flagellar reduction led to the subsequent evolution of the non-flagellar T3SS (NF-T3SS, here shortened to T3SS). The T3SS has many genes homologous to the flagellum, even if it has fewer components (flagella are encoded by ~50 genes, T3SS by less than half) [36-38]. The analysis of the components of the T3SS, and of their phylogeny, shows that most core components of the T3SS were directly derived from the ancestral flagellum, including the filament and the F-T3SS. This process was accompanied by the loss of many flagellar genes, which may have led to intermediate systems involved in protein transport to the periplasm or to the extracellular space [39]. A few systems of unknown function in Myxococcales may be representative of such intermediate systems. Some crucial gene gains then led to the T3SS, a machine that secretes a plethora of effectors directly into eukaryotic cells (Figure 2a). The only recognizably ubiquitous component of the T3SS that lacks a homolog in the flagellum is the pore-forming secretin, which was shown by phylogenetic inference to have been acquired at least three times from different origins (see Box 2 and Figure 3). Other key components found in the T3SS and lacking in the flagellum are the device puncturing the eukaryotic cells (translocon and tip of the needle). They provide functions relevant for the

T3SS, but not for the F-T3SS. Their relations of homology within the different sub-types of T3SS cannot always be ascertained because they evolve very fast and the analogous components often lack sequence similarity. Rapid evolution of these components may be the result of diversifying selection because they are in direct contact with the host and are targeted by immune responses [40]. Hence, following the initial crucial steps of gains and losses, T3SS evolved into systems adapted to the peculiarities of eukaryotic cells, notably into variants dedicated to puncture either animal, plant or fungi cells [41]. T3SS concomitantly spread by horizontal gene transfer among Bacteria, events that are rarely observed among flagella [42, 43]. Some Bacteria (*e.g. Burkholderia*) encode multiple T3SS to interact with multiple types of eukaryotic hosts [44]. The natural history of T3SS shows that complex systems can be co-opted following the loss of some of their functional modules and then lead, by accretion of novel components, to system with novel characteristics.

Functional diversity using a common set of components

The T2SS resulted from the co-option of a type IVa pilus (T4aP) presumably involved in twitching motility. Together with other systems of Archaea and Bacteria, these two systems constitute the super-family of type IV filaments (TFF) [45, 46]. TFFs typically have, among others, five to six core components: AAA+ ATPase(s), major and minor pilins, cytoplasmic membrane platform(s), a prepilin peptidase and sometimes a secretin (see Figure 4). The phylogeny of the ATPase family places the root of the tree between the bacterial and the archaeal groups of TFF, suggesting that the system pre-dated the last common ancestor of all cellular life forms [46]. Most other key components have recognizable homology across TFF, suggesting that they were already present in the ancestor of all TFF. Extant TFFs then evolved by a succession of mutations, gene deletions, duplications, fissions and fusions, resulting in systems involved in adhesion, protein secretion, twitching motility, flagellar motility (in Archaea, unrelated with the F-T3SS) and DNA uptake [27, 47-50]. For example, T2SS have evolved pseudo-pili producing shorter structures than those found in the T4aP from which they seem to derive. These specialized components are involved in protein secretion by transporting effectors across the outer membrane [51]. The traces of the evolution of some types of components show complex scenarios. Soon after the divergence of the ancestral TFF

into an archaeal and a bacterial branch, the ATPase of the latter systems was duplicated into a pair of proteins specialized in pilus extension (PilB) and retraction (PilT). This event was thought to have endowed the T4aP with the ability to provide twitching motility by rounds of pilus extension and contraction. Recent results have shown that the single homologous ATPase of the tight adherence (Tad) pilus systems can perform both functions [19]. Hence, the PilB/PilT duplication seems to have allowed a specialization of two ancestral functions (sub-functionalization). When the T2SS evolved from a T4aP, the retraction function was unnecessary and the corresponding specialized ATPase was lost.

TFFs specialized in the secretion of protein effectors (T2SS-like) have emerged multiple times along the evolution of TFF [46]. Many components — including the major pilin of the T2SS-like system of Chlamydiales — were recruited from different TFFs and have phylogenies different from that of the T2SS. The T2SS-like systems from Bacteroidetes, *e.g.* those of *Cytophaga*, constitute independent co-options of (different) T4aP for protein secretion [46]. This suggests that a specialized protein secretion system may require relatively few steps to evolve from the T4aP, possibly because the latter already secretes a specific type of protein (pilins). Once the T2SS evolved into a specialized secretion system it became able to secrete many different types of proteins [52, 53] (Box 1).

The spread of TFF among Prokaryotes was promoted by horizontal gene transfers. For example, the Tad pilus originated from an archaeal pilus that was transferred to diderm Bacteria, where it recruited a secretin, and then spread to many phyla [46] (see Box 2). In such situations, horizontal transfer not only spreads functions across communities, but it spurs innovation because the systems may be maladapted to the novel genetic background. Yet, horizontal transfer of a system can only take at high rates if its genetic organization is compact. Systems encoded in a single locus, such as the Tad and the T2SS, are more frequently transferred horizontally than those encoded in several loci scattered on the genome [46], presumably because one single event of transfer is enough to provide a novel function to the recipient [54].

Co-options leading to radical inventions

The T6SS exemplifies how functional innovations can result from co-option of machines with radically different functions. The T6SS delivers effectors directly into eukaryotic or prokaryotic cells and is implicated in inter-specific competition, virulence, resource scavenging, and genetic exchanges [55-57]. The T6SS has a baseplate-like platform bound to the membrane, which anchors a contractile sheathed tube decorated by a puncturing device that allows toxins to penetrate the target cells (see Figure 2). The structure of many of these components resemble bacteriophage proteins and the overall T6SS resembles an inverted phage tail [58-62], which suggests that a large part of the T6SS derived from a phage. Unfortunately, sequence similarities between phage and T6SS homologs are very low and preclude the study of the initial processes of co-option using standard phylogenetic approaches. There are also similarities between T6SS and phages in terms of assembly/disassembly dynamics. In particular, T6SS sheaths are contractile in order to project the tube towards target cells, in a similar fashion that phage tails contract for DNA delivery into target cells [63]. This activity is dependent on the ClpV ATPase [63], which is an AAA+ ATPase, homologous to ATPases of T2SS and T4SS, but not related to any phagic components (see Figure 1). The T6SS also has two components homologous to the type IVb pilus (T4bP) of the TFF super-family, which led to its initial misnaming as a T4bP [64]. This suggests that T6SS arose by integration and co-option of a tail and a baseplate from a phage with components present in the genome that were implicated in other processes.

Some genomes encode many different T6SS. For example, *Burkholderia thailandensis* have five different systems, one of which is specialized in the interaction with the host and another with competing Bacteria [65]. As the structural components of the T6SS are also effectors (Box 1), different systems may be associated with different effectors. There are four known major variants of T6SS. Variants of the most widespread T6SS (T6SS^{i}) were initially found in *Francisella* (T6SS^{ii}) and in Bacteroidetes (T6SS^{iii}) [66-68]. In spite of having specific components, these systems include most of the T6SS^{i} core components and are presumed to have derived from the same ancestor. More remarkably, the T6SS^{iv} is present in at least six phyla, it has most of the core components of other T6SS, but seems to lack some components of non-phagic origins, such as the ATPase or the trans-envelope complex [69]. It was speculated that it could have emerged from phages independently from other T6SS or,

alternatively, it could have given rise to other bacterial contractile-injection systems (including T6SS) and phages [69]. These results suggest that even the evolution of a very complex machine with functions very different from the original one – protein secretion versus DNA injection – may occur several times in parallel.

Concluding Remarks and Perspectives

The frequency of co-option events and its determinants are unknown, but the study of secretion systems provides interesting clues on both aspects. The evolution of T4SS exemplifies how this process may require relatively minor changes when there is *a priori* some functional promiscuity (the original system is inherently capable of protein secretion). The key step of this co-option is probably the acquisition of the ability to recognize a variety of protein effectors, explaining why it evolved many times independently in natural history [18]. Similarly, all TFF secrete to the extracellular space components producing some sort of filament, which may have facilitated the multiple independent evolution of T2SS-like systems. The history of TFF illustrates how a common set of components can diversify into systems with very different functions. The evolution of the T3SS required more steps, including gene losses and gains, and the ability to interact specifically with diverse effectors. Some of these steps occurred several times, notably the loss of parts of the flagellar component, the acquisition of secretins (Box 2), and possibly the acquisition of translocons [39]. However, all known T3SS are monophyletic, suggesting that the whole path to T3SS took place only once. Finally, the T6SS has components from phages and from other systems and the function of protein secretion it performs is very different from the original function of phage tails. The complexity of this evolutionary process might suggest it occurred only once in natural history, but at least another T6SS-like system may have emerged independently [69]. If true, this unexpected observation may stem from the constant influx of temperate phages in bacterial genomes that provide abundant genetic material for evolutionary tinkering [70]. Based on these works, it is thus tempting to speculate that the frequency of co-option events is shaped by the availability of machinery that can be tinkered by natural selection, the magnitude of the functional differences between the original machines and the final protein secretion system, and the complexity and plasticity of the mechanisms of effector recognition.

The evolution of a novel secretion system requires the existence of components for the key functions: ATPases for assembly and eventually secretion, membrane channels, effector recognition, and eventually a pilus. Analogous components between different systems are not simply interchangeable, but genomes show many putative systems that share only a fraction of the expected homologs with known systems. Many of them may be non-functional, but there is the distinct possibility that they correspond to unknown co-options of membrane nanomachineries. This is illustrated by systems that are chimeras of different secretion systems and other related machines. One clade of T4SS — type I, also called T4SSb — includes a T4bP that is essential for conjugative transfer of plasmid R64 in liquid media and for adherence to host cells by the T4SS of *Legionella* [71]. Some *Haemophilus spp*. export DNA to the environment, favoring biofilm formation, using components of a T4SS and a TFF secretin [72]. The core export apparatus of the T3SS was co-opted to produce nanotubes that allow bacteria to extract nutrients from infected host cells [73]. Finally, the T1SS is itself a system composed of two more ancient components: an outer membrane porin and an ABC transporter connected by a membrane fusion protein [74]. The frequency with which components were combined to produce novel systems suggests that humans could engineer secretion systems with novel features by recruiting components of membrane-associated machines.

This review focused on well-known secretion systems, but many relations of homology between protein secretion systems and other machines have been reported. For example, the ESX inner membrane transport system in monoderms has components homologous to the T4SS, the autotransporters (T5SS) are structurally homologous to the porins of the novel FAP system that exports amyloid subunits in *Pseudomonas* [75], and there are intriguing homologies between components of several secretion systems and a system putatively involved in the transport of proteins between mother and forespore cells in Firmicutes [76]. A TonB-dependent transporter (TBDT) of *Myxococcus xanthus*, a widespread family of systems usually involved in protein and nutrient import, was recently shown to be involved in the two-step protein secretion of a protein [77]. Finally, little is known about the evolution of the other secretion systems. Notably, the T9SS is present in many bacteria of the Fibrobacteres-Chlorobi-Bacteroidetes super-phylum and is involved in both gliding motility and protein secretion [78, 79], but its evolutionary associations with other cellular

components has not been studied. The T7SS is present in Actinobacteria and shares homology to systems involved in protein transport across the cytoplasmic membrane of Firmicutes, both types of systems containing AAA+ ATPases like T4SS and T2SS [80, 81]. To the list of known secretion systems one can't, but would like to, add all systems we don’t know yet. Research programs to identify novel systems can now rely on the abundance of genomes to identify components homologous of membrane-associated machineries, metagenomics data to investigate conditions where they are expressed, structural and microbial cell biology to understand their function and evolutionary biology to help integrate this information. This will likely accelerate the discovery of novel systems.

Box 1: Where do all effectors come from?

At a given moment in evolution, protein secretion systems acquired the key ability to recognize novel effectors. This is particularly remarkable for systems able to secrete many different effectors, such as T2SS, T3SS, T4SS, T6SS, or T9SS, many of which evolved from machines able to secrete at least one protein (pilins, flagellins, etc). Systems able to secrete many different effectors may be more efficient than systems secreting a single one, because the production of one single machine is sufficient to secrete a large panel of proteins. Also, sometimes these effectors must interact to be efficient, which means they must be secreted at the same time. For example, T3SS effectors require co-secretion of specific protein translocases to traverse the membrane of eukaryotic cells [82]. Selection for the ability to secrete many different effectors in a single cell may have been a driver of the evolution of the complex machineries of T2SS, T3SS, T4SS, and T6SS [38]. On the other hand, systems with fewer components, but usually associated with one or a few effectors like the T1SS and the T5SS, can be horizontally transferred with their effectors across bacteria.

Secretion systems discriminate effectors using mechanisms that provide some clues on how effector recognition evolved. Some effectors of T6SS are structural proteins of the system making them more recognizable [83], and some T4SS effectors have domains that are structural homologs of the recognition domain of the relaxase or that interact with partners of the T4CP [84]. In both cases, gene fusions could be simple mechanisms driving the evolution of novel effectors. Gene fusions are frequent in secreted proteins, as revealed by the combinatorial variation of domains of secreted polymorphic toxins [85]. The mechanisms of effector recognition for some systems are poorly known, which complicates the study of their evolution. To further blur this picture, some effectors can be recognized by multiple systems. For example, a toxin was recently shown to be secreted by both T2SS and T3SS in *Vibrio* [86].

Some effectors might evolve by co-option, just like their cognate secretion systems. A large fraction of the repertoire of *Legionella* T4SS effectors may have been co-opted from proteins of Eukaryotes [28], and *Burkholderia* deploy a T3SS anti-fungal protein that may have been co-opted from a prophage tail-like protein [87]. Proteins that evolved to become secreted by a protein secretion system can subsequently endure processes of gene duplication and transfer,

and diversify into novel functions. Genes encoding secreted proteins are very often on mobile genetic elements, such as plasmids [9] and temperate phages [88].

BOX 2. Sharing and recruiting components

The transfer of a system to another genetic background may require novel functions already present there, in which case the latter may be recruited from other systems and initially shared with them. An interesting example is provided by the T1SS, a system composed of an ABC transporter, a porin, and a fusion protein connecting the two. In several T1SS of *E. coli*, the porin is TolC, a protein also involved in the transport of small molecules [89]. In this case, a single gene, distant from the other genes of the T1SS, provides multiple functions to the cell. Similar gene sharing is observed in TFF, where some prepilin peptidases contribute to the assembly of both T4P and T2SS [90]. However, a component that is shared by several systems accumulates functional, structural and regulatory constraints [91]. Hence, one expects that subsequent gene duplication and sub-functionalization may eventually result in the presence of multiple homologous genes in the genome.

The initial gene sharing of a residing secretin — outer-membrane porins — by systems in diderm Bacteria lacking an outer membrane porin may have been common. It was probably the mechanism leading to the proto-Tad system when it was transferred from an archaeon to a diderm bacterium. It also probably took place three times independently in the evolution of the T3SS, and occurred in some filamentous phages that use a secretin to secrete virions from living cells (Figure 3). The grafting of secretins into secretion systems requires a remarkable structural flexibility. Secretins assemble independently of the rest of the translocons in T2SS and in filamentous phages, and the substitution of a single amino-acid in some T4P secretins was shown to make them capable of self-assembly [92]. Yet this structure is not stable in the absence of the inner-membrane components with which its N-terminus domains interact [93, 94]. It is tempting to speculate that the ability of secretins to assemble autonomously from the rest of the system for a short period of time has facilitated the initial recruitment of the secretin by so many different systems.

Acknowledgements

We thank the many people with whom we have collaborated on the topic of protein secretion systems and related machineries along the years, notably Fernando de la Cruz, Julien Guglielmini, Bertrand Néron, Olivera Francetic and Elie Dassa. We are grateful to Claude Parsot and Laura Gomez-Valero for comments and suggestions on an earlier version of this manuscript.

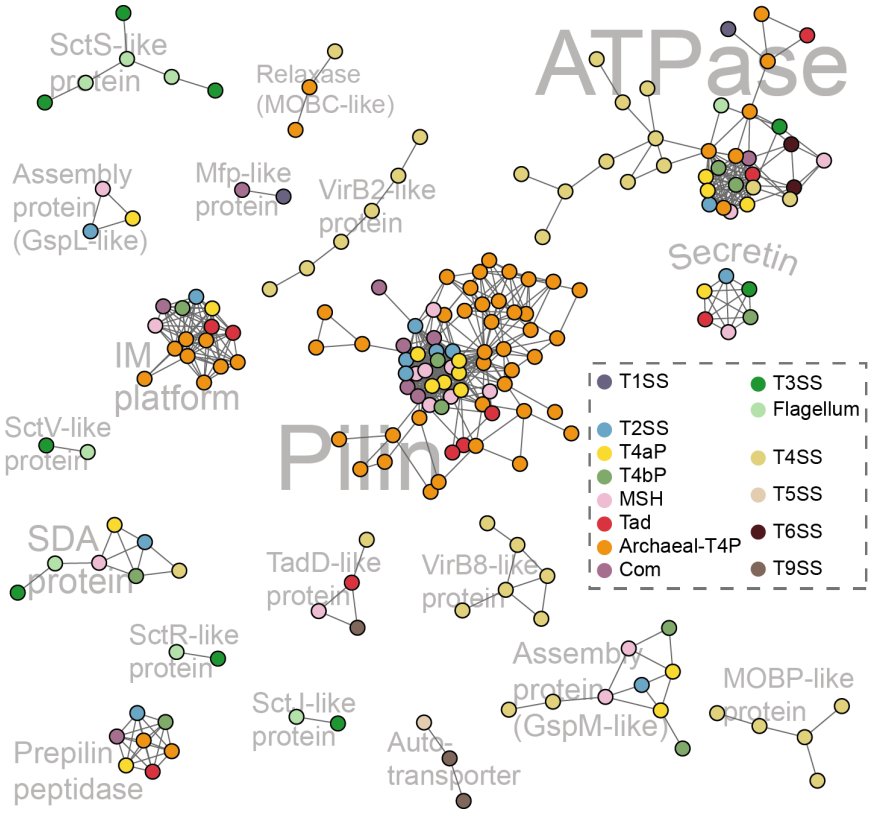


Figure 1 – Pairwise HMM profile alignments between all the proteins of the TXSS-related systems. The HMM profiles were obtained from TXSScan [46, 95]. The color of nodes represents systems in which the proteins were found. To establish relationships of homology between the components of the different systems, *i.e.* to draw edges between nodes, we made pairwise alignments of their HMM protein profiles using HHSearch v3.0.3 (p-value threshold of 0.001). Groups of proteins that gathered more than two components from at least two systems are displayed. The function attributed to each group is written in grey in its background. Given the current difficulty in precisely delineating the functions of the TFF of Archaea, they were put together under “Archaeal-T4P”.

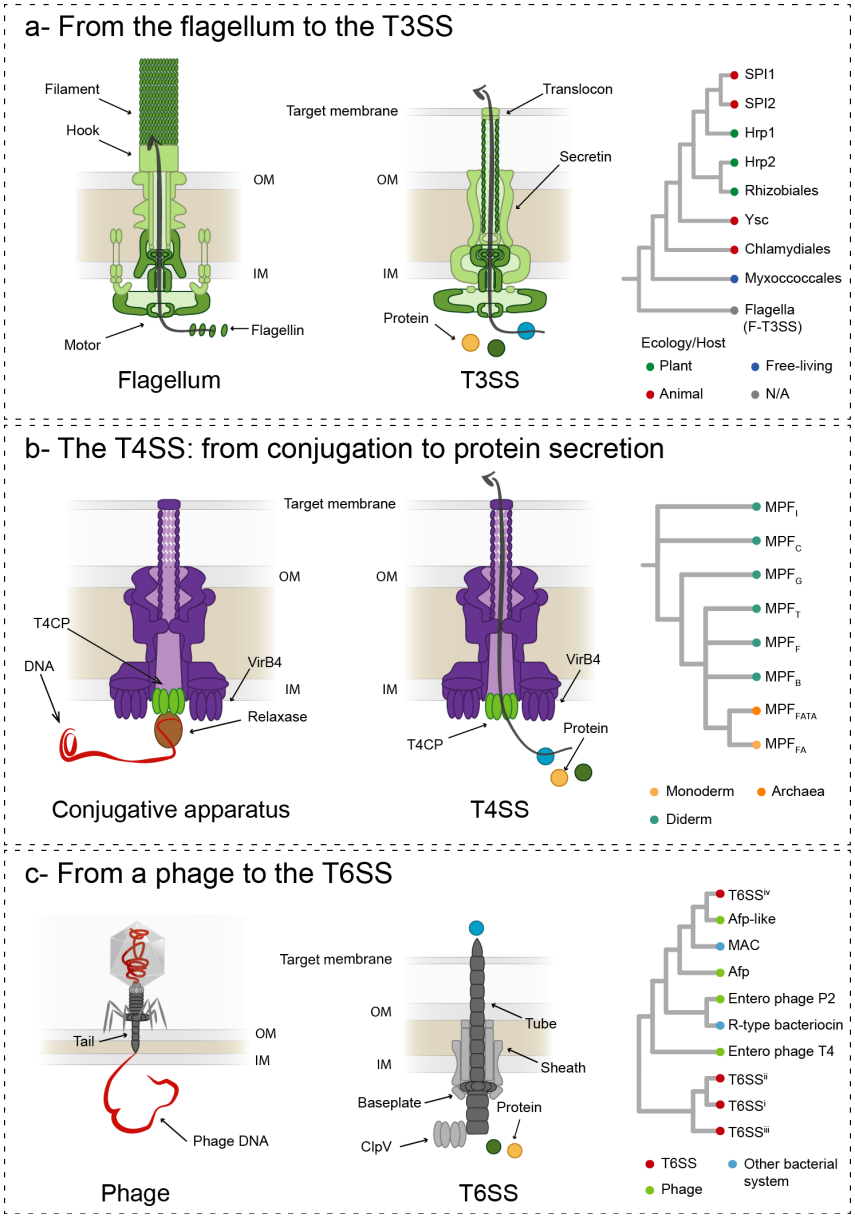


Figure 2 – The evolution of protein secretion systems delivering effectors directly into other cells. (a) The diversification of the T3SS from the bacterial flagellum involved initially the loss of flagellum-specific genes and the motility function. The subsequent multiple acquisitions of pore-forming secretins, translocons, and a few other genes led to the extant T3SS. On the right, the rooted cladogram represents the history of the T3SS [39]. The *Myxococcales* system is not a genuine T3SS since, to the best of our knowledge, it lacks an outer membrane porin. (b) Conjugative apparatuses, involved in ssDNA conjugation, were co-opted multiple times independently into T4SS secreting proteins into other cells. On the right, the rooted cladogram

represents the evolution of T4SS [18]. MPF stands for mating pair formation and includes the T4SS. (c) The T6SS resulted from the co-option of contractile tail phage genes (and their integration with other genes). This resulted in a contractile structure able to puncture eukaryotic or prokaryotic cells and deliver effectors, often toxins. On the right is presented a non-rooted cladogram of the history of T6SS [69]. MAC stands for metamorphosis- inducing structures. Afps stands for insecticidal anti-feeding prophages. The drawings of the systems are based on [38]. OM stands for outer-membrane, IM for inner (cytoplasmic) membrane; the periplasm is shown in brown between the IM and OM; “Entero” stands for enterobacteria.

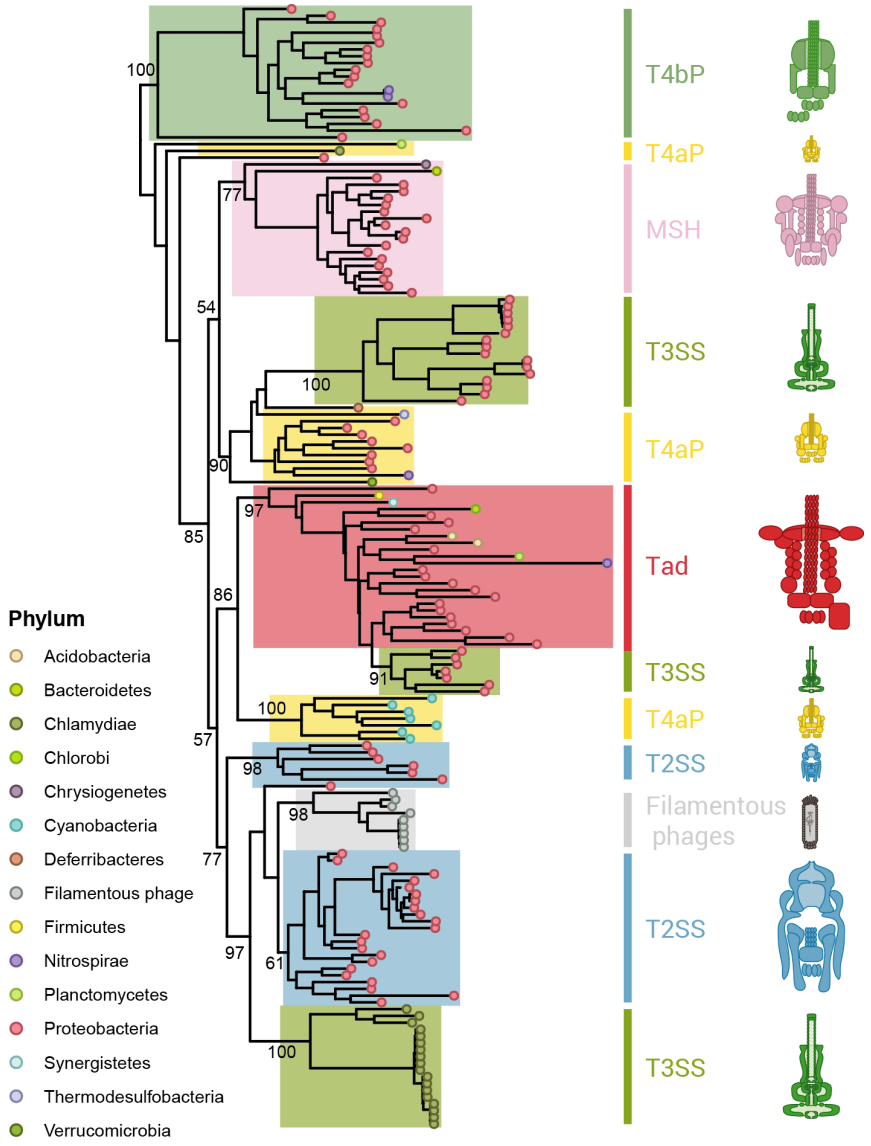


Figure 3 – Phylogeny of secretin proteins. The tree was built using the secretin domain of the protein sequences from [46], with the addition of T3SS and phage sequences from [39]. We aligned the sequences using MAFFT v7.273 (einsi algorithm) [96], selected informative sites in the multiple alignment using Noisy v1.5.12 [97] (default parameters), and inferred the maximum-likelihood tree from these alignments with IQ-TREE v1.6.7.2 [98] (using the best evolutionary model, options -MF, BIC criterion, -allnni, -ntop 1000, -nm 10000). Node supports displayed at nodes were estimated using the option -bb 1000 for ultrafast bootstraps [99]. The tree is consistent with the results of several previous studies. The root was positioned between T4bP and the remaining clades, as suggested elsewhere [39, 46, 92]. If correct, secretins were first components of the TFF superfamily and subsequently co-opted by phages

and T3SS (three times independently from Tad for Rhizobiales, from T2SS for Chlamydiae, and from T4aP for other proteobacterial T3SS). They were also recruited by the Tad upon the transfer of the ancestor of this system from Archaea. The color of circles at the tip of the tree corresponds to different bacterial phyla. The colors of groups and drawings on the right depict the different systems where secretins have been identified.

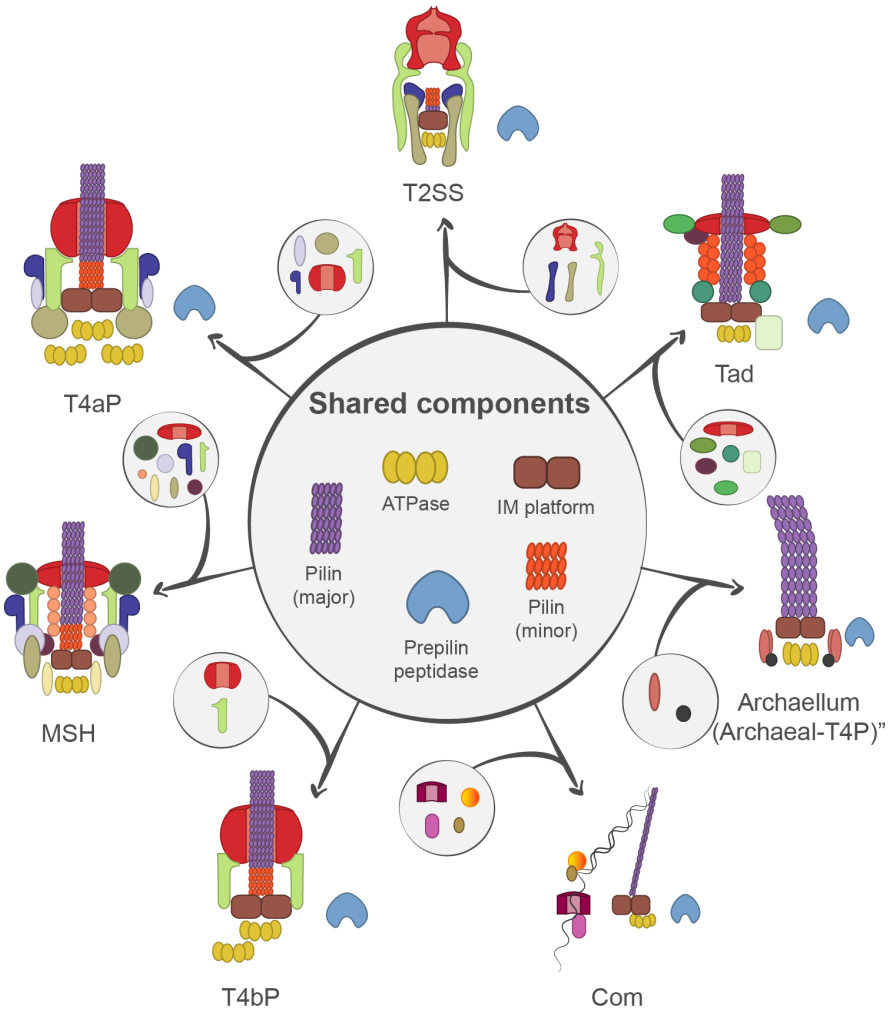


Figure 4 & KEY FIGURE – Diversification of the type IV filament (TFF) superfamily around a common set of components. The TFF superfamily diversified into many different systems using a few homologous core components (in the middle) and integrating some new ones. The different machines were able to diversify into functions as different as secretion of toxins, DNA uptake, motility, adhesion to surface. Components colored in the same way correspond to homologs. The drawings of the systems are based on [100].

References

1. Pal, C. and Papp, B. (2017) Evolution of complex adaptations in molecular systems. Nat Ecol Evol 1 (8), 1084-1092.

2. Lenski, R.E. (2017) Convergence and Divergence in a Long-Term Experiment with Bacteria. Am Nat 190 (S1), S57-S68.

3. Sorek, R. et al. (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. Science 318 (5855), 1449-52.

4. Jacob, F. (1977) Evolution and tinkering. Science 196, 1161-1166.

5. Ochman, H. et al. (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405 (6784), 299-304.

6. Wiedenbeck, J. and Cohan, F.M. (2011) Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev 35 (5), 957-76.

7. Touchon, M. et al. (2014) The chromosomal accommodation and domestication of mobile genetic elements. Curr Opin Microbiol 22, 22-29.

8. Gould, S.J. and Vrba, E.S. (1982) Exaptation-A Missing Term in the Science of Form. Paleobiology 8, 4-15.

9. Nogueira, T. et al. (2009) Horizontal Gene Transfer of the Secretome Drives the Evolution of Bacterial Cooperation and Virulence. Curr Biol 19, 1683-91.

10. Granato, E.T. et al. (2019) The Evolution and Ecology of Bacterial Warfare. Curr Biol 29 (11), R521-R537.

11. Desvaux, M. et al. (2009) Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. Trends Microbiol 17 (4), 139-45.

12. Dalbey, R.E. and Kuhn, A. (2012) Protein traffic in Gram-negative bacteria--how exported and secreted proteins find their way. FEMS Microbiology Reviews 36 (6), 1023-45.

13. Costa, T.R. et al. (2015) Secretion systems in Gram-negative bacteria: structural and mechanistic insights. Nat Rev Microbiol 13 (6), 343-59.

14. Xu, Q. et al. (2016) A Distinct Type of Pilus from the Human Microbiome. Cell 165 (3), 690-703.

15. Alvarez-Martinez, C.E. and Christie, P.J. (2009) Biological diversity of prokaryotic type IV secretion systems. Microbiol Mol Biol Rev 73, 775-808.

16. de la Cruz, F. et al. (2010) Conjugative DNA Metabolism in Gram-negative Bacteria. FEMS Microbiol Rev 34, 18-40.

17. Grohmann, E. et al. (2018) Type IV secretion in Gram-negative and Gram-positive bacteria. Mol Microbiol 107 (4), 455-471.

18. Guglielmini, J. et al. (2013) Evolution of Conjugation and Type IV Secretion Systems. Mol Biol Evol 30 (2), 315-331.

19. Ellison, C.K. et al. (2017) Obstruction of pilus retraction stimulates bacterial surface sensing. Science 358 (6362), 535-538.

20. Blesa, A. et al. (2017) The transjugation machinery of Thermus thermophilus: Identification of TdtA, an ATPase involved in DNA donation. PLoS Genet 13 (3), e1006669.

21. Ghinet, M.G. et al. (2011) Uncovering the Prevalence and Diversity of Integrating Conjugative Elements in Actinobacteria. PLoS ONE 6, e27846.

22. Trokter, M. and Waksman, G. (2018) Translocation through the Conjugative Type IV Secretion System Requires Unfolding of Its Protein Substrate. J Bacteriol 200 (6).

23. Souza, D.P. et al. (2015) Bacterial killing via a type IV secretion system. Nat Commun 6, 6453.

24. Suzuki, H. et al. (2010) Predicting plasmid promiscuity based on genomic signature. J Bacteriol 192 (22), 6045-55.

25. Draper, O. et al. (2005) Site-specific recombinase and integrase activities of a conjugative relaxase in recipient cells. Proc Natl Acad Sci U S A 102 (45), 16385-90.

26. Vogel, J. et al. (1998) Conjugative transfer by the virulence system of Legionella pneumophila. Science 279, 873-6.

27. Johnston, C. et al. (2014) Bacterial transformation: distribution, shared mechanisms and divergent control. Nat Rev Microbiol 12 (3), 181-96.

28. Gomez-Valero, L. et al. (2019) More than 18,000 effectors in the Legionella genus genome provide multiple, independent combinations for replication in human cells. Proc Natl Acad Sci U S A 116 (6), 2265-2273.

29. Christie, P.J. et al. (2017) Biological Diversity and Evolution of Type IV Secretion Systems. Curr Top Microbiol Immunol 413, 1-30.

30. Chaban, B. et al. (2015) The flagellum in bacterial pathogens: For motility and a whole lot more. Semin Cell Dev Biol 46, 91-103.

31. Young, G.M. et al. (1999) A new pathway for the secretion of virulence factors by bacteria: the flagellar export apparatus functions as a protein-secretion system. Proc Natl Acad Sci U S A 96 (11), 6456-61.

32. Konkel, M.E. et al. (2004) Secretion of virulence proteins from Campylobacter jejuni is dependent on a functional flagellar export apparatus. J Bacteriol 186 (11), 3296-303.

33. Scanlan, E. et al. (2017) A quantitative proteomic screen of the Campylobacter jejuni flagellar-dependent secretome. J Proteomics 152, 181-187.

34. Maezawa, K. et al. (2006) Hundreds of flagellar basal bodies cover the cell surface of the endosymbiotic bacterium Buchnera aphidicola sp. strain APS. J Bacteriol 188 (18), 6539-43.

35. Ferreira, J.L. et al. (2019) gamma-proteobacteria eject their polar flagella under nutrient depletion, retaining flagellar motor relic structures. PLoS Biol 17 (3), e3000165.

36. Ginocchio, C.C. et al. (1994) Contact with epithelial cells induces the formation of surface appendages on Salmonella typhimurium. Cell 76 (4), 717-24.

37. Pallen, M.J. and Matzke, N.J. (2006) From The Origin of Species to the origin of bacterial flagella. Nature Reviews Microbiology 4 (10), 784-790.

38. Galan, J.E. and Waksman, G. (2018) Protein-Injection Machines in Bacteria. Cell 172 (6), 1306-1318.

39. Abby, S.S. and Rocha, E.P. (2012) The non-flagellar type III secretion system evolved from the bacterial flagellum and diversified into host-cell adapted systems. PLoS Genetics 8 (9), e1002983.

40. Guttman, D.S. et al. (2006) Diversifying selection drives the evolution of the type III secretion system pilus of Pseudomonas syringae. Mol Biol Evol 23 (12), 2342-2354.

41. Troisfontaines, P. and Cornelis, G.R. (2005) Type III secretion: more systems than you think. Physiology 20, 326-39.

42. Nguyen, L. et al. (2000) Phylogenetic analyses of the constituents of Type III protein secretion systems. Journal of molecular microbiology and biotechnology 2 (2), 125-44.

43. Gophna, U. et al. (2003) Bacterial type III secretion systems are ancient and evolved by multiple horizontal-transfer events. Gene 312, 151-163.

44. Sun, G.W. and Gan, Y.H. (2010) Unraveling type III secretion systems in the highly versatile Burkholderia pseudomallei. Trends Microbiol 18 (12), 561-8.

45. Berry, J.L. and Pelicic, V. (2015) Exceptionally widespread nanomachines composed of type IV pilins: the prokaryotic Swiss Army knives. FEMS Microbiol Rev 39 (1), 134-54.

46. Denise, R. et al. (2019) Diversification of the type IV filament superfamily into machines for adhesion, protein secretion, DNA uptake, and motility. PLoS Biol 17 (7), e3000390.

47. Tomich, M. et al. (2007) The tad locus: postcards from the widespread colonization island. Nature Reviews. Microbiology 5 (5), 363-375.

48. Cianciotto, N.P. and White, R.C. (2017) Expanding Role of Type II Secretion in Bacterial Pathogenesis and Beyond. Infect Immun 85 (5).

49. Makarova, K.S. et al. (2016) Diversity and Evolution of Type IV pili Systems in Archaea. Front Microbiol 7, 667.

50. Roux, N. et al. (2012) Neglected but amazingly diverse type IVb pili. Res Microbiol 163 (9-10), 659-73.

51. Lopez-Castilla, A. et al. (2017) Structure of the calcium-dependent type 2 secretion pseudopilus. Nat Microbiol 2 (12), 1686-1695.

52. DebRoy, S. et al. (2006) Legionella pneumophila type II secretome reveals unique exoproteins and a chitinase that promotes bacterial persistence in the lung. Proc Natl Acad Sci U S A 103 (50), 19146-51.

53. Korotkov, K.V. and Sandkvist, M. (2019) Architecture, Function, and Substrates of the Type II Secretion System. EcoSal Plus 8 (2).

54. Lawrence, J.G. and Roth, J.R. (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics 143, 1843-1860.

55. Mougous, J.D. et al. (2006) A virulence locus of Pseudomonas aeruginosa encodes a protein secretion apparatus. Science 312 (5779), 1526-30.

56. Hood, R.D. et al. (2010) A type VI secretion system of Pseudomonas aeruginosa targets a toxin to bacteria. Cell host & microbe 7 (1), 25-37.

57. Borgeaud, S. et al. (2015) The type VI secretion system of Vibrio cholerae fosters horizontal gene transfer. Science 347 (6217), 63-7.

58. Pukatzki, S. et al. (2007) Type VI secretion system translocates a phage tail spike-like protein into target cells where it cross-links actin. Proc Natl Acad Sci U S A 104 (39), 15508-15513.

59. Leiman, P.G. et al. (2009) Type VI secretion apparatus and phage tail-associated protein complexes share a common evolutionary origin. Proc Natl Acad Sci U S A 106 (11), 4154-9.

60. Pell, L.G. et al. (2009) The phage lambda major tail protein structure reveals a common evolution for long-tailed phages and the type VI bacterial secretion system. Proc Natl Acad Sci U S A 106 (11), 4160-5.

61. Logger, L. et al. (2017) Fusion Reporter Approaches to Monitoring Transmembrane Helix Interactions in Bacterial Membranes. Methods Mol Biol 1615, 199-210.

62. Lossi, N.S. et al. (2013) The HsiB1C1 (TssB-TssC) complex of the Pseudomonas aeruginosa type VI secretion system forms a bacteriophage tail sheathlike structure. J Biol Chem 288 (11), 7536-48.

63. Basler, M. et al. (2012) Type VI secretion requires a dynamic contractile phage tail-like structure. Nature 483 (7388), 182-6.

64. Cascales, E. (2008) The type VI secretion toolkit. EMBO Rep 9 (8), 735-41.

65. Schwarz, S. et al. (2010) Burkholderia type VI secretion systems have distinct roles in eukaryotic and bacterial cell interactions. PLoS Pathogens 6 (8), e1001068.

66. Ludu, J.S. et al. (2008) The Francisella pathogenicity island protein PdpD is required for full virulence and associates with homologues of the type VI secretion system. J Bacteriol 190 (13), 4584-95.

67. Russell, A.B. et al. (2014) Type VI secretion system effectors: poisons with a purpose. Nat Rev Microbiol 12 (2), 137-48.

68. Russell, A.B. et al. (2014) A type VI secretion-related pathway in Bacteroidetes mediates interbacterial antagonism. Cell Host Microbe 16 (2), 227-236.

69. Bock, D. et al. (2017) In situ architecture, function, and evolution of a contractile injection system. Science 357 (6352), 713-717.

70. Bobay, L.M. et al. (2014) Pervasive domestication of defective prophages by bacteria. Proc Natl Acad Sci U S A 111, 12127–12132.

71. Komano, T. et al. (2000) The transfer region of IncI1 plasmid R64: similarities between R64 tra and legionella icm/dot genes. Mol Microbiol 35 (6), 1348-59.

72. Jurcisek, J.A. et al. (2017) Nontypeable Haemophilus influenzae releases DNA and DNABII proteins via a T4SS-like complex and ComE of the type IV pilus machinery. Proc Natl Acad Sci U S A 114 (32), E6632-E6641.

73. Pal, R.R. et al. (2019) Pathogenic E. coli Extracts Nutrients from Infected Host Cells Utilizing Injectisome Components. Cell 177 (3), 683-696 e18.

74. Spitz, O. et al. (2019) Type I Secretion Systems-One Mechanism for All? Microbiol Spectr 7 (2).

75. Rouse, S.L. et al. (2017) A new class of hybrid secretion system is employed in Pseudomonas amyloid biogenesis. Nat Commun 8 (1), 263.

76. Morlot, C. and Rodrigues, C.D.A. (2018) The New Kid on the Block: A Specialized Secretion System during Bacterial Sporulation. Trends Microbiol 26 (8), 663-676.

77. Gomez-Santos, N. et al. (2019) A TonB-dependent transporter is required for secretion of protease PopC across the bacterial outer membrane. Nat Commun 10 (1), 1360.

78. Lauber, F. et al. (2018) Type 9 secretion system structures reveal a new protein transport mechanism. Nature 564 (7734), 77-82.

79. McBride, M.J. (2019) Bacteroidetes Gliding Motility and the Type IX Secretion System. Microbiol Spectr 7 (1).

80. Groschel, M.I. et al. (2016) ESX secretion systems: mycobacterial evolution to counter host immunity. Nat Rev Microbiol 14 (11), 677-691.

81. Pallen, M. et al. (2002) Bacterial FHA domains: neglected players in the phospho-threonine signalling game? Trends Microbiol 10 (12), 556-63.

82. Blocker, A. et al. (1999) The tripartite type III secreton of Shigella flexneri inserts IpaB and IpaC into host membranes. The Journal of Cell Biology 147 (3), 683-693.

83. Hachani, A. et al. (2016) Type VI secretion and anti-host effectors. Curr Opin Microbiol 29, 81-93.

84. Kwak, M.J. et al. (2017) Architecture of the type IV coupling protein complex of Legionella pneumophila. Nat Microbiol 2, 17114.

85. Jamet, A. and Nassif, X. (2015) New players in the toxin field: polymorphic toxin systems in bacteria. MBio 6 (3), e00285-15.

86. Matsuda, S. et al. (2019) Export of a Vibrio parahaemolyticus toxin by the Sec and type III secretion machineries in tandem. Nat Microbiol 4 (5), 781-788.

87. Swain, D.M. et al. (2017) A prophage tail-like protein is deployed by Burkholderia bacteria to feed on fungi. Nat Commun 8 (1), 404.

88. Tobe, T. et al. (2006) An extensive repertoire of type III secretion effectors in Escherichia coli O157 and the role of lambdoid phages in their dissemination. Proc Natl Acad Sci U S A 103 (40), 14941-6.

89. Wandersman, C. and Delepelaire, P. (1990) TolC, an Escherichia coli outer membrane protein required for hemolysin secretion. Proc Natl Acad Sci U S A 87 (12), 4776-80.

90. Marsh, J.W. and Taylor, R.K. (1998) Identification of the Vibrio cholerae type 4 prepilin peptidase required for cholera toxin secretion and pilus formation. Molecular Microbiology 29 (6), 1481-92.

91. Duboule, D. and Wilkins, A.S. (1998) The evolution of 'bricolage'. Trends Genet 14 (2), 54-9.

92. Nickerson, N.N. et al. (2012) A Single Amino Acid Substitution Changes the Self-Assembly Status of a Type IV Piliation Secretin. Journal of Bacteriology 194 (18), 4951-4958.

93. Korotkov, K.V. et al. (2011) Secretins: dynamic channels for protein transport across membranes. Trends Biochem Sci 36 (8), 433-43.

94. Crago, A.M. and Koronakis, V. (1998) Salmonella InvG forms a ring-like multimer that requires the InvH lipoprotein for outer membrane localization. Mol Microbiol 30 (1), 47-56.

95. Abby, S.S. et al. (2016) Identification of protein secretion systems in bacterial genomes. Sci Rep 6, 23080.

96. Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30 (4), 772-80.

97. Dress, A.W. et al. (2008) Noisy: identification of problematic columns in multiple sequence alignments. Algorithms Mol Biol 3, 7.

98. Nguyen, L.T. et al. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32 (1), 268-74.

99. Hoang, D.T. et al. (2018) UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol 35 (2), 518-522.

100. Korotkov, K.V. et al. (2012) The type II secretion system: biogenesis, molecular architecture and mechanism. Nature Reviews. Microbiology 10 (5), 336-51.

Troisième partie

Diversification de la TFF-SF

CHAPITRE 2

INTRODUCTION

2.1 Contexte

Les procaryotes produisent à la surface de leur membrane différents types de systèmes macromoléculaires. Ces appendices sont utilisés par la cellule dans différentes fonctions (motilité, sécrétion de toxines, échange d'ADN...). Du fait de leur grande diversité, ils permettent à la cellule de s'adapter au changement de leur environnement. Les appendices bactériens sont des exemples frappants de diversification fonctionnelle. Il s'agit de machineries macromoléculaires complexes, qui dépendent de nombreux gènes, englobant plusieurs compartiments cellulaires et qui peuvent évoluer vers de nouvelles fonctions. Par exemple, le système de sécrétion de protéines de type III (T3SS) a évolué à partir de l'appareil de sécrétion du flagelle bactérien [19], le T4SS à partir de l'appareil de conjugaison [242], et le T6SS a possiblement évolué à partir de la co-option des structures phagiques [243, 244].

Une illustration particulièrement remarquable de ces processus est fournie par la super-famille des systèmes bactériens et archéens qui comprend le système de sécrétion de protéines de type II (T2SS), le pilus de type IVa (T4aP), le pilus de type IVb (T4bP), le pilus mannose-sensitive hemagglutinin (MSH), le pilus à adhésion serrée (Tad), l'appareil de compétence (Com) et les pili de type IV dans les archées (Archaeal-T4P). Ces systèmes partagent plusieurs protéines homologues notamment des AAA+ ATPases (parmi lesquelles le T4aP PilT est le moteur moléculaire le plus puissant connu [245]), une plateforme de la membrane interne (IM plateforme) et une peptidase de prépiline qui permet la maturation un ensemble de pilines ou pseudo-pilines spécifiques [95]. Les bactéries didermes codent également une sécrétine qui forme un pore dans la membrane externe [246]. D'autres protéines de ces systèmes sont spécifiques de chaque système ou évoluent trop rapidement pour permettre d'inférer des relations d'homologie.

2.2 Diversité

La superfamille des filaments de type IV a été identifiée et étudiée expérimentalement chez de nombreuses bactéries et archées. Il a été montré que cette superfamille est présente chez virtuellement tous les clades de bactéries et d'archées du fait que l'on trouve des homologues des protéines composant les TFFs au sein de tous ces clades [83]. Cette analyse montre que les TFFs sont bien plus répandus qu'étudié précédemment [112]. Ceci est probablement la conséquence de leur extrême polyvalence fonctionnelle et du fait qu'un ancêtre des TFFs pourrait déjà être présent dans l'ancêtre commun aux bactéries et aux archées.

Il a été suggéré que les membres des TFFs sont encodées dans les génomes d'environ 1800 espèces différentes [83] et qu'on trouve des T2SS, des T4aP et des Tad chacun dans environ un tiers des génomes de la base de données du NCBI Refseq (Novembre 2013) [152]. Ces espèces couvrent la plupart des phyla bactériens et archéens. Ces résultats sont renforcés par le fait que ces espèces présentent simultanément des gènes qui codent pour une peptidase de prépiline, une ATPase de trafic et une protéine de membrane cytoplasmique.

La diversification fonctionnelle de la super-famille n'est donc pas spécifique aux clades puisque différents types de systèmes sont présents dans les mêmes clades. Cela suggère un transfert horizontal fréquent et/ou une origine ancienne de la super-famille. Le T4aP et le Tad pilus se trouvent dans la plupart des phyla [112, 154] et les Archaeal-T4Ps dans la plupart des archées [99]. Les T2SS, T4bP et MSH n'ont été décrits que dans les didermes [152, 164]. La répartition de l'appareil de compétence est mal connue car dans les didermes un T4aP est souvent nécessaire mais pas suffisant pour la compétence [76]. En résumé, le TFF-SF s'est diversifié en plusieurs fonctions différentes par des procédés de co-option utilisant un ensemble commun de composants homologues identifiables dans les procaryotes.

2.3 Objectifs

Des adaptations complexes peuvent se produire par une série de petites étapes adaptatives. Par exemple, les réseaux métaboliques évoluent par étapes pour s'adapter à de nouvelles réactions [247]. L'innovation peut également résulter de processus de néo-fonctionnalisation ou de sous-fonctionnalisation suite à la duplication de gènes codant des protéines à fonctions multiples ou à l'acquisition, par transfert horizontal, de systèmes génétiques homologues (ces fonctions acquises par transfert peuvent ou non avoir des homologues dans le génome receveur). La façon dont ces processus façonnent l'évolution des complexes macromoléculaires demeure mal connue.

Les études consacrées à l'évolution des AAA+ ATPases, Tad, T4aP et T2SS datent de la décennie précédente [20,154,248,249], lorsque les données étaient rares et les méthodes phylogénétiques moins sophistiquées. Les systèmes archéens ont été étudiés en détail récemment [99, 223], mais indépendamment de l'évolution des

systèmes bactériens. Des travaux plus récents n'ont étudié que brièvement les phylogénies de certains des composants de ces systèmes [46]. Il est important de noter le manque d'études intégrant tous les systèmes et toutes les données génomiques disponibles, qui sont des prérequis pour comprendre les processus de diversification fonctionnelle de la super-famille. Ici, nous avons identifié les systèmes typiques de la TFF-SF et leurs variantes en utilisant des outils d'annotation spécifiques sur tous les génomes complets de procaryotes. Ces systèmes ont été analysés à l'aide de techniques phylogénétiques pour caractériser l'histoire de la TFF-SF, clarifier les relations entre ses membres et déchiffrer les mécanismes d'évolution moléculaire sous-jacents à sa diversification fonctionnelle. Enfin, nous avons caractérisé leurs organisations génétiques et leurs relations avec les taux de transfert horizontal de gènes. Cette analyse intégrative a fourni un scénario complet pour la diversification de la super-famille impliquant des processus de duplication, de fission, de transfert, d'accrétion et de diversification des séquences.

CHAPITRE 3

MÉTHODES ET RESULTATS

3.1 Article 2 : Diversification of the type IV filament super-family into machines for adhesion, protein secretion, DNA uptake and motility

PLOS BIOLOGY

RESEARCH ARTICLE

Diversification of the type IV filament superfamily into machines for adhesion, protein secretion, DNA uptake, and motility

Rémi Denise[1,2]*, **Sophie S. Abby**[3‡], **Eduardo P. C. Rocha**[1‡]

1 Microbial Evolutionary Genomics, Institut Pasteur, CNRS, UMR3525, Paris, France, **2** Sorbonne Université, Collège doctoral, Paris, France, **3** Université Grenoble Alpes, CNRS, Grenoble INP, TIMC-IMAG, Grenoble, France

‡ These authors are joint senior authors on this work.
* remi.denise@gmail.com

OPEN ACCESS

Citation: Denise R, Abby SS, Rocha EPC (2019) Diversification of the type IV filament superfamily into machines for adhesion, protein secretion, DNA uptake, and motility. PLoS Biol 17(7): e3000390. https://doi.org/10.1371/journal.pbio.3000390

Academic Editor: Morgan Beeby, Imperial College London, UNITED KINGDOM

Received: March 13, 2019

Accepted: July 3, 2019

Published: July 19, 2019

Data Availability Statement: The data produced by these analyses are available at https://gitlab.pasteur.fr/rdenise/diversification-of-tff-sf-data, accessed March 2019.

Funding: Doctoral school Complexité du vivant (ED515) (contract number 2449/2016) (http://www.ed515.upmc.fr/fr/index.php); INCEPTION project (PIA/ANR-16-CONV-0005) (https://research.pasteur.fr/en/program_project/inception/) The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Processes of molecular innovation require tinkering and shifting in the function of existing genes. How this occurs in terms of molecular evolution at long evolutionary scales remains poorly understood. Here, we analyse the natural history of a vast group of membrane-associated molecular systems in Bacteria and Archaea—the type IV filament (TFF) superfamily—that diversified in systems involved in flagellar or twitching motility, adhesion, protein secretion, and DNA uptake. The phylogeny of the thousands of detected systems suggests they may have been present in the last universal common ancestor. From there, two lineages—a bacterial and an archaeal—diversified by multiple gene duplications, gene fissions and deletions, and accretion of novel components. Surprisingly, we find that the 'tight adherence' (Tad) systems originated from the interkingdom transfer from Archaea to Bacteria of a system resembling the 'EppA-dependent' (Epd) pilus and were associated with the acquisition of a secretin. The phylogeny and content of ancestral systems suggest that initial bacterial pili were engaged in cell motility and/or DNA uptake. In contrast, specialised protein secretion systems arose several times independently and much later in natural history. The functional diversification of the TFF superfamily was accompanied by genetic rearrangements with implications for genetic regulation and horizontal gene transfer: systems encoded in fewer loci were more frequently exchanged between taxa. This may have contributed to their rapid evolution and spread across Bacteria and Archaea. Hence, the evolutionary history of the superfamily reveals an impressive catalogue of molecular evolution mechanisms that resulted in remarkable functional innovation and specialisation from a relatively small set of components.

Introduction

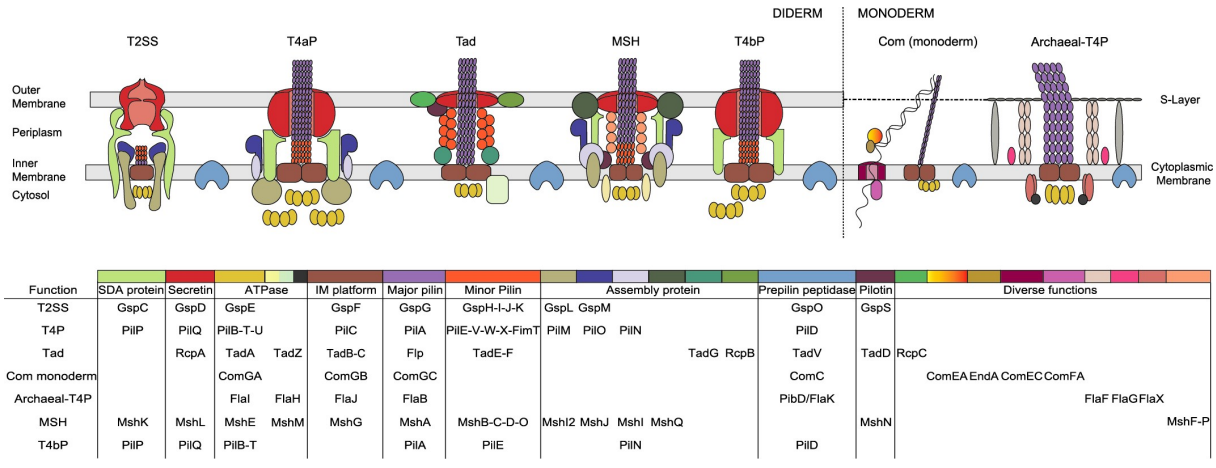
New complex forms, functions, and molecular systems arise by the shift in function (co-option) of elements that may have evolved to tackle different adaptive needs [1]. At the molecular level, this involves tinkering with pre-existing molecular structures by diverse processes,

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

Abbreviations: Aap, Archaeal adhesive pilus; ABC, ATP-binding cassette; Archaeal-T4P, type IV-related pili in Archaea; arCOG, archaeal Cluster of Orthologous Genes; AU, Approximately Unbiased; Bas, Bindosome; Ced, Crenarchaeal system for exchange of DNA; Com, competence pilus; ComM, Com of monoderms; DTL, duplication, transfer, or loss; Epd, EppA dependent; HMM, hidden Markov model; IM, integral membrane; MGR, minimum genes required; MMGR, minimum mandatory genes required; MSH, mannose-sensitive hemagglutinin pilus; SDA, secretin-dynamic–associated; Tad, tight adherence; TFF, type IV filament; T2SS, type II protein secretion system; T3SS, type III protein secretion system; T4aP, type IVa pilus; T4bP, type IVb pilus; Ups, UV-inducible pilus of *Sulfolobus*; UFBoot, Ultrafast Bootstrap Approximation.

including mutation, recombination, and gene fusion and fission [2]. These variants are ultimately subject to natural selection and may eventually become fixed in populations [3]. In Bacteria and Archaea, this is facilitated by the constant income of novel genetic information by horizontal gene transfer [4–6]. Complex adaptations can evolve through a series of small adaptive steps. E.g., metabolic networks evolve stepwise to accommodate novel reactions at their edges [7]. Innovation may also arise by processes of neofunctionalisation or subfunctionalisation following the duplication of genes encoding proteins with multiple functions or the acquisition by horizontal transfer of homologous genetic systems. How these processes shape the evolution of macromolecular complexes remains poorly known.

The appendages of Bacteria and Archaea are striking examples of functional diversification. They are complex macromolecular machineries encoded by many genes and spanning several cellular compartments that can evolve towards novel functions. E.g., the type III protein secretion system (T3SS) evolved from the secretion apparatus of the bacterial flagellum [8], the type IV secretion system (T4SS) from the conjugation apparatus [9], and the type VI secretion system (T6SS) possibly from co-option of phage structures [10,11]. A particularly remarkable illustration of these processes is provided by the type IV filament (TFF) superfamily of bacterial and archaeal systems that include the type II secretion system (T2SS), the type IVa pilus (T4aP), the type IVb pilus (T4bP), the mannose-sensitive hemagglutinin pilus (MSH), the tight adherence (Tad) pilus, the competence pilus (Com), and the type IV-related pili in Archaea (Archaeal-T4P), which includes the archaeal flagella (archaellum). These systems have core homologous components, sometimes in multiple copies, and present similarities in terms of macromolecular architecture throughout Bacteria and Archaea (Fig 1) [12–14]. They include AAA+ ATPases, among which the T4aP PilT is the most powerful molecular motor known [15]; an integral (cytoplasmic) membrane (IM) platform; and a prepilin peptidase that matures a set of specific pilins or pseudopilins (in T2SS) [16]. Bacteria with two cell membranes (diderms) also encode a secretin that forms an outer-membrane pore [17]. Other



Function	SDA protein	Secretin	ATPase		IM platform	Major pilin	Minor Pilin	Assembly protein								Prepilin peptidase	Pilotin	Diverse functions								
T2SS	GspC	GspD	GspE		GspF	GspG	GspH-I-J-K	GspL	GspM							GspO	GspS								
T4P	PilP	PilQ	PilB-T-U		PilC	PilA	PilE-V-W-X-FimT	PilM	PilO	PilN						PilD									
Tad		RcpA	TadA	TadZ	TadB-C	Flp	TadE-F							TadG	RcpB	TadV	TadD	RcpC							
Com monoderm			ComGA		ComGB	ComGC										ComC		ComEA	EndA	ComEC	ComFA				
Archaeal-T4P			FlaI	FlaH	FlaJ	FlaB										PibD/FlaK							FlaF	FlaG	FlaX
MSH	MshK	MshL	MshE	MshM	MshG	MshA	MshB-C-D-O	MshI2	MshJ	MshI	MshQ						MshN								MshF-P
T4bP	PilP	PilQ	PilB-T			PilA	PilE			PilN						PilD									

Fig 1. Schematic representation of the different systems and associated genes. Homologous components are represented in the same colour. The table below the drawing indicates the colour code and the name of the different components in each type of system. For the Archaeal-T4P, the representation of the systems is based on the representation of the archaellum, and the genes mentioned in the legend are the names of the genes used in the literature (not the arCOG database's names). Some systems have multiple homologues of the ATPase, and these are shown as multiple clusters in the figure (with same shape and colour). Archael-T4P, type IV-related pili in Archaea; arCOG, archaeal Cluster of Orthologous Genes; Com, competence pilus; IM, integral membrane; MSH, mannose-sensitive hemagglutinin pilus; SDA, secretin-dynamic–associated; Tad, tight adherence; TFF, type IV filament; T2SS, type II protein secretion system; T4aP, type IVa pilus; T4bP, type IVb pilus.

https://doi.org/10.1371/journal.pbio.3000390.g001

proteins of these systems are either specific for each system or evolve too fast to allow the inference of homology among all variants.

The TFF nanomachines assemble filaments composed of subunits with an N-terminal sequence motif named class III signal peptide, generically named type IV pilins [14]. These systems are involved in functions typically associated with extracellular pili in Bacteria and Archaea, including adherence, cell–cell attachment, and the formation of biofilms [18–20]. They are exploited by phages for cell infection [21]. T4aP, T2SS, T4bP, and Tad are also important virulence factors in pathogenic Bacteria [22–27]. Nevertheless, and in spite of their homology, TFFs have evolved specific biological functions. T4aP and T4bP allow Bacteria to move by twitching motility (a form of surface movement promoted by repeated cycles of extension–retraction of the pilus) [28,29]. T2SS secrete proteins from the periplasm across the outer membrane [16]. Some Com, T4aP, and Archaeal-T4P facilitate the uptake of extracellular DNA into the cell [30,31]. In Bacteria, these systems are by far the most frequent appendages involved in natural transformation [31], the exception being *Helicobacter*, which use a system derived from a T4SS [32]. Archaeal-T4P include the archaellum involved in motility by rotation of the appendage, extracellular structures involved in sugar uptake (Bindosome or Bas), the UV-inducible pilus of *Sulfolobus* (Ups) involved in establishing cell–cell contacts to enable DNA repair under stress conditions, and several pili with poorly characterised functions [33–35].

The functional diversification of the superfamily is not clade-specific because different types of systems are present in the same clades. This suggests frequent horizontal transfer and/or an ancient origin of the superfamily. T4aP and the Tad pilus can be found in most bacterial phyla [36,37], and Archaeal-T4P in most Archaea [35]. The T2SS, T4bP, and MSH have only been described in diderms [38,39]. The distribution of the TFFs involved in competence is poorly known because different types may be involved in the process, in which they have a necessary but not sufficient role [31], and still keep additional functions associated with motility or adhesion. In summary, the TFF superfamily has diversified into several different functions by co-option processes using a common set of homologous components identifiable across Bacteria and Archaea.

Previous studies dedicated to the evolution of the AAA+ ATPases, Tad, T4aP, and T2SS date from the previous decade [12,37,40,41], when data were scarce and phylogenetic methods less sophisticated. Archaeal systems were studied in detail recently [35,42] but independently of the evolution of bacterial systems. More recent works only briefly studied the phylogenies of some of the components of these systems [43]. Importantly, there is a lack of studies integrating all the systems and all available genomic data, a prerequisite to understand the processes of functional diversification of the superfamily. Here, we identified the typical TFFs and their variants using specific annotation tools on all complete genomes of Bacteria and Archaea. These systems were analysed using phylogenetic techniques to characterise the history of the TFF superfamily, clarify the relationships among its members, and decipher the molecular evolution mechanisms underlying its functional diversification. Finally, we characterised their genetic organisations and how they relate to the rates of horizontal gene transfer. This integrative analysis provided a consistent scenario for the diversification of the superfamily involving processes of gene duplication, fission, transfer, accretion, and mutation.

Results

Relations of homology between the key components of the machineries

We started our study by building MacSyFinder models [44] for the identification of TFFs in the complete genomes of Bacteria and Archaea. Briefly, these models give a detailed account of

the genetic composition and organisation of the systems. We adapted previously published models of T4aP (including the Com systems of diderms), T2SS, and Tad [39,45], in which we incorporated additional components and stricter rules in terms of genetic composition and organisation to identify TFFs with high stringency for the initial phylogenetic and genomics analyses (S1 Fig). We used the literature to produce equivalent models and associated hidden Markov model (HMM) protein profiles for the Com of monoderms (ComM) and for the Archaeal-T4P. For the latter, we used 66 archaeal Cluster of Orthologous Genes (arCOGs) identified from [35] after a step of reanalysis of the initial 191 arCOGs to remove redundancy. We could not build models for T4bP and MSH systems at this point because too few systems were described in the literature. This work resulted in five initial models, including 154 HMM protein profiles, of which 17 are novel (S1 Table).

To establish the relations of homology between the components of the different systems in a precise and homogeneous manner, we made pairwise profile–profile alignments of their HMM protein profiles using HHsearch v3.0.3 [46] (p-value threshold of 0.001). These alignments are very sensitive and highlight more distant relations of homology than typical sequence alignment methods [46]. We obtained a graph with 10 components (sets of connected nodes), representing the significant relations of reciprocal similarity between the profiles (Fig 2). The five largest components include the proteins known to be homologous and represent each individual key function: secretins, prepilin peptidases, ATPases, IM platforms, and pilins (major and minor). The ATPase component includes TadZ, a protein from another subfamily of P-loop ATPases (FleN) with an atypical Walker-A motif that retains ATP binding capacity while displaying low ATPase activity [47,48]. It localises at the pole at early stages of pili biogenesis and functions as a hub for recruiting other Tad pili components, contrary to the ATPases involved in pilus assembly or retraction.

These results establish a precise and extensive network of sequence similarity between the key components of TFFs, systematising previous descriptions. The largest component of the graph includes the major and minor pilins, which are small and very diverse across the TFF superfamily. Their profile–profile alignments suggest they are all evolutionarily related. The remaining components were smaller and usually revealed at most one component per TFF family.

The phylogenies of the components of the TFF superfamily

The presence of homologues of the major functional components of the TFFs across most types of systems raises the question of how their functional diversification took place from a common ancestor. To study this, we added to the models described above a very simple generic model to identify all systems with three key components (the ATPase, the IM platform, and a major pilin). Accessorily, it also searches for a secretin, absent in monoderms, and a prepilin peptidase, sometimes shared between systems [49–51] (S1 Fig). The search for systems using the MacSyFinder models resulted in the identification of 6,652 systems in 3,700 genomes (1,486 species) (S2 Fig), of which 1,584 were classed as generic systems, reflecting the conservative character of the initial models. This data set was too large to analyse using sophisticated phylogenetic methods and included many systems that were very similar, e.g., from different strains of the same species. We reduced this redundancy by clustering very similar systems. We then picked one representative per cluster, thus preserving most of the diversity of the data set. In this process, we prioritised the inclusion of experimentally validated systems, including MSH (1) and T4bP (5), for which models were not available (see Methods). This nonredundant set contains 309 representative systems (33 T4aP, 47 Archaeal-T4P, 29 T2SS, 5 T4bP, 1 MSH, 31 ComM, 72 Tad, 101 generic) (S2 Table). Hence, the systems used in the subsequent

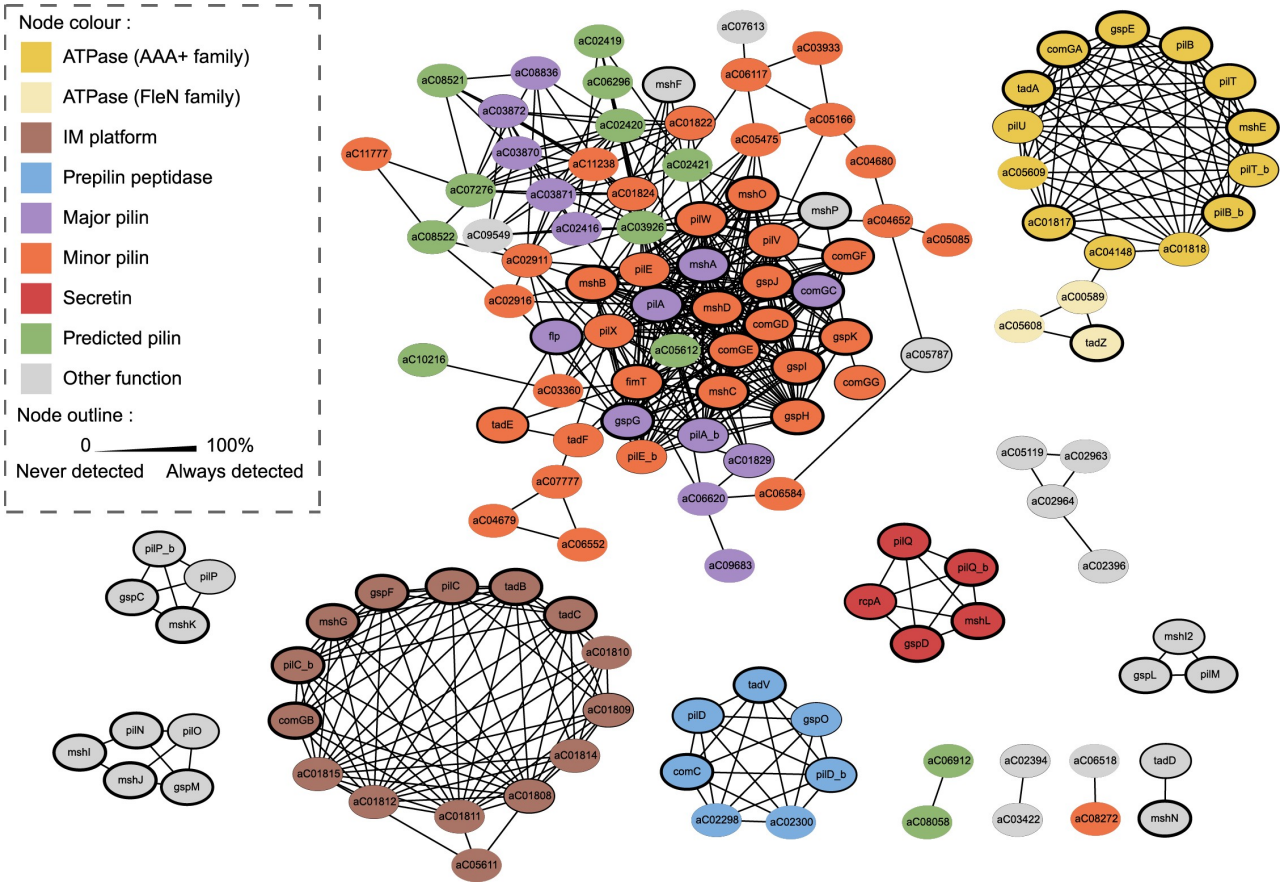


Fig 2. Results of the HMM–HMM alignments (HHSearch) between all the components of the TFF superfamily. The colour of the nodes represents the known or predicted function of the protein. The size of the outlines is proportional to the frequency of the profiles in the detected systems (thicker outlines indicate higher frequencies). Com, competence pilus; HMM, hidden Markov model; IM, integral membrane; MSH, mannose-sensitive hemagglutinin pilus; Tad, tight adherence; TFF, type IV filament.

https://doi.org/10.1371/journal.pbio.3000390.g002

analyses are associated with a (sometimes large) number of other very similar systems that are from the same cluster.

We inferred the phylogeny of each of the five core protein components (AAA+ ATPase, IM platform, major pilin, secretin, and prepilin peptidase) by maximum likelihood with IQ-Tree [52]. We made 10 reconstructions per component with the most thorough mode of topological search to account for the stochasticity of the method. The detailed analysis of key events revealed by these trees can be found in S3 Table (the trees themselves are in S4 Table). The ATPase trees are very well supported at most of the key nodes, they are consistent across replicated inferences, and they clearly separate the different types of systems (S3 Fig). The trees include two system-specific duplication events of the ATPases, one ancestral to the large clade —including T4aP, T4bP, MSH, T2SS, and ComM (PilT/PilB)—and another within a clade of T4aP (PilT/PilU). The IM platform tree also discriminates between types of systems and includes pairs of homologues in Tad (TadB/TadC) and some archaeal systems (S6 Fig). The prepilin peptidase tree is poorly supported and shows scattered distribution of the different

types of systems (S8 Fig). Because prepilin peptidases can be exchanged between systems [49–51], we have excluded them from further analyses. The secretin and major pilin trees have some poorly supported branches, but they separate the different systems well. Overall, the protein components' trees show that the ATPase, the IM platform, and the major pilin are good phylogenetic markers for the evolution of the TFF superfamily. The secretin tree, even if relatively well supported, is less informative for inferring the global evolutionary scenario because the component is absent from monoderms.

The root of the TFF superfamily

The ATPase tree is the only one that can be rooted because this is the only ubiquitous component with well-conserved homologues in distinct machineries that can serve as external groups [40,41]. We used FtsK as an outgroup to root the tree because it is very conserved, single-copy, present and essential in most bacterial phyla [53,54], and shows little evidence of horizontal transfer [41]. Its closest homologue, HerA, is an archaeal protein from which it diverged concomitantly with the archaeal–bacterial division after the last universal common ancestor [41]. We retrieved the sequences of FtsK from a previous study [9], aligned them with the ATPase sequences of the investigated systems, and inferred a maximum likelihood tree. This tree shows that the FtsK sequences are monophyletic (100% Ultrafast Bootstrap Approximation [UFBoot] support) and branch between two large clades: Tad and Archaeal-T4P on one side (100% UFBoot) and a clade grouping the T2SS, T4aP, ComM, and T4bP on the other side (100% UFBoot) (S4 Fig). The overall rooted topology is very similar to that of the unrooted tree in 8 out of 10 trees (S3 Table). The inclusion of the ubiquitous ATPase of the T4SS (VirB4) as an outgroup with FtsK also showed a split between the archaeal and the bacterial branches of the tree (S5 Fig). This confirms that this ATPase family is also an outgroup of the TFF superfamily. We rooted the trees of the IM platform and major pilin using the root of the ATPase trees because all three proteins showed a consistent split between Tad/Archaeal-T4P on one side and the remaining systems on the other (S6 and S7 Figs).

The analysis of gene duplications provides additional information on the possible roots of the superfamily phylogenetic tree. Placing a duplication event on a tree corresponds to setting as anterior the branch in which the duplication occurred, and as posterior, those of the two paralogues [55,56]. The duplications of the ATPases therefore exclude the root from the group T4aP, T4bP, ComM, MSH, and T2SS. The duplication of the IM platform in the Tad system, also present in some Archaeal-T4P, excludes the root from within these groups. Hence, the analyses of duplication events are consistent with the root as defined above by the tree of the ATPases.

Producing a concatenate tree

Because the ATPase and the IM platform have phylogenetic trees that are broadly consistent (S3 Table) and are the most informative markers of the phylogeny, we computed a phylogenetic tree of their concatenate using a partition model (best model for each gene partition, as computed by IQ-Tree). The major pilin was excluded from the concatenate because it shows less-consistent and less-supported topologies. Concatenation required the use of a procedure to deal with multiple homologues in the same system (to have one marker per component per system). For those present in a few taxa, we chose in each system the protein most similar in sequence to the most closely related systems lacking paralogues (see Methods). For the ATPases, we used PilB because this ATPase is responsible for the assembly of the pilus, which is an essential function in all families, contrary to the function of PilT/PilU (retraction). There was no good argument to pick TadB or TadC platform proteins, and we therefore made 10

phylogenetic reconstructions with each of them in parallel in ATPase/IM platform concatenates. As expected, the best trees (highest log likelihood) for the two TadB/PilB and TadC/PilB concatenates were never rejected by the individual proteins' alignments ($p > 0.05$, Approximately Unbiased [AU test], S7 Table). Furthermore, after correction for multiple comparisons, only two of the 40 comparisons between the individual proteins and the concatenate trees were significantly incongruent (TadB versus the two trees with lowest likelihood obtained for the TadB/PilB concatenate). We present in Fig 3 the highest log-likelihood tree obtained for the TadC/PilB concatenate (the combination of markers with no significant conflict with the gene trees). Overall, these concatenate trees show that the TFF families derived from an ancestral system, which diversified initially into an archaeal system ancestor of Tad/Archaeal-T4P and a bacterial system ancestor of the remaining TFFs.

The archaeal systems and the emergence of Tad

The ATPase, IM platform, and concatenate trees are broadly consistent with five or six groups within Archaea (Fig 3, S3, S6 and S10), several of which replicate previous findings (groups archaellum, Halo pilus, Epd, 'Adhesive archaeal pilus' [Aap], Bas/Ups [35]). All experimentally validated archaella are part of a highly supported clade (100% UFBoot, group 3 in [35]) that is the sister clade to another highly supported clade containing two pili involved in surface adhesion in Halobacteria (Halo pilus, group 2 in [35]). They are sister groups of a clade gathering the Bas, the Ups, and noncharacterised pili from Crenarchaeota and Thaumarchaeaota (group 4 in [35]). Aap cluster with the Halo pilus in the concatenate TadC/PilB and group apart closer to the Bas and the Ups pilus in other trees. The rooted tree shows two basal clades of Archaeal-T4P systems of unknown function, mostly found in methanogens (group 1 from [35], which is separated by the root in our tree).

Unexpectedly, the position of the root places Tad as a system derived from Archaeal-T4P systems. This feature is found in the trees of ATPase, IM platform, and major pilin with high confidence. Furthermore, all these trees showed a monophyletic clade, including the Tad and the 'EppA-dependent' (Epd) pilus (clade 'Epd-like'), whose major pilins have similarly short sequence lengths when compared to the others from Archaeal-T4P (S7 Fig). Both Epd-like pili and Tad have two homologous genes encoding the IM platform, suggesting that their common ancestor already contained them both. We examined the domain structure of these two genes and found that each has one 'T2SSF' domain (PFAM domain PF00482), whereas most other Archaeal-T4Ps have two such domains and longer IM platform proteins. This strongly suggests that TadB and TadC were derived from an ancestral event of gene fission and not a duplication as previously suggested. To confirm this observation, we aligned the TadB and TadC profiles with the archaeal IM platforms containing two T2SSF domains. In these cases, TadC aligned best with the N-terminal domain, while TadB aligned best with the C-terminal domain of the archaeal proteins. To further test the gene fission scenario, we made a tree using the concatenate of TadC and TadB, and this tree was similar to the tree of the concatenate (S4 Table). Finally, the Tad systems have a protein, TadZ, that has significant HMM–HMM profile alignments with Archaeal-T4P components (arCOG00589 and arCOG05608), including those from the Epd-like clade (group 1 from [35]), but not with profiles from the bacterial systems. Altogether, these results strongly suggest that an ancestral Archaeal-T4P harbouring two genes encoding the IM platform diversified into Epd-like systems in Archaea and was transferred horizontally, apparently only once, to Bacteria, leading to the extant Tad systems.

The transfer of the system from Archaea to Bacteria was very ancient. Tad systems were frequently transferred among Bacteria since then (see below), and it is not possible to infer the precise bacterial taxa that acquired the original system. However, the Tad systems at the basis

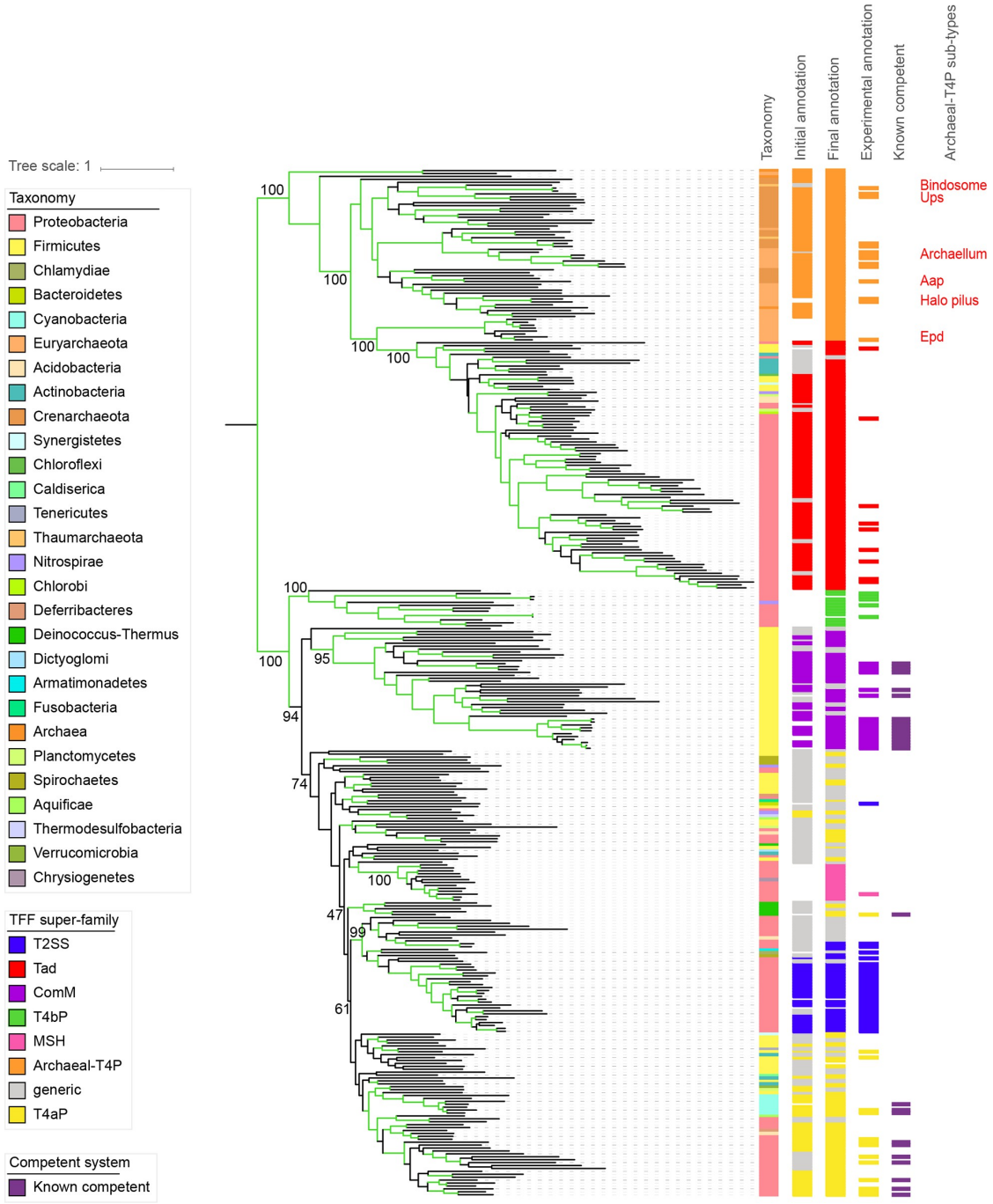


Fig 3. Rooted phylogeny of the TFF superfamily. The tree was built with the concatenate of the IM platform (using TadC) and the AAA+ ATPase (using PilB). The branches are in green if the ultrafast bootstrap is >95%. The supports of the significant nodes are indicated in text. The different coloured strips indicate the classification of the systems with the MacSyFinder annotation (with the initial model and with the final one) and the annotation of the systems in the literature. The systems known to be implicated in natural transformation are indicated in dark purple. Known subtypes of Archaeal-T4P are indicate by text in red. The tree was built using IQ-Tree, 10,000 replicates of UFBoot, with a partition model. Halo pilus

indicates two pili characterised in Halobacteria. Aap, adhesive archaeal pilus; Archaeal-T4P, type IV-related pili in Archaea; ComM, competence pilus of monoderms; Epd, EppA dependent; IM, integral membrane; MSH, mannose-sensitive hemagglutinin pilus; Tad, tight adherence; TFF, type IV filament; T2SS, type II protein secretion system; T4aP, type IVa pilus; T4bP, type IVb pilus; UFBoot, Ultrafast Bootstrap Approximation; Ups, UV-inducible pilus of *Sulfolobus*.

https://doi.org/10.1371/journal.pbio.3000390.g003

of the clade are from Proteobacteria in 18 out of 20 concatenate trees, often with very good support (S3 Table). The two odd concatenate trees place Firmicutes at the base of the Tad clade but with very low support. This suggests that the ancestor of the Tad system was acquired by a diderm bacterium, and the accretion of the outer-membrane, pore-forming secretin to the Tad system may have been the founding event of these systems.

The diversification of the bacterial TFF superfamily

The other major clade of the TFF superfamily only has bacterial systems (T4aP, T4bP, ComM, MSH, T2SS). The vast majority of the concatenate and component trees place T4bP at the basal position in the clade (in the others, some generic systems take this position). This is followed by a split between ComM on one side and T4aP, MSH, and T2SS on the other. T4aP are polyphyletic in all the phylogenetic reconstructions, showing a few clusters with the experimentally validated systems (Fig 3). Some of these systems are in monoderms such as Firmicutes and Actinobacteria, as previously observed [14,57]. MSH and T2SS are both clearly distinct and derived from the T4aP. The MSH system falls in a highly supported clade (100% UFBoot) with other systems of very similar gene composition. Intriguingly, all MSH loci lack a prepilin peptidase. They may use a protein from another system because MSH were systematically present in genomes with T2SS, T4aP, or Tad, which encode a prepilin peptidase. Systems previously identified as T2SS show two exceptions to monophyly. First, the position of chlamydial T2SS next to the other T2SS is highly supported in the ATPase and in the concatenate tree (>95% UFBoot) but not in the trees of the secretin, major pilin, and IM platform. This suggests a chimeric origin for this system in which different components were recruited from different types of systems. Second, the so-called T2SS of Bacteroidetes (represented by *Cytophaga*, [58]) always cluster with T4aP and away from the remaining T2SS.

The key early event in the ATPase trees of the Bacteria-only TFF large clade was the amplification leading to the paralogues PilB (the assembly ATPase) and PilT (the retraction ATPase). This event appears as a simple duplication at the base of the tree in certain of the ATPase trees but also shows more complex scenarios in others (S3 Table). In the PilB part of the ATPase tree, T4bP is basal, and the other systems are regrouped with T4aP. This scenario is consistent with that of the secretin tree, in which if one places the root between T4aP and T4bP, one finds T2SS deriving from a T4aP system, as in the PilB trees. This is also sustained, albeit with low support, by the major pilin tree, in which one finds at basal positions T4aP and T4bP. The presence of PilT in the early stages of evolution of the TFF superfamily could be an indication that the most ancient systems already had ATPases specialised in pilus retraction.

One of the most interesting functions of the superfamily, from the evolutionary point of view, is the involvement of some of its systems in natural transformation. The ComM system is commonly found in Firmicutes, even if it is unclear whether it is always involved in transformation. It is monophyletic in all the phylogenetic reconstructions we made, usually with very high support (≥95%). In the concatenate trees, ComM branches apart from a group gathering T4aP, MSH, and T2SS after the divergence with T4bP. The trees of individual components show similar scenarios once one accounts for the effects of the ATPase paralogues and for the low support of some parts of the IM platform trees. In summary, these results suggest that ComM arose early and only once in the history of the TFF superfamily. The T4aP systems

experimentally linked to natural transformation in diderms were systematically identified as T4aP and also tend to cluster together in the tree.

TFFs are ubiquitous in the prokaryotic world

We used the best concatenate tree, rooted using the information of the rooted ATPase tree, to class the numerous generic systems that we had previously identified. We assumed that clades in which all systems were either generic or of a single type (of which at least one was validated experimentally) could be tentatively assigned to that type. Generic systems in clades lacking experimentally validated systems were left unassigned. Only two types of systems were paraphyletic in the tree—T4aP and Archaeal-T4P—and were thus treated differently. T4aP was split in a few monophyletic clades, and systems within each clade were reassigned using the method above. The Archaeal-T4P systems, from which Tad derives, can be easily distinguished from the latter and thus reassigned using a taxonomic criterion. This analysis significantly clarified the systems' assignment (compare S2 Fig with S11 Fig): 1,795 out of the 2,031 generic systems were reassigned to classical systems, mostly T2SS (479) and T4aP (748).

We used these tentatively assigned systems to produce more sensitive MacSyFinder models. First, we changed the HMM profiles to account for the genetic diversity introduced by the reassigned systems. Second, we created models to detect T4bP and MSH because we now had a much larger number of examples of these systems. Finally, we searched for genes systematically associated with the systems' loci in a neighbourhood of ±20 genes that were not matched by any of the HMM profiles of the models. We clustered the proteins by sequence similarity and analysed the largest families. This 'guilt-by-association' approach failed to show other proteins systematically associated with a particular type of system (S5 Table), suggesting that our models already encompass their most frequent components. This process resulted in more sensitive models that accounted for all known types of systems and correctly identified the 94 experimentally validated systems of Bacteria analysed in S2 Table, except the T2SS of Chlamydia and Bacteroidetes (shown above to be peculiar).

Using the improved models, we found 9,026 systems within 4,610 genomes, including 1,728 T2SS, 2,021 Tad, 2,558 T4aP, 908 ComM, 559 Archaeal-T4P, 177 T4bP, 191 MSH, and 884 generic systems (Fig 4, S6 Table). A few systems classed in a given type with the initial conservative models—14 T2SS, 10 T4aP, 5 Tad, 1 ComM—are classed as generic with the new models. However, the inverse is much more frequent because we reclassified 1,114 generic systems as 1,408 T4aP, 338 T2SS, 670 Tad, 4 ComM, and 226 Archaeal-T4P. The large number of generic systems reassigned to T4aP is not surprising because these systems are encoded in multiple loci, are very diverse, and are present in several clades in the tree. This makes them harder to detect using the initial model. The many reassignments of generic systems as T2SS reflect a posteriori the excessive stringency of our initial model based on existing knowledge of systems in Proteobacteria and the existence of T2SS with little or no experimental evidence in other phyla. The reassignment led to identification of T2SS in a much broader set of taxa, including Armatimonadetes, Deferribacteres, Clostridia (from a clade known to contain diderms, further supported by the presence of a secretin), Spirochaetes [59], and Aquificae. We also observe many new Tad systems in Elusimicrobia, Actinobacteria, Bacilli, and Clostridia (Fig 4 versus S2 Fig). Our phylogenetics-driven approach for designing new models allowed us to detect diverse putative MSH and T4bP. These systems were so far only described as such in Gamma-proteobacteria, but we identified them also in Chrysiogenetes and Epsilonproteobacteria for MSH and in Acidithiobacillia and Nitrospirae for T4bP.

In certain cases, the phylogenetic annotation identified some systems that we missed using the improved, models and provides information to explain the large number of generic

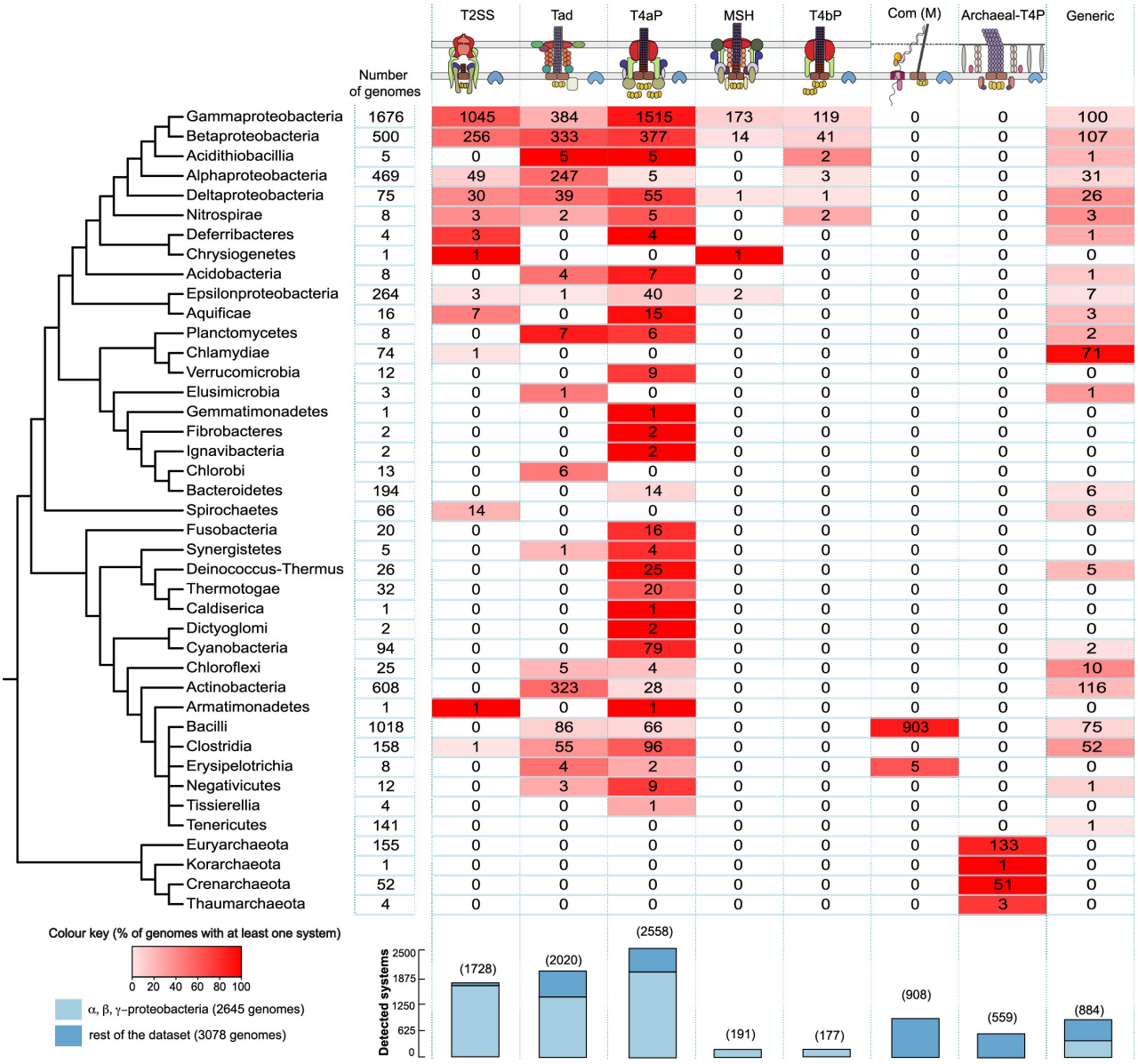


Fig 4. Taxonomic distribution of the systems in Bacteria and Archaea obtained using the final models. Cells indicate the number of genomes with at least one detected system. The cell's colour gradient represents the proportion of genomes with at least one system in the clade. The bar plot shows the total number of detected systems. The bars are separated in two categories: Alpha-, Beta-, and Gamma-proteobacteria versus the other clades. The cladogram symbolises approximated relationships between the bacterial and archaeal taxa analysed in this study. Archaeal-T4P, type IV-related pili in Archaea; Com, competence pilus; ComM, Com in monoderms; MSH, mannose-sensitive hemagglutinin pilus; Tad, tight adherence; T2SS, type II protein secretion system; T4aP, type IVa pilus; T4bP, type IVb pilus.

https://doi.org/10.1371/journal.pbio.3000390.g004

systems in certain clades. The T2SS in Chlamydiae [60] are close to the other T2SS for several phylogenetic markers but are classed as generic because they apparently lack homologues of the minor pilins and the assembly proteins GspLM [60]. Many of the systems of Actinobacteria remain classed as generic systems. A large fraction of them could be classed as Tad by

proximity to experimentally validated systems in the phylogeny, but they lack identifiable homologues of some usual components such as the minor pilins and TadC (their TadB does not contain two domains like those of homologues in some Archaea, showing this is not the result of a gene fusion).

Genetic organisation is associated with differences in rates of horizontal transfer

The systems differ strikingly in terms of genetic organisation (Fig 5). ComM and T4aP are usually found in multiple loci, whereas MSH and Tad are almost exclusively encoded in a single locus. This characteristic further contributes to set MSH apart from the remaining T4aP. Hence, as systems diverged, their genetic organisation also changed. To detail the prototypical genetic organisations of each type of system, we built a graph on which nodes represent components and edges link components that are encoded contiguously in the genome. The edges are weighted by the frequency of contiguity: genes that are systematically contiguous are linked by thick edges. This graph quantifies the prevailing genetic organisations for most types of

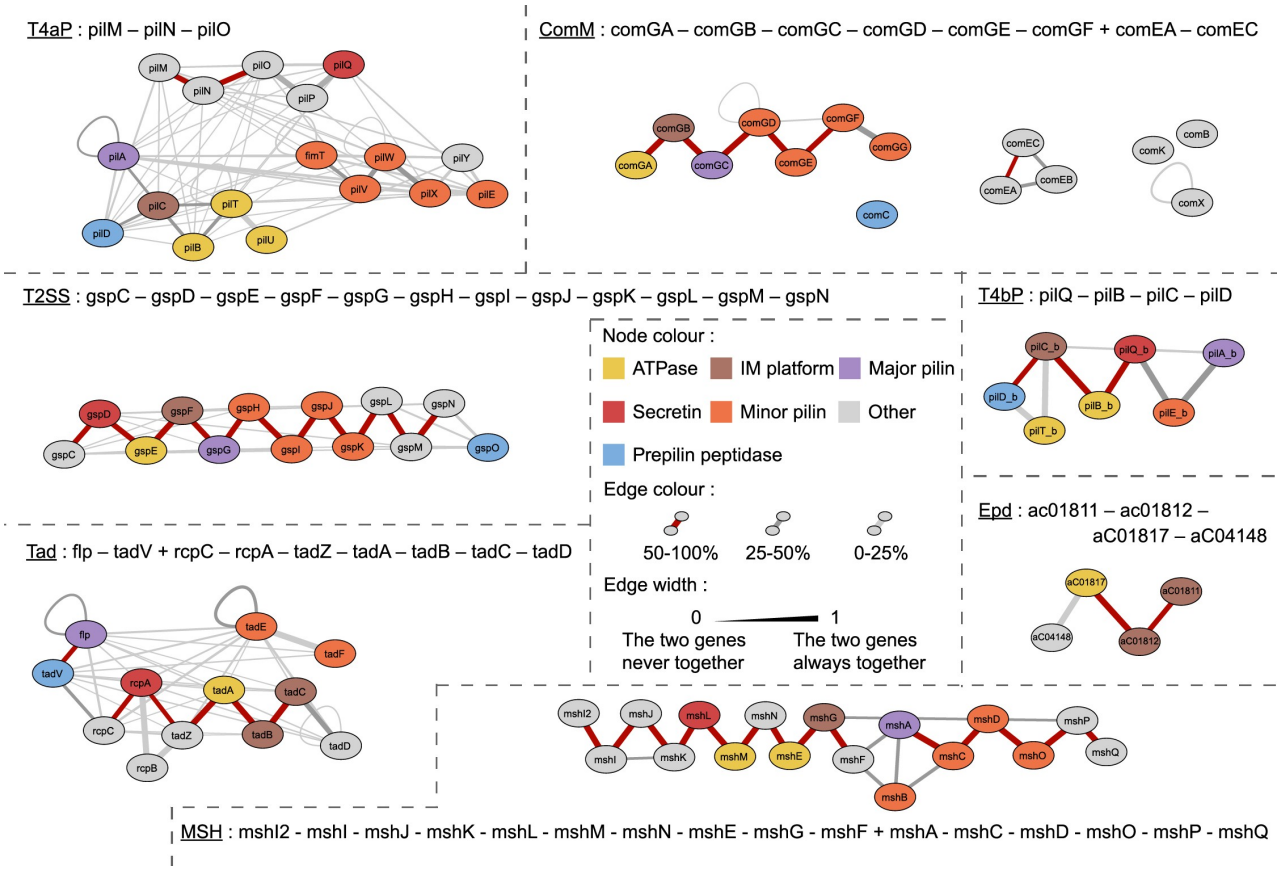


Fig 5. Genetic organisation of the detected systems. For each detected system (those indicated in Fig 4), the edge width represents the number of times the two genes are contiguous divided by the number of times the rarest gene is present in the system. The colour of the edge represents the number of times the two genes are contiguous in the system divided by the number of systems. Com, competence pilus; ComM, Com in monoderms; Epd, EppA dependent; IM, integral membrane; MSH, mannose-sensitive hemagglutinin pilus; Tad, tight adherence; T2SS, type II protein secretion system; T4aP, type IVa pilus; T4bP, type IVb pilus.

https://doi.org/10.1371/journal.pbio.3000390.g005

systems (Fig 5). The Archaeal-T4P show very diverse genetic organisation, presumably because they include very different systems (S12 Fig). The representative archaella systems show more conserved genetic organisation [35,42] (S13 Fig). Interestingly, the genetic organization of the key components of Epd is very similar to the Tad, presumably pre-dating Tad's ancestor transfer to Bacteria: the two IM platform genes are contiguous and followed by the major ATPase and the secondary one (TadZ in Tad and FlaH [arCOG04148] in Epd) (see Fig 5).

The patterns of genetic organisation of the homologous components differ between systems. In general, pilins are encoded in a single locus but can vary in their colocalisation with the rest of the genes: they can be apart (T4aP), at the edge of the locus (ComM, Tad), or in the middle (T2SS, T4bP). In Archaea, all cases were found. Interestingly, many duplicated genes tend to be contiguous, e.g., *pilTU* (ATPases). This is consistent with models suggesting that duplication processes often produce tandem duplicates [61]. The variability between types of systems and the conservation within types suggest that genetic organisation is under selection within types but changes rapidly upon functional innovation.

The genetic organisation of the loci can also reflect the action of horizontal gene transfer. If the systems are often gained or lost within lineages, as was shown for Tad [62] but much less so for the archaellum [42], then systems encoded in a single locus are much more likely to be successfully transferred because all the necessary genetic information can be transferred in one event [63]. Systems scattered across the genomes cannot be transferred in a single event (although parts of the system can presumably be exchanged if the recipient genome encodes a system with similar genetic organisation). We thus hypothesised that single-locus systems are more likely to undergo horizontal gene transfer. To test this hypothesis, we compared the phylogenetic tree of each system, i.e., a subtree of the larger phylogenetic reconstruction, with a maximum likelihood tree of the 16S rRNA sequences of the species carrying the systems (S14 Fig). We excluded the archaeal systems from these analyses because their loci are harder to define precisely (sometimes scattered and multiple systems per genome) and their functions are still poorly delimited in most cases (complicating the definition of the clade to use in the analysis). We found that systems encoded systematically in a single locus are more frequently transferred than those encoded in several loci (Fig 6). These results are reinforced by the analysis of the frequency with which systems are encoded in plasmids, which closely follows the trends observed for the frequency of transfer (highest in Tad and lowest in ComM; Fig 6). The contrast is especially interesting between the Tad and T4aP systems that are both present in many different clades and are encoded almost exclusively in one locus (Tad) or many loci (T4aP). This association between rates of transfer and organisation suggests that systems that are frequently gained and lost endure a selective pressure for being encoded in a single locus.

Discussion

We used comparative genomics and phylogenetics to produce models and protein profiles that identify TFFs in the genomes of Bacteria and Archaea. The final models classify most systems and assign them classifications that are consistent with the phylogenetic analysis. Some discrepancies persist. They can be due to systems very divergent from the models (see below) or to the presence of inactive and partly deleted loci (remnants of formerly functional systems). The models are publicly available and provide a significant advance relative to our previous work because they are more sensitive and cover more types of systems (Archaeal-T4P, ComM, MSH, and T4bP). We used them to quantify the frequency and taxonomic distribution of the different systems and found that every inspected phylum of Bacteria and Archaea has TFFs from one or several families. Some of these are widespread (e.g., T4aP, Tad), whereas others (MSH, T4bP) are abundant in Proteobacteria but absent from most other phyla. With the

Proportion of systems at a single locus	0%	16%	85%	95%*
Proportion of systems on chromosomes	100%	99.6%	98%	94%

*With 10% with one gene apart from the locus in the genome

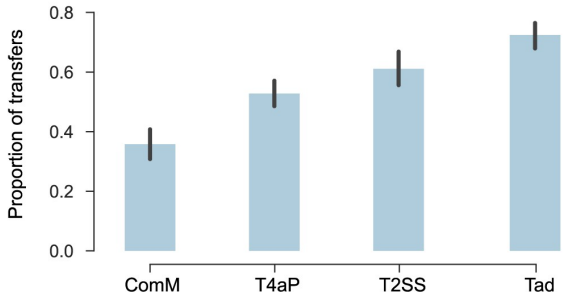


Fig 6. Association between organisation and horizontal transfer of the different systems. For each system, we compared the subtree of the systems with the 16S tree of the same species using ALE v0.4 to obtain the proportion of transfers. The panel above the graphic indicates the proportion of systems in a single locus and the proportion of systems on chromosomes (the others being found on plasmids). ComM, competence pilus of monoderms; Tad, tight adherence; T2SS, type II protein secretion system; T4aP, type IVa pilus.

https://doi.org/10.1371/journal.pbio.3000390.g006

exception of archaella, most archaeal systems are poorly characterized. When known, they tend to have diverse functions, components and genetic organization. Further experimental study of these systems is required to produce reliable MacSyFinder models for each of them.

Our approach may be regarded as conservative. First, some components of the systems were excluded from phylogenetic analyses because they were not sufficiently conserved in terms of amino acid sequence. The minor pilins are a particularly important set of proteins that were ignored because they produced short and very poor multiple alignments across systems. Second, the models were built based on experimentally validated systems and using information from monophyletic clades of a given type. If the systems were described in few species or in a small number of phyla, our ability to identify them is limited, especially when they are very different from known systems in terms of gene repertoires and protein sequences.

These limitations may explain why our improved models classed the T2SS of Chlamydiae as generic: they carry few components, and they have different origins. This may result from the impact of the peculiar developmental cycle and intracellular lifestyle of *Chlamydia* on its envelope [60]. In other cases, systems may actually differ from the descriptions in the literature. This is probably the case of the so-called T2SS of Bacteroidetes. This system is involved in protein secretion [58] but consistently branches apart from T2SS in all analyses of the phylogenetic markers. The major pilin of this system is very divergent compared to major pseudopilins from Proteobacteria. Our analysis raises the exciting possibility that it might represent a novel type of protein secretion system derived from the T4aP independently of the T2SS.

All trees show that the widely studied T4aP systems are very diverse and form several different clades in the tree, whereas the one with the PilU ATPase, the most widely studied, accounts for a minority of the identified systems. Most of the other T4aP are poorly characterised and may represent systems with novel properties. Finally, the results obtained with the final improved models showed few systems identified as generic. This suggests that there may be few novel families of systems to be discovered in the superfamily that contain the three key components (ATPase, IM platform, major pilin) and are present in the phyla represented in the genome database. On the other hand, the diversity of certain types of systems—such as the T4aP and the Archaeal-T4P—may still reveal surprising novel functionalities.

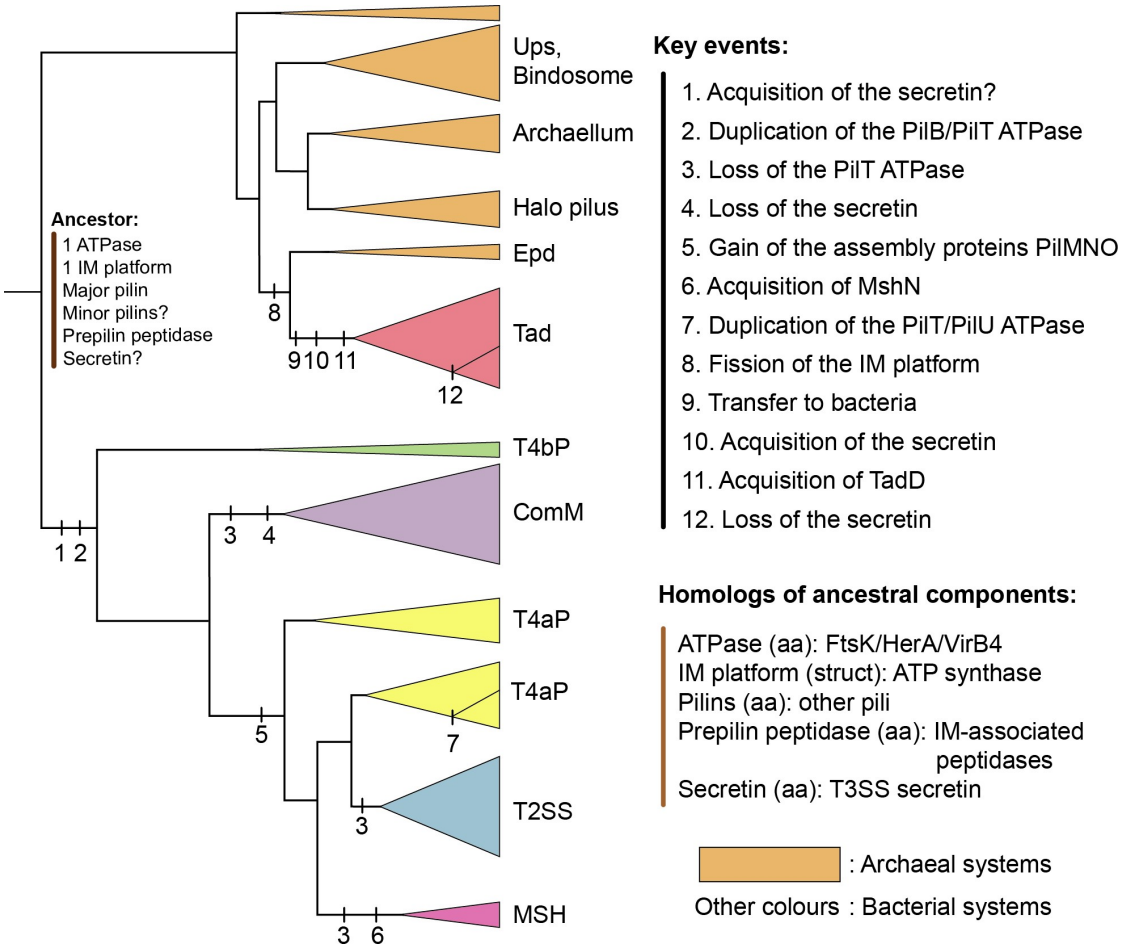


Fig 7. Evolutionary scenario of the TFF superfamily. The tree was based on the information of the trees of the concatenate and simplified to highlight the key clades and events. The colour of the triangles indicates the type of the systems. Each vertical bar on the branch indicates a numbered evolutionary event, whose details are specified under the corresponding number in the list 'Key events'. The hypotheses for the composition of the last common ancestor of the TFF superfamily are indicated at the root, and the distant homologues of these systems are indicated in the list 'Homologous of ancestral components', in which homology was observed by sequence ('aa') or structural ('struct') similarity. Halo pilus indicates two pili characterised in Halobacteria. Aap, adhesive archaeal pilus; Epd, EppA dependent; IM, integral membrane; MSH, mannose-sensitive hemagglutinin pilus; Tad, tight adherence; TFF, type IV filament; T2SS, type II protein secretion system; T3SS, type III protein secretion system; T4aP, type IVa pilus; T4bP, type IVb pilus; Ups, UV-inducible pilus of *Sulfolobus*.

https://doi.org/10.1371/journal.pbio.3000390.g007

The phylogeny of the key components of the TFFs revealed an initial split between archaeal and bacterial systems, suggesting that these structures may have pre-dated the last common ancestor of all cellular organisms (Fig 7; see also S8 Table). This ancestral system presumably had one ATPase for its assembly (the function performed by PilB in T4aP), an IM platform, pilins, and a prepilin peptidase. Among these key components, only the ATPase has identifiable sequence homologues outside the superfamily, but the other components have distant sequence or structural homologues that suggest they may pre-date the last common ancestor of all TFFs. The PFAM domain of the prepilin peptidase of TFF belongs to the PFAM clan CL0130 with other signal-peptide inner-membrane–associated peptidases, several of which are found in Bacteria, Eukaryotes (the presenillin family proteases), and Archaea [64]. The protein

profiles of the integral membrane platform match those of some ATP-binding cassette (ABC) transporters, and the protein is structurally very similar to one of the V-type ATP synthase subunits [65]. The platform may thus have been co-opted from these ubiquitous membrane-associated systems. The small size and rapid evolution of the pilins preclude the tracing of their evolution at deep time scales. Fast evolution of pilin globular domains may be associated with the variability of essential inner-membrane components that promote pilin targeting to the assembly site or connect the inner- and outer-membrane subcomplexes [66,67]. It is also difficult to determine whether there were other components in the ancestral system of the superfamily because they either evolve swiftly or are present in only a small number of systems. A recent study showed that a minimal set of eight genes was sufficient for the assembly of the T4aP of *Neisseria meningitidis* [68]. Four of them, PilMNOP, are essential for the assembly but are lacking in our list of ancestral genes because their homologs were lacking, or very rare, in genes neighboring T4bP, ComM, Tad, and Archaeal-T4P. They were found in MSH and T2SS (Fig 2, S5 Table), suggesting that they arose more recently and that other systems do not require these proteins for assembly (Fig 7; S8 Table). In short, our results are consistent with the idea that the ancestral system was able to energise its assembly and build up a pilus with matured pilins on top of an assembly platform, the basic molecular architecture of extant systems.

Our results and previous data on the genetic composition and organisation of archaeal systems [35] reveal processes of functional diversification leading to families of different functions. The *Sulfolobus* genus alone counts systems from four of the seven different Archaeal-T4P types studied experimentally (Aap, Bas, Ups, and archaellum). Even though horizontal transfers might be frequent among Archaea, our approach places the root of Archaeal-T4P within systems of methanogens from the Euryarchaeota phylum (group 1 of Makarova and colleagues [35]), and this is consistent with a proposed rooting for the archaeal tree of life within methanogens [69]. Further experimental work is needed to elucidate the functions of these Archaeal-T4P.

The archaeal origin of Tad is consistently suggested by the rooted phylogenetic analyses and the specific shared characteristics of pilins, the IM platform, and TadZ-like proteins in Tad and Archaeal-T4P (S8 Table). The literature often classes Tad pilus as T4bP [70]. Our study shows that these systems are very different in terms of components, genetic organisation, and evolutionary origins. This is in accordance with recent works proposing to clearly separate Tad from T4bP and to name them as T4cP [43]. The Epd-like systems share the closest ancestry with Tad systems among the entire TFF superfamily and are the ones with more similar genetic organization of the key components. They were only characterised in *Methanococcus maripaludis*, in which they are involved in surface attachment, a trait they share with the Tad pilus [71]. A striking trait of Tad (and Epd) is the systematic presence of two genes (*tadB* and *tadC*) encoding the IM platform. This has been regarded as the result of a gene duplication [37], but the size, domain content, and sequence similarity of these genes are more parsimoniously explained by a gene fission event, e.g., by a mutation integrating a stop codon within the ancestral gene. This produces a complex evolutionary scenario: the original IM platform was probably the result of an internal gene duplication event that pre-dated the last common ancestor of the TFF superfamily and is present in most systems. In the Epd and Tad clades, this was followed by a fission event that resulted in two tandem homologous genes. In some Tad systems of Actinobacteria, one of these components (TadC) was lost. The adaptive relevance of these successive events in the light of emerging structural data could be an interesting topic of future research.

The secretin tree provides some information about the process of transfer of the ancestral Tad to Bacteria. It places Tad's secretin within those of T4aP systems with high confidence and

typically close to Proteobacteria. This suggests that the co-option of the secretin upon transfer of the ancestor to a diderm was the founding event of Tad systems. It occurred at a time when most types of systems (T4aP, T4bP, ComM, and possibly MSH and T2SS) were already in place. Tad's secretin makes a monophyletic clade in the tree, suggesting that the accretion of the secretin to this system only happened once. Interestingly, it has been shown that TadD is essential to the assembly of the Tad secretin in *Aggregatibacter actinomycetemcomitans* [72]. While it was originally thought that TadD had no homologues in the other TFFs, we observed that it has a homologue in MSH systems (MshN, Fig 2). Further work will be needed to determine if the acquisitions of the secretin and TadD are linked or result from independent co-option events. If the scenario of a single, ancestral secretin acquisition in Tad is correct, then the adaptation of Tad to monoderms, which occurs at several places independently in the tree, involved the loss of the secretin. This event of loss seems very common because it is also found once in the initial evolution of ComM and several times at the emergence of T4aP of monoderms. Finally, the large taxonomic distribution of Tad, in spite of its relatively recent origin, is in agreement with the high frequency of horizontal transfer observed for this system.

Secretins were co-opted on multiple occasions in the TFF superfamily. Co-options of a secretin from other systems are very common. They were observed multiple times in the evolution of the T3SS (e.g., from Tad and from T2SS) and in filamentous phages [8]. In this respect, it is interesting to analyse the six TFFs with secretins (3 T4aP, 2 Tad, and 1 T2SS) in Firmicutes with an outer membrane (Halanaerobiales and Negativicutes). Four of these systems group with the TFFs of Firmicutes, and two were more closely related to systems from diderms. Their secretins were placed in the tree with proteins from the same TFF family of diderms and lacked distinctive domain architecture. There is thus no evidence that secretins were co-opted to adapt to the membrane of diderm Firmicutes, possibly because the ancestors of Firmicutes were diderm [73].

The T4bP is the most basal system among Bacteria. Subsequently, a split separated ComM from T4aP, and the latter then diversified into T2SS and MSH. Recent works suggest that the last common ancestor of Bacteria was a diderm [73]. Our analysis shows that Tad, T2SS, and T4aP are monophyletic clades in the phylogeny of the secretin, in agreement with previous works [74], suggesting that there was little transfer of the secretin between systems. If one roots the secretin tree between T4bP and T4aP, as in the concatenate trees, then it largely recapitulates the tree of the TFFs concatenate (except for the position of Tad). This is in line with a very ancient acquisition of the secretin by the TFF superfamily. Because T4bP are the most basal systems in the tree and are only found in diderms, this strongly suggests that the original bacterial system had a secretin and was present in a diderm.

Until recently, it was thought that only systems encoding PilT were capable of pilus retraction. This would suggest that the ability to retract the pilus resulted from the neofunctionalisation of one of the copies (PilB in T4aP) of the protein at the moment of the duplication of the ATPase, leading to PilB/PilT in T4P. Surprisingly, it was recently shown that the Tad system of *Caulobacter crescentus* is also able to retract the pilus [43], and there is evidence that the PilB ATPase is implicated in the process [75]. As these authors, we could not identify any orthologue of PilT in this system, in agreement with a PilT-independent retraction of the pilus. It is possible that the original ATPase of the ancestor system of T4bP and T4aP could perform both activities—assembly and retraction/disassembly—and that the duplication resulted in subfunctionalisation of these functions when both PilB and PilT were present. Interestingly, the T4aP in *Vibrio cholerae* was shown to retract with low speed in the absence of PilT [76], even if *pilT* mutants have extremely low [76] or undetectable rates [77] of DNA uptake. This would contribute to explain how ComM, devoid of PilT, could retract a filament carrying DNA in *Streptococcus* [78] and how nontypeable *Haemophilus* are able to uptake DNA using a T4aP that

lacks PilT [79], whereas PilT mutants are defective in transformation in several Bacteria [77]. It was also previously suggested that pilus retraction of the Archaeal-T4P might be required to explain some archaeal communities' behaviour, like transition from sessile to swimming stages [80]. Some overlap between the functions of PilT and PilB might also explain why PilT is frequently lost (e.g., in T2SS and ComM). In line with the specialisation of the roles of this family of ATPases following duplication, a recent study showed that PilU improves retraction in high friction environments, whereas PilT is sufficient for motility in free solution [81]. It is tempting to speculate that other, rarer duplications of these ATPases are involved in further specialisations of retraction functions.

The increase in the diversity of systems able to retract the pilus opens the possibility that many other pili could be involved in natural transformation because the role of the pilus is to attach the DNA and bring it to the cell surface. Actually, a number of arguments are in favour of the hypothesis that the ancestor of the bacterial systems, or even the last common ancestor of the superfamily, might have been able to facilitate natural transformation. First, ComM, Tad, and T4aP have been associated with this mechanism. Second, the predicted repertoire of genes of the last common ancestor of these types of systems could suffice for DNA attachment and retraction towards the cell envelope necessary for transformation. Third, systems that emerged within Archaea are able to facilitate transformation. The Ups pilus in *Sulfolobus* is highly expressed under UV light, mediates cell aggregation, and facilitates natural transformation mediated by the independent Crenarchaeal system for exchange of DNA (Ced) [82,83]. A Tad locus (archaea-derived, like all Tad) from *Micrococcus luteus* has recently been shown to be required for natural transformation [84]. We identified this system within the Tad clade, and its gene repertoire includes a single ATPase. Fourth, it has been previously shown that other key components of the transformation machinery—DprA and ComEC—are widespread across Bacteria [31,85]. The DNA uptake machinery required for transformation is encoded in many Bacteria that were never shown to be naturally transformable [86,87]. E.g., *Escherichia coli* and other enterobacteria contain functional T4aP genes coregulated with the competence machinery [66,88], which are required for natural transformation [89]. If many TFFs have the same ability, then a vast majority of Bacteria could potentially be naturally transformable. Interestingly, ComM and T4aP systems known to be involved in natural transformation tend to cluster (apart) in the phylogenetic tree of the concatenate. This suggests that even though many T4aP might facilitate transformation, those effectively involved in transformation have evolved certain traits improving this function. One such feature is the presence of two disulphide bonds in pilins, which may stabilise the structure and improve retraction-force resistance of *Acinetobacter* [90] and enterobacterial T4aP major pilins [66] or of the competence-specific minor pilins in *Neisseria* [91].

Our study has revealed how a small set of proteins with different functions evolved to produce different adaptive functions involved in different types of motility, adherence, DNA uptake, and protein secretion. This process involved 1) accretion of accessory proteins, such as the secretin and secretin-associated proteins, to cope with the existence of an outer membrane in diderms; 2) duplication and subfunctionalisation of key components, such as pilins and ATPases; 3) internal gene duplication in the IM platform, followed by gene fission in TadBC; 4) several cases of gene loss, notably for some of the IM platform homologues in Tad, for the PilT ATPase, and for the secretin in monoderms; 5) gene transfer between distant clades, including a rare example of a large macromolecular system (Tad) transferred from Archaea to Bacteria; and 6) these events being accompanied by rearrangements of the genetic loci. TFFs were frequently transferred horizontally, which certainly accelerated their evolution because genetic exchanges break clonal interference and accelerate innovation processes by recombination [92]. Interestingly, we observed that genetic organisation and horizontal transfer were

intimately associated, with systems encoded in one single locus showing higher rates of transfer. This may be a general pattern in the molecular evolution of complex systems in Bacteria and Archaea. Strong genetic linkage facilitates positive selection in physically interacting proteins [93] and the spread of the system to other species [63]. Novel genetic contexts may, in turn, select for further changes in the systems. Once functions remain for a long period of time in the lineage, as seems to be the case for ComM and some T4aP, major adaptive changes in the systems may become rare, and rearrangements splitting the initial locus may be eventually fixed. Radical changes in systems encoded in split loci are less likely to be spread by horizontal transfer, unless the recipient cell has already a copy of the system with similar genetic organisation. As a result, a tight association is established between genetic organisation and the ability of a system to evolve and spread by the action of horizontal gene transfer.

Methods

Data

We analysed 5,768 complete bacterial and archaeal genomes from NCBI RefSeq (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/, last accessed in November 2016), representing 2,268 species of Bacteria and 168 species of Archaea.

Detection of the TFF superfamily

All the systems of the family were detected using MacSyFinder v1.0.2 [44]. This program uses a model to identify a type of system in a DNA sequence (typically a replicon). The model specifies the components of the system, each represented by an HMM profile, and how their systems are organised in the sequence. A full description of the program and the models can be found in http://macsyfinder.readthedocs.io/en/latest/index.html. Briefly, the components can be mandatory, accessory, or forbidden. This does not represent a biological classification. The classification is made to distinguish between components that are ubiquitous and easy to identify (mandatory) and those that are either frequently absent or easily missed (accessory). A system is only validated if it fulfils a quorum of mandatory (minimum mandatory genes required [MMGR]) and/or mandatory + accessory genes (minimum genes required [MGR]). A locus is excluded if it contains a forbidden gene (these are useful to discriminate between closely related systems with a few specific components). Components are expected to be clustered in the genome at a short distance (defined in the model). Yet, some components can be defined as 'loners' and encoded apart. A component can be set as 'exchangeable', in which case several HMM profiles can be used to detect it (e.g., the same prepilin peptidase is used by T2SS and T4aP in some cases [49–51], and both profiles can be used to identify the prepilin peptidase of each of the two types of systems).

For this work, we could use the models previously proposed by TXSScan [39] for T2SS, T4aP, and Tad, but we wished to add a few components that were missing there. For the Archaeal-T4P and for ComM, we did not have an initial model. We proceeded in two steps. First, we made conservative initial models that matched the archetypal systems but sometimes were too strict for some atypical systems. This resulted in a list of systems in which we had strong confidence. However, it also missed many systems. To identify these systems, we built a model called 'generic' that had only the basic building blocks of these systems, with all the homologous proteins set as 'exchangeable'. Following the comparative and phylogenetic analyses, we redefined all the models to make less-initial models that could identify a larger number of systems. Both sets of models are made available. The table with all protein profiles is given in S4 Table.

Generic. We defined the model called 'generic' to search for variants of the TFF superfamily that include the key components but do not fit the strict definitions of the more specific models (T4aP,T2SS, etc.). This model assumes that all the HMM profiles of the same connected

component in the profile–profile graph of similarity can fill in for the function. Hence, it can identify very divergent or minimalistic systems, as well as chimeric systems with components that match profiles from different types. A cluster of components is classed as generic if it does not fit any of the more specific models and contains an ATPase, an IM platform, and a major pilin. In addition to these three proteins, the generic model also includes the prepilin peptidase and the secretin that are not deemed essential for the system because the former may be recruited from other systems in the genome [51], and the latter is specific to diderms.

Tad. The initial model of Tad closely followed the definitions proposed in [39]. This model includes all the known key components of the system and assumes that they are all encoded together, with the exception of *tadV*, the prepilin peptidase gene that can be encoded apart (loner) and be exchangeable with a number of homologous components from the T4aP (*pilD*) and ComM (*comC*).

The final model includes *tadD*, *rcpB*, and *rcpC* as new accessory components. The prepilin peptidase TadV is no longer exchangeable. The model defines the Tad pilus as multi_loci to allow for the existence of systems encoded in loci scattered in the genome (even if this is very rare).

T4aP. The initial model of T4aP was significantly improved from the model in [39]. It is more precise in the annotation of the retraction ATPases (*pilT* and *pilU*) and the major (*pilA*) and minor pilins (*pilE*, *pilX*, and *fimT*), and now accounting for five further components: *pilT*, *pilE*, *pilA*, and *fimT* set as mandatory and *pilU* and *pilX* set as accessory, according to their occurrence in the systems. Accordingly, the number of MGR and MMGR was increased to 8. The prepilin peptidase *pilD* was changed to mandatory, loner, and exchangeable with a number of homologous components from T2SS (*gspO*) and ComM (*comC*) according to its localisation, which could be found alone in the genome, and the fact that the HMM profiles of these two genes often have better e-value than the one of the T4aP.

The final model of T4aP includes *pilW*, *pilX*, and *pilY* as new accessory components. We decreased MMGR to 4 and MGR to 5, which better fit the data. We set *fimT*, *pilM*, *pilP*, and *pilA* as accessory to help MacSyFinder to search more complete T4aP in the genome. We also removed the forbidden genes *gspN*, *tadZ*, and *gspC*.

T2SS. The initial model of T2SS followed closely the definitions proposed in [39], in which we increased the MGR to 8 and set the prepilin peptidase *gspO* as mandatory, loner, and exchangeable with a number of homologous components from the T4aP (*pilD*).

The final model was relaxed to identify a larger fraction of the systems. We reduced the MMGR to 4 and the MGR to 5. To fit the data better, we added the prepilin peptidase of ComM (*comC*) as another exchangeable gene of *gspO*. We set the *gspC* gene as mandatory and *gspM* as accessory, and *gspD* was set as a loner to better fit the data.

ComM. In this initial model, only the genes that compose the pilus were used in the model, not the genes that encode DNA uptake system, such as *comEA*, *comEB*, and *comEC* [31,78,94–96]. The minimal distance between genes was set to 5. The MMGR was set to 3 and the MGR to 5, and the system was set as multi_loci because some genes are loners. The genes *comC*, *comGA*, *comGB*, *comGC*, and *comGD* were set as mandatory, and the other ones were set as accessory in relation with their presence in experimentally validated systems, curated with an exploratory phase to know the relative abundance of the genes in the systems (the genes with more than 80% of presence in the detected systems were set as mandatory and the others as accessory). *comC* was set as a loner and exchangeable with *pilD* of the T4aP because we found a case in which the HMM profile of *pilD* was better in e-value than that of *comC*, and for the same reason, *comGA* was set as exchangeable with *pilB* of T4aP. The genes *comB*, *comK*, and *comX* were set as loners because they are often found alone in the genome.

In the final model, we changed the number of genes for the MMGR and MGR to 4. We also added the genes encoding the DNA uptake system in the plasma membrane (*comEC*, *comEB*,

and *comEA*). *comEC* was set as mandatory, *comEB* and *comEA* were set as accessory, and *comEC* and *comEB* were set as loners. Changing the gene *comGD* to accessory allowed us to search for loner genes without changing the MGR number.

Archaeal. Initial model. We here describe the first tool, to our knowledge, to detect Archaeal-T4P. We extracted the sequences from the 200 arCOG families (2014 version, [97]) deemed to be associated with Archaeal-T4P by Makarova and colleagues [35]. We built HMM profiles for each of these families: sequences were aligned with MAFFT v7.273 and linsi algorithm, and the alignment extremities were trimmed based on the results of BMGE with BLOSUM40 matrix [98,99]. HMM profiles were generated using HMMER version 3.1b2 [100]. These profiles were compared to profiles of Tad, T4aP, and T2SS from TXSScan [39] using HHsearch (e-value and *p*-value threshold of 0.001 for the family cutoff) in order to define supra-families of components [46]. Core 'mandatory' components were defined based on the literature and experimentally validated systems. Other components were set as 'accessory'. The arCOG families that matched on the same component were defined as exchangeable. The prepilin peptidase was set as a loner gene that can be part of multiple systems. This initial model asked for a minimal number of mandatory genes and overall number of genes of 4. Of 14 experimentally validated systems found in the literature, 10 were detected with this initial model (S1 Table). After counting the occurrence of the different arCOGs in the detected Archaeal-T4P, we removed those without any occurrence to reduce the number to 109 arCOG families. Final model. The number of genes for MMGR and MGR was reduced to 3, which better fits the data.

T4bP. Final Model. This class includes the R64 thin pilus, toxin-coregulated pilus, bundle-forming pilus, longus pilus, and Cof pilus [27]. Because we do not have many experimentally validated systems for the T4bP, we used the phylogenetic information of the TFF superfamily trees to have a set of T4bP-related proteins to create the HMM profiles and the definition of the model. We created 8 HMM profiles; the MMGR was set to 4 and the MGR to 4. The system was set as multi_loci because some genes are loners. The genes *pilD*, *pilB*, *pilA*, *pilC*, and *pilQ* were set as mandatory, and the other ones were set as accessory, according to their occurrence in the detected systems. The prepilin peptidase *pilD* was set as a loner. *pilA* was set as exchangeable with *pilA* of T4aP because we found cases in which the HMM profile of *pilA* of T4aP had a better e-value in matches to T4bP than the *pilA* of T4bP.

MSH. Final Model. Because we do not have many experimentally validated systems for the MSH, we used the phylogenetic information of the TFF superfamily tree to have a set of MSH-related proteins to create the HMM profiles and the definition of the model. We created 20 HMM profiles; the MMGR was set to 3 and the MGR to 4. The system was set as multi_loci because some genes are loners. The genes *mshA*, *mshE*, *mshG*, *mshL*, and *mshM* were set as mandatory, and the others were set as accessory, according to their occurrence in the detected systems. The gene *mshA* was set as a loner, according to detected systems found in the genomes. *mshB* was set as exchangeable with *pilA* of T4bP because we found cases in which the HMM profile of *pilA* of T4bP provided better e-values when matching MSH systems than the *mshB* of MSH. For similar reasons, *mshC* was set as exchangeable with *fimT* of T4aP. The model for MSH does not include a prepilin peptidase because such a gene could not be identified in the locus.

Retrieval and construction of protein profiles

We retrieved 37 profiles for T2SS, T4aP, and Tad from TXSScan [39]. For two HMM profiles of T4aP that combine the detection of two protein paralogues (T4P_pilT_pilU and T4P_pilAE), we decided to separate the sequence of the different proteins from the original alignment of this profile to generate five separate HMM profiles (T4P_pilT, T4P_pilU, T4P_pilA, T4P_fimT, and T4P_pilE).

To create the HMM profiles, we used the following methodology. For the genes that had few representatives in the experimentally validated data set, we used BLASTP v 2.5.0+ (default settings, e-value $<1 \times 10^{-6}$) [101] to search for homologues among complete genomes. To remove very closely related proteins, we performed an all-against-all BLASTP v2.5.0+ analysis and clustered the proteins with at least 80% sequence similarity using SiLiX v1.2.10-p1 (default settings) [102]. We selected one sequence from each family as a representative. We aligned all the representatives using MAFFT v7.273 (—auto, automatic selection of the parameters depending of the size of the alignment, default values for the other parameters) [98]. With SEAVIEW [103], the poorly aligned regions at the extremities were manually trimmed in the alignment. The trimmed alignment was used to build the HMM profile using hmmbuild (default parameters) from HMMER package v3.1b2 [104].

For the HMM profiles of the final model, we used the sequences of the profiles described above. Using the information of the phylogeny of the systems, we added the sequences of the systems that were annotated as generic but that clustered in a group of experimentally validated systems. We aligned all the representatives using MAFFT v7.273 (—auto, automatic selection of the parameters depending of the size of the alignment, default values for the other parameters) [98]. With SEAVIEW [103], the poorly aligned regions at the extremities were manually trimmed in the alignment. The trimmed alignment was used to build the HMM profile using hmmbuild (default parameters) from HMMER package v3.1b2 [104].

Phylogenetic inference

Phylogenetic analyses based on protein sequences involved an initial alignment of the sequences using MAFFT v7.273 (linsi algorithm) [98]. Multiple alignments were analysed using Noisy v1.5.12 (default parameters) [105] to select the informative sites. We inferred maximum likelihood trees from the curated alignments or their concatenates, using IQ-TREE v 1.6.7.2 [52] (options -allnni, -nstop 1,000, -nm 100,000). We evaluated the node supports using the options -bb 1,000 for ultrafast bootstraps and -alrt 1,000 for SH-aLRT [106]. The best evolutionary model was selected with ModelFinder (option -MF, BIC criterion) [107]. We used the option -wbtl to conserve all optimal trees and their branch lengths.

The phylogenetic trees of 16S rRNA sequences were built from a data set including one sequence per genome of 5,776 genomes. The 16S sequences were retrieved from genome sequences using RNammer v1.2 [108] (options -S set to bac and the -m to ssu). We aligned archaeal and bacterial 16S rRNA separately using the secondary structure models with SSU_Align v0.1.1 (http://eddylab.org/software/ssu-align/, default options). Poorly aligned positions were masked with ssu-mask. The alignment was trimmed with trimAl v1.4rev15 [109] (-noallgaps, which allows for removing regions that are only composed of gaps from the alignment). The maximum likelihood trees were inferred using IQ-TREE v1.6.7.2 [52] (using the best-selected model SYM + R6 for the archaeal tree and SYM + R10 for the bacterial tree, -bb 10,000 ultrafast bootstrap [106], -wbtl to conserve all optimal trees and their branch lengths).

Reference systems data set

The data set with all the systems identified in the genomes is too large to make phylogenetic inferences. It also contains many very closely related systems that may provide little additional information to infer the deeper nodes of the tree. Hence, we developed a method to remove redundancy in the data set while maximising its genetic diversity. The method prioritises the inclusion of systems that were experimentally validated to facilitate the analysis of the results. The method consists of several sequential steps (S15 Fig).

1. We inferred the maximum likelihood tree for each key components' family as mentioned above and extracted the matrices of patristic distances (using the R function 'cophenetic_phylo' of the package ape) between all leaves of the trees. This resulted in a set of distance matrices between systems.
2. When there were multiple copies of a family of clade-specific paralogues, the system was represented multiple times in the phylogeny and in the distance matrix. To solve the problem and to have only one distance between two systems, we chose the minimal distance between the paralogues of the systems.
3. Each core protein family has a different rate of evolution. To compare them, we normalised each matrix by the sum of all the branch lengths in the tree of the family. We then built a matrix that is the average of all normalised matrices. This average matrix was used to infer a tree with bioNJ [110]. The tree was rooted at the midpoint.
4. We used the bioNJ tree to define monophyletic groups of similar systems. We iteratively used the R function 'cutree' from the stats package by gradually decreasing or increasing the heights at which the tree should be 'cut' until we obtained between 200 and 300 groups.
5. At this stage in the method, we had obtained a set of monophyletic groups of closely related systems. To pick the representative system of each group, we had the following order of priorities: i) inclusion of systems validated experimentally, ii) inclusion of the systems with fewest paralogues. In some rare cases, a given type of systems (e.g., T2SS) had less than 20 instances after this procedure. In this case, and to increase the statistical power of the analyses, we modified the height of the 'cutree' function for the specific subtree of the systems lacking representative to obtain a minimum of 20 systems for each group if possible. The systems selected to make the phylogeny are named 'representative systems'.
6. We removed some complex systems from the reference ones (6/39 Archaeal-T4P and 10/101 generic) because they had two paralogues of all the genes or were generic systems with components from different types (e.g., T2SS_gspE and T4P_pilB).

Dereplicated data set

To reduce the number of paralogues in each system, we used the following method (S16 Fig).

1. We inferred the maximum likelihood tree for each key components' family of the representative data set as mentioned above and extracted the matrices of patristic distances (using the R function 'cophenetic_phylo' of the package ape) between all leaves of the trees. This resulted in a set of distance matrices between proteins.
2. For each system with more than one copy per gene, we found the nearest system, based on patristic distances extracted from the ATPase or the IM platform tree (depending on the number of copies of the ATPase), that had only one copy of this gene.
3. We use this nearest system to choose the copy of the duplicate gene with the smallest distance to its homologue in this nearest system.
4. In the end, each system is represented by a single instance of each of the core proteins, and we called this set of selected sequences and systems the 'dereplicated data set'.

Concatenate trees and ML topology tests

The Tad pilus and T4aP show cases of system-wide duplications: some of their gene families have several members in the same systems. That is the case of the IM platform for the Tad

pilus (TadB and TadC are homologues that resulted from an initial gene internal duplication before the last common ancestor of TFFs, followed by a gene fission event) and for the ATPase of the T4aP (PilB and PilT/PilU are paralogues). Candidate systems' trees were generated based on the concatenation of all possible combinations of putative sets of orthologues, i.e., each paralogue was picked one after the other to represent their system in phylogenetic analyses. E.g., for the IM platform, a first set of orthologues would consist of TadB sequences for Tad systems together with IM platform sequences from all other systems, and another would consist of TadC sequences for Tad systems together with IM platform sequences from all other systems. For the ATPase, we decided to only focus on the functional orthologous gene, the PilB sequences for T4aP systems.

Therefore, there were two possible combinations of the mandatory genes to generate concatenates of the ATPase IM platform. In total, we generated two concatenates and used IQ-Tree to compute maximum likelihood phylogenetic trees, using partition models (option -spp, the location of the genes in the concatenation defines the partitions, the model for each partition corresponds to the model found previously for the individual analyses).

In order to assess the congruence between the concatenate trees and the individual protein trees, a maximum likelihood topology test (AU for 'Approximately Unbiased' [111]) was performed using IQ-Tree. Each protein alignment was used as an input to assess the congruence of its ML tree with those of the 10 concatenate trees. The parameters of the model were estimated on the initial parsimony tree (option -n 0). A correction for multiple tests was applied to the *p*-values (sequential Bonferroni per batch of 10 concatenate trees).

Analysis of the neighbourhood of the systems

We searched for genes systematically associated with a given type of systems by analysing the neighbourhood of each system. For each locus of a system, we identified its first gene (position X_{First}) and last gene (position X_{End}). We then took all the genes in a neighbourhood of 10 (i.e., between $X_{First-10}$ and X_{End+10}). When a system was encoded in multiple loci in the genome, each locus was analysed in the same way. We then clustered all these proteins by sequence similarity using BLASTP v. 2.5.0+ (default settings, e-value $< 1 \times 10^{-6}$) [101] and SiLiX v1.2.10-p1 (minimal percentage of identity to accept blast hits for building families at 50%) [102]. We kept the clusters if they had proteins represented in systems of different leaves in the tree. The proteins of each cluster were aligned with MAFFT and used to build HMM protein profiles as described above.

To annotate the protein clusters, we used two methods. First, we searched for similarities of their HMM profiles with the profiles used to identify the TFFs' components using HHsearch (v3.0.3, *p*-value $< 1 \times 10^{-5}$). Second, we searched for homologies between the remaining clusters and the profiles of the PFAM database (v31.0, same method).

To test whether a given cluster is significantly positively associated with a given type of system, we made the following analysis. We counted the occurrences of the elements of the cluster associated with a given type and made a contingency table in which the columns are the type versus all other types and the lines are presence or absence of an element from the cluster. To test statistical significance, we used a Fisher's exact test on the contingency table. Because this implicates many statistical tests—one test per type per cluster and this for many clusters and several types—we adjusted the *p*-values for multiple tests using the Bonferroni correction. We kept the association between a given cluster and a given type of system if the number of elements in the cluster neighbouring systems of that type was higher than expected by chance and if the corrected *p*-value < 0.05. The resulting matrix of presence/absence for genes positively associated with the systems can be found in S5 Table.

Inferring transfers of systems

We took the phylogenetic tree of all systems and picked the subtrees of each type of system. For each of these trees, we pruned the 16S rRNA tree such that it only includes species present in the system tree. We used ALE v0.4 (default parameters), a reconciliation program that introduces events of duplication, transfer, and loss (DTL) in a gene tree, and amalgamated the most frequent subtrees in a sample distribution of the gene tree to improve it and make it congruent with the species (reference) tree in a maximum likelihood framework [112]. Using ALE, we computed a number of DTL events introduced in each system's trees given the 16S rRNA tree as a reference. We then computed the proportion of transfers by collecting the number of transfers for each type of system and dividing it by the number of branches in the subtrees of each type of system.

Analysis of genetic organisation

We identified all pairs of contiguous components of the systems (for a gene X_p in the cluster, we look at the genes X_{p-1} and X_{p+1}). We constructed an adjacency matrix using this information, and we used it to construct a graph of the genetic organisation of the systems. We normalised the association between two genes to represent two different types of information:

1. To know how frequently two genes are contiguous, we divided the number of contiguous occurrences by the number of occurrences of the rarest of the two genes. This corresponds to the edge widths in Fig 5.
2. To know how many times the contiguity is found in the system, we divided the number of times the contiguity is observed by the number of systems detected. This corresponds to the edge colours in Fig 5.

Supporting information

S1 Fig. Representation of the initial and final models of the systems. Homologous genes are indicated by the same colour. Mandatory genes are indicated with a full outline, accessory genes are indicated with a dash outline, and forbidden genes are indicated with a red cross. The exchangeable genes are indicated by an arrow. The loner genes are indicated by a star below the gene. For the Archaeal-T4P, the aCXXX name indicates that all the homologous arCOGs for this function (they are set as exchangeable). The empty box in the genetic model indicates that the genes are exchangeable with all the homologous genes of the other models. Archaeal-T4P, type IV-related pili in Archaea; arCOG, archaeal Cluster of Orthologous Genes. (PDF)

S2 Fig. Taxonomic distribution of the systems in Bacteria and Archaea with the initial models. Cells indicate the number of genomes with at least one detected system. The cell's colour gradient represents the proportion of genomes with at least one system in the clade. The bar plot shows the total number of detected systems. The bars are separated in two categories: Alpha-, Beta-, and Gamma-proteobacteria versus the other clades. The cladogram symbolises approximated relationships between the bacterial and archaeal taxa analysed in this study. (PDF)

S3 Fig. Rooted phylogeny of the ATPase. The colour of the label of the leaves indicates the taxonomic group of the species. The different coloured strips indicate the classification of the systems with the MacSyFinder annotation (with the initial model and with the final one) and the annotation of the systems in the literature. The systems known to be implicated in natural

transformation are indicated in dark purple. Known subtypes of Archaeal-T4P are indicate by text in red. The annotation of the domains of the proteins using are also added. The tree was built using IQ-Tree, 10,000 replicates of UFBoot, model LG + 10. Halo pilus indicates two pili characterised in Halobacteria. Archaeal-T4P, type IV-related pili in Archaea; UFBoot, Ultrafast Bootstrap Approximation.
(PDF)

S4 Fig. Rooted phylogeny of the ATPase with FtsK as external group. The colour of the label of the leaves indicates the taxonomic group of the species. The different coloured strips indicate the classification of the systems with the MacSyFinder annotation (with the initial model and with the final one) and the annotation of the systems in the literature. The systems known to be implicated in natural transformation are indicated in dark purple. Known subtypes of Archaeal-T4P are indicate by text in red. The annotation of the domains of the proteins using are also added. The tree was built using IQ-Tree, 10,000 replicates of UFBoot, model LG + R10. Halo pilus indicates two pili characterised in Halobacteria. Archaeal-T4P, type IV-related pili in Archaea; UFBoot, Ultrafast Bootstrap Approximation.
(PDF)

S5 Fig. Rooted phylogeny of the ATPase with FtsK and virB4 as external group. The colour of the label of the leaves indicates the taxonomic group of the species. The different coloured strips indicate the classification of the systems with the MacSyFinder annotation (with the initial model and with the final one) and the annotation of the systems in the literature. The systems known to be implicated in natural transformation are indicated in dark purple. Known subtypes of Archaeal-T4P are indicate by text in red. The annotation of the domains of the proteins using are also added. The tree was built using IQ-Tree, 10,000 replicates of UFBoot, model LG + R9. Halo pilus indicates two pili characterised in Halobacteria. Archaeal-T4P, type IV-related pili in Archaea; UFBoot, Ultrafast Bootstrap Approximation.
(PDF)

S6 Fig. Rooted phylogeny of the IM platform. The colour of the label of the leaves indicates the taxonomic group of the species. The different coloured strips indicate the classification of the systems with the MacSyFinder annotation (with the initial model and with the final one) and the annotation of the systems in the literature. The systems known to be implicated in natural transformation are indicated in dark purple. Known subtypes of Archaeal-T4P are indicate by text in red. The annotation of the domains of the proteins using are also added. The tree was built using IQ-Tree, 10,000 replicates of UFBoot, model LG + F + R8. Halo pilus indicates two pili characterised in Halobacteria. Archaeal-T4P, type IV-related pili in Archaea; IM, integral membrane; UFBoot, Ultrafast Bootstrap Approximation.
(PDF)

S7 Fig. Rooted phylogeny of the major pilin. The colour of the label of the leaves indicates the taxonomic group of the species. The different coloured strips indicate the classification of the systems with the MacSyFinder annotation (with the initial model and with the final one) and the annotation of the systems in the literature. The systems known to be implicated in natural transformation are indicated in dark purple. Known subtypes of Archaeal-T4P are indicate by text in red. The annotation of the domains of the proteins using are also added. The tree was built using IQ-Tree, 10,000 replicates of UFBoot, model LG + F + R7. Halo pilus indicates two pili characterised in Halobacteria. Archaeal-T4P, type IV-related pili in Archaea; UFBoot, Ultrafast Bootstrap Approximation.
(PDF)

S8 Fig. Unrooted phylogeny of the prepilin peptidase. The colour of the label of the leaves indicates the taxonomic group of the species. The different coloured strips indicate the classification of the systems with the MacSyFinder annotation (with the initial model and with the final one) and the annotation of the systems in the literature. The systems known to be implicated in natural transformation are indicated in dark purple. Known subtypes of Archaeal-T4P are indicate by text in red. The annotation of the domains of the proteins used are also added. The tree was built using IQ-Tree, 10,000 replicates of UFBoot, model VT + F + R6. Archaeal-T4P, type IV-related pili in Archaea; UFBoot, Ultrafast Bootstrap Approximation.
(PDF)

S9 Fig. Unrooted phylogeny of the secretin. The colour of the label of the leaves indicates the taxonomic group of the species. The different coloured strips indicate the classification of the systems with the MacSyFinder annotation (with the initial model and with the final one) and the annotation of the systems in the literature. The systems known to be implicated in natural transformation are indicated in dark purple. Known subtypes of Archaeal-T4P are indicate by text in red. The annotation of the domains of the proteins used are also added. The tree was built using IQ-Tree, 10,000 replicates of UFBoot, model LG + F + R8. Archaeal-T4P, type IV-related pili in Archaea; UFBoot, Ultrafast Bootstrap Approximation.
(PDF)

S10 Fig. Rooted phylogeny of the TFF superfamily. The tree was built with the concatenate of the IM platform (using TadB) and the AAA+ ATPase (using PilB). The colour of the label of the leaves indicates the taxonomic group of the species. The different coloured strips indicate the classification of the systems with the MacSyFinder annotation (with the initial model and with the final one) and the annotation of the systems in the literature. The systems known to be implicated in natural transformation are indicated in dark purple. Known subtypes of Archaeal-T4P are indicate by text in red. The tree was built using IQ-Tree, 10,000 replicates of UFBoot, with a partition model. Halo pilus indicates two pili characterised in Halobacteria. Archaeal-T4P, type IV-related pili in Archaea; Tad, tight adherence; TFF, type IV filament; UF, Ultrafast Bootstrap Approximation.
(PDF)

S11 Fig. Taxonomic distribution of the systems in Bacteria and Archaea using the phylogenetic clustering to annotate generic systems. Cells indicate the number of genomes with at least one detected system. The cell's colour gradient represents the proportion of genomes with at least one system in the clade. The bar plot shows the total number of detected systems. The bars are separated into two categories: Alpha-, Beta-, and Gamma-proteobacteria versus the other clades. The cladogram symbolises approximated relationships between the bacterial and archaeal taxa analysed in this study.
(PDF)

S12 Fig. Genetic organisation of the Archaeal-T4P in genomes. The edge width represents the number of times the two genes are contiguous divided by the number of times the rarest gene is present in the system. The colour of the edge represents the number of times the two genes are contiguous in the system divided by the number of systems. Archaeal-T4P, type IV-related pili in Archaea.
(PDF)

S13 Fig. Genetic organisation of the archaellum in genomes. The edge width represents the number of times the two genes are contiguous divided by the number of times the rarest gene is present in the system. The colour of the edge represents the number of times the two genes

are contiguous in the system divided by the number of systems.
(PDF)

S14 Fig. 16S tree used to infer horizontal transfers. The colour of the leaves represents the phyla of the Bacteria. The tree was built using IQ-Tree, 1,000 replicates of UFBoot, model SYM + R10. UFBoot, Ultrafast Bootstrap Approximation.
(PDF)

S15 Fig. Schema of the workflow used to choose the representative systems.
(PDF)

S16 Fig. Schema of the workflow used to choose the species-specific paralogues.
(PDF)

S1 Table. List of all the profiles of the TFF superfamily used in the analysis. TFF, type IV filament.
(XLSX)

S2 Table. Experimentally validated systems used in the analysis.
(XLSX)

S3 Table. Description of all the genes and concatenate trees inferred in this study.
(XLSX)

S4 Table. All trees inferred in this study in newick format.
(TXT)

S5 Table. Matrix of presence/absence of neighbouring genes positively associated with the systems (family of genes is in columns and systems are in rows).
(XLSX)

S6 Table. All the systems detected by MacSyFinder with the search using the final models.
(XLSX)

S7 Table. Tree topology tests (AU) using IQ-TREE between concatenated trees and the genes that compose the concatenate. AU, Approximately Unbiased.
(XLSX)

S8 Table. Global scenario of TFF evolution: Summary of the evolutionary events presented in Fig 7. TFF, type IV filament.
(PDF)

Acknowledgments

We are much indebted to Olivera Francetic for comments, encouragement, and guidance during the development of this study. We thank Simonetta Gribaldo for comments and suggestions on a previous version of the manuscript and Bertrand Néron for continuous support of MacSyFinder. We thank Simonetta Gribaldo and Panagiotis Adam for providing a cladogram symbolising the relationships between the clades analysed in this study, which we used as support for Figs 4, S2 and S11.

Author Contributions

Conceptualization: Rémi Denise, Sophie S. Abby, Eduardo P. C. Rocha.

Data curation: Rémi Denise.

Funding acquisition: Rémi Denise, Eduardo P. C. Rocha.

Investigation: Rémi Denise, Sophie S. Abby, Eduardo P. C. Rocha.

Methodology: Rémi Denise, Sophie S. Abby, Eduardo P. C. Rocha.

Software: Rémi Denise, Sophie S. Abby.

Supervision: Sophie S. Abby, Eduardo P. C. Rocha.

Writing – original draft: Rémi Denise, Sophie S. Abby, Eduardo P. C. Rocha.

Writing – review & editing: Rémi Denise, Sophie S. Abby, Eduardo P. C. Rocha.

References

1. Gould SJ, Vrba ES. Exaptation—a Missing Term in the Science of Form. Paleobiology. 1982; 8(1):4–15.
2. Pal C, Papp B, Lercher MJ. An integrated view of protein evolution. Nat Rev Genet. 2006; 7:337–48. https://doi.org/10.1038/nrg1838 PMID: 16619049
3. Jacob F. Evolution and tinkering. Science. 1977; 196(4295):1161–6. https://doi.org/10.1126/science.860134 PMID: 860134
4. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature. 2000; 405(6784):299–304. https://doi.org/10.1038/35012500 PMID: 10830951
5. Gogarten JP, Townsend JP. Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol. 2005; 3(9):679–87. https://doi.org/10.1038/nrmicro1204 PMID: 16138096
6. Jain R, Rivera MC, Moore JE, Lake JA. Horizontal gene transfer accelerates genome innovation and evolution. Mol Biol Evol. 2003; 20(10):1598–602. https://doi.org/10.1093/molbev/msg154 PMID: 12777514
7. Szappanos B, Fritzemeier J, Csorgo B, Lazar V, Lu X, Fekete G, et al. Adaptive evolution of complex innovations through stepwise metabolic niche expansion. Nat Commun. 2016; 7:11607. https://doi.org/10.1038/ncomms11607 PMID: 27197754
8. Abby SS, Rocha EP. The non-flagellar type III secretion system evolved from the bacterial flagellum and diversified into host-cell adapted systems. PLoS Genet. 2012; 8(9):e1002983. https://doi.org/10.1371/journal.pgen.1002983 PMID: 23028376
9. Guglielmini J, de la Cruz F, Rocha EP. Evolution of conjugation and type IV secretion systems. Mol Biol Evol. 2013; 30(2):315–31. https://doi.org/10.1093/molbev/mss221 PMID: 22977114
10. Pell LG, Kanelis V, Donaldson LW, Howell PL, Davidson AR. The phage lambda major tail protein structure reveals a common evolution for long-tailed phages and the type VI bacterial secretion system. Proc Natl Acad Sci U S A. 2009; 106(11):4160–5. https://doi.org/10.1073/pnas.0900044106 PMID: 19251647
11. Leiman PG, Basler M, Ramagopal UA, Bonanno JB, Sauder JM, Pukatzki S, et al. Type VI secretion apparatus and phage tail-associated protein complexes share a common evolutionary origin. Proc Natl Acad Sci U S A. 2009; 106(11):4154–9. https://doi.org/10.1073/pnas.0813360106 PMID: 19251641
12. Peabody CR, Chung YJ, Yen MR, Vidal-Ingigliardi D, Pugsley AP, Saier MH Jr. Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella. Microbiology. 2003; 149(Pt 11):3051–72. https://doi.org/10.1099/mic.0.26364-0 PMID: 14600218
13. Jarrell KF, Albers SV. The archaellum: an old motility structure with a new name. Trends Microbiol. 2012; 20(7):307–12. https://doi.org/10.1016/j.tim.2012.04.007 PMID: 22613456
14. Berry JL, Pelicic V. Exceptionally widespread nanomachines composed of type IV pilins: the prokaryotic Swiss Army knives. FEMS Microbiol Rev. 2015; 39(1):134–54. https://doi.org/10.1093/femsre/fuu001 PMID: 25793961
15. Gold V, Kudryashev M. Recent progress in structure and dynamics of dual-membrane-spanning bacterial nanomachines. Curr Opin Struct Biol. 2016; 39:1–7. https://doi.org/10.1016/j.sbi.2016.03.001 PMID: 26995496
16. Korotkov KV, Sandkvist M, Hol WG. The type II secretion system: biogenesis, molecular architecture and mechanism. Nat Rev Microbiol. 2012; 10(5):336–51. https://doi.org/10.1038/nrmicro2762 PMID: 22466878

17. Korotkov KV, Gonen T, Hol WG. Secretins: dynamic channels for protein transport across membranes. Trends Biochem Sci. 2011; 36(8):433–43. https://doi.org/10.1016/j.tibs.2011.04.002 PMID: 21565514

18. Jouravleva EA, McDonald GA, Marsh JW, Taylor RK, Boesman-Finkelstein M, Finkelstein RA. The Vibrio cholerae mannose-sensitive hemagglutinin is the receptor for a filamentous bacteriophage from V. cholerae O139. Infect Immun. 1998; 66(6):2535–9. PMID: 9596713

19. Marsh JW, Taylor RK. Genetic and transcriptional analyses of the Vibrio cholerae mannose-sensitive hemagglutinin type 4 pilus gene locus. J Bacteriol. 1999; 181(4):1110–7. PMID: 9973335

20. Fitzgerald LA, Petersen ER, Ray RI, Little BJ, Cooper CJ, Howard EC, et al. Shewanella oneidensis MR-1 Msh pilin proteins are involved in extracellular electron transfer in microbial fuel cells. Process Biochemistry. 2012; 47(1):170–4.

21. Rakhuba DV, Kolomiets EI, Dey ES, Novik GI. Bacteriophage receptors, mechanisms of phage adsorption and penetration into host cell. Pol J Microbiol. 2010; 59(3):145–55. PMID: 21033576

22. Wairuri CK, van der Waals JE, van Schalkwyk A, Theron J. Ralstonia solanacearum needs Flp pili for virulence on potato. Mol Plant Microbe Interact. 2012; 25(4):546–56. https://doi.org/10.1094/MPMI-06-11-0166 PMID: 22168446

23. Hirst TR, Sanchez J, Kaper JB, Hardy SJ, Holmgren J. Mechanism of toxin secretion by Vibrio cholerae investigated in strains harboring plasmids that encode heat-labile enterotoxins of Escherichia coli. Proc Natl Acad Sci U S A. 1984; 81(24):7752–6. https://doi.org/10.1073/pnas.81.24.7752 PMID: 6393126

24. Sikora AE, Zielke RA, Lawrence DA, Andrews PC, Sandkvist M. Proteomic analysis of the Vibrio cholerae type II secretome reveals new proteins, including three related serine proteases. J Biol Chem. 2011; 286(19):16555–66. https://doi.org/10.1074/jbc.M110.211078 PMID: 21385872

25. Cadoret F, Ball G, Douzi B, Voulhoux R. Txc, a new type II secretion system of Pseudomonas aeruginosa strain PA7, is regulated by the TtsS/TtsR two-component system and directs specific secretion of the CbpE chitin-binding protein. J Bacteriol. 2014; 196(13):2376–86. https://doi.org/10.1128/JB.01563-14 PMID: 24748613

26. DebRoy S, Dao J, Soderberg M, Rossier O, Cianciotto NP. Legionella pneumophila type II secretome reveals unique exoproteins and a chitinase that promotes bacterial persistence in the lung. Proc Natl Acad Sci U S A. 2006; 103(50):19146–51. https://doi.org/10.1073/pnas.0608279103 PMID: 17148602

27. Roux N, Spagnolo J, de Bentzmann S. Neglected but amazingly diverse type IVb pili. Res Microbiol. 2012; 163(9–10):659–73. https://doi.org/10.1016/j.resmic.2012.10.015 PMID: 23103334

28. Mazariego-Espinosa K, Cruz A, Ledesma MA, Ochoa SA, Xicohtencatl-Cortes J. Longus, a type IV pilus of enterotoxigenic Escherichia coli, is involved in adherence to intestinal epithelial cells. J Bacteriol. 2010; 192(11):2791–800. https://doi.org/10.1128/JB.01595-09 PMID: 20348256

29. Skerker JM, Berg HC. Direct observation of extension and retraction of type IV pili. Proc Natl Acad Sci U S A. 2001; 98(12):6901–4. https://doi.org/10.1073/pnas.121171698 PMID: 11381130

30. Wagner A, Whitaker RJ, Krause DJ, Heilers JH, van Wolferen M, van der Does C, et al. Mechanisms of gene flow in archaea. Nat Rev Microbiol. 2017; 15(8):492–501. https://doi.org/10.1038/nrmicro.2017.41 PMID: 28502981

31. Johnston C, Martin B, Fichant G, Polard P, Claverys JP. Bacterial transformation: distribution, shared mechanisms and divergent control. Nat Rev Microbiol. 2014; 12(3):181–96. https://doi.org/10.1038/nrmicro3199 PMID: 24509783

32. Hofreuter D, Odenbreit S, Haas R. Natural transformation competence in Helicobacter pylori is mediated by the basic components of a type IV secretion system. Mol Microbiol. 2001; 41(2):379–91. PMID: 11489125

33. Patenge N, Berendes A, Engelhardt H, Schuster SC, Oesterhelt D. The fla gene cluster is involved in the biogenesis of flagella in Halobacterium salinarum. Mol Microbiol. 2001; 41(3):653–63. PMID: 11532133

34. Ng SY, Chaban B, Jarrell KF. Archaeal flagella, bacterial flagella and type IV pili: a comparison of genes and posttranslational modifications. J Mol Microbiol Biotechnol. 2006; 11(3–5):167–91. https://doi.org/10.1159/000094053 PMID: 16983194

35. Makarova KS, Koonin EV, Albers SV. Diversity and Evolution of Type IV pili Systems in Archaea. Front Microbiol. 2016; 7:667. https://doi.org/10.3389/fmicb.2016.00667 PMID: 27199977

36. Pelicic V. Type IV pili: e pluribus unum? Mol Microbiol. 2008; 68(4):827–37. https://doi.org/10.1111/j.1365-2958.2008.06197.x PMID: 18399938

37. Tomich M, Planet PJ, Figurski DH. The tad locus: postcards from the widespread colonization island. Nat Rev Microbiol. 2007; 5(5):363–75. https://doi.org/10.1038/nrmicro1636 PMID: 17435791

38. Pugsley AP. The complete general secretory pathway in gram-negative bacteria. Microbiol Rev. 1993; 57(1):50–108. PMID: 8096622

39. Abby SS, Cury J, Guglielmini J, Neron B, Touchon M, Rocha EP. Identification of protein secretion systems in bacterial genomes. Sci Rep. 2016; 6:23080. https://doi.org/10.1038/srep23080 PMID: 26979785

40. Planet PJ, Kachlany SC, DeSalle R, Figurski DH. Phylogeny of genes for secretion NTPases: identification of the widespread tadA subfamily and development of a diagnostic key for gene classification. Proc Natl Acad Sci U S A. 2001; 98(5):2503–8. https://doi.org/10.1073/pnas.051436598 PMID: 11226268

41. Iyer LM, Makarova KS, Koonin EV, Aravind L. Comparative genomics of the FtsK-HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging. Nucleic Acids Res. 2004; 32(17):5260–79. https://doi.org/10.1093/nar/gkh828 PMID: 15466593

42. Desmond E, Brochier-Armanet C, Gribaldo S. Phylogenomics of the archaeal flagellum: rare horizontal gene transfer in a unique motility structure. BMC Evol Biol. 2007; 7:106. https://doi.org/10.1186/1471-2148-7-106 PMID: 17605801

43. Ellison CK, Kan J, Dillard RS, Kysela DT, Ducret A, Berne C, et al. Obstruction of pilus retraction stimulates bacterial surface sensing. Science. 2017; 358(6362):535–8. https://doi.org/10.1126/science.aan5706 PMID: 29074778

44. Abby SS, Neron B, Menager H, Touchon M, Rocha EP. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. PLoS ONE. 2014; 9(10):e110726. https://doi.org/10.1371/journal.pone.0110726 PMID: 25330359

45. Abby SS, Rocha EPC. Identification of Protein Secretion Systems in Bacterial Genomes Using MacSyFinder. Methods Mol Biol. 2017; 1615:1–21. https://doi.org/10.1007/978-1-4939-7033-9_1 PMID: 28667599

46. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics. 2005; 21(7):951–60. https://doi.org/10.1093/bioinformatics/bti125 PMID: 15531603

47. Xu Q, Christen B, Chiu HJ, Jaroszewski L, Klock HE, Knuth MW, et al. Structure of the pilus assembly protein TadZ from Eubacterium rectale: implications for polar localization. Mol Microbiol. 2012; 83(4):712–27. https://doi.org/10.1111/j.1365-2958.2011.07954.x PMID: 22211578

48. Perez-Cheeks BA, Planet PJ, Sarkar IN, Clock SA, Xu Q, Figurski DH. The product of tadZ, a new member of the parA/minD superfamily, localizes to a pole in Aggregatibacter actinomycetemcomitans. Mol Microbiol. 2012; 83(4):694–711. https://doi.org/10.1111/j.1365-2958.2011.07955.x PMID: 22239271

49. Nunn DN, Lory S. Product of the Pseudomonas aeruginosa gene pilD is a prepilin leader peptidase. Proc Natl Acad Sci U S A. 1991; 88(8):3281–5. https://doi.org/10.1073/pnas.88.8.3281 PMID: 1901657

50. Pepe CM, Eklund MW, Strom MS. Cloning of an Aeromonas hydrophila type IV pilus biogenesis gene cluster: complementation of pilus assembly functions and characterization of a type IV leader peptidase/N-methyltransferase required for extracellular protein secretion. Mol Microbiol. 1996; 19(4):857–69. PMID: 8820654

51. Marsh JW, Taylor RK. Identification of the Vibrio cholerae type 4 prepilin peptidase required for cholera toxin secretion and pilus formation. Mol Microbiol. 1998; 29(6):1481–92. PMID: 9781884

52. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015; 32(1):268–74. https://doi.org/10.1093/molbev/msu300 PMID: 25371430

53. Sherratt DJ, Arciszewska LK, Crozat E, Graham JE, Grainge I. The Escherichia coli DNA translocase FtsK. Biochem Soc Trans. 2010; 38(2):395–8. https://doi.org/10.1042/BST0380395 PMID: 20298190

54. Nolivos S, Touzain F, Pages C, Coddeville M, Rousseau P, El Karoui M, et al. Co-evolution of segregation guide DNA motifs and the FtsK translocase in bacteria: identification of the atypical Lactococcus lactis KOPS motif. Nucleic Acids Res. 2012; 40(12):5535–45. https://doi.org/10.1093/nar/gks171 PMID: 22373923

55. Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci U S A. 1989; 86(23):9355–9. https://doi.org/10.1073/pnas.86.23.9355 PMID: 2531898

56. Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, et al. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. Proc Natl Acad Sci U S A. 1989; 86(17):6661–5. https://doi.org/10.1073/pnas.86.17.6661 PMID: 2528146

57. Imam S, Chen Z, Roos DS, Pohlschroder M. Identification of surprisingly diverse type IV pili, across a broad range of gram-positive bacteria. PLoS ONE. 2011; 6(12):e28919. https://doi.org/10.1371/journal.pone.0028919 PMID: 22216142

58. Wang X, Han Q, Chen G, Zhang W, Liu W. A Putative Type II Secretion System Is Involved in Cellulose Utilization in Cytophaga hutchisonii. Front Microbiol. 2017; 8:1482. https://doi.org/10.3389/fmicb.2017.01482 PMID: 28848505

59. Zeng L, Zhang Y, Zhu Y, Yin H, Zhuang X, Zhu W, et al. Extracellular proteome analysis of Leptospira interrogans serovar Lai. OMICS. 2013; 17(10):527–35. https://doi.org/10.1089/omi.2013.0043 PMID: 23895271

60. Nguyen BD, Valdivia RH. Virulence determinants in the obligate intracellular pathogen Chlamydia trachomatis revealed by forward genetic approaches. Proc Natl Acad Sci U S A. 2012; 109(4):1263–8. https://doi.org/10.1073/pnas.1117884109 PMID: 22232666

61. Achaz G, Rocha EP, Netter P, Coissac E. Origin and fate of repeats in bacteria. Nucleic Acids Res. 2002; 30(13):2987–94. https://doi.org/10.1093/nar/gkf391 PMID: 12087185

62. Planet PJ, Kachlany SC, Fine DH, DeSalle R, Figurski DH. The Widespread Colonization Island of Actinobacillus actinomycetemcomitans. Nat Genet. 2003; 34(2):193–8. https://doi.org/10.1038/ng1154 PMID: 12717435

63. Lawrence JG, Roth JR. Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics. 1996; 143(4):1843–60. PMID: 8844169

64. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019; 47(D1):D427–D32. https://doi.org/10.1093/nar/gky995 PMID: 30357350

65. Iwata M, Imamura H, Stambouli E, Ikeda C, Tamakoshi M, Nagata K, et al. Crystal structure of a central stalk subunit C and reversible association/dissociation of vacuole-type ATPase. Proc Natl Acad Sci U S A. 2004; 101(1):59–64. https://doi.org/10.1073/pnas.0305165101 PMID: 14684831

66. Luna Rico A, Zheng W, Petiot N, Egelman EH, Francetic O. Functional reconstitution of the type IVa pilus assembly system from enterohaemorrhagic Escherichia coli. Mol Microbiol. 2019; 111(3):732–749. https://doi.org/10.1111/mmi.14188 PMID: 30561149

67. Nivaskumar M, Santos-Moreno J, Malosse C, Nadeau N, Chamot-Rooke J, Tran Van Nhieu G, et al. Pseudopilin residue E5 is essential for recruitment by the type 2 secretion system assembly platform. Mol Microbiol. 2016; 101(6):924–41. https://doi.org/10.1111/mmi.13432 PMID: 27260845

68. Goosens VJ, Busch A, Georgiadou M, Castagnini M, Forest KT, Waksman G, et al. Reconstitution of a minimal machinery capable of assembling periplasmic type IV pili. Proc Natl Acad Sci U S A. 2017; 114(25):E4978–E86. https://doi.org/10.1073/pnas.1618539114 PMID: 28588140

69. Raymann K, Brochier-Armanet C, Gribaldo S. The two-domain tree of life is linked to a new root for the Archaea. Proc Natl Acad Sci U S A. 2015; 112(21):6670–5. https://doi.org/10.1073/pnas.1420858112 PMID: 25964353

70. O'Connell Motherway M, Zomer A, Leahy SC, Reunanen J, Bottacini F, Claesson MJ, et al. Functional genome analysis of Bifidobacterium breve UCC2003 reveals type IVb tight adherence (Tad) pili as an essential and conserved host-colonization factor. Proc Natl Acad Sci U S A. 2011; 108(27):11217–22. https://doi.org/10.1073/pnas.1105380108 PMID: 21690406

71. Nair DB, Uchida K, Aizawa S, Jarrell KF. Genetic analysis of a type IV pili-like locus in the archaeon Methanococcus maripaludis. Arch Microbiol. 2014; 196(3):179–91. https://doi.org/10.1007/s00203-014-0956-4 PMID: 24493292

72. Clock SA, Planet PJ, Perez BA, Figurski DH. Outer membrane components of the Tad (tight adherence) secreton of Aggregatibacter actinomycetemcomitans. J Bacteriol. 2008; 190(3):980–90. https://doi.org/10.1128/JB.01347-07 PMID: 18055598

73. Antunes LC, Poppleton D, Klingl A, Criscuolo A, Dupuy B, Brochier-Armanet C, et al. Phylogenomic analysis supports the ancestral presence of LPS-outer membranes in the Firmicutes. Elife. 2016; 5: e14589. https://doi.org/10.7554/eLife.14589 PMID: 27580370

74. Nickerson NN, Abby SS, Rocha EP, Chami M, Pugsley AP. A Single Amino Acid Substitution Changes the Self-Assembly Status of a Type IV Piliation Secretin. J Bacteriol. 2012; 194(18):4951–8. https://doi.org/10.1128/JB.00798-12 PMID: 22773793

75. Ellison CK, Kan J, Chlebek JL, Hummels KR, Panis G, Viollier PH, et al. A bifunctional ATPase drives tad pilus extension and retraction. bioRxiv. 616128 [Preprint] [cited 2019 Jun 3]. Available from: https://www.biorxiv.org/content/10.1101/616128v1.

76. Ellison CK, Dalia TN, Vidal Ceballos A, Wang JC, Biais N, Brun YV, et al. Retraction of DNA-bound type IV competence pili initiates DNA uptake during natural transformation in Vibrio cholerae. Nat Microbiol. 2018; 3(7):773–80. https://doi.org/10.1038/s41564-018-0174-y PMID: 29891864

77. Seitz P, Blokesch M. DNA-uptake machinery of naturally competent Vibrio cholerae. Proc Natl Acad Sci U S A. 2013; 110(44):17987–92. https://doi.org/10.1073/pnas.1315647110 PMID: 24127573

78. Laurenceau R, Pehau-Arnaudet G, Baconnais S, Gault J, Malosse C, Dujeancourt A, et al. A type IV pilus mediates DNA binding during natural transformation in Streptococcus pneumoniae. PLoS Pathog. 2013; 9(6):e1003473. https://doi.org/10.1371/journal.ppat.1003473 PMID: 23825953

79. Carruthers MD, Tracy EN, Dickson AC, Ganser KB, Munson RS Jr., Bakaletz LO. Biological roles of nontypeable Haemophilus influenzae type IV pilus proteins encoded by the pil and com operons. J Bacteriol. 2012; 194(8):1927–33. https://doi.org/10.1128/JB.06540-11 PMID: 22328674

80. Pohlschroder M, Esquivel RN. Archaeal type IV pili and their involvement in biofilm formation. Front Microbiol. 2015; 6:190. https://doi.org/10.3389/fmicb.2015.00190 PMID: 25852657

81. Tala L, Fineberg A, Kukura P, Persat A. Pseudomonas aeruginosa orchestrates twitching motility by sequential control of type IV pili movements. Nat Microbiol. 2019; 4(5):774–80. https://doi.org/10.1038/s41564-019-0378-9 PMID: 30804544

82. van Wolferen M, Wagner A, van der Does C, Albers SV. The archaeal Ced system imports DNA. Proc Natl Acad Sci U S A. 2016; 113(9):2496–501. https://doi.org/10.1073/pnas.1513740113 PMID: 26884154

83. Frols S, Ajon M, Wagner M, Teichmann D, Zolghadr B, Folea M, et al. UV-inducible cellular aggregation of the hyperthermophilic archaeon Sulfolobus solfataricus is mediated by pili formation. Mol Microbiol. 2008; 70(4):938–52. https://doi.org/10.1111/j.1365-2958.2008.06459.x PMID: 18990182

84. Angelov A, Bergen P, Nadler F, Hornburg P, Lichev A, Ubelacker M, et al. Novel Flp pilus biogenesis-dependent natural transformation. Front Microbiol. 2015; 6:84. https://doi.org/10.3389/fmicb.2015.00084 PMID: 25713572

85. Pimentel ZT, Zhang Y. Evolution of the Natural Transformation Protein, ComEC, in Bacteria. Front Microbiol. 2018; 9:2980. https://doi.org/10.3389/fmicb.2018.02980 PMID: 30627116

86. Claverys J-P, Martin B. Bacterial "competence" genes: signatures of active transformation, or only remnants? Trends in Microbiology. 2003; 11(4):161–5. PMID: 12706993

87. Oliveira PH, Touchon M, Rocha EP. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. Nucleic Acids Res. 2014; 42(16):10618–31. https://doi.org/10.1093/nar/gku734 PMID: 25120263

88. Sinha S, Cameron AD, Redfield RJ. Sxy induces a CRP-S regulon in Escherichia coli. J Bacteriol. 2009; 191(16):5180–95. https://doi.org/10.1128/JB.00476-09 PMID: 19502395

89. Sinha S, Redfield RJ. Natural DNA uptake by Escherichia coli. PLoS ONE. 2012; 7(4):e35620. https://doi.org/10.1371/journal.pone.0035620 PMID: 22532864

90. Ronish LA, Lillehoj E, Fields JK, Sundberg EJ, Piepenbrink KH. The structure of PilA from Acinetobacter baumannii AB5075 suggests a mechanism for functional specialization in Acinetobacter type IV pili. J Biol Chem. 2019; 294(1):218–30. https://doi.org/10.1074/jbc.RA118.005814 PMID: 30413536

91. Cehovin A, Simpson PJ, McDowell MA, Brown DR, Noschese R, Pallett M, et al. Specific DNA recognition mediated by a type IV pilin. Proc Natl Acad Sci U S A. 2013; 110(8):3065–70. https://doi.org/10.1073/pnas.1218832110 PMID: 23386723

92. Cooper TF. Recombination speeds adaptation by reducing competition between beneficial mutations in populations of Escherichia coli. PLoS Biol. 2007; 5(9):e225. https://doi.org/10.1371/journal.pbio.0050225 PMID: 17713986

93. Stahl FW, Murray NE. The evolution of gene clusters and genetic circularity in microorganisms. Genetics. 1966; 53(3):569–76. PMID: 5331527

94. Provvedi R, Dubnau D. ComEA is a DNA receptor for transformation of competent Bacillus subtilis. Mol Microbiol. 1999; 31(1):271–80. PMID: 9987128

95. Burrows LL. Pseudomonas aeruginosa twitching motility: type IV pili in action. Annu Rev Microbiol. 2012; 66:493–520. https://doi.org/10.1146/annurev-micro-092611-150055 PMID: 22746331

96. Craig L, Li J. Type IV pili: paradoxes in form and function. Curr Opin Struct Biol. 2008; 18(2):267–77. https://doi.org/10.1016/j.sbi.2007.12.009 PMID: 18249533

97. Makarova KS, Wolf YI, Forterre P, Prangishvili D, Krupovic M, Koonin EV. Dark matter in archaeal genomes: a rich source of novel mobile elements, defense systems and secretory complexes. Extremophiles. 2014; 18(5):877–93. https://doi.org/10.1007/s00792-014-0672-7 PMID: 25113822

98. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30(4):772–80. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

99. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol. 2010; 10:210. https://doi.org/10.1186/1471-2148-10-210 PMID: 20626897

100. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011; 39(Web Server issue):W29–37. https://doi.org/10.1093/nar/gkr367 PMID: 21593126

101. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10:421. https://doi.org/10.1186/1471-2105-10-421 PMID: 20003500

102. Miele V, Penel S, Duret L. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics. 2011; 12:116. https://doi.org/10.1186/1471-2105-12-116 PMID: 21513511

103. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 2010; 27(2):221–4. https://doi.org/10.1093/molbev/msp259 PMID: 19854763

104. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011; 7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195 PMID: 22039361

105. Dress AW, Flamm C, Fritzsch G, Grunewald S, Kruspe M, Prohaska SJ, et al. Noisy: identification of problematic columns in multiple sequence alignments. Algorithms Mol Biol. 2008; 3:7. https://doi.org/10.1186/1748-7188-3-7 PMID: 18577231

106. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol. 2018; 35(2):518–22. https://doi.org/10.1093/molbev/msx281 PMID: 29077904

107. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017; 14(6):587–9. https://doi.org/10.1038/nmeth.4285 PMID: 28481363

108. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007; 35(9):3100–8. https://doi.org/10.1093/nar/gkm160 PMID: 17452365

109. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009; 25(15):1972–3. https://doi.org/10.1093/bioinformatics/btp348 PMID: 19505945

110. Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol. 1997; 14(7):685–95. https://doi.org/10.1093/oxfordjournals.molbev.a025808 PMID: 9254330

111. Shimodaira H. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 2002; 51(3):492–508. https://doi.org/10.1080/10635150290069913 PMID: 12079646

112. Szollosi GJ, Davin AA, Tannier E, Daubin V, Boussau B. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. Philos Trans R Soc Lond B Biol Sci. 2015; 370(1678):20140335. https://doi.org/10.1098/rstb.2014.0335 PMID: 26323765

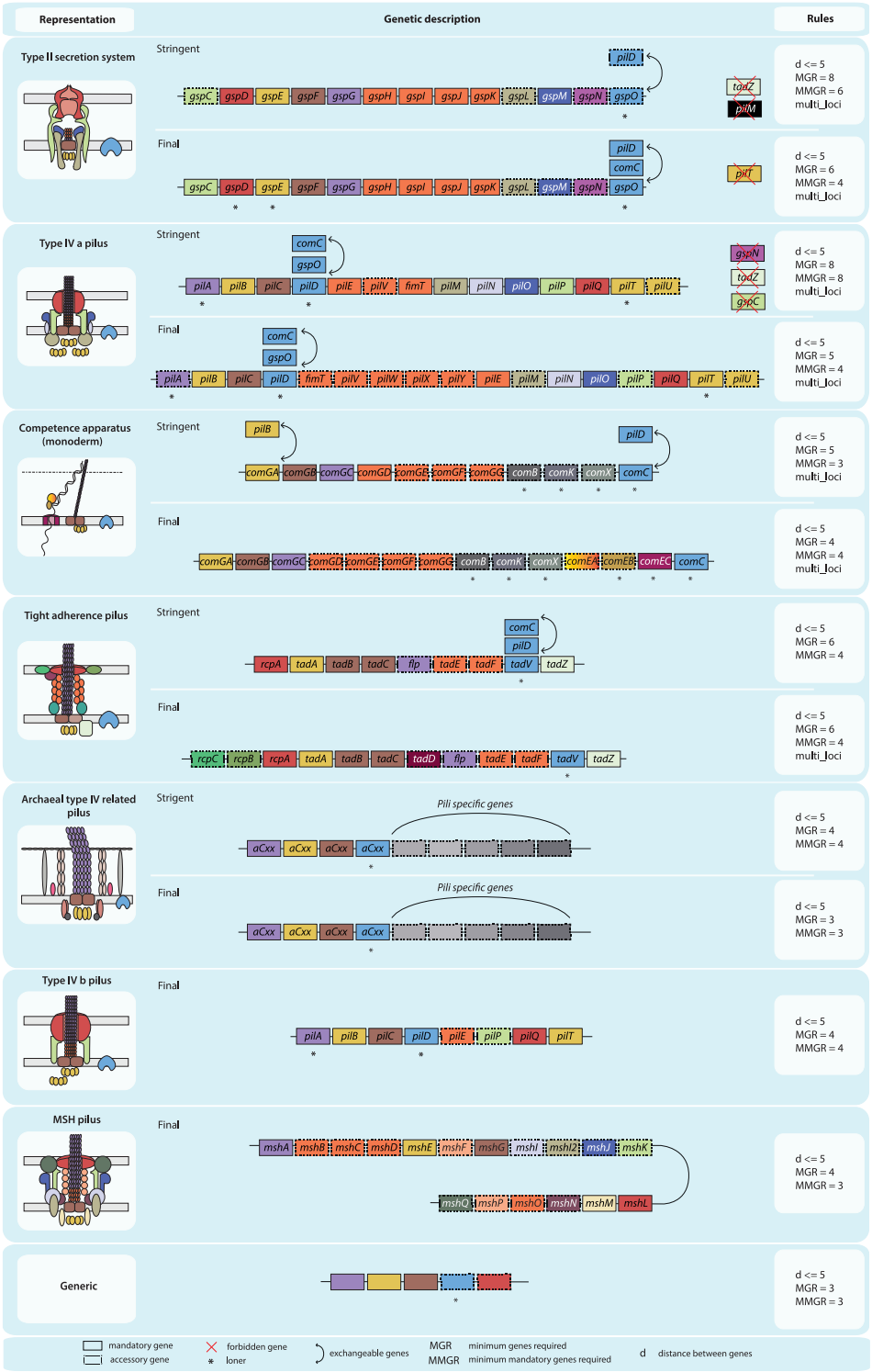


Figure 3.1 – S1 Fig. Representation of the initial and final models of the systems. Homologous genes are indicated by a same colour. Mandatory genes are indicated with a full outline, accessory genes are indicated with a dash outline and forbidden genes are indicated with a red cross. The exchangeable genes are indicated by an arrow. The loner genes are indicated by a star below the gene. For the Archaeal-T4P the aCXXX name indicates that all the homologous arCOGs for this function (they are set as exchangeable). The empty box in the genetic model indicates that the genes are exchangeable with all the homologous genes of the other models.

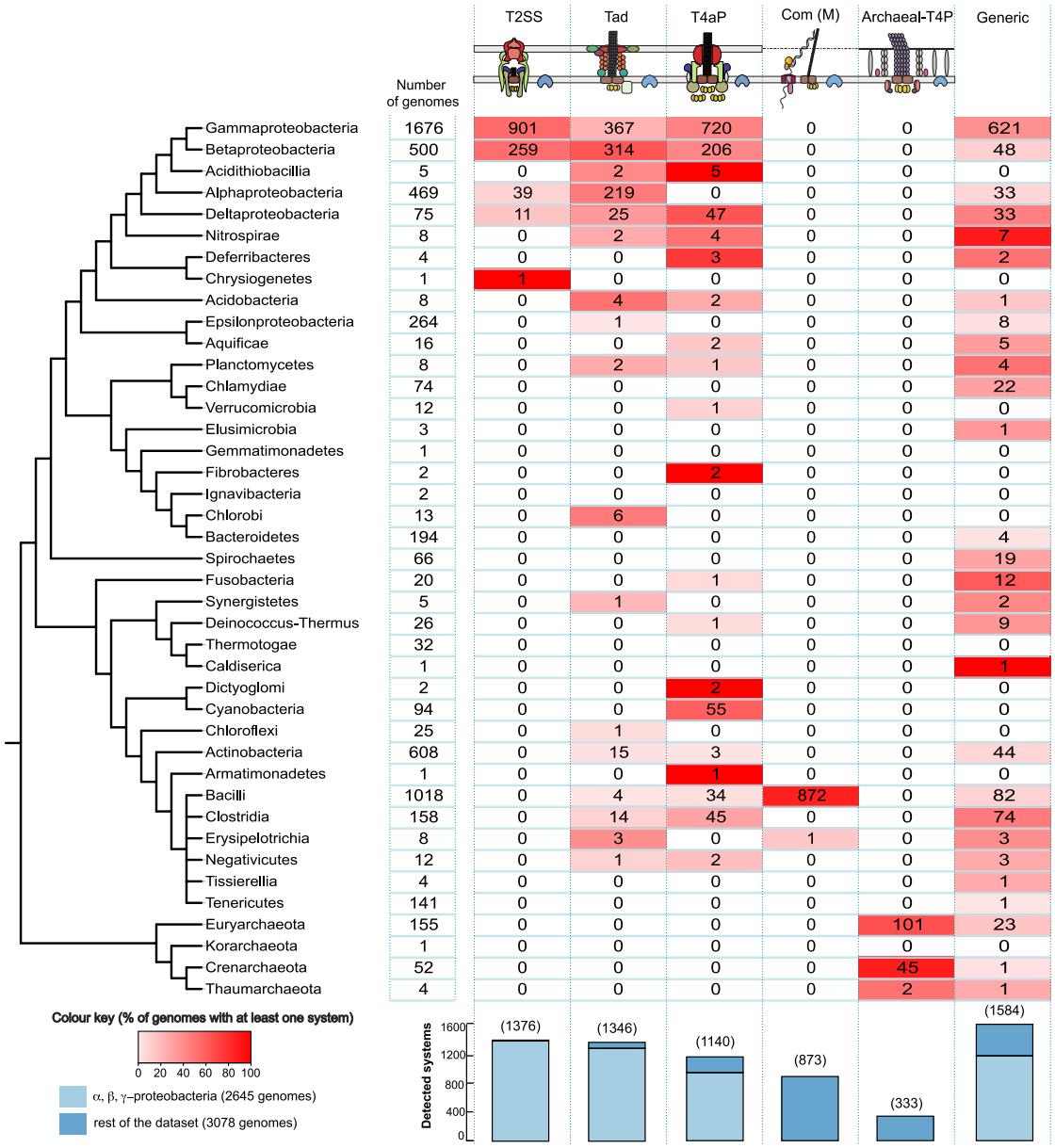


Figure 3.2 – S2 Fig. Taxonomic distribution of the systems in Bacteria and Archaea with the initial models. Cells indicate the number of genomes with at least one detected system. The cell's colour gradient represents the proportion of genomes with at least one system in the clade. The bar plot shows the total number of detected systems. The bars are separated in two categories : Alpha-, Beta-, Gamma-proteobacteria versus the other clades. The cladogram symbolizes approximated relationships between the bacterial and archaeal taxa analysed in this study.

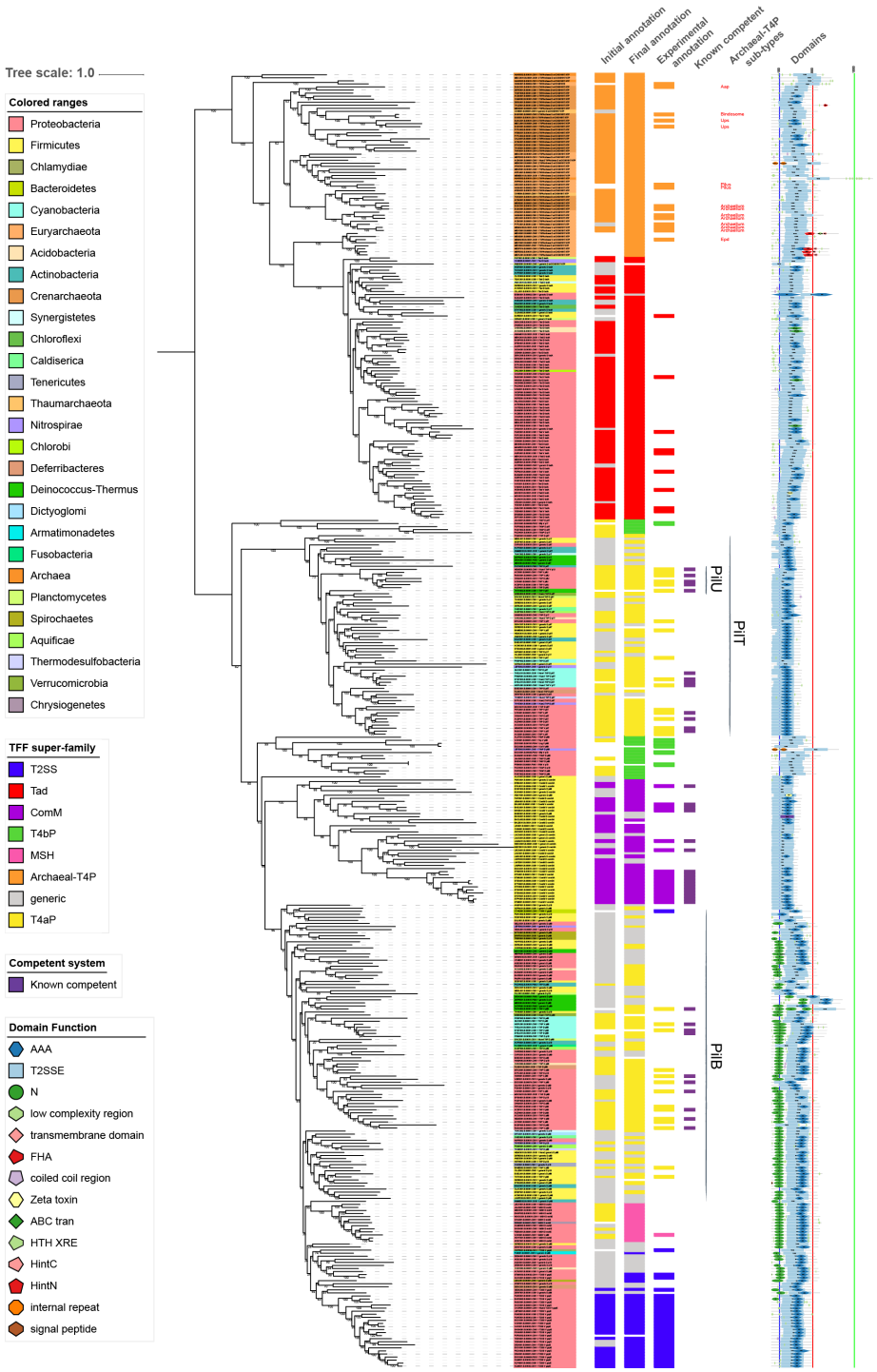


Figure 3.3 – S3 Fig. Rooted phylogeny of the ATPase. The colour of the label of the leaves indicates the taxonomic group of the species. The different coloured strips indicate the classification of the systems with the MacSyFinder annotation (with the initial model and with the final one) and the annotation of the systems in the literature. The systems known to be implicated in natural transformation are indicated in dark purple. Known sub-types of Archaeal-T4P are indicate by a text in red. The annotation of the domains of the proteins using are also added. The tree was built using IQ-Tree, 10000 replicates of UF-Boot, model LG+10.

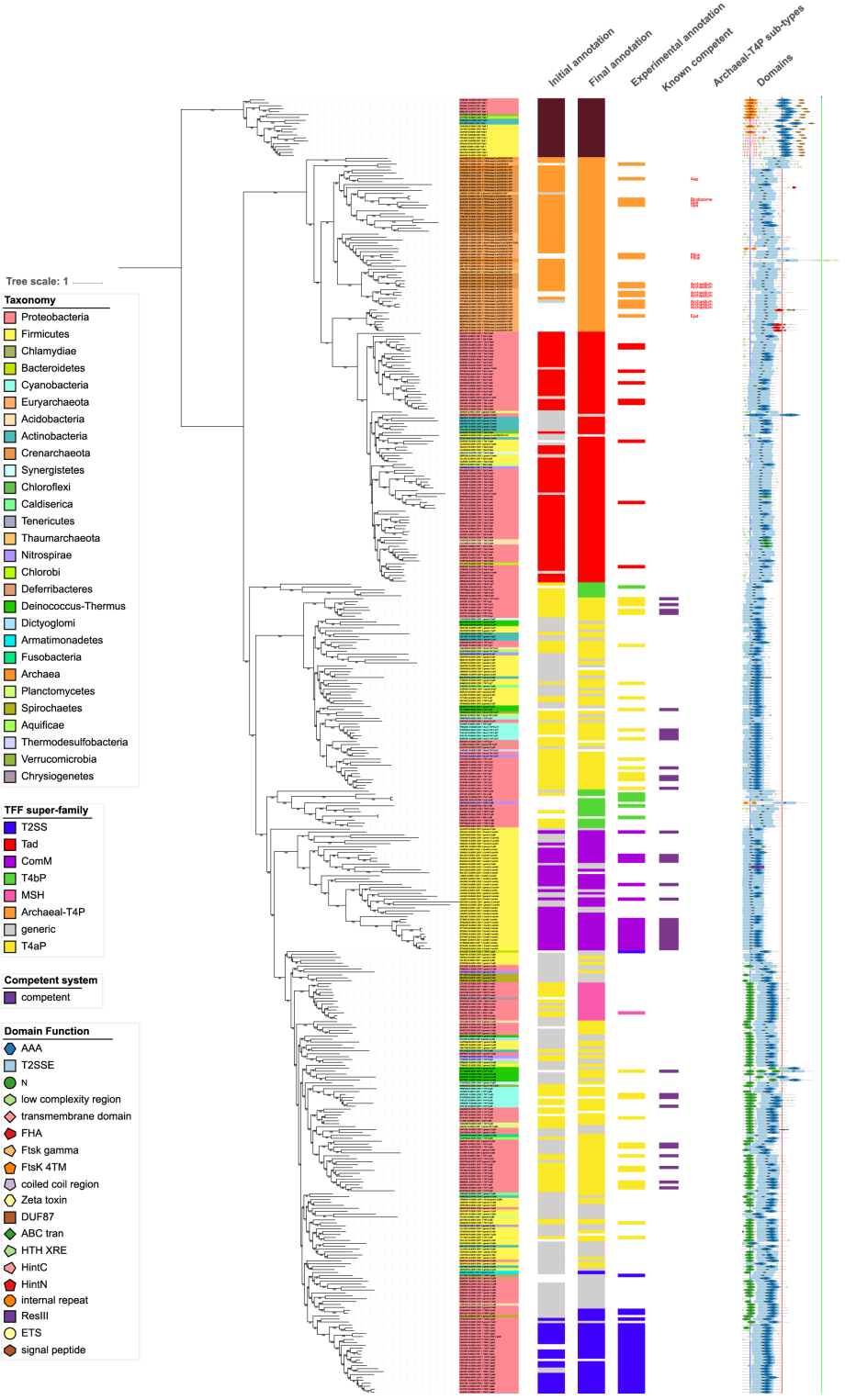


Figure 3.4 – S4 Fig. Rooted phylogeny of the ATPase with FtsK as external group. The colour of the label of the leaves indicates the taxonomic group of the species. The different coloured strips indicate the classification of the systems with the MacSyFinder annotation (with the initial model and with the final one) and the annotation of the systems in the literature. The systems known to be implicated in natural transformation are indicated in dark purple. Known sub-types of Archaeal-T4P are indicate by a text in red. The annotation of the domains of the proteins using are also added. The tree was built using IQ-Tree, 10000 replicates of UF-Boot, model LG+R10.

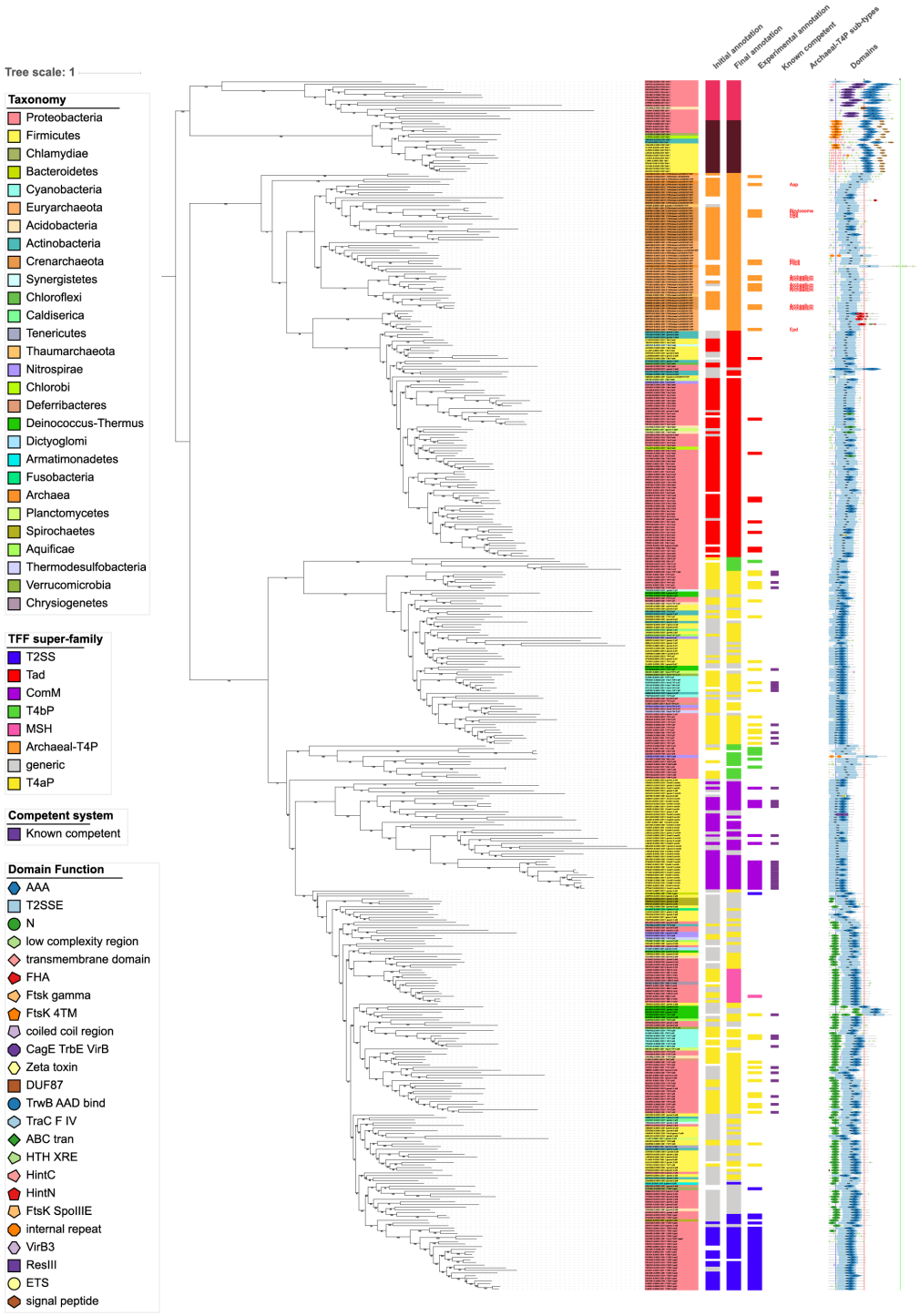


Figure 3.5 – S5 Fig. Rooted phylogeny of the ATPase with FtsK and virB4 as external group. The colour of the label of the leaves indicates the taxonomic group of the species. The different coloured strips indicate the classification of the systems with the MacSyFinder annotation (with the initial model and with the final one) and the annotation of the systems in the literature. The systems known to be implicated in natural transformation are indicated in dark purple. Known sub-types of Archaeal-T4P are indicate by a text in red. The annotation of the domains of the proteins using are also added. The tree was built using IQ-Tree, 10000 replicates of UF-Boot, model LG+R9.

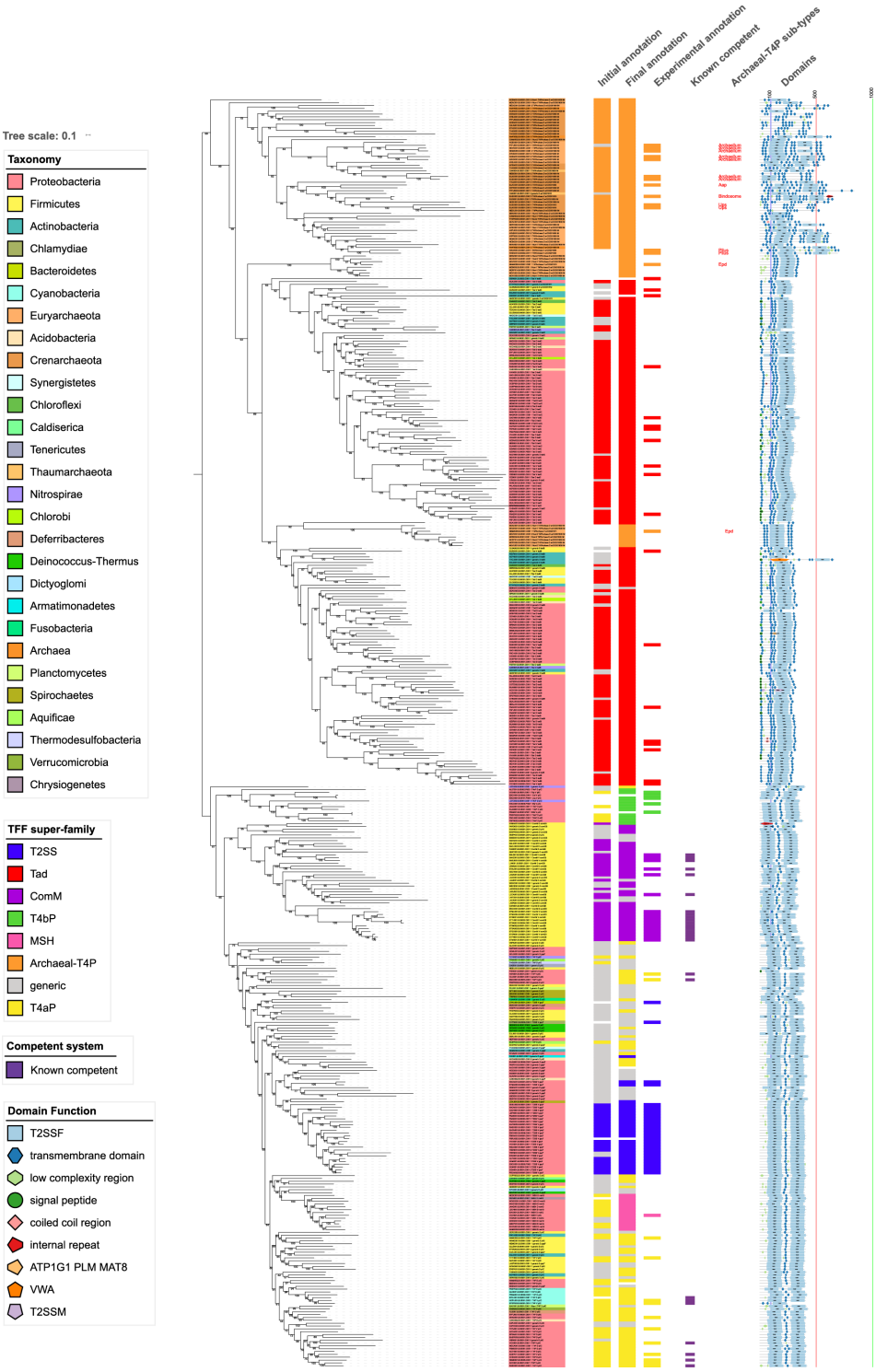


Figure 3.6 – S6 Fig. Rooted phylogeny of the integral membrane platform. The colour of the label of the leaves indicates the taxonomic group of the species. The different coloured strips indicate the classification of the systems with the MacSyFinder annotation (with the initial model and with the final one) and the annotation of the systems in the literature. The systems known to be implicated in natural transformation are indicated in dark purple. Known sub-types of Archaeal-T4P are indicate by a text in red. The annotation of the domains of the proteins using are also added. The tree was built using IQ-Tree, 10000 replicates of UF-Boot, model LG+F+R8.

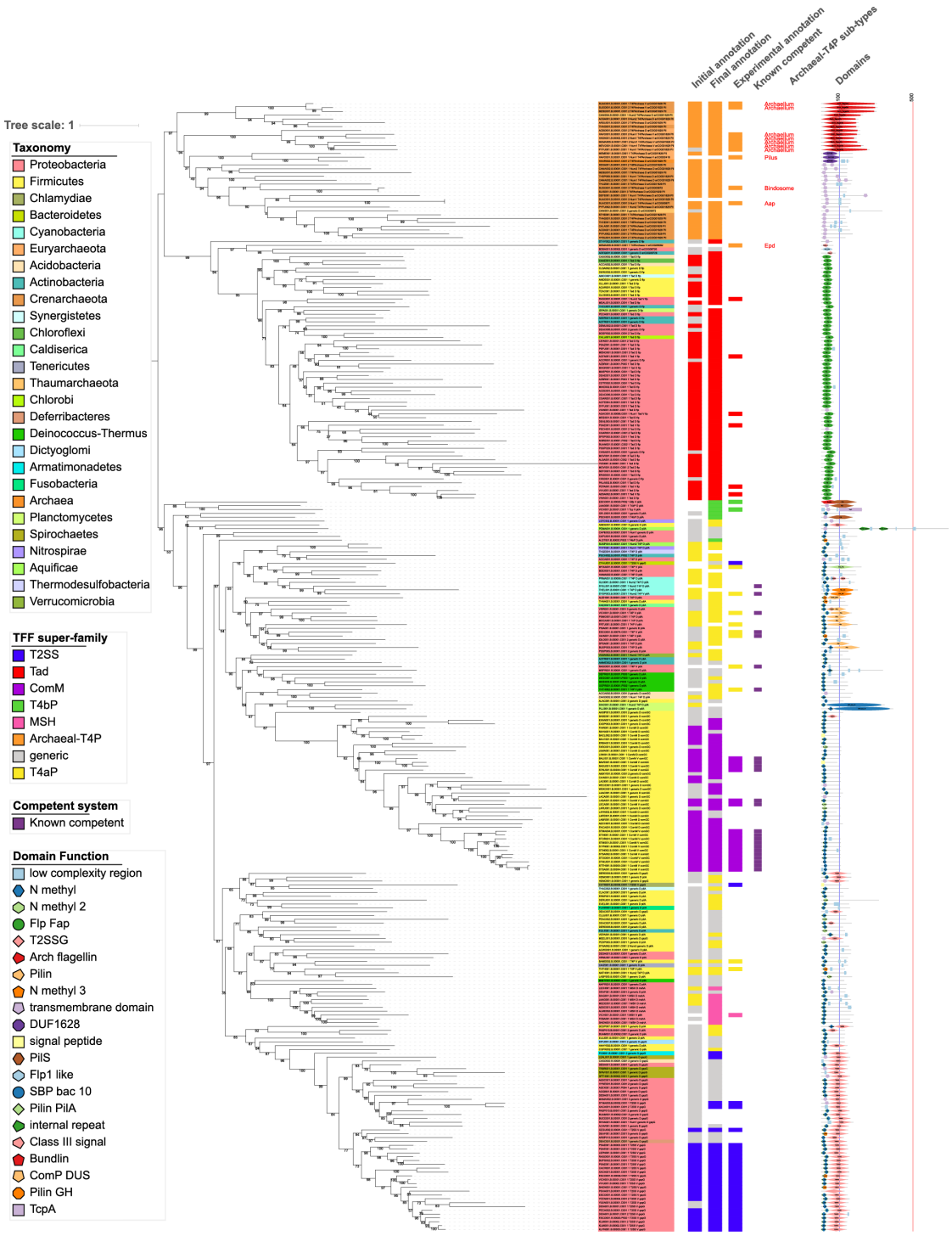


Figure 3.7 – S7 Fig. Rooted phylogeny of the major pilin. The colour of the label of the leaves indicates the taxonomic group of the species. The different coloured strips indicate the classification of the systems with the MacSyFinder annotation (with the initial model and with the final one) and the annotation of the systems in the literature. The systems known to be implicated in natural transformation are indicated in dark purple. Known sub-types of Archaeal-T4P are indicate by a text in red. The annotation of the domains of the proteins using are also added. The tree was built using IQ-Tree, 10000 replicates of UF-Boot, model LG+F+R7.

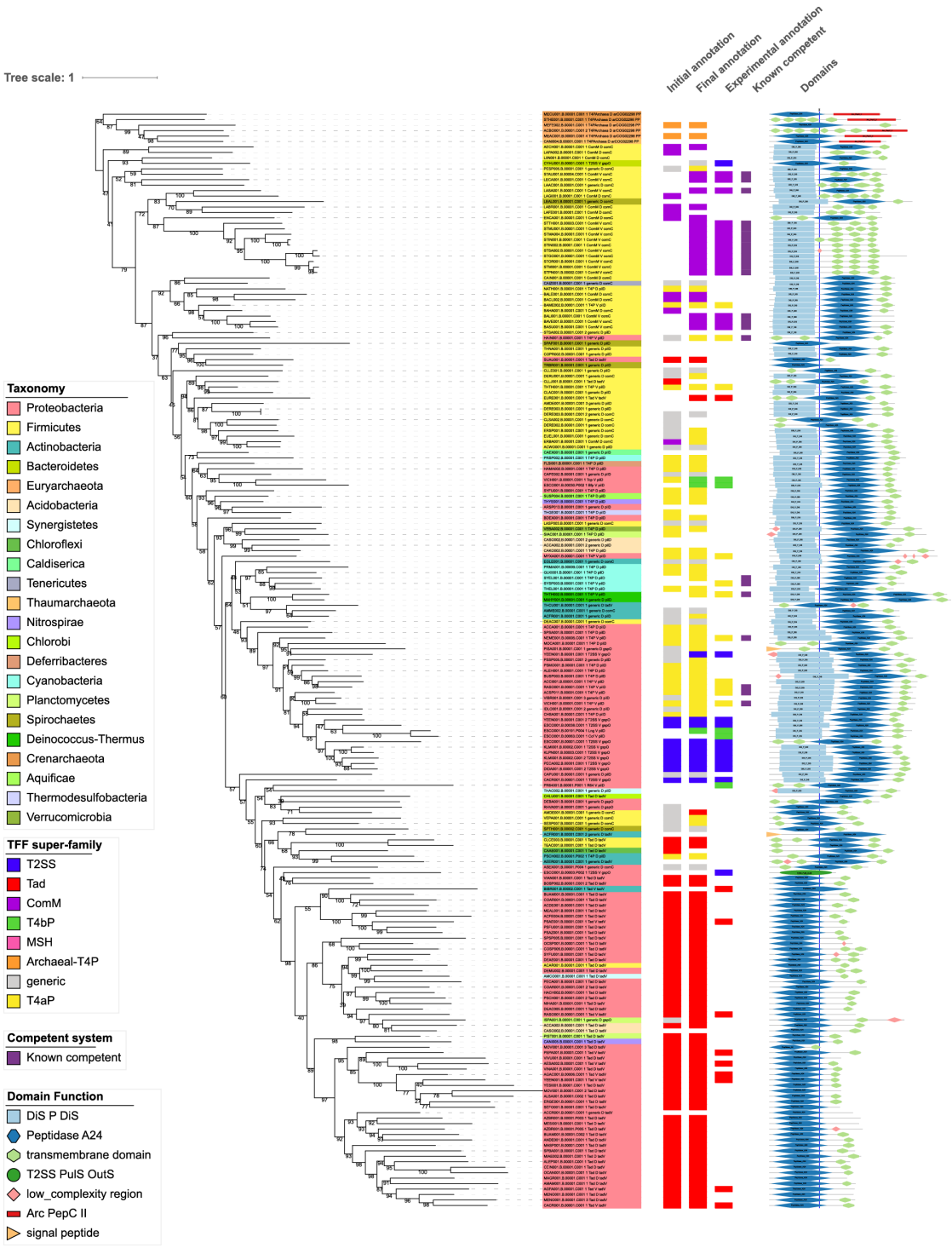


Figure 3.8 – S8 Fig. Unrooted phylogeny of the prepilin peptidase. The colour of the label of the leaves indicates the taxonomic group of the species. The different coloured strips indicate the classification of the systems with the MacSyFinder annotation (with the initial model and with the final one) and the annotation of the systems in the literature. The systems known to be implicated in natural transformation are indicated in dark purple. Known sub-types of Archaeal-T4P are indicate by a text in red. The annotation of the domains of the proteins using are also added. The tree was built using IQ-Tree, 10000 replicates of UF-Boot, model VT+F+R6.

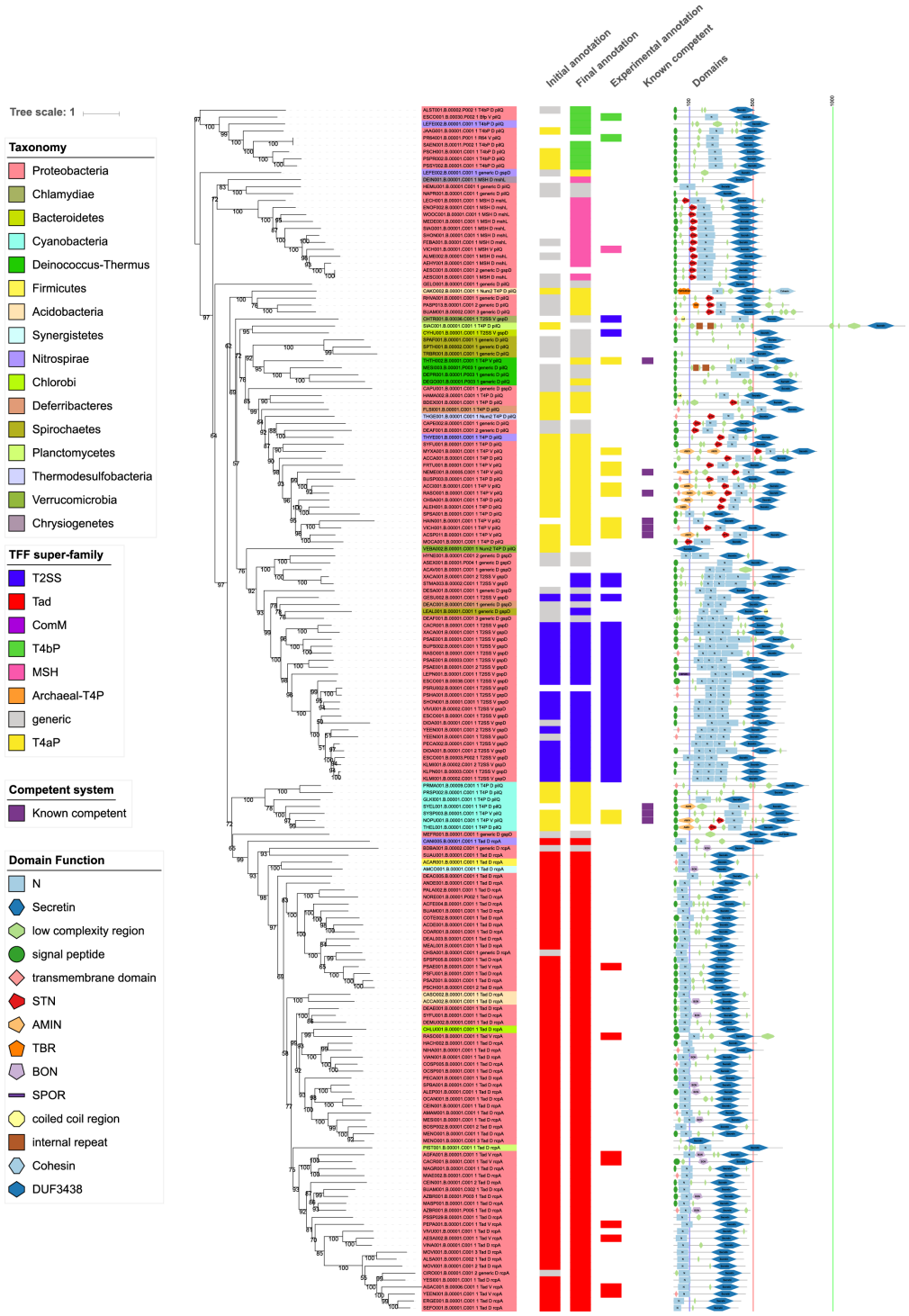


Figure 3.9 – S9 Fig. Unrooted phylogeny of the secretin. The colour of the label of the leaves indicates the taxonomic group of the species. The different coloured strips indicate the classification of the systems with the MacSyFinder annotation (with the initial model and with the final one) and the annotation of the systems in the literature. The systems known to be implicated in natural transformation are indicated in dark purple. Known sub-types of Archaeal-T4P are indicate by a text in red. The annotation of the domains of the proteins using are also added. The tree was built using IQ-Tree, 10000 replicates of UF-Boot, model LG+F+R8.

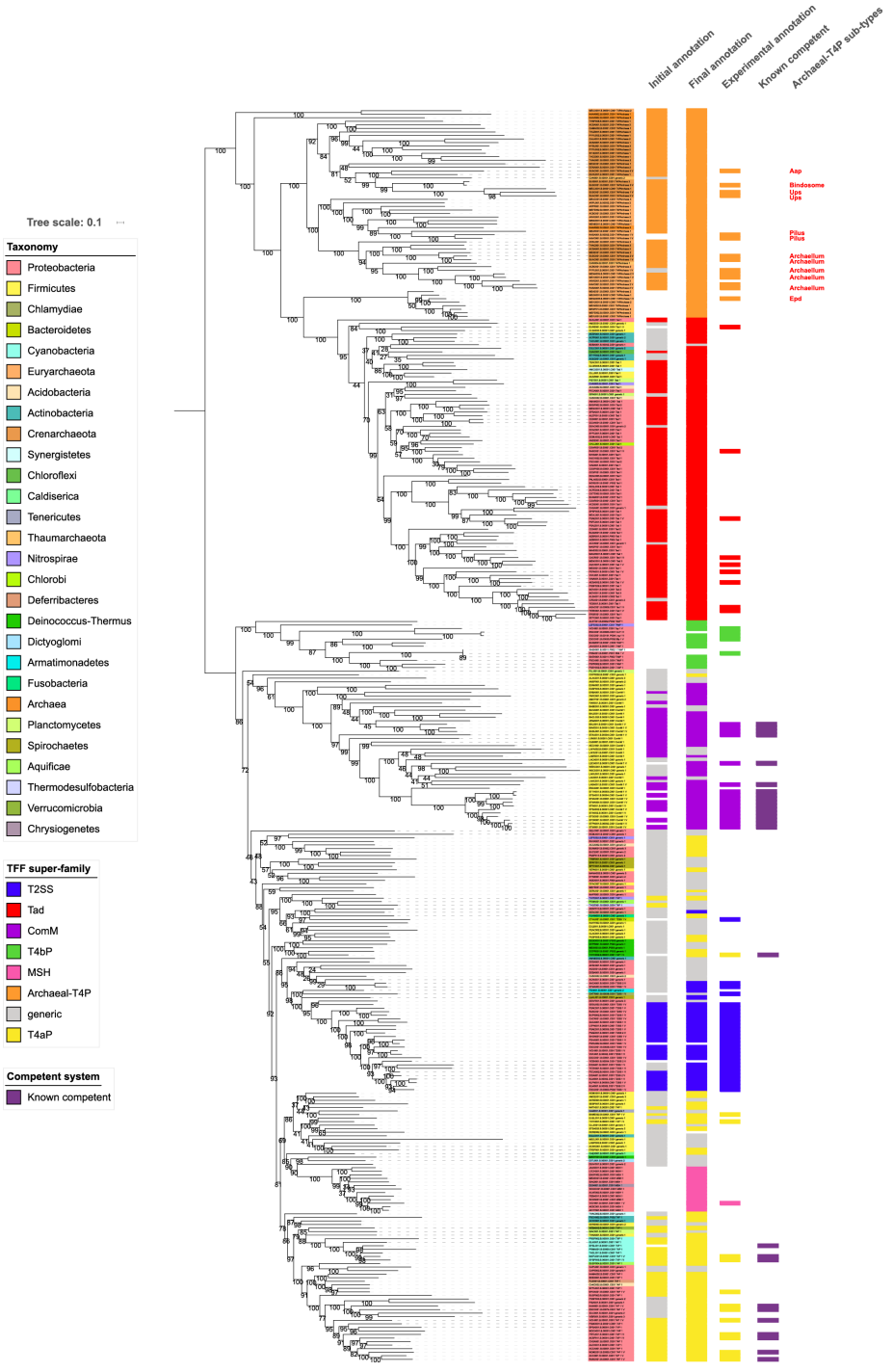


Figure 3.10 – S10 Fig. Rooted phylogeny of the TFF super-family. The was built with the concatenate of the IM platform (using TadB) and the AAA+ ATPase (using PilB). The colour of the label of the leaves indicates the taxonomic group of the species. The different coloured strips indicate the classification of the systems with the MacSyFinder annotation (with the initial model and with the final one) and the annotation of the systems in the literature. The systems known to be implicated in natural transformation are indicated in dark purple. Known sub-types of Archaeal-T4P are indicate by a text in red. The tree was built using IQ-Tree, 10000 replicates of UF-Boot, with a partition model.

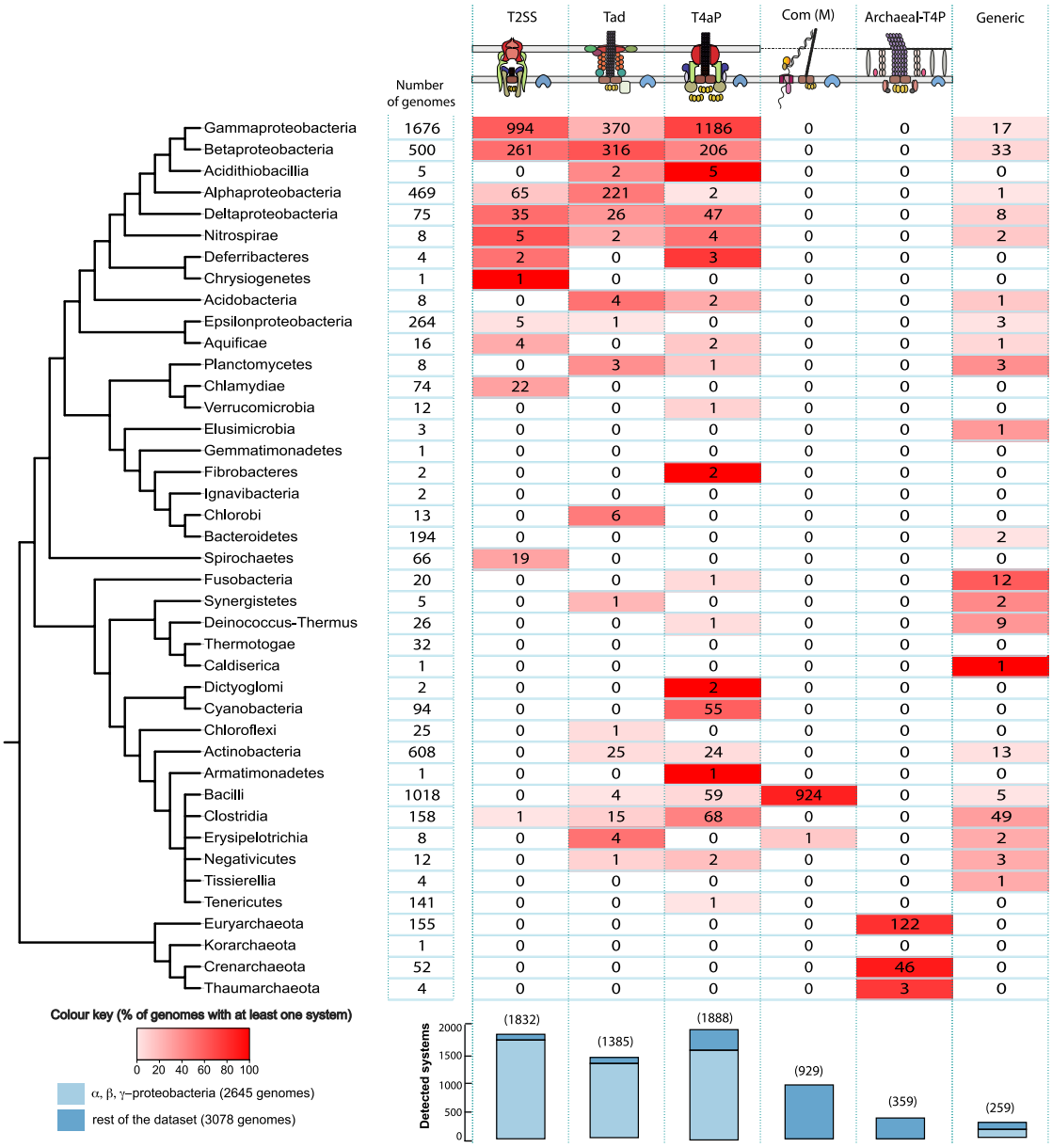


Figure 3.11 – S11 Fig. Taxonomic distribution of the systems in Bacteria and Archaea using the phylogenetic clustering to annotate generic systems. Cells indicate the number of genomes with at least one detected system. The cell's colour gradient represents the proportion of genomes with at least one system in the clade. The bar plot shows the total number of detected systems. The bars are separated in two categories : Alpha-, Beta-, Gamma-proteobacteria versus the other clades. The cladogram symbolizes approximated relationships between the bacterial and archaeal taxa analysed in this study.

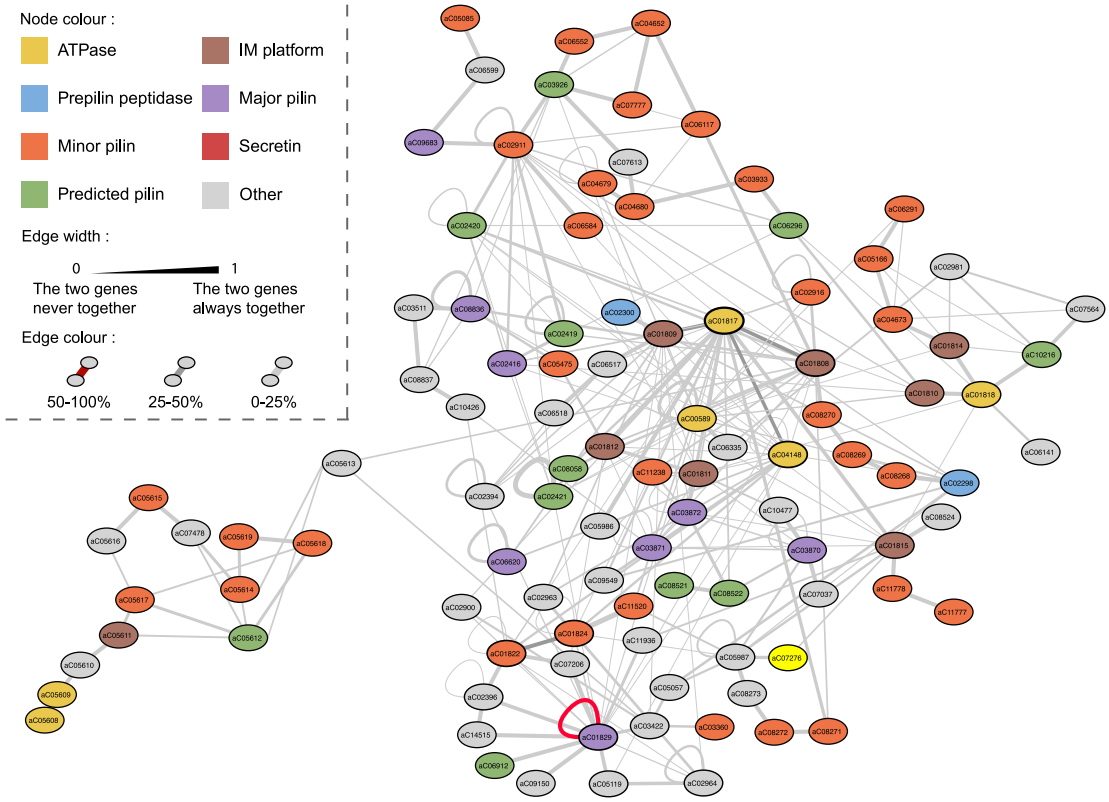


Figure 3.12 – S12 Fig. Genetic organization of the Archaeal-T4P in genomes. The edge width represents the number of times the two genes are contiguous divided by the number of times the rarest gene is present in the system. The colour of the edge represents the number of times the two genes are contiguous in the system divided by the number of systems.

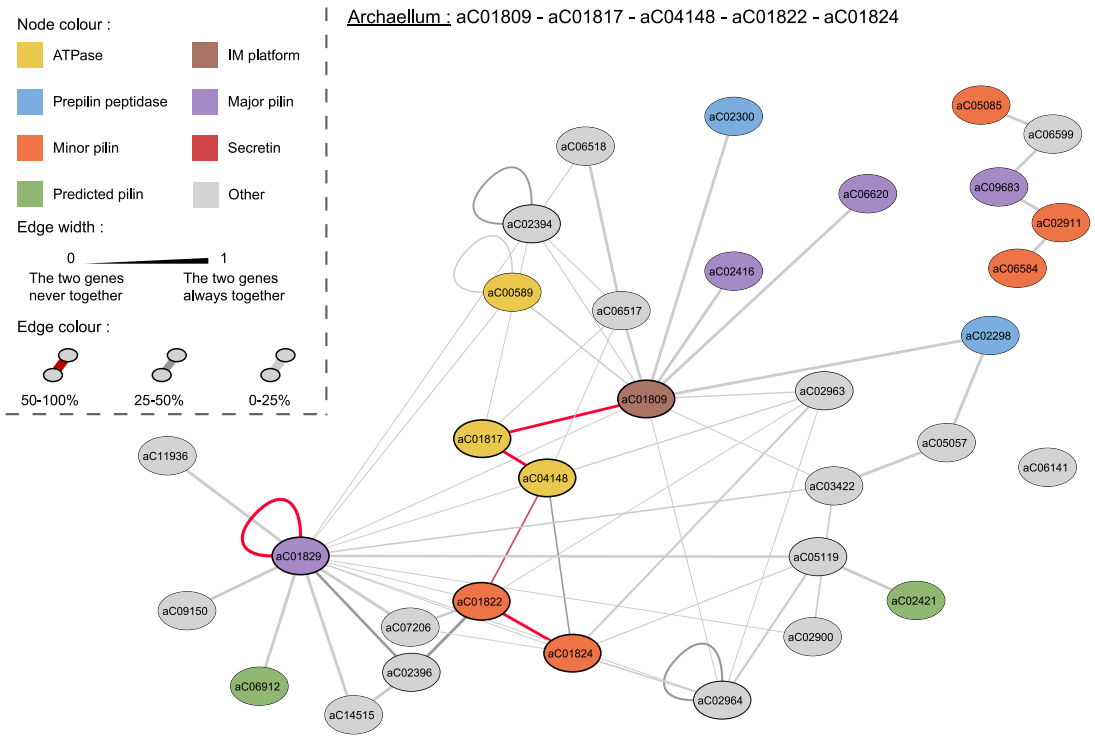


Figure 3.13 – S13 Fig. Genetic organization of the Archaellum in genomes. The edge width represents the number of times the two genes are contiguous divided by the number of times the rarest gene is present in the system. The colour of the edge represents the number of times the two genes are contiguous in the system divided by the number of systems.

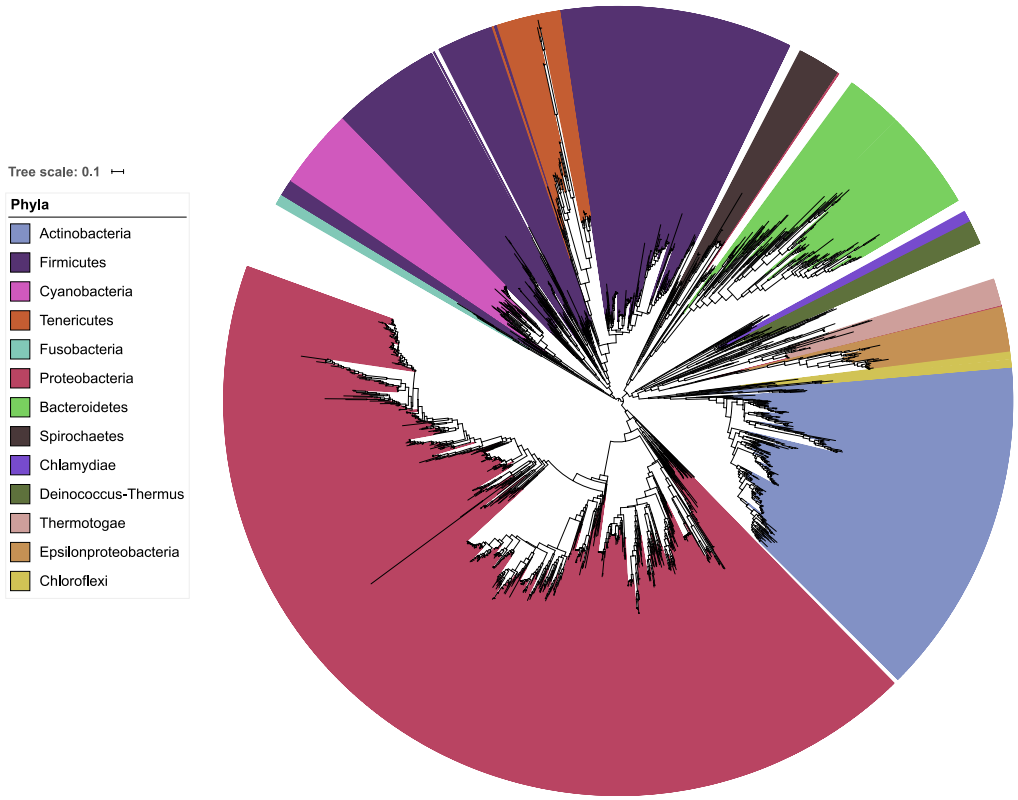


Figure 3.14 – S14 Fig. 16S tree used to infer horizontal transfers. The colour of the leaves represents the phyla of the bacteria. The tree was built using IQ-Tree, 1000 replicates of UF-Boot, model SYM+R10.

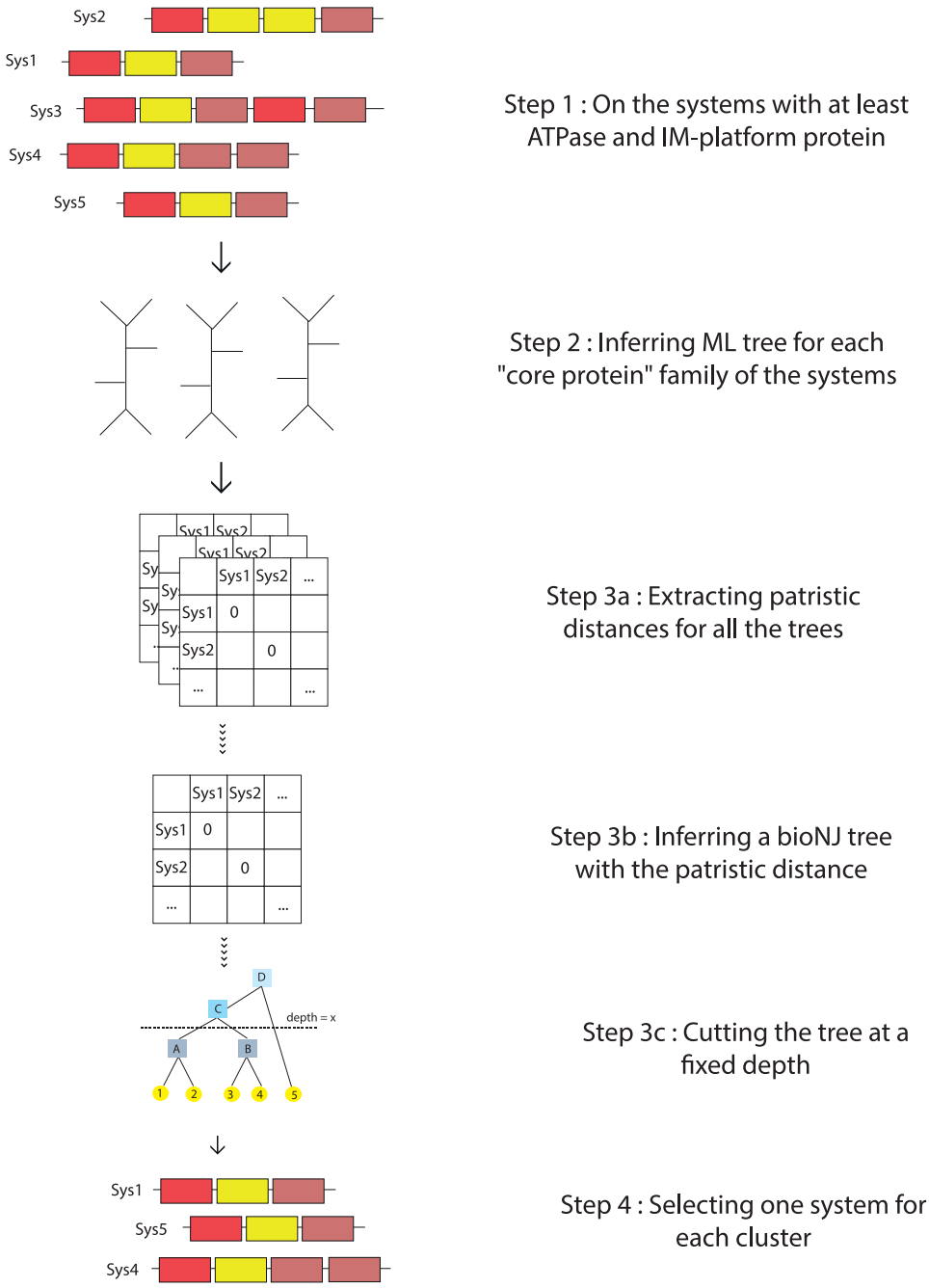


Figure 3.15 – S15 Fig. Schema of the workflow used to choose the representative systems.

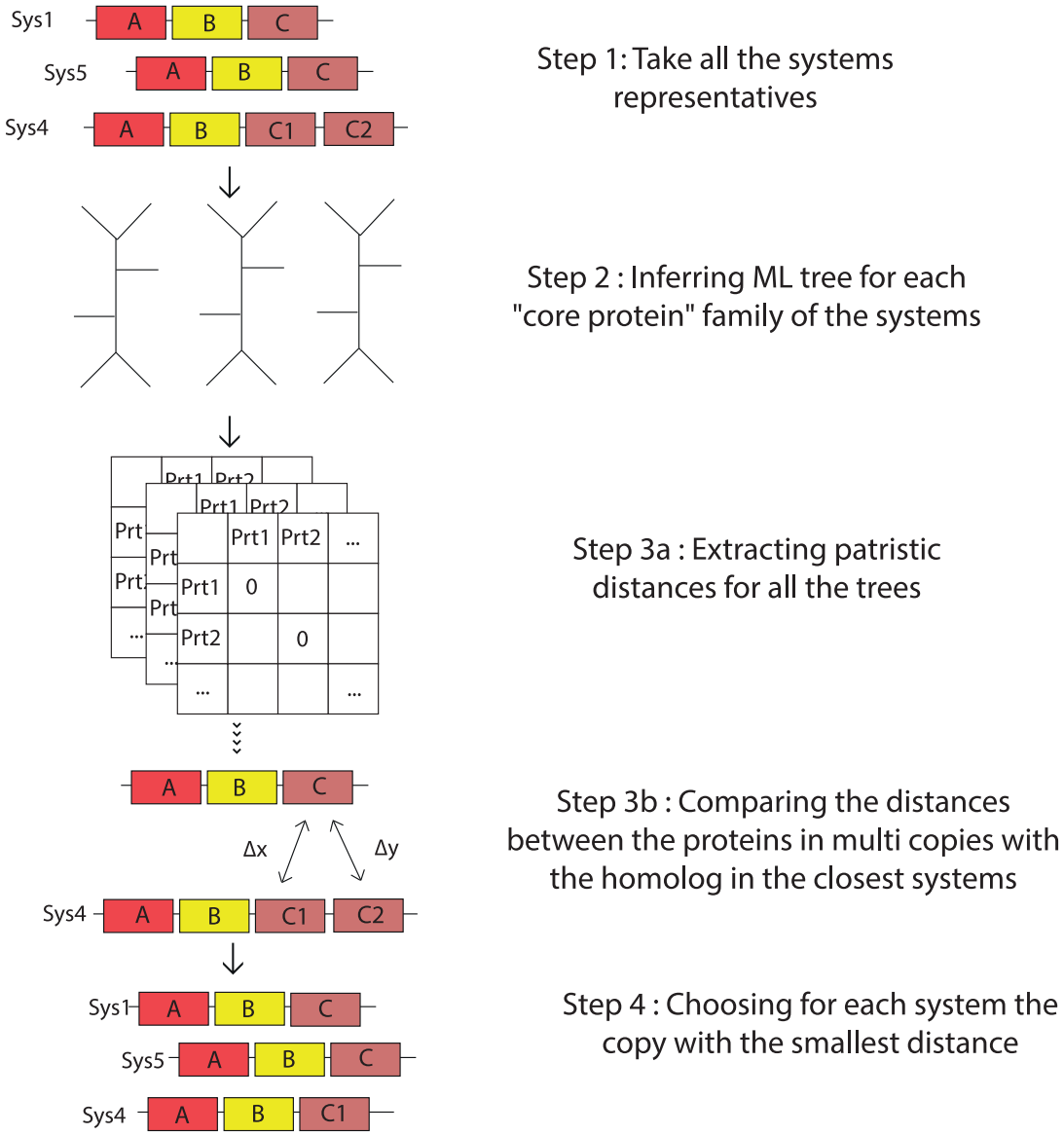


Figure 3.16 – S16 Fig. Schema of the workflow used to choose the species-specific paralogs.

CHAPITRE 4

CONCLUSION

Dans ce chapitre, nous avons développé une méthode permettant de détecter, à large échelle, les systèmes composant la superfamille des filaments de type IV. Cette méthode de détection est basée sur deux analyses précédentes [152, 250]. Ces deux précédentes analyses ont permis d'obtenir des profils HMM et un outil pour commencer la détection des TFFs. Leur combinaison a permis de créer un outil efficace pour l'analyse génomique des filaments de type IV dans les génomes bactériens et archéens.

L'utilisation de cet outil et de cette méthode a permis l'analyse la plus exhaustive des filaments de type IV à ce jour. Cette étude constitue une intéressante base pour de futures analyses sur les filaments de type IV, principalement par la confirmation ou l'infirmation d'observations précédentes. Par exemple, nous avons confirmé la grande modularité des TFFs nous intéressant à leur organisation génétique. Nous avons permis d'avoir une vue plus générale de la distribution des TFFs suivant le type de système. Nous avons également montré que l'histoire phylogénétique des TFFs laisse apercevoir un transfert entre archées et bactéries qui a permis l'émergence le Tad, ainsi que l'acquisition de gènes suivi de la perte de ceux-ci au cours de l'évolution (p. ex. l'ATPase PilT perdu chez les T2SS et ComM). La paraphilie des T4aP montre que ce système semble avoir été à l'origine d'autres systèmes (T2SS, MSH), mais aussi qu'il existe probablement d'autres systèmes appartenant à la TFF-SF ressemblant à des T4aP.

Pour comprendre la relation qu'il existe entre tous les membres de la TFF-SF, nous nous sommes concentrés sur les gènes partagés entre les systèmes (ATPase, IM plateforme, piline majeure et mineure, sécrétine et prépiline peptidase). Grâce à ces composants, nous avons inféré des phylogénies pour chacun des composants. Par la suite, en utilisant les composants dont l'histoire était congruente, nous avons pu inférer une phylogénie des systèmes pour mieux comprendre l'histoire générale des TFFs.

Ce travail a permis de voir que la majorité des membres de la superfamille des TFFs sont monophylétiques, à l'exception du T4aP et de l'Archaeal-T4P. Dans le

cas des TFFs chez les archées, cela est sûrement dû en grande partie du fait que dans cette analyse, nous n'avons pas pu faire de modèles permettant de sous-typer les systèmes (aucun modèle spécifique pour l'archaellum, l'Epd, l'Ups...). Quand on regarde les TFFs chez les archées, on peut se rendre compte que l'archaellum semble former un groupe à part, de même pour l'Epd mais cela semble plus difficile pour les autres TFFs archéens. On a aussi pu voir que la sécrétine semble avoir été acquise tôt dans l'ancêtre des TFFs. Cette remarque est soutenue par le fait que les TFFs dans les espèces monodermes se trouvent inclus dans des groupes de TFFs présents chez les didermes. Une étude des clades didermes chez les firmicutes a aussi pu montrer que la sécrétine a probablement suivi des gains et pertes successifs au sein des firmicutes.

Il est intéressant de noter que le Tad qui était jusqu'alors considéré comme faisant parti du groupe des T4bP a en fait une histoire différente de celle des autres T4bP et plus généralement des TFFs bactériens. Celui-ci est vraisemblablement issu du transfert, des archées aux bactéries, d'un système proche du système archéen Epd. Le Tad partage en effet des similarités avec ce système, comme la présence de deux gènes d'IM plateforme. De plus ces deux TFFs sont groupes frères dans toutes les reconstructions phylogénétiques. Ces résultats permettent donc de mieux comprendre pourquoi le Tad comporte des gènes qui ne sont pas présent dans les autres TFFs bactériens, comme par exemple l'ATPase TadZ qui ressemble aux ATPases arCOG00589 et arCOG05608 présentes chez les Epd.

Ce travail s'est basé sur le postulat que si les gènes conservés sont détectables, alors le système détecté est probablement fonctionnel. Cependant cela doit être confirmé avec des validations expérimentales pour attester de la fonctionnalité des machineries. Cela est impossible à large échelle et nécessite un effort au niveau de la communauté. Quoi qu'il en soit, des analyses bioinformatiques adaptées peuvent donner des indices sur le fait qu'un système est vraisemblablement fonctionnel ou non. Dans le cas d'un système défectif, on pourrait s'attendre à ce que des parties importantes de la machinerie manquent. À titre d'exemple, il a été montré, chez *Neisseria meningitidis*, qu'un minimum de huit gènes spécifiques est requis pour qu'un T4aP soit fonctionnel [118]. Quatre d'entre eux, *pilM*, *pilN*, *pilO* et *pilP*, sont essentiels à l'assemblage mais il ne semble pas exister d'homologues dans la grande majorité des TFFs (T4bP, ComM, Tad et Archaeal-T4P). Cela suggère qu'il pourrait y avoir un nombre de gène minimal, inférieur à huit, permettant d'obtenir un système fonctionnel.

Finalement ce travail peut permettre de mieux définir les différents systèmes et ainsi de les discriminer ou grouper plus facilement. On pourrait ainsi penser à utiliser l'arbre des systèmes. Par exemple, si on inclut dans l'arbre un nouveau système expérimentalement validé comme fonctionnel mais non typé, grâce à la position relative de ses gènes par rapport aux systèmes connus, cela peut permettre de typer ce système non plus seulement par la fonction qu'il semble procurer à la bactérie, mais en ajoutant l'information de sa position dans l'arbre par rapport aux autres systèmes. Ainsi des systèmes comme le pseudo-T2SS de *Cytophaga hutchinsonii* [251], pourraient être plus étudiés comme étant un nouveau type de

système et non pas comme assimilés à des systèmes décrits précédemment qui ont une fonction similaire. Ou encore l'étude plus poussé des pilus MSH qui ont tendance, à être classés comme T4aP et donc ne sont pas étudiés pour ce qu'ils sont vraiment à savoir un TFF qui a vraisemblablement évolué d'un T4aP pour se spécialiser dans certaines fonctions. Mais cela peut aussi permettre de grouper des systèmes qui sont étudié de façon indépendante alors qu'ils semblent faire partie de la même famille et donc ne devraient plus être étudiés comme des systèmes isolés (p. ex. les différents T4bP, ou certains Archaeal-T4P).

Quatrième partie

Conclusions et perspectives

Le but de ce travail était de mieux comprendre l'histoire évolutive de la superfamille des filaments de type IV, connue pour son rôle dans la pathogénicité des bactéries mais également pour son rôle lors de la transformation naturelle. Une partie de ce travail s'est concentré sur la détection des systèmes. J'ai développé et rendu disponible pour la communauté un outil et des méthodes permettant de révéler l'étendue de la diversité des filaments de type IV chez les procaryotes. L'analyse de ces filaments a permis de révéler les différences et similarités entre eux.

Parmi les filaments de type IV, il y a une répartition parmi les bactéries et archées qui peut aller de la présence uniquement chez un phylum particulier à la présence parmi tous les phyla bactériens. Une mise à jour de la distribution est présentée dans la figure 4.1. Ce nouveau jeu de données [1] est toujours assez biaisé par rapport aux bactéries pathogènes (l'abondance de bactéries pathogènes y est plus grande que dans la nature), comme le montre la forte abondance de génomes de protéobactéries (divisé dans ses principales classes) et de firmicutes. Les TFFs (T4aPs et Tads) sont, comme c'est le cas avec la base de données utilisées dans la partie précédente (partie III), fréquents dans tous les clades majeurs et ces pili sont absents dans peu de clades. Il est important de noter que bien que je ne détecte pas de TFFs, les clades peuvent en posséder. En particulier, beaucoup de TFFs connus dans les *Chlamydiae* (p. ex. T2SS) restent non détectés du fait de la difficulté à détecter leurs pilines, ainsi que les potentiels autres gènes du système. Cependant, on peut quand même observer un certain nombre de systèmes « génériques » laissant sous-entendre l'existence de TFFs dans les clades où ils ne sont pas détectés (p. ex. *Chlamydiae*, *Tenericutes*, *Tissierellia*...) (voir fig. 4.1). L'augmentation du nombre de génomes de nouvelles espèces séquencées permet également d'accroitre le nombre de systèmes détectés, en effet on peut observer dans la figure 4.1 que l'on détecte des systèmes dans des clades pour lesquels je n'avais pas de systèmes lors de l'analyse de la troisième partie (cf. fig. 4 de l'article 2). Ces nouveaux systèmes, dans des clades plus divers, pourront être utilisé pour augmenter la diversité des séquences présentes dans les différent profils HMM ce qui peut permettre ainsi de détecter des séquences dans des clades plus divers phylogénétiquement. Il est cependant à noter que certaines annotations taxonomiques des génomes de NCBI RefSeq sont à prendre avec précaution. En effet, on peut remarquer dans la figure 4.1 que l'on détecte deux ComM au seins des alphaprotéobatéries, cependant, lorsque l'on étudie plus attentivement ces génomes, on peut observer qu'ils ne possèdent pas de sécrétines, ce qui laisse à penser que ces bactéries sont monodermes et qu'il y a possiblement eu une mauvaise annotation de ces génomes lors de la soumission de ceux-ci.

En plus de leur distribution non homogène au sein des génomes (certains étant présents dans des clades précis et d'autres présent dans presque tous les clades), les systèmes ne montrent pas tout à fait la même distribution lorsque l'on regarde

1. Ce jeu de données est composé de 13299 génomes de bactéries et 283 génomes d'archées obtenus de NCBI RefSeq en avril 2019

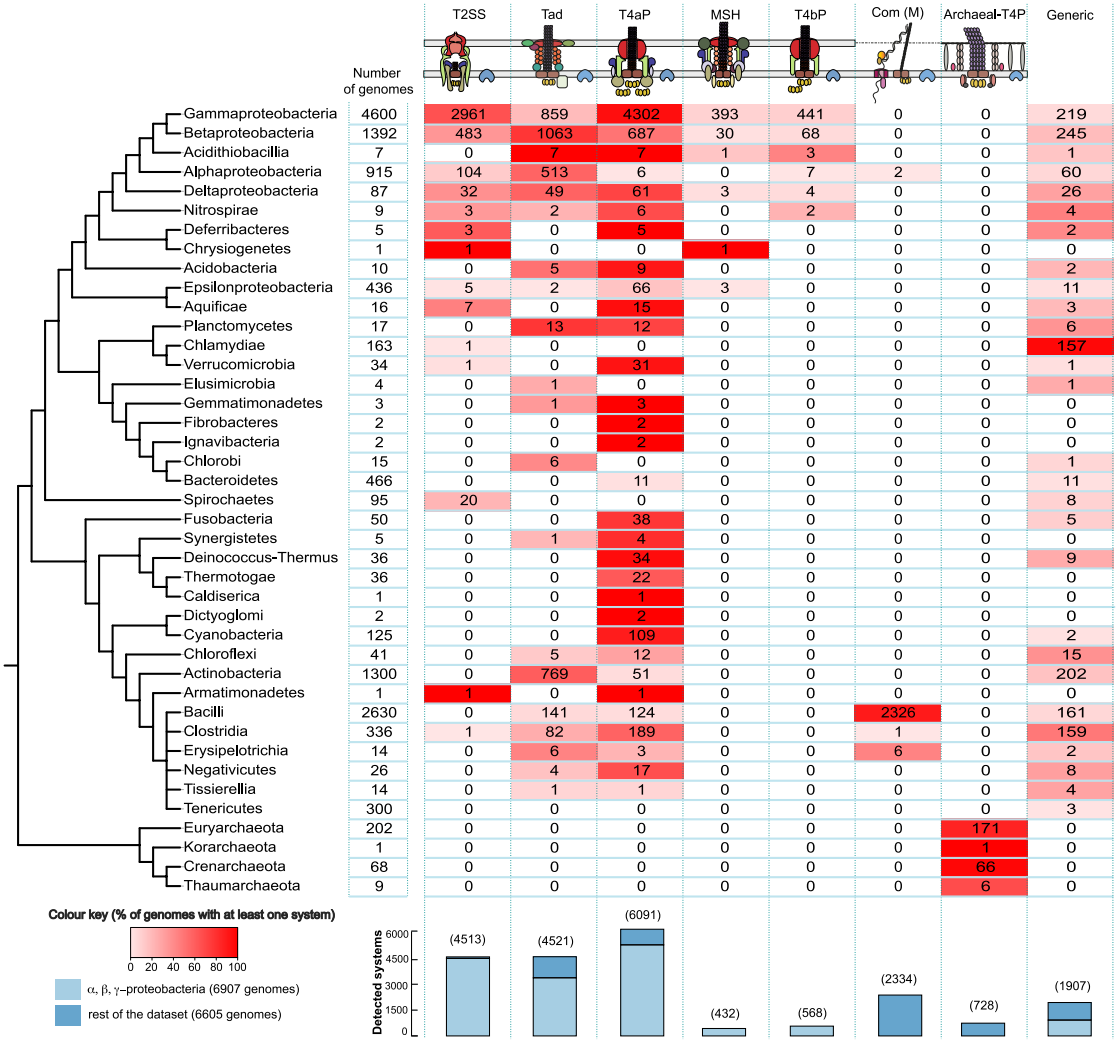
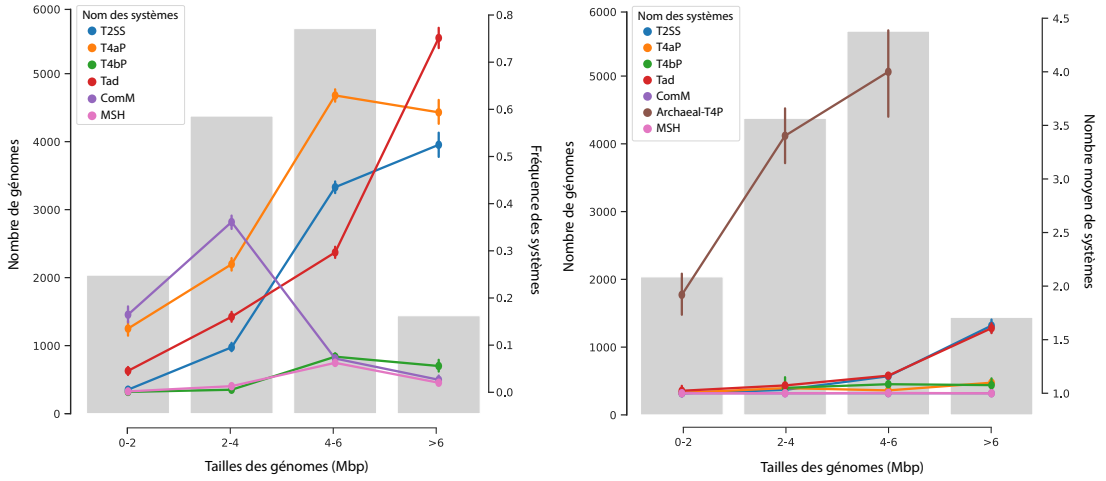


Figure 4.1 – Distribution taxonomique des systèmes dans les bactéries et les archées avec les modèles finaux sur la database d'avril 2019. Les cellules indiquent le nombre de génomes possédant au moins un système détecté. Le gradient de couleur de la cellule représente la proportion de génomes avec au moins un système dans le clade. Le diagramme en bâtons indique le nombre total de systèmes détectés. Les barres sont séparées en deux catégories : Alpha-, Beta-, Gamma-protéobactéries contre les autres clades. Le cladogramme symbolise les relations approximatives entre les taxons bactériens et archéens analysés dans cette étude.

la taille des génomes dans lesquels on les détecte (fig. 4.2a). La corrélation positive de la fréquence des T2SS, Tad, et T4aP avec la taille de leur « hôte » peut être expliqué par le fait que les grands génomes participent plus souvent dans les transferts horizontaux (bien que l'on puisse également soutenir que ces génomes font plus de transferts parce qu'ils ont plus d'éléments mobiles) [252, 253]. Les génomes les plus petits subissent probablement peu de transferts horizontaux de gènes, parce que ces espèces ont majoritairement un style de vie endosymbiotique qui les conduit à l'isolation [254]. On observe également pour les T2SS et les Tad que le nombre moyen de systèmes est nettement supérieur à 1 (fig. 4.2b) ce qui implique que les grands génomes ont tendance à accumuler ces systèmes, permettant probablement une meilleure flexibilité par rapport à l'environnement. On n'observe

pas cette accumulation chez les T4aP malgré le fait qu'ils soient eux aussi assez fréquent dans les grands génomes. En ce qui concerne les TFFs présent chez les archées, on observe (données non présentées) également une corrélation entre la taille des génomes et la fréquence des systèmes. Cela peut s'expliquer par le fait que les grands génomes sont plus souvent sujet aux transferts horizontaux. On peut aussi observer, dans la figure 4.2b, que plus le génome est grand plus le nombre d'Archaeal-T4P dans le génome l'est aussi. Cependant, dans cette analyse, je ne distinguais pas les différents types de TFFs chez les archées, et donc la distribution des différents types de TFFs archéens ne peut être observé ici. En ce qui concerne ComM, ce système est très présent chez les génomes de petites tailles et très peu dans les génomes des grandes tailles. Ceci est probablement dû au fait qu'il n'est présent que chez les firmicutes, qui ont un génome plus petit que les protéobactérie (la taille moyenne des firmicutes est de 3,2 Mbp et celle des protéobactéries est de 4,5 Mbp).



(a) Fréquence **(b)** Nombre moyen

Figure 4.2 – Filaments de type IV en fonction de la taille du génome « hôte ». Le diagramme en bâton représente le nombre de génomes pour chaque catégorie de taille de génomes. (a) Les tracés linéaires représentent la fréquence (axe de droite) d'un élément donné dans pour une catégorie de taille de génomes. Les barres d'erreur représentent l'intervalle de confiance de 95% de la moyenne, représentée par un point. Ici les génomes sont réduits aux bactéries pour réduire les biais. (b) Les tracés linéaires représentent le nombre moyen (axe de droite) d'un élément donné pour une catégorie de taille de génomes dans les génomes avec au moins un élément donné détecté. Les barres d'erreur représentent l'intervalle de confiance de 95% de la moyenne, représentée par un point.

Au cours de ce travail de recherche, je ne me suis pas intéressé à trouver un moyen d'annoter les différents TFFs présent chez les archées. Cependant, l'étude phylogénétique de ces systèmes a permis d'observer différents groupes de pili en cartographiant les pili archéens expérimentalement validés sur les différents arbres inférés. Ainsi parmi ces pili on a pu observer, cinq groupes différents de pili (Epd, Aap, Bindosome + Ups, archaellum et les pili des Halobacteria) (fig. 4.3). Phylogénétiquement, ces systèmes sont donc aussi diversifiés que leur cousins bactériens et la diversification de ces TFFs est un pan d'étude qui mérite plus d'investigations. On pourrait ainsi facilement, en utilisant les données phylogénétiques, développer

des modèles MacSyFinder permettant d'annoter plus précisément les systèmes. Par exemple, le fait que le bindosome et l'Ups soient finalement compris dans le même groupe phylogénétique suggère la possibilité de regrouper ces deux pili sous le même nom. Des études plus précises de ces pili permettraient aussi d'avoir plus d'exemples des différents pili qui sont, à l'instar du T4aP, en plusieurs loci dans le génome. En effet les systèmes en plusieurs loci sont difficilement caractérisable avec les outils actuels car la majorité des protéines entre ces TFFs archéens sont homologues.

Une autre découverte intéressante de ce travail est le proche lien de parenté qui est observé entre les pili Tad et Epd. Il montre, d'un point de vue évolutifs, un fait particulier concernant l'interaction entre les bactéries et archées. En effet, le Tad est issue du transfert d'une archée à une bactérie d'un système homologue à l'Epd. Par la suite, ce proto-Tad a continué à être transféré comme le montre la grande dispersion du Tad parmi tous les différents phyla bactériens. Cette dispersion montre l'importance et l'utilité de ce pilus chez les bactéries pour s'adapter à leur environnement.

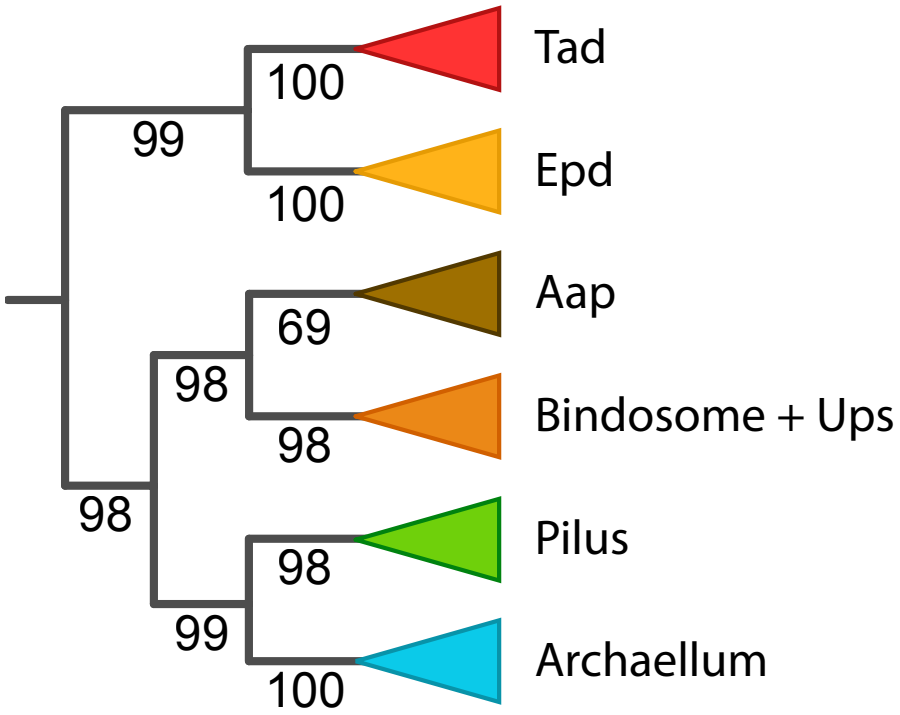


Figure 4.3 – Arbre schématique de la distribution des différents TFFs chez les archées.

Pour finir, la mission d'identifier les filaments de type IV peut être améliorée, surtout pour les T2SS et Archaeal-T4P. Durant cette thèse, j'ai essayé de m'attaquer à ce problème à l'aide de trois approches :

- Premièrement, j'ai essayé d'utiliser la position phylogénétique des systèmes pour me permettre d'annoter précisément les systèmes dont les modèles initiaux n'avaient pas réussi à annoter comme un membre précis des TFFs. Ceci

m'a permis d'ajouter des séquences de systèmes non validés expérimentalement aux séquences initiales de mes profils HMM et de moduler le nombre de gènes requis pour typer le système. Cependant, certain locus de ces systèmes ne sont pas bien définis ou il manque certains gènes (p. ex. chez *Chlamydia trachomatis* on a un T2SS atypique qui n'a que quatre gènes [255]), ce qui ne permet pas de proprement le discriminer des autres TFFs. De plus, même dans les cas où le cluster est bien défini, il existe des cas où les profils HMM ne permettent pas de discriminer avec précision la protéine et donc de pouvoir utiliser la protéine pour enrichir un profil HMM avec cette séquence.

- Pour résoudre ce problème de discrimination de protéines, j'ai essayé une deuxième approche complémentaire consistant à utiliser l'arbre phylogénétique des protéines homologues afin d'arriver à voir s'il était possible de les discriminer par rapport à leur position dans l'arbre. Cette approche a ainsi permis pour certaine protéines (comme l'ATPase) de départager les séquences entre l'homologue PilT et PilU facilement. Cependant, cette approche n'a pas permis une discrimination efficace pour toutes les protéines, c'est en effet le cas pour les pilines dont les profils HMM résultant de cette méthode n'ont pas permis d'obtenir une discrimination parfaite des différentes pilines. Cela est peut-être dû au fait que les pilines sont des protéines qui possèdent deux domaines, un petit domaine très conservé parmi les différentes pilines, et un grand domaine très variable. De ce fait,la discrimination est difficile à faire entre les pilines.
- Enfin, j'ai également utilisé une dernière approche permettant d'avoir des profils HMM plus précis. Dans celle-ci, j'ai construit les profils HMM des protéines homologues en passant d'abord par un alignement de toutes les protéines homologues entre elle, puis à partir de cet alignement j'ai construit les profils HMM pour les différents homologues. De cette manière j'ai pu obtenir des profils HMM basés sur des alignements de même longueur, ce qui m'a permis une plus précise discrimination de ces protéines. Cependant les alignements des différentes protéines homologues ne sont pas « parfait » (les alignements possèdent beaucoup d'indel bien que la protéine soit homologue). Ceci est dû au fait que la diversification des protéines s'est faite depuis longtemps (l'ancêtre des TFFs étant probablement présent chez LUCA).

Un autre problème dans la construction des modèles est cette construction est faite manuellement. Elle se base sur la description de la composition en protéines des TFFs dans la littérature. De ce fait il est souvent difficile d'ajuster les paramètres pour avoir un modèle permettant de détecter le maximum de systèmes sans pour autant ne pas augmenter le nombre de mauvaise annotation. Un autre problème de la construction des modèles est que l'évaluation des modèles n'est pas basée sur des approches statistiques, mais sur des seuils imposés basés sur quelques statistiques simples. La procédure permettant de quantifier la sensitivité du modèle (c-à-d de savoir comment le modèle est capable de bien identifier un TFF et ses différents composants) est testé sur des systèmes expérimentalement

validés qui ne représentent pas la grande diversité des systèmes. C'est encore plus difficile d'évaluer sa spécificité (c-à-d sa capacité à séparer le vrai du faux), car il n'existe pas d'informations qui permettent de déterminer un TFF invalide.

Par ailleurs, il existe de nos jours des outils statistiques plus puissants qui n'étaient pas concevable pour des applications biologiques il y a quelques années. Ces méthodes d'apprentissage profond (deep learning) commencent maintenant à être utilisés en biologie pour la détection de motif d'ADN [256, 257]. De plus ces outils requièrent une grosse quantité de séquences annotées pour entraîner le réseau neuronal avant toute prédiction. Pour ce faire, ils s'appuient sur des expériences haut débit [256] ou des données simulées [258]. Les méthodes de détection de TFFs que j'ai développées pourraient permettre de donner assez de labels pour remplir ce manque qui pourrait être utilisé pour entraîner des méthodes plus complexes de machine learning.

Ces outils, méthodes et données que j'ai produites durant cette thèse peuvent être utiles pour la génomique microbienne et la microbiologie. Premièrement, cela permet une analyse détaillée de filaments de type IV qui peut amener de nouvelles questions dans le domaine. Deuxièmement, l'étude de l'évolution de la super-famille des filaments de type IV a montré qu'il y a eu du « bricolage » moléculaire à l'aide des outils disponibles : aucun mécanisme ou composant radicalement nouveau n'a été introduit pour permettre la diversité fonctionnelle des TFFs. Au contraire, les fonctions existantes des composants présents ont été exaptés pour changer la fonction des machineries existantes. Cela a permis d'obtenir des systèmes divers et variés à partir d'un même système ancestral.

Bibliographie

[1] Van Leeuwenhoek Antoni. Observations, communicated to the publisher by mr. antony van leewenhoeck, in a dutch letter of the 9th octob. 1676. here english'd : concerning little animals by him observed in rain-well-sea- and snow water ; as also in water wherein pepper had lain infused. Philosophical Transactions of the Royal Society of London, 12(133) :821–831, 1677.

[2] Christian Gottfried Ehrenberg. Symbolae physicae, volume Zoologica II, Animalia evertebrata. Biodiversity library, 1828.

[3] R. Sender, S. Fuchs, and R. Milo. Revised estimates for the number of human and bacteria cells in the body. PLoS Biol, 14(8) :e1002533, 2016.

[4] J. Carlet. The gut is the epicentre of antibiotic resistance. Antimicrob Resist Infect Control, 1(1) :39, 2012.

[5] C. R. Woese and G. E. Fox. Phylogenetic structure of the prokaryotic domain : the primary kingdoms. Proceedings of the National Academy of Sciences of the United States of America, 74(11) :5088–5090, 1977.

[6] G. E. Fox, L. J. Magrum, W. E. Balch, R. S. Wolfe, and C. R. Woese. Classification of methanogenic bacteria by 16s ribosomal rna characterization. Proceedings of the National Academy of Sciences of the United States of America, 74(10) :4537–4541, 1977.

[7] Laura Eme and W. Ford Doolittle. Archaea. Current Biology, 25(19) :R851 – R855, 2015.

[8] Corinna Bang and Ruth A. Schmitz. Archaea associated with human surfaces : not to be underestimated. FEMS Microbiology Reviews, 39(5) :631–648, 04 2015.

[9] Sonja-Verena Albers and Benjamin H. Meyer. The archaeal cell envelope. Nature Reviews Microbiology, 9 :414, 2011.

[10] Uwe B. Sleytr and Margit Sara. Bacterial and archaeal s-layer proteins : structure-function relationships and their biotechnological applications. Trends in Biotechnology, 15(1) :20–26, 1997.

[11] T. J. Silhavy, D. Kahne, and S. Walker. The bacterial cell envelope. Cold Spring Harb Perspect Biol, 2(5) :a000414, 2010.

[12] Iain C Sutcliffe. A phylum level perspective on bacterial cell envelope architecture. Trends Microbiol, 18(10) :464–470, 2010.

[13] Wikipedia gram-negative bacteria. `https://en.wikipedia.org/wiki/Gram-negative_bacteria`. Accessed : 2019-05-09.

[14] H. Nikaido. Molecular basis of bacterial outer membrane permeability revisited. Microbiol Mol Biol Rev, 67(4) :593–656, 2003.

[15] L. Brown, J. M. Wolf, R. Prados-Rosales, and A. Casadevall. Through the wall : extracellular vesicles in gram-positive bacteria, mycobacteria and fungi. Nat Rev Microbiol, 13(10) :620–30, 2015.

[16] W. Vollmer. Structural variation in the glycan strands of bacterial peptidoglycan. FEMS Microbiol Rev, 32(2) :287–306, 2008.

[17] C. R. Raetz and C. Whitfield. Lipopolysaccharide endotoxins. Annu Rev Biochem, 71 :635–700, 2002.

[18] C. M. Khursigara, S. F. Koval, D. M. Moyles, and R. J. Harris. Inroads through the bacterial cell envelope : seeing is believing. Can J Microbiol, 64(9) :601–617, 2018.

[19] S. S. Abby and E. P. Rocha. The non-flagellar type iii secretion system evolved from the bacterial flagellum and diversified into host-cell adapted systems. PLoS Genetics, 8(9) :e1002983, 2012.

[20] C. R. Peabody, Y. J. Chung, M. R. Yen, D. Vidal-Ingigliardi, A. P. Pugsley, and Jr. Saier, M. H. Type ii protein secretion and its relationship to bacterial type iv pili and archaeal flagella. Microbiology, 149(Pt 11) :3051–72, 2003.

[21] B. Daum and V. Gold. Twitch or swim : towards the understanding of prokaryotic motion based on the type iv pilus blueprint. Biol Chem, 399(7) :799–808, 2018.

[22] Y. Kinosita, N. Uchida, D. Nakane, and T. Nishizaka. Direct observation of rotation and steps of the archaellum in the swimming halophilic archaeon halobacterium salinarum. Nat Microbiol, 1(11) :16148, 2016.

[23] Dael Wolfle. Mcgraw-hill encyclopedia of science and technology. mcgraw-hill, new york, 1960. 15 vols. 175. Science, 133(3450) :374–375, 1961.

[24] K. F. Jarrell and M. J. McBride. The surprisingly diverse ways that prokaryotes move. Nat Rev Microbiol, 6(6) :466–76, 2008.

[25] J. M. Skerker and H. C. Berg. Direct observation of extension and retraction of type iv pili. Proc Natl Acad Sci U S A, 98(12) :6901–4, 2001.

[26] C. L. Giltner, Y. Nguyen, and L. L. Burrows. Type iv pilin proteins : versatile molecular modules. Microbiol Mol Biol Rev, 76(4) :740–72, 2012.

[27] J. Henrichsen. Twitching motility. Annu Rev Microbiol, 37 :81–93, 1983.

[28] John S. Mattick. Type iv pili and twitching motility. Annual Review of Microbiology, 56(1) :289–314, 2002.

[29] N. Biais, B. Ladoux, D. Higashi, M. So, and M. Sheetz. Cooperative retraction of bundled type iv pili enables nanonewton force generation. PLoS Biol, 6(4) :e87, 2008.

[30] N. Biais, D. L. Higashi, J. Brujic, M. So, and M. P. Sheetz. Force-dependent polymorphism in type iv pili reveals hidden epitopes. Proc Natl Acad Sci U S A, 107(25) :11358–63, 2010.

[31] R. Marathe, C. Meel, N. C. Schmidt, L. Dewenter, R. Kurre, L. Greune, M. A. Schmidt, M. J. Muller, R. Lipowsky, B. Maier, and S. Klumpp. Bacterial twitching motility is coordinated by a two-dimensional tug-of-war with directional memory. Nat Commun, 5 :3759, 2014.

[32] A. J. Merz, M. So, and M. P. Sheetz. Pilus retraction powers bacterial twitching motility. Nature, 407(6800) :98–102, 2000.

[33] M. J. Muller, S. Klumpp, and R. Lipowsky. Bidirectional transport by molecular motors : enhanced processivity and response to external forces. Biophys J, 98(11) :2610–8, 2010.

[34] L. Tala, A. Fineberg, P. Kukura, and A. Persat. Pseudomonas aeruginosa orchestrates twitching motility by sequential control of type iv pili movements. Nat Microbiol, 2019.

[35] J. W. Costerton, Z. Lewandowski, D. E. Caldwell, D. R. Korber, and H. M. Lappin-Scott. Microbial biofilms. Annu Rev Microbiol, 49 :711–45, 1995.

[36] O. E. Petrova and K. Sauer. Sticky situations : key components that control bacterial surface attachment. J Bacteriol, 194(10) :2413–25, 2012.

[37] T. Danhorn, M. Hentzer, M. Givskov, M. R. Parsek, and C. Fuqua. Phosphorus limitation enhances biofilm formation of the plant pathogen agrobacterium tumefaciens through the phor-phob regulatory system. J Bacteriol, 186(14) :4492–501, 2004.

[38] J. E. Heindl, M. E. Hibbing, J. Xu, R. Natarajan, A. M. Buechlein, and C. Fuqua. Discrete responses to limitation for iron and manganese in agrobacterium tumefaciens : Influence on attachment and biofilm formation. J Bacteriol, 198(5) :816–29, 2015.

[39] C. Berne, A. Ducret, G. G. Hardy, and Y. V. Brun. Adhesins involved in attachment to abiotic surfaces by gram-negative bacteria. Microbiol Spectr, 3(4), 2015.

[40] C. Berne, C. K. Ellison, A. Ducret, and Y. V. Brun. Bacterial adhesion at the single-cell level. Nat Rev Microbiol, 16(10) :616–627, 2018.

[41] Y. Ren, C. Wang, Z. Chen, E. Allan, H. C. van der Mei, and H. J. Busscher. Emergent heterogeneous microenvironments in biofilms : substratum surface heterogeneity and bacterial adhesion force-sensing. FEMS Microbiol Rev, 42(3) :259–272, 2018.

[42] S. Bagherifard, D. J. Hickey, A. C. de Luca, V. N. Malheiro, A. E. Markaki, M. Guagliano, and T. J. Webster. The influence of nanostructured features on bacterial adhesion and bone cell functions on severely shot peened 316l stainless steel. Biomaterials, 73 :185–97, 2015.

[43] Jr. Dunne, W. M. Bacterial adhesion : seen any good biofilms lately ? Clin Microbiol Rev, 15(2) :155–66, 2002.

[44] J. Palmer, S. Flint, and J. Brooks. Bacterial cell attachment, the beginning of a biofilm. J Ind Microbiol Biotechnol, 34(9) :577–88, 2007.

[45] L. A. Pratt and R. Kolter. Genetic analysis of escherichia coli biofilm formation : roles of flagella, motility, chemotaxis and type i pili. Mol Microbiol, 30(2) :285–93, 1998.

[46] C. K. Ellison, J. Kan, R. S. Dillard, D. T. Kysela, A. Ducret, C. Berne, C. M. Hampton, Z. Ke, E. R. Wright, N. Biais, A. B. Dalia, and Y. V. Brun.

Obstruction of pilus retraction stimulates bacterial surface sensing. Science, 358(6362) :535–538, 2017.

[47] Y. Wang, C. H. Haitjema, and C. Fuqua. The ctp type ivb pilus locus of agrobacterium tumefaciens directs formation of the common pili and contributes to reversible surface attachment. J Bacteriol, 196(16) :2979–88, 2014.

[48] L. L. Burrows. Pseudomonas aeruginosa twitching motility : type iv pili in action. Annu Rev Microbiol, 66 :493–520, 2012.

[49] J. Maynard Smith and J. Haig. The hitch-hiking effect of a favourable gene. Genet Res, 23 :23–35, 1974.

[50] K. N. Laland, T. Uller, M. W. Feldman, K. Sterelny, G. B. Muller, A. Moczek, E. Jablonka, and J. Odling-Smee. The extended evolutionary synthesis : its structure, assumptions and predictions. Proc Biol Sci, 282(1813) :20151019, 2015.

[51] A. Stoltzfus and D. M. McCandlish. Mutational biases influence parallel adaptation. Mol Biol Evol, 34(9) :2163–2172, 2017.

[52] A. Wagner. Distributed robustness versus redundancy as causes of mutational robustness. Bioessays, 27(2) :176–88, 2005.

[53] I. Martincorena, A. S. Seshasayee, and N. M. Luscombe. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. Nature, 485(7396) :95–8, 2012.

[54] M. Lynch. Evolution of the mutation rate. Trends Genet, 26(8) :345–52, 2010.

[55] Michael Lynch, John Conery, and Reinhard Burger. Mutation accumulation and the extinction of small populations. The American Naturalist, 146(4) :489–518, 1995.

[56] D. Berger, J. Stangberg, K. Grieshop, I. Martinossi-Allibert, and G. Arnqvist. Temperature effects on life-history trade-offs, germline maintenance and mutation rate under simulated climate warming. Proc Biol Sci, 284(1866), 2017.

[57] J. B. S. Haldane. The part played by recurrent mutation in evolution. The American Naturalist, 67(708) :5–19, 1933.

[58] S. Ohno. Evolution by gene duplication. Springer-Verlag, Berlin, 1970.

[59] T. J. Treangen and EPC Rocha. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet, 7 :e1001284, 2011.

[60] S. K. Kummerfeld and S. A. Teichmann. Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet, 21 :25–30, 2005.

[61] E.L. Tatum and J. Lederberg. Gene recombination in the bacterium escherichia coli. J Bacteriol, 53 :673–684, 1947.

[62] N. D. Zinder and J. Lederberg. Genetic exchange in salmonella. J Bacteriol, 64(5) :679–99, 1952.

[63] F. Griffith. The significance of pneumococcal types. The Journal of hygiene, 27(2) :113–59, 1928.

[64] O. T. Avery, C. M. Macleod, and M. McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types : Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. The Journal of experimental medicine, 79(2) :137–58, 1944.

[65] H. Shimizu, Y. Maruyama, M. Hashiba, S. Muraki, and K. Shimoji. Effect of althesin on human central and peripheral nervous system - with special reference to the mecanism of development of involuntary movement. Masui, 26(4) :442–8, 1977.

[66] A. Kerr. Transfer of virulence between isolates of agrobacterium. Nature, 223(5211) :1175–1176, 1969.

[67] J. Huang, Y. Xu, and J. P. Gogarten. The presence of a haloarchaeal type tyrosyl-trna synthetase marks the opisthokonts as monophyletic. Mol Biol Evol, 22(11) :2142–6, 2005.

[68] E. F. Smith and C. O. Townsend. A plant-tumor of bacterial origin. Science, 25(643) :671–3, 1907.

[69] M. Touchon, J. A. Moura de Sousa, and E. P. Rocha. Embracing the enemy : the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. Curr Opin Microbiol, 38 :66–73, 2017.

[70] K. M. Derbyshire and T. A. Gray. Distributive conjugal transfer : New insights into horizontal gene transfer and genetic exchange in mycobacteria. Microbiol Spectr, 2(1) :MGM2–0022–2013, 2014.

[71] J. Guglielmini, L. Quintais, M. P. Garcillan-Barcia, F. de la Cruz, and E. P. Rocha. The repertoire of ice in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. PLoS Genetics, 7(8) :e1002222, 2011.

[72] E. M. te Poele, H. Bolhuis, and L. Dijkhuizen. Actinomycete integrative and conjugative elements. Antonie Van Leeuwenhoek, 94(1) :127–43, 2008.

[73] A. Campbell. The future of bacteriophage biology. Nat Rev Genet, 4(6) :471–7, 2003.

[74] J. Ebel-Tsipis, D. Botstein, and M. S. Fox. Generalized transduction by phage p22 in salmonella typhimurium. i. molecular origin of transducing dna. J Mol Biol, 71(2) :433–48, 1972.

[75] T. Kenzaka, K. Tani, and M. Nasu. High-frequency phage-mediated gene transfer in freshwater environments determined at single-cell level. ISME J, 4(5) :648–59, 2010.

[76] Calum Johnston, Bernard Martin, Gwennaele Fichant, Patrice Polard, and Jean-Pierre Claverys. Bacterial transformation : distribution, shared mechanisms and divergent control. Nat Rev Micro, 12(3) :181–196, 2014.

[77] D. Hofreuter, S. Odenbreit, and R. Haas. Natural transformation competence in helicobacter pylori is mediated by the basic components of a type iv secretion system. Mol Microbiol, 41(2) :379–91, 2001.

[78] H. L. Hamilton and J. P. Dillard. Natural transformation of neisseria gonorrhoeae : from dna donation to homologous recombination. Mol Microbiol, 59 :376–85, 2006.

[79] C. K. Ellison, T. N. Dalia, A. Vidal Ceballos, J. C. Wang, N. Biais, Y. V. Brun, and A. B. Dalia. Retraction of dna-bound type iv competence pili initiates dna uptake during natural transformation in vibrio cholerae. Nat Microbiol, 3(7) :773–780, 2018.

[80] I. Chen and D. Dubnau. Dna uptake during bacterial transformation. Nat Rev Microbiol, 2 :241–9, 2004.

[81] J. C. Mell and R. J. Redfield. Natural competence and the evolution of dna uptake specificity. J Bacteriol, 196(8) :1471–83, 2014.

[82] K. F. Jarrell and S. V. Albers. The archaellum : an old motility structure with a new name. Trends Microbiol, 20(7) :307–12, 2012.

[83] J. L. Berry and V. Pelicic. Exceptionally widespread nanomachines composed of type iv pilins : the prokaryotic swiss army knives. FEMS Microbiol Rev, 39(1) :134–54, 2015.

[84] E. A. Jouravleva, G. A. McDonald, J. W. Marsh, R. K. Taylor, M. Boesman-Finkelstein, and R. A. Finkelstein. The vibrio cholerae mannose-sensitive hemagglutinin is the receptor for a filamentous bacteriophage from v. cholerae o139. Infect Immun, 66(6) :2535–9, 1998.

[85] J. W. Marsh and R. K. Taylor. Genetic and transcriptional analyses of the vibrio cholerae mannose-sensitive hemagglutinin type 4 pilus gene locus. Journal of Bacteriology, 181(4) :1110–7, 1999.

[86] Lisa A. Fitzgerald, Emily R. Petersen, Richard I. Ray, Brenda J. Little, Candace J. Cooper, Erinn C. Howard, Bradley R. Ringeisen, and Justin C. Biffinger. Shewanella oneidensis mr-1 msh pilin proteins are involved in extracellular electron transfer in microbial fuel cells. Process Biochemistry, 47(1) :170–174, 2012.

[87] C. K. Wairuri, J. E. van der Waals, A. van Schalkwyk, and J. Theron. Ralstonia solanacearum needs flp pili for virulence on potato. Molecular Plant-Microbe Interactions, 25(4) :546–56, 2012.

[88] T. R. Hirst, J. Sanchez, J. B. Kaper, S. J. Hardy, and J. Holmgren. Mechanism of toxin secretion by vibrio cholerae investigated in strains harboring plasmids that encode heat-labile enterotoxins of escherichia coli. Proc Natl Acad Sci U S A, 81(24) :7752–6, 1984.

[89] Aleksandra E Sikora, Ryszard A Zielke, Daniel A Lawrence, Philip C Andrews, and Maria Sandkvist. Proteomic analysis of the vibrio cholerae type ii secretome reveals new proteins, including three related serine proteases. J Biol Chem, 286(19) :16555–16566, 2011.

[90] F. Cadoret, G. Ball, B. Douzi, and R. Voulhoux. Txc, a new type ii secretion system of pseudomonas aeruginosa strain pa7, is regulated by the ttss/ttsr two-component system and directs specific secretion of the cbpe chitin-binding protein. Journal of Bacteriology, 196(13) :2376–2386, 2014.

[91] S. DebRoy, J. Dao, M. Soderberg, O. Rossier, and N. P. Cianciotto. Legionella pneumophila type ii secretome reveals unique exoproteins and a chitinase that promotes bacterial persistence in the lung. Proc Natl Acad Sci U S A, 103(50) :19146–51, 2006.

[92] N. Roux, J. Spagnolo, and S. de Bentzmann. Neglected but amazingly diverse type ivb pili. Res Microbiol, 163(9-10) :659–73, 2012.

[93] K. Mazariego-Espinosa, A. Cruz, M. A. Ledesma, S. A. Ochoa, and J. Xicohtencatl-Cortes. Longus, a type iv pilus of enterotoxigenic escherichia coli, is involved in adherence to intestinal epithelial cells. J Bacteriol, 192(11) :2791–800, 2010.

[94] S C Kachlany, P J Planet, M K Bhattacharjee, E Kollia, R DeSalle, D H Fine, and D H Figurski. Nonspecific adherence by actinobacillus actinomycetemcomitans requires genes widespread in bacteria and archaea. Journal of Bacteriology, 182(21) :6169–6176, 2000.

[95] K. V. Korotkov, M. Sandkvist, and W. G. Hol. The type ii secretion system : biogenesis, molecular architecture and mechanism. Nature Reviews. Microbiology, 10(5) :336–51, 2012.

[96] A. Wagner, R. J. Whitaker, D. J. Krause, J. H. Heilers, M. van Wolferen, C. van der Does, and S. V. Albers. Mechanisms of gene flow in archaea. Nat Rev Microbiol, 15(8) :492–501, 2017.

[97] N. Patenge, A. Berendes, H. Engelhardt, S. C. Schuster, and D. Oesterhelt. The fla gene cluster is involved in the biogenesis of flagella in halobacterium salinarum. Mol Microbiol, 41(3) :653–63, 2001.

[98] Sandy Y M Ng, Bonnie Chaban, and Ken F Jarrell. Archaeal flagella, bacterial flagella and type iv pili : a comparison of genes and posttranslational modifications. Journal of molecular microbiology and biotechnology, 11(3-5) :167–191, 2006.

[99] K. S. Makarova, E. V. Koonin, and S. V. Albers. Diversity and evolution of type iv pili systems in archaea. Front Microbiol, 7 :667, 2016.

[100] Y. W. Chang, L. A. Rettberg, A. Treuner-Lange, J. Iwasa, L. Sogaard-Andersen, and G. J. Jensen. Architecture of the type iva pilus machine. Science, 351(6278), 2016.

[101] Mark S. Strom and Stephen Lory. Structure-function and biogenesis of the type iv pili. Annual Review of Microbiology, 47(1) :565–596, 1993.

[102] M. Coureuil, H. Lecuyer, M. G. Scott, C. Boularan, H. Enslen, M. Soyer, G. Mikaty, S. Bourdoulous, X. Nassif, and S. Marullo. Meningococcus hijacks a beta2-adrenoceptor/beta-arrestin pathway to cross brain microvasculature endothelium. Cell, 143(7) :1149–60, 2010.

[103] M. D. Johnson, C. K. Garrett, J. E. Bond, K. A. Coggan, M. C. Wolfgang, and M. R. Redinbo. Pseudomonas aeruginosa pily1 binds integrin in an rgd- and calcium-dependent manner. PLoS One, 6(12) :e29629, 2011.

[104] S. Helaine, E. Carbonnelle, L. Prouvensier, J. L. Beretti, X. Nassif, and V. Pelicic. Pilx, a pilus-associated protein essential for bacterial aggregation, is a key to pilus-facilitated attachment of neisseria meningitidis to human cells. Mol Microbiol, 55(1) :65–77, 2005.

[105] L. Craig, M. E. Pique, and J. A. Tainer. Type iv pilus structure and bacterial pathogenicity. Nat Rev Microbiol, 2(5) :363–78, 2004.

[106] S. J. Krebs and R. K. Taylor. Protection and attachment of vibrio cholerae mediated by the toxin-coregulated pilus in the infant mouse model. J Bacteriol, 193(19) :5260–70, 2011.

[107] G. A. O'Toole and R. Kolter. Flagellar and twitching motility are necessary for pseudomonas aeruginosa biofilm development. Mol Microbiol, 30(2) :295–304, 1998.

[108] R. Salzer, F. Joos, and B. Averhoff. Type iv pilus biogenesis, twitching motility, and dna uptake in thermus thermophilus : discrete roles of antagonistic atpases pilf, pilt1, and pilt2. Appl Environ Microbiol, 80(2) :644–52, 2014.

[109] X. Han, R. M. Kennan, J. K. Davies, L. A. Reddacliff, O. P. Dhungyel, R. J. Whittington, L. Turnbull, C. B. Whitchurch, and J. I. Rood. Twitching motility is essential for virulence in dichelobacter nodosus. J Bacteriol, 190(9) :3323–35, 2008.

[110] S. Chakraborty, M. Monfett, T. M. Maier, J. L. Benach, D. W. Frank, and D. G. Thanassi. Type iv pili in francisella tularensis : roles of pilf and pilt in fiber assembly, host cell adherence, and virulence. Infect Immun, 76(7) :2852–61, 2008.

[111] V. A. Gold, R. Salzer, B. Averhoff, and W. Kuhlbrandt. Structure of a type iv pilus machinery in the open and closed state. Elife, 4, 2015.

[112] V. Pelicic. Type iv pili : e pluribus unum ? Molecular Microbiology, 68(4) :827–37, 2008.

[113] J. Koo, S. Tammam, S. Y. Ku, L. M. Sampaleanu, L. L. Burrows, and P. L. Howell. Pilf is an outer membrane lipoprotein required for multimerization and localization of the pseudomonas aeruginosa type iv pilus secretin. J Bacteriol, 190(21) :6961–9, 2008.

[114] M. Ayers, L. M. Sampaleanu, S. Tammam, J. Koo, H. Harvey, P. L. Howell, and L. L. Burrows. Pilm/n/o/p proteins form an inner membrane complex that affects the stability of the pseudomonas aeruginosa type iv pilus secretin. J Mol Biol, 394(1) :128–42, 2009.

[115] M. Georgiadou, M. Castagnini, G. Karimova, D. Ladant, and V. Pelicic. Large-scale study of the interactions between proteins involved in type iv pilus biology in neisseria meningitidis : characterization of a subcomplex involved in pilus assembly. Mol Microbiol, 84(5) :857–73, 2012.

[116] Mangayarkarasi Nivaskumar, Javier Santos-Moreno, Christian Malosse, Nathalie Nadeau, Julia Chamot-Rooke, Guy Tran Van Nhieu, and Olivera Francetic. Pseudopilin residue e5 is essential for recruitment by the type 2 secretion system assembly platform. Molecular Microbiology, 101(6) :924–941, 2016.

[117] H. K. Takhar, K. Kemp, M. Kim, P. L. Howell, and L. L. Burrows. The platform protein is essential for type iv pilus biogenesis. J Biol Chem, 288(14) :9721–8, 2013.

[118] V. J. Goosens, A. Busch, M. Georgiadou, M. Castagnini, K. T. Forest, G. Waksman, and V. Pelicic. Reconstitution of a minimal machinery capable of assembling periplasmic type iv pili. Proc Natl Acad Sci U S A, 114(25) :E4978–E4986, 2017.

[119] M. McCallum, S. Tammam, D. J. Little, H. Robinson, J. Koo, M. Shah, C. Calmettes, T. F. Moraes, L. L. Burrows, and P. L. Howell. Piln binding modulates the structure and binding partners of the pseudomonas aeruginosa type iva pilus protein pilm. J Biol Chem, 291(21) :11003–15, 2016.

[120] Mangayarkarasi Nivaskumar, Guillaume Bouvier, Manuel Campos, Nathalie Nadeau, Xiong Yu, Edward H. Egelman, Michael Nilges, and Olivera Francetic. Distinct docking and stabilization steps of the pseudopilus conformational transition path suggest rotational assembly of type iv pilus-like fibers. Structure, 22(5) :685 – 696, 2014.

[121] V. Karuppiah, R. F. Collins, A. Thistlethwaite, Y. Gao, and J. P. Derrick. Structure and assembly of an inner membrane platform for initiation of type iv pilus biogenesis. Proc Natl Acad Sci U S A, 110(48) :E4638–47, 2013.

[122] P. Chiang, M. Habash, and L. L. Burrows. Disparate subcellular localization patterns of pseudomonas aeruginosa type iv pilus atpases involved in twitching motility. J Bacteriol, 187(3) :829–39, 2005.

[123] M. Wolfgang, H. S. Park, S. F. Hayes, J. P. van Putten, and M. Koomey. Suppression of an absolute defect in type iv pilus biogenesis by loss-of-function mutations in pilt, a twitching motility gene in neisseria gonorrhoeae. Proc Natl Acad Sci U S A, 95(25) :14973–8, 1998.

[124] B. Maier, L. Potter, M. So, C. D. Long, H. S. Seifert, and M. P. Sheetz. Single pilus motor forces exceed 100 pn. Proc Natl Acad Sci U S A, 99(25) :16012–7, 2002.

[125] M. McCallum, S. Tammam, A. Khan, L. L. Burrows, and P. L. Howell. The molecular mechanism of the type iva pilus motors. Nat Commun, 8 :15091, 2017.

[126] Y. Nguyen, S. Sugiman-Marangos, H. Harvey, S. D. Bell, C. L. Charlton, M. S. Junop, and L. L. Burrows. Pseudomonas aeruginosa minor pilins prime type iva pilus assembly and promote surface display of the pily1 adhesin. J Biol Chem, 290(1) :601–11, 2015.

[127] S. Kolappan, M. Coureuil, X. Yu, X. Nassif, E. H. Egelman, and L. Craig. Structure of the neisseria meningitidis type iv pilus. Nat Commun, 7 :13015, 2016.

[128] L. Craig, R. K. Taylor, M. E. Pique, B. D. Adair, A. S. Arvai, M. Singh, S. J. Lloyd, D. S. Shin, E. D. Getzoff, M. Yeager, K. T. Forest, and J. A. Tainer. Type iv pilin structure and assembly : X-ray and em analyses of vibrio cholerae toxin-coregulated pilus and pseudomonas aeruginosa pak pilin. Mol Cell, 11(5) :1139–50, 2003.

[129] G. Reguera, K. D. McCarthy, T. Mehta, J. S. Nicoll, M. T. Tuominen, and D. R. Lovley. Extracellular electron transfer via microbial nanowires. Nature, 435(7045) :1098–101, 2005.

[130] M. Vargas, N. S. Malvankar, P. L. Tremblay, C. Leang, J. A. Smith, P. Patel, O. Snoeyenbos-West, K. P. Nevin, and D. R. Lovley. Aromatic amino acids required for pili conductivity and long-range extracellular electron transport in geobacter sulfurreducens. MBio, 4(2) :e00105–13, 2013.

[131] A. Siryaporn, S. L. Kuchma, G. A. O'Toole, and Z. Gitai. Surface attachment induces pseudomonas aeruginosa virulence. Proc Natl Acad Sci U S A, 111(47) :16860–5, 2014.

[132] R. W. Heiniger, H. C. Winther-Larsen, R. J. Pickles, M. Koomey, and M. C. Wolfgang. Infection of human mucosal tissue by pseudomonas aeruginosa requires sequential and mutually dependent virulence factors and a novel pilus-associated adhesin. Cell Microbiol, 12(8) :1158–73, 2010.

[133] T. Rudel, I. Scheurerpflug, and T. F. Meyer. Neisseria pilc protein identified as type-4 pilus tip-located adhesin. Nature, 373(6512) :357–9, 1995.

[134] E. Carbonnelle, S. Helaine, X. Nassif, and V. Pelicic. A systematic genetic analysis in neisseria meningitidis defines the pil proteins required for assembly, functionality, stabilization and export of type iv pili. Mol Microbiol, 61(6) :1510–22, 2006.

[135] C. L. Giltner, M. Habash, and L. L. Burrows. Pseudomonas aeruginosa minor pilins are incorporated into type iv pili. J Mol Biol, 398(3) :444–61, 2010.

[136] H. C. Winther-Larsen, M. Wolfgang, S. Dunham, J. P. van Putten, D. Dorward, C. Lovold, F. E. Aas, and M. Koomey. A conserved set of pilin-like molecules controls type iv pilus dynamics and organelle-associated functions in neisseria gonorrhoeae. Mol Microbiol, 56(4) :903–17, 2005.

[137] S. C. Kachlany, P. J. Planet, R. Desalle, D. H. Fine, D. H. Figurski, and J. B. Kaplan. flp-1, the first representative of a new pilin gene subfamily, is required for non-specific adherence of actinobacillus actinomycetemcomitans. Mol Microbiol, 40(3) :542–54, 2001.

[138] S. R. Kim and T. Komano. The plasmid r64 thin pilus identified as a type iv pilus. J Bacteriol, 179(11) :3594–603, 1997.

[139] Indira Sohel, Jose Luis Puente, William J. Murray, Jaana Vuopio-Varkila, and Gary K. Schoolnik. Cloning and characterization of the bundle-forming pilin gene of enteropathogenic escherichia coli and its distribution in salmonella serotypes. Molecular Microbiology, 7(4) :563–575, 1993.

[140] C E Shaw and R K Taylor. Vibrio cholerae o395 tcpa pilin gene sequence and comparison of predicted protein structural features to those of type 4 pilins. Infection and Immunity, 58(9) :3042–3049, 1990.

[141] Y. W. Chang, A. Kjaer, D. R. Ortega, G. Kovacikova, J. A. Sutherland, L. A. Rettberg, R. K. Taylor, and G. J. Jensen. Architecture of the vibrio cholerae toxin-coregulated pilus machine revealed by electron cryotomography. Nature Microbiol, 2 :16269, 2017.

[142] M. A. De la Cruz, A. Ruiz-Tagle, M. A. Ares, S. Pacheco, J. A. Yanez, L. Cedillo, J. Torres, and J. A. Giron. The expression of longus type 4 pilus of enterotoxigenic escherichia coli is regulated by lngr and lngs and by h-ns, cpxr and crp global regulators. Environ Microbiol, 19(5) :1761–1775, 2017.

[143] C. F. Martinez de la Pena, L. De Masi, S. Nisa, G. Mulvey, J. Tong, M. S. Donnenberg, and G. D. Armstrong. Bfpi, bfpj, and bfpk minor pilins are important for the function and biogenesis of bundle-forming pili expressed by enteropathogenic escherichia coli. J Bacteriol, 198(5) :846–56, 2015.

[144] R. M. Hyland, J. Sun, T. P. Griener, G. L. Mulvey, J. S. Klassen, M. S. Donnenberg, and G. D. Armstrong. The bundlin pilin protein of enteropathogenic escherichia coli is an n-acetyllactosamine-specific lectin. Cell Microbiol, 10(1) :177–87, 2008.

[145] P. A. Manning. The tcp gene cluster of vibrio cholerae. Gene, 192(1) :63–70, 1997.

[146] S. L. Chiang, R. K. Taylor, M. Koomey, and J. J. Mekalanos. Single amino acid substitutions in the n-terminus of vibrio cholerae tcpa affect colonization, autoagglutination, and serum resistance. Mol Microbiol, 17(6) :1133–42, 1995.

[147] T. J. Kirn, N. Bose, and R. K. Taylor. Secretion of a soluble colonization factor by the tcp type 4 pilus biogenesis pathway in vibrio cholerae. Mol Microbiol, 49(1) :81–92, 2003.

[148] A. Gyohda, N. Furuya, A. Ishiwa, S. Zhu, and T. Komano. Structure and function of the shufflon in plasmid r64. Adv Biophys, 38 :183–213, 2004.

[149] H. Oki, K. Kawahara, T. Maruno, T. Imai, Y. Muroga, S. Fukakusa, T. Iwashita, Y. Kobayashi, S. Matsuda, T. Kodama, T. Iida, T. Yoshida, T. Ohkubo, and S. Nakamura. Interplay of a secreted protein with type ivb pilus for efficient enterotoxigenic escherichia coli colonization. Proc Natl Acad Sci U S A, 115(28) :7422–7427, 2018.

[150] B. Rosan, J. Slots, R. J. Lamont, M. A. Listgarten, and G. M. Nelson. Actinobacillus actinomycetemcomitans fimbriae. Oral Microbiol Immunol, 3(2) :58–63, 1988.

[151] F. A. Scannapieco, K. S. Kornman, and A. L. Coykendall. Observation of fimbriae and flagella in dispersed subgingival dental plaque and fresh bacterial isolates from periodontal disease. J Periodontal Res, 18(6) :620–33, 1983.

[152] S. S. Abby, J. Cury, J. Guglielmini, B. Neron, M. Touchon, and E. P. Rocha. Identification of protein secretion systems in bacterial genomes. Sci Rep, 6 :23080, 2016.

[153] J. L. Telford, M. A. Barocchi, I. Margarit, R. Rappuoli, and G. Grandi. Pili in gram-positive pathogens. Nature Reviews. Microbiology, 4(7) :509–19, 2006.

[154] Mladen Tomich, Paul J Planet, and David H Figurski. The tad locus : postcards from the widespread colonization island. Nature Reviews. Microbiology, 5(5) :363–375, 2007.

[155] Y. Wang, C. H. Haitjema, and C. Fuqua. The ctp type ivb pilus locus of agrobacterium tumefaciens directs formation of the common pili and contributes to reversible surface attachment. Journal of Bacteriology, 196(16) :2979–88, 2014.

[156] P. Entcheva-Dimitrov and A. M. Spormann. Dynamics and control of biofilms of the oligotrophic bacterium caulobacter crescentus. J Bacteriol, 186(24) :8254–66, 2004.

[157] J. M. Skerker and L. Shapiro. Identification and cell cycle control of a novel pilus system in caulobacter crescentus. EMBO Journal, 19(13) :3223–34, 2000.

[158] B. A. Perez, P. J. Planet, S. C. Kachlany, M. Tomich, D. H. Fine, and D. H. Figurski. Genetic analysis of the requirement for flp-2, tadv, and rcpb in actinobacillus actinomycetemcomitans biofilm formation. J Bacteriol, 188(17) :6361–75, 2006.

[159] Y. Wang, A. Liu, and C. Chen. Genetic basis for conversion of rough-to-smooth colony morphology in actinobacillus actinomycetemcomitans. Infect Immun, 73(6) :3749–53, 2005.

[160] J. Mignolet, G. Panis, and P. H. Viollier. More than a tad : spatiotemporal control of caulobacter pili. Curr Opin Microbiol, 42 :79–86, 2018.

[161] Paul J Planet, Scott C Kachlany, Daniel H Fine, Rob DeSalle, and David H Figurski. The widespread colonization island of actinobacillus actinomycetemcomitans. Nat Genet, 34(2) :193–198, 2003.

[162] E. M. Haase, J. L. Zmuda, and F. A. Scannapieco. Identification and molecular analysis of rough-colony-specific outer membrane proteins of actinobacillus actinomycetemcomitans. Infection and Immunity, 67(6) :2901–8, 1999.

[163] Tiago R. D. Costa, Catarina Felisberto-Rodrigues, Amit Meir, Marie S. Prevost, Adam Redzej, Martina Trokter, and Gabriel Waksman. Secretion systems in gram-negative bacteria : structural and mechanistic insights. Nat Rev Micro, 13(6) :343–359, 2015.

[164] A. P. Pugsley. The complete general secretory pathway in gram-negative bacteria. Microbiology and Molecular Biology Reviews, 57 :50–108, 1993.

[165] Y. Ferrandez and G. Condemine. Novel mechanism of outer membrane targeting of proteins in gram-negative bacteria. Molecular Microbiology, 69(6) :1349–57, 2008.

[166] R. Voulhoux, G. Ball, B. Ize, M. L. Vasil, A. Lazdunski, L. F. Wu, and A. Filloux. Involvement of the twin-arginine translocation system in protein secretion via the type ii pathway. The EMBO journal, 20(23) :6735–6741, 2001.

[167] Jenny-Lee Thomassin, Javier Santos Moreno, Ingrid Guilvout, Guy Tran Van Nhieu, and Olivera Francetic. The trans-envelope architecture and function of the type 2 secretion system : new insights raising new questions. Molecular Microbiology, 105(2) :211–226, 2017.

[168] Béatrice Py, Laurent Loiseau, and Frédéric Barras. An inner membrane platform in the type ii secretion machinery of gram-negative bacteria. EMBO reports, 2(3) :244–248, 2001.

[169] Jan Abendroth, Daniel D. Mitchell, Konstantin V. Korotkov, Tanya L. Johnson, Allison Kreger, Maria Sandkvist, and Wim G. J. Hol. The three-dimensional structure of the cytoplasmic domains of epsf from the type 2 secretion system of vibrio cholerae. Journal of structural biology, 166(3) :303–315, 2009.

[170] O. M. Possot, M. Gérard-Vincent, and A. P. Pugsley. Membrane association and multimerization of secreton component pulc. Journal of bacteriology, 181(13) :4004–4011, 1999.

[171] O. M. Possot, G. Vignon, N. Bomchil, F. Ebel, and A. P. Pugsley. Multiple interactions between pullulanase secreton components involved in stabilization and cytoplasmic membrane association of pule. Journal of bacteriology, 182(8) :2142–2152, 2000.

[172] Jan Abendroth, Allison C. Kreger, and Wim G. J. Hol. The dimer formed by the periplasmic domain of epsl from the type 2 secretion system of vibrio parahaemolyticus. Journal of structural biology, 168(2) :313–322, 2009.

[173] Jan Abendroth, Adrian E. Rice, Karen McLuskey, Michael Bagdasarian, and Wim G. J. Hol. The crystal structure of the periplasmic domain of the type ii secretion system protein epsm from vibrio cholerae : The simplest version of the ferredoxin fold. Journal of Molecular Biology, 338(3) :585–596, 2004.

[174] Jorik Arts, Arjan de Groot, Geneviève Ball, Eric Durand, Mohammed El Khattabi, Alain Filloux, Jan Tommassen, and Margot Koster. Interaction domains in the pseudomonas aeruginosa type ii secretory apparatus component xcps (gspf). Microbiology, 153(5) :1582–1592, 2007.

[175] Eric Durand, Gerard Michel, Romé Voulhoux, Julia Kurner, Alain Bernadac, and Alain Filloux. Xcpx controls biogenesis of the pseudomonas aeruginosa xcpt-containing pseudopilus. Journal of Biological Chemistry, 280(36) :31378–31389, 2005.

[176] Nathalie Sauvonnet, Guillaume Vignon, Anthony P. Pugsley, and Pierre Gounon. Pilus formation and protein secretion by the same machinery in escherichia coli. The EMBO Journal, 19(10) :2221, 2000.

[177] M. Campos, M. Nilges, D. A. Cisneros, and O. Francetic. Detailed structural and assembly model of the type ii secretion pilus from sparse data. Proc Natl Acad Sci U S A, 107(29) :13081–6, 2010.

[178] Manuel Campos, Olivera Francetic, and Michael Nilges. Modeling pilus structures from sparse data. Journal of Structural Biology, 173(3) :436–444, 2011.

[179] D. A. Cisneros, G. Pehau-Arnaudet, and O. Francetic. Heterologous assembly of type iv pili by a type ii secretion system reveals the role of minor pilins in assembly initiation. Mol Microbiol, 86(4) :805–18, 2012.

[180] David A. Cisneros, Peter J. Bond, Anthony P. Pugsley, Manuel Campos, and Olivera Francetic. Minor pseudopilin self-assembly primes type ii secretion pseudopilus elongation. The EMBO Journal, 31(4) :1041–1053, 2012.

[181] Rolf Kohler, Karsten Schäfer, Shirley Muller, Guillaume Vignon, Kay Diederichs, Ansgar Philippsen, Philippe Ringler, Anthony P. Pugsley, Andreas Engel, and Wolfram Welte. Structure and assembly of the pseudopilin pulg. Molecular Microbiology, 54(3) :647–664, 2004.

[182] Hans E. Parge, Katrina T. Forest, Michael J. Hickey, Deborah A. Christensen, Elizabeth D. Getzoff, and John A. Tainer. Structure of the fibre-forming protein pilin at 2.6 Å resolution. Nature, 378(6552) :32–38, 1995.

[183] Lucienne Letellier, S. Peter Howard, and J. Thomas Buckley. Studies on the energetics of proaerolysin secretion across the outer membrane of aeromonas species : Evidence for a requirement for both the protonmotive force and atp. Journal of Biological Chemistry, 272(17) :11109–11113, 1997.

[184] Marcella Patrick, Konstantin V. Korotkov, Wim G. J. Hol, and Maria Sandkvist. Oligomerization of epse coordinates residues from multiple subunits to facilitate atpase activity. The Journal of biological chemistry, 286(12) :10378–10386, 2011.

[185] Odile M. Possot, Lucienne Letellier, and Anthony P. Pugsley. Energy requirement for pullulanase secretion by the main terminal branch of the general secretory pathway. Molecular Microbiology, 24(3) :457–464, 1997.

[186] Jan Abendroth, Paul Murphy, Maria Sandkvist, Michael Bagdasarian, and Wim G. J. Hol. The x-ray structure of the type ii secretion system complex formed by the n-terminal domain of epse and the cytoplasmic domain of epsl of vibrio cholerae. Journal of Molecular Biology, 348(4) :845–855, 2005.

[187] M. Sandkvist, M. Bagdasarian, S. P. Howard, and V. J. DiRita. Interaction between the autokinase epse and epsl in the cytoplasmic membrane is required for extracellular secretion in vibrio cholerae. The EMBO journal, 14(8) :1664–1673, 1995.

[188] Mohamed Chami, Ingrid Guilvout, Marco Gregorini, Hervé W. Rémigy, Shirley A. Muller, Marielle Valerio, Andreas Engel, Anthony P. Pugsley, and Nicolas Bayan. Structural insights into the secretin puld and its trypsin-resistant core. Journal of Biological Chemistry, 280(45) :37732–37741, 2005.

[189] Steve L. Reichow, Konstantin V. Korotkov, Wim G. J. Hol, and Tamir Gonen. Structure of the cholera toxin secretion channel in its closed state. Nature structural & molecular biology, 17(10) :1226–1232, 2010.

[190] Ingrid Guilvout, Mohamed Chami, Catherine Berrier, Alexandre Ghazi, Andreas Engel, Anthony P. Pugsley, and Nicolas Bayan. In vitro multimerization and membrane insertion of bacterial outer membrane secretin puld. Journal of Molecular Biology, 382(1) :13–23, 2008.

[191] K. R. Hardie, S. Lory, and A. P. Pugsley. Insertion of an outer membrane protein in escherichia coli requires a chaperone-like protein. EMBO J, 15(5) :978–88, 1996.

[192] Justin A. Lemkul and David R. Bevan. Characterization of interactions between pila from pseudomonas aeruginosa strain k and a model membrane. The Journal of Physical Chemistry B, 115(24) :8004–8008, 2011.

[193] Elena Disconzi, Ingrid Guilvout, Mohamed Chami, Muriel Masi, Gerard H. M. Huysmans, Anthony P. Pugsley, and Nicolas Bayan. Bacterial secretins form constitutively open pores akin to general porins. Journal of bacteriology, 196(1) :121–128, 2014.

[194] R. A. Adegbola and D. C. Old. New fimbrial hemagglutinin in serratia species. Infect Immun, 38(1) :306–15, 1982.

[195] G. Jonson, J. Holmgren, and A. M. Svennerholm. Identification of a mannose-binding pilus on vibrio cholerae el tor. Microb Pathog, 11(6) :433–41, 1991.

[196] P. I. Watnick, K. J. Fullner, and R. Kolter. A role for the mannose-sensitive hemagglutinin in biofilm formation by vibrio cholerae el tor. J Bacteriol, 181(11) :3606–9, 1999.

[197] J. P. Claverys and B. Martin. Bacterial "competence" genes : signatures of active transformation, or only remnants ? Trends Microbiol, 11(4) :161–5, 2003.

[198] L. L. Burrows. Prime time for minor subunits of the type ii secretion and type iv pilus systems. Mol Microbiol, 86(4) :765–9, 2012.

[199] R. J. Redfield, A. D. Cameron, Q. Qian, J. Hinds, T. R. Ali, J. S. Kroll, and P. R. Langford. A novel crp-dependent regulon controls expression of competence genes in haemophilus influenzae. J Mol Biol, 347(4) :735–47, 2005.

[200] J. P. Claverys, B. Martin, and P. Polard. The genetic transformation machinery : composition, localization, and mechanism. FEMS Microbiol Rev, 33(3) :643–56, 2009.

[201] I. Chen, R. Provvedi, and D. Dubnau. A macromolecular complex formed by a pilin-like protein in competent bacillus subtilis. J Biol Chem, 281(31) :21720–7, 2006.

[202] R. Laurenceau, G. Pehau-Arnaudet, S. Baconnais, J. Gault, C. Malosse, A. Dujeancourt, N. Campo, J. Chamot-Rooke, E. Le Cam, J. P. Claverys, and R. Fronzes. A type iv pilus mediates dna binding during natural transformation in streptococcus pneumoniae. PLoS Pathog, 9(6) :e1003473, 2013.

[203] K. H. Piepenbrink. Dna uptake by type iv filaments. Front Mol Biosci, 6 :1, 2019.

[204] E. S. Antonova and B. K. Hammer. Genetics of natural competence in vibrio cholerae and other vibrios. Microbiology Spectrum, 3(3), 2015.

[205] G. S. Inamine and D. Dubnau. Comea, a bacillus subtilis integral membrane protein required for genetic transformation, is needed for both dna binding and transport. J Bacteriol, 177(11) :3045–51, 1995.

[206] D. Dubnau and R. Provvedi. Internalizing dna. Res Microbiol, 151(6) :475–80, 2000.

[207] J. L. Berry, Y. Xu, P. N. Ward, S. M. Lea, S. J. Matthews, and V. Pelicic. A comparative structure/function analysis of two type iv pilin dna receptors defines a novel mode of dna binding. Structure, 24(6) :926–34, 2016.

[208] S. L. Drake and M. Koomey. The product of the pilq gene is essential for the biogenesis of type iv pili in neisseria gonorrhoeae. Mol Microbiol, 18(5) :975–86, 1995.

[209] I. Chen and E. C. Gotschlich. Come, a competence protein from neisseria gonorrhoeae with dna-binding activity. J Bacteriol, 183(10) :3160–8, 2001.

[210] D. Facius and T. F. Meyer. A novel determinant (coma) essential for natural transformation competence in neisseria gonorrhoeae and the effect of a coma defect on pilin variation. Mol Microbiol, 10(4) :699–712, 1993.

[211] N. Matthey and M. Blokesch. The dna-uptake process of naturally competent vibrio cholerae. Trends Microbiol, 24(2) :98–110, 2016.

[212] K. L. Meibom, M. Blokesch, N. A. Dolganov, C. Y. Wu, and G. K. Schoolnik. Chitin induces natural competence in vibrio cholerae. Science, 310(5755) :1824–7, 2005.

[213] S. Yamamoto, M. Morita, H. Izumiya, and H. Watanabe. Chitin disaccharide (glcnac)2 induces natural competence in vibrio cholerae through transcriptional and translational activation of a positive regulatory gene tfoxvc. Gene, 457(1-2) :42–9, 2010.

[214] P. Seitz and M. Blokesch. Dna-uptake machinery of naturally competent vibrio cholerae. Proceedings of the National Academy of Sciences of the United States of America, 110(44) :17987–92, 2013.

[215] Paushali Chaudhury, Tessa E. F. Quax, and Sonja-Verena Albers. Versatile cell surface structures of archaea. Molecular Microbiology, 107(3) :298–311, 2018.

[216] D. M. Faguy, K. F. Jarrell, J. Kuzio, and M. L. Kalmokoff. Molecular analysis of archael flagellins : similarity to the type iv pilin-transport superfamily widespread in bacteria. Can J Microbiol, 40(1) :67–71, 1994.

[217] S. Reindl, A. Ghosh, G. J. Williams, K. Lassak, T. Neiner, A. L. Henche, S. V. Albers, and J. A. Tainer. Insights into flai functions in archaeal motor assembly and motility from structures, conformations, and genetics. Mol Cell, 49(6) :1069–82, 2013.

[218] Paushali Chaudhury, Tomasz Neiner, Edoardo D'Imprima, Ankan Banerjee, Sophia Reindl, Abhrajyoti Ghosh, Andrew S. Arvai, Deryck J. Mills, Chris van der Does, John A. Tainer, Janet Vonck, and Sonja-Verena Albers. The nucleotide-dependent interaction of flah and flai is essential for assembly and function of the archaellum motor. Molecular Microbiology, 99(4) :674–685, 2016.

[219] Ankan Banerjee, Chi-Lin Tsai, Paushali Chaudhury, Patrick Tripp, Andrew S Arvai, Justin P Ishida, John A Tainer, and Sonja-Verena Albers. Flaf is a beta-sandwich protein that anchors the archaellum in the archaeal cell envelope by binding the s-layer protein. Structure(London, England :1993), 23(5) :863–872, 2015.

[220] Reinhard Wirth. Colonization of black smokers by hyperthermophilic microorganisms. Trends in Microbiology, 25(2) :92 – 99, 2017.

[221] Rajesh Shahapure, Rosalie P.C. Driessen, M. Florencia Haurat, Sonja-Verena Albers, and Remus Th. Dame. The archaellum : a rotating type iv pilus. Molecular Microbiology, 91(4) :716–723, 2014.

[222] K. Lassak, A. Ghosh, and S. V. Albers. Diversity, assembly and regulation of archaeal type iv pili-like and non-type-iv pili-like surface structures. Res Microbiol, 163(9-10) :630–44, 2012.

[223] Elie Desmond, Celine Brochier-Armanet, and Simonetta Gribaldo. Phylogenomics of the archaeal flagellum : rare horizontal gene transfer in a unique motility structure. BMC Evol Biol, 7 :106, 2007.

[224] Lydia Gerl, Rainer Deutzmann, and Manfred Sumper. Halobacterial flagellins are encoded by a multigene family identification of all five gene products. FEBS Letters, 244(1) :137–140, 1989.

[225] Sonja-Verena Albers and Ken F. Jarrell. The archaellum : how archaea swim. Frontiers in Microbiology, 6 :23, 2015.

[226] Abhrajyoti Ghosh and Sonja-Verena Albers. Assembly and function of the archaeal flagellum. Biochem Soc Trans, 39(1) :64–69, 2011.

[227] Stefan Streif, Wilfried Franz Staudinger, Wolfgang Marwan, and Dieter Oesterhelt. Flagellar rotation in the archaeon halobacterium salinarum depends on atp. Journal of Molecular Biology, 384(1) :1 – 8, 2008.

[228] Ankan Banerjee, Tomasz Neiner, Patrick Tripp, and Sonja-Verena Albers. Insights into subunit interactions in the sulfolobus acidocaldarius archaellum cytoplasmic complex. The FEBS Journal, 280(23) :6141–6149, 2013.

[229] Bonnie Chaban, Sandy Y. M. Ng, Masaomi Kanbe, Ilana Saltzman, Graeme Nimmo, Shin-Ichi Aizawa, and Ken F. Jarrell. Systematic deletion analyses of the fla genes in the flagella operon identify several genes essential for proper assembly and function of flagella in the archaeon, methanococcus maripaludis. Molecular Microbiology, 66(3) :596–609, 2007.

[230] Anna-Lena Henche, Andrea Koerdt, Abhrajyoti Ghosh, and Sonja-Verena Albers. Influence of cell surface structures on crenarchaeal biofilm formation using a thermostable green fluorescent protein. Environmental Microbiology, 14(3) :779–793, 2012.

[231] Anna-Lena Henche, Abhrajyoti Ghosh, Xiong Yu, Torsten Jeske, Edward Egelman, and Sonja-Verena Albers. Structure and function of the adhesive type iv pilus of sulfolobus acidocaldarius. Environmental Microbiology, 14(12) :3188–3202, 2012.

[232] Benham Zolghadr, Andreas Klingl, Reinard Rachel, Arnold J. M. Driessen, and Sonja-Verena Albers. The bindosome is a structural component of the sulfolobus solfataricus cell envelope. Extremophiles, 15(2) :235–244, Mar 2011.

[233] Sabrina Fröls, Malgorzata Ajon, Michaela Wagner, Daniela Teichmann, Behnam Zolghadr, Mihaela Folea, Egbert J Boekema, Arnold J M Driessen, Christa Schleper, and Sonja-Verena Albers. Uv-inducible cellular aggregation of the hyperthermophilic archaeon sulfolobus solfataricus is mediated by pili formation. Mol Microbiol, 70(4) :938–952, 2008.

[234] Zalan Szabo, Adriana Stahl, Sonja-V. Albers, Jessica Kissinger, Arnold Driessen, and Mechthild Pohlschroder. Identification of diverse archaeal proteins with class III signal peptides cleaved by distinct archaeal prepilin peptidases. J Bacteriol, 189(3) :772–778, 2007.

[235] Divya Nair, Kaoru Uchida, Shin-Ichi Aizawa, and Ken Jarrell. Genetic analysis of a type IV pili-like locus in the archaeon methanococcus maripaludis. Arch Microbiol, 196(3) :179–191, 2014.

[236] Rémi Denise, Sophie S Abby, and Eduardo PC Rocha. Diversification of the type iv filament super-family into machines for adhesion, secretion, dna transformation and motility. bioRxiv, 2019.

[237] M. Basler, M. Pilhofer, G. P. Henderson, G. J. Jensen, and J. J. Mekalanos. Type vi secretion requires a dynamic contractile phage tail-like structure. Nature, 483(7388) :182–6, 2012.

[238] V. S. Nguyen, L. Logger, S. Spinelli, P. Legrand, T. T. Huyen Pham, T. T. Nhung Trinh, Y. Cherrak, A. Zoued, A. Desmyter, E. Durand, A. Roussel, C. Kellenberger, E. Cascales, and C. Cambillau. Type vi secretion tssk baseplate protein exhibits structural similarity with phage receptor-binding proteins and evolved to bind the membrane complex. Nat Microbiol, 2 :17103, 2017.

[239] E. C. Garcia, A. I. Perault, S. A. Marlatt, and P. A. Cotter. Interbacterial signaling via burkholderia contact-dependent growth inhibition system proteins. Proc Natl Acad Sci U S A, 113(29) :8296–301, 2016.

[240] W. Zhao, F. Caro, W. Robins, and J. J. Mekalanos. Antagonism toward the intestinal microbiota and its effect on vibrio cholerae virulence. Science, 359(6372) :210–213, 2018.

[241] C. E. Alvarez-Martinez and P. J. Christie. Biological diversity of prokaryotic type iv secretion systems. Microbiology and Molecular Biology Reviews, 73 :775–808, 2009.

[242] J. Guglielmini, F. de la Cruz, and E. P. Rocha. Evolution of conjugation and type iv secretion systems. Molecular Biology and Evolution, 30(2) :315–31, 2013.

[243] L. G. Pell, V. Kanelis, L. W. Donaldson, P. L. Howell, and A. R. Davidson. The phage lambda major tail protein structure reveals a common evolution for long-tailed phages and the type vi bacterial secretion system. Proc Natl Acad Sci U S A, 106(11) :4160–5, 2009.

[244] P. G. Leiman, M. Basler, U. A. Ramagopal, J. B. Bonanno, J. M. Sauder, S. Pukatzki, S. K. Burley, S. C. Almo, and J. J. Mekalanos. Type vi secretion apparatus and phage tail-associated protein complexes share a common evolutionary origin. Proceedings of the National Academy of Sciences of the United States of America, 106(11) :4154–9, 2009.

[245] V. Gold and M. Kudryashev. Recent progress in structure and dynamics of dual-membrane-spanning bacterial nanomachines. Curr Opin Struct Biol, 39 :1–7, 2016.

[246] K. V. Korotkov, T. Gonen, and W. G. Hol. Secretins : dynamic channels for protein transport across membranes. Trends Biochem Sci, 36(8) :433–43, 2011.

[247] B. Szappanos, J. Fritzemeier, B. Csorgo, V. Lazar, X. Lu, G. Fekete, B. Balint, R. Herczeg, I. Nagy, R. A. Notebaart, M. J. Lercher, C. Pal, and B. Papp.

Adaptive evolution of complex innovations through stepwise metabolic niche expansion. Nat Commun, 7 :11607, 2016.

[248] P. J. Planet, S. C. Kachlany, R. DeSalle, and D. H. Figurski. Phylogeny of genes for secretion ntpases : identification of the widespread tada subfamily and development of a diagnostic key for gene classification. Proceedings of the National Academy of Sciences of the United States of America, 98(5) :2503–8, 2001.

[249] L. M. Iyer, K. S. Makarova, E. V. Koonin, and L. Aravind. Comparative genomics of the ftsk-hera superfamily of pumping atpases : implications for the origins of chromosome segregation, cell division and viral capsid packaging. Nucleic Acids Res, 32(17) :5260–79, 2004.

[250] S. S. Abby, B. Neron, H. Menager, M. Touchon, and E. P. Rocha. Macsyfinder : A program to mine genomes for molecular systems with an application to crispr-cas systems. PLoS ONE, 9(10) :e110726, 2014.

[251] Xia Wang, Qingqing Han, Guanjun Chen, Weixin Zhang, and Weifeng Liu. A putative type ii secretion system is involved in cellulose utilization in cytophaga hutchisonii. Frontiers in Microbiology, 8, 2017.

[252] O. X. Cordero and P. Hogeweg. The impact of long-distance horizontal gene transfer on prokaryotic genome size. Proc Natl Acad Sci U S A, 106(51) :21748–53, 2009.

[253] Howard Ochman, Jeffrey G. Lawrence, and Eduardo A. Groisman. Lateral gene transfer and the nature of bacterial innovation. Nature, 405 :299, 2000.

[254] J. P. McCutcheon and N. A. Moran. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol, 10(1) :13–26, 2011.

[255] B. D. Nguyen and R. H. Valdivia. Virulence determinants in the obligate intracellular pathogen chlamydia trachomatis revealed by forward genetic approaches. Proceedings of the National Academy of Sciences of the United States of America, 109(4) :1263–8, 2012.

[256] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey. Predicting the sequence specificities of dna- and rna-binding proteins by deep learning. Nat Biotechnol, 33(8) :831–8, 2015.

[257] C. Angermueller, T. Parnamaa, L. Parts, and O. Stegle. Deep learning for computational biology. Mol Syst Biol, 12(7) :878, 2016.

[258] S. Sheehan and Y. S. Song. Deep learning for population genetic inference. PLoS Comput Biol, 12(3) :e1004845, 2016.