Mécanismes moléculaires régissant la reconnaissance sélective et l'encapsidation du génome d'un virus bactérien

Mehdi El Sadek Fadel

université
PARIS-SACLAY

Mécanismes moléculaires régissant la reconnaissance sélective et l'encapsidation du génome d'un virus bactérien

Molecular mechanisms of selective viral DNA recognition and packaging by a bacterial virus

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 569, Innovation thérapeutique du fondamental à l'appliqué (IFTA)
Spécialité de doctorat: Biochimie et biologie structurale
Graduate school : Santé et médicaments. Référent : Faculté de pharmacie

Thèse préparée dans l'unité de recherche **I2BC (Université Paris-Saclay, CEA, CNRS)**, sous la direction de **Paulo TAVARES**, Directeur de recherche

Thèse soutenue à Paris-Saclay, le 20 Décembre 2021, par

Mehdi EL SADEK FADEL

Composition du Jury

Stéphanie BURY-MONÉ Professeure, CNRS, Université Paris-Saclay	Présidente
Claire LE HÉNAFF-LE MARREC Professeure, INRAE, Université de Bordeaux	Rapporteur & Examinatrice
Jean-Yves BOUET Directeur de recherche, CNRS, Université Paul Sabatier	Rapporteur & Examinateur
Mireille ANSALDI Directrice de recherche, CNRS, Aix Marseille Université	Examinatrice
Patrice POLARD Directeur de recherche, CNRS, Université Paul Sabatier	Examinateur
Paulo TAVARES Directeur de recherche, CNRS, Université Paris-Saclay	Directeur de thèse

Thèse de doctorat
NNT : 2021UPASQ054

Remerciements

Je tiens tout d'abord à remercier Paulo Tavares qui m'a encadré durant cette thèse de doctorat, il a toujours su se rendre disponible et répondre avec justesse et pédagogie à toutes mes interrogations. Il a brillamment su transmettre son savoir-faire. Je remercie également Sandrine Brasilès qui fut la première à m'initier aux différentes techniques de biochimie, de biologie moléculaire et de microbiologie qui allaient par la suite me servir tout au long de mon doctorat. J'exprime toute ma reconnaissance à Jack Dorling dont le soutien m'a été très précieux ces trois dernières années passées au laboratoire. Il a toujours fait preuve d'une grande curiosité et nos conversations ont toujours été très stimulantes.

Je remercie également Delphine Naquin, Yves d'Aubenton Carafa, Kévin Gorrichon et l'ensemble des autres membres de la plateforme de séquençage haut débit de l'i2bc sans lesquels certaines des expériences présentées dans ce manuscrit n'auraient pas été réalisables. De même, je remercie Éric Jacquet et Naima Nhiri pour leur aide avec les expériences qRT-PCR, et Malika Ould Ali pour les expériences de microscopie électronique.
Je tiens bien évidement à remercier l'ensemble des autres membres de mon équipe passés et présents avec lesquels j'ai quotidiennement travaillé : Karima Djacem, Andrés Corral-Lugo, Hounay'iraa Chadouli, Lia Marques Godinho, Leonor Oliveira, Stéphane Roche, Audrey Labarde et Isabelle Auzat.

Je remercie également les membres de mon jury de thèse, Claire Le Hénaff-Le Marrec, Jean-Yves Bouet, Mireille Ansaldi, Stéphanie Bury-Moné ainsi que Patrice Polard qui ont accepté d'évaluer mes travaux.

J'exprime mes plus sincères remerciements à Maxime Allioux pour son soutien précieux, sa grande lucidité et ses conseils avisés qui furent très profitables. Enfin, je voudrai exprimer toute ma gratitude à ma conjointe Marick Esberard qui a toujours été présente à mes côtés. Elle m'a constamment soutenu et encouragé.

Sommaire

Liste des figures et tableaux

Liste des abréviations

ADN	Acide DésoxyriboNucléique
ARN(m)	Acide RiboNucléique (messager)
ATP	Adénosine TriPhosphate
ATPase	Adénosine TriPhosphatase
cfu	colony forming unit
Ct	*Cycle Threshold* - cycle de seuil
DNase	Désoxyribonucléase
DO (600nm)	Densité optique à 600nm
EDTA	Acide éthylène-diamine-tétraacétique
GFP	*Green Fluorescent Protein* - protéine fluorescente verte
gp	*Gene Product* -Produit du gène
GTA	*Gene Tranfer Agent* – Agent de transfert de gènes
HSV1	Herpès Simplex Virus I
kDa	Kilo Dalton
LS	*Large Scale liquid lysate* – Lysat liquide à large échelle
NGS	*Next Generation Sequencing* - Séquençage de nouvelle génération
NAP	*Nucleoid associated proteins* – Protéines associées au nucléoïde
pb	paire de base
PCR	*Polymerase Chain Reaction* - Réaction en chaîne par polymérase
pfu	plage forming unit
PL	*Plate Lysate* – Lysat sur boîte
RNase	Ribonucléase
rpm	rotations par minute
SP	*Single Plaque* – Plage isolée
SSLL	*Small Scale Liquid Lysate* – Lysat liquide à petite échelle
SPP1	*Subtilis* Pavia Phage 1
sus	*supressor sensitive mutant* – mutant suppresseur
TerS	*TERminase Small subunit* - petite sous-unité de la terminase
TerL	*TERminase Large subunit* - grande sous-unité de la terminase
wt	*wild type* - sauvage

Introduction

I. Introduction

I.I. Généralités sur les virus

Un virus est un agent infectieux qui ne peut se répliquer qu'en parasitant une cellule vivante. Le virus détourne le métabolisme de la cellule qu'il infecte pour générer de nouveaux virus. Ces agents biologiques sont capables d'infecter tous les domaines du vivant ; des êtres pluricellulaires comme les animaux ou les végétaux, aux unicellulaires comme les bactéries ou les archées. Chez l'humain, ils sont responsables de nombreuses maladies souvent anodines mais parfois très graves comme le SIDA, la grippe ou encore la maladie à virus Ebola.

I.I.I. Rappels historiques

Le terme virus signifie poison en latin, car pendant l'antiquité déjà, ce mot était utilisé pour définir la cause de maladies graves. À cette époque, il était bien sûr difficile de déduire que ces maladies étaient produites par un agent microscopique. On utilise donc ce mot pour désigner des affections diverses et variées qui étaient souvent transmises par l'intermédiaire de liquides contaminés.

À la fin du XIXe siècle, Dimitri Ivanovski et Martinus Willem Beijerinck découvrent que l'agent responsable de la maladie de la mosaïque du tabac est capable de traverser un filtre-chandelle de Chamberland (Blos, 2014). Ce filtre est réputé suffisamment fin pour ne laisser passer aucune bactérie. On dispose dorénavant d'un critère de taille permettant de différencier ce que Beijerinck appelle virus en référence au terme utilisé depuis l'antiquité, d'autres agents pathogènes comme les bactéries dont on sait déjà, notamment grâce aux travaux de Robert Koch, qu'elles provoquent des maladies. Les virus étant trop petits pour être observés au microscope optique, on pense qu'il s'agit de très petites bactéries (Grmek, 1994). Cependant, pour Beijerinck, les virus sont fondamentalement différents des bactéries. Il démontre notamment que, contrairement aux bactéries, le virus de la mosaïque du tabac ne semble pas se multiplier en absence de tissus végétaux métaboliquement actifs (Beijerinck, 1898). Au début du XXe siècle Félix d'Hérelle et Frederick William Twort découvrent également que des virus, que d'Hérelle appelle bactériophages, peuvent infecter et détruire des bactéries (d'Hérelle, 1919 ; Dublanchet et al, 2008 ; Ansaldi et al, 2018), ce qui laisse planer le doute sur la véritable nature des virus. Il faut attendre l'invention du microscope électronique pour qu'il soit enfin possible en 1939 d'observer pour la première fois le virus de la mosaïque du tabac (Kausche et al,1939, figure 1). Depuis, de nombreux virus ont pu être observés et

caractérisés, grâce notamment aux évolutions de la biochimie, de la génétique et de la biologie moléculaire.

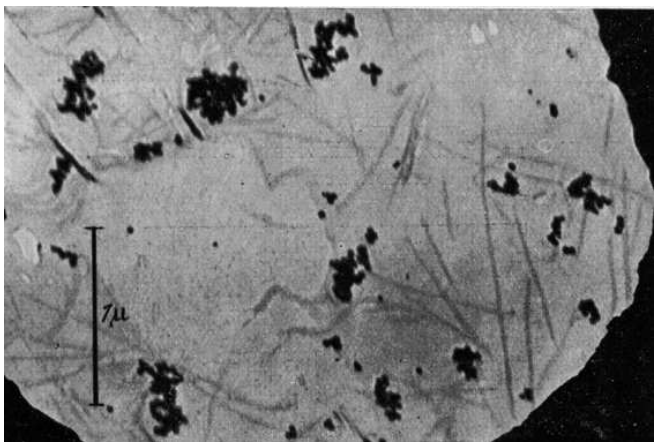


Figure 1. Virus de la mosaïque du tabac observé en microscopie électronique par Helmut Ruska (Ruska. 1939).

I.I.II. Le cycle viral

Les virus sont caractérisés par deux phases : une phase extracellulaire et une phase intracellulaire. La forme extracellulaire d'un virus est inerte. Il s'agit de la particule virale proprement dite. Elle est composée d'un génome à ARN ou/et ADN qui peut être simple ou double brin. Il est parfois divisé en plusieurs segments. Ce génome est protégé par une structure protéique appelée capside. Celle-ci est parfois enveloppée d'une bicouche lipidique provenant de la cellule hôte dans laquelle s'insèrent des protéines d'enveloppe. Lorsque le virus entre en contact avec sa cellule hôte, des protéines présentes à sa surface interagissent avec un récepteur à la membrane ou paroi de l'hôte et tout ou partie du virus pénètre dans la cellule. Chaque virus a un spectre d'hôte assez restreint, mais il n'est pas rare qu'un virus puisse infecter des hôtes d'espèces ou de types cellulaires différents (Coyette et Mergeay, 2013). Lorsque le virus infecte la cellule, il entre dans sa phase intracellulaire. Il s'agit de l'étape où le virus utilise le métabolisme cellulaire de l'hôte pour répliquer son génome et produire les éléments nécessaires à la génération de nouveaux agents infectieux. Cependant, le virus persiste parfois dans la cellule hôte où il reste quiescent. Certains virus s'intègrent même dans le génome de leur hôte comme le VIH dans les lymphocytes T4 ou certains virus de bactéries qui persistent sous forme de prophages (voir partie II).

I.I.III. La reconnaissance du génome viral

Certaines étapes clés nécessaires à la génération de particules infectieuses sont retrouvées chez l'ensemble des virus. L'une d'entre-elles est l'incorporation spécifique du génome viral dans une particule nucléoprotéique. Le génome viral, qu'il soit à ADN ou ARN, doit être reconnu puis protégé par une capside protéique au cours de l'assemblage de la particule virale. Or, ce génome viral se retrouve souvent dans le même compartiment cellulaire que le génome et/ou les ARNs de l'hôte. Il faut alors que le virus discrimine son génome du matériel génétique de l'hôte, sauf dans les cas où l'infection virale conduit à la dégradation de l'ADN/ARN cellulaire.

Chez les virus, plusieurs structures de nucléocapsides existent. En fonction de la nature de la nucléocapside le génome est reconnu et encapsidé différemment (Comas-Garcia, 2019).

Dans le cas de nombreux virus dont notamment ceux à capside hélicoïdale (Green et al, 2014), et de certains virus à capside icosaédrique (Schneemann, 2006), une nucléocapside s'assemble directement autour du génome (figure 2 A.B). Il faut donc qu'il y ait une interaction spécifique, directe ou avec l'aide d'un intermédiaire, entre le génome et la protéine de la nucléocapside. À titre d'exemple, chez les rétrovirus comme le VIH I, c'est la partie 5' non traduite de l'ARN génomique en amont du gène *gag* appelée site Ψ qui est reconnue spécifiquement par la polyprotéine Gag au niveau du domaine correspondant à la nucléoprotéine (D'Souza et al, 2005). Le clivage endoprotéolytique de Gag donne les protéines nécessaires à la formation d'une capside conique contenant 2 génomes protégés par la nucléoprotéine.

Chez les virus à ARN négatif segmenté, dont le génome est constitué de plusieurs molécules d'ARN distinctes, l'encapsidation est beaucoup plus complexe car chaque segment doit être reconnu séparément mais tous sont incorporés dans un seul et même virion. Dans le cas virus de la grippe A, dont le génome est composé de 8 segments, chaque ARN possède une séquence d'encapsidation spécifique qui est reconnue différemment par la nucléoprotéine dans un réseau d'interactions très complexe. Des erreurs dans la reconnaissance de ces segments conduisent à des réassortiments qui jouent un rôle important dans l'évolution de ces virus (Bolte et al, 2019 ; Giese et al, 2016).

Chez de nombreux virus à capside icosaédrique, le génome est reconnu par un complexe protéique spécifique qui encapside l'acide nucléique à l'intérieur d'une procapside préformée composée d'une ou plusieurs protéine(s) majeure(s) de capside soutenue(s) par une ou des

protéines(s) d'échafaudage (Figure 2 C). Chez les adénovirus, herpèsvirus et bactériophages caudés qui sont des virus à ADN double brin, la reconnaissance du génome est assurée par un complexe protéique qui interagit avec une séquence spécifique (Fujisawa et al, 2016 ; Abedon et al, 2009). Dans la suite du manuscrit, nous étudierons comment la reconnaissance du génome a lieu chez ces virus et plus précisément chez les bactériophages caudés.

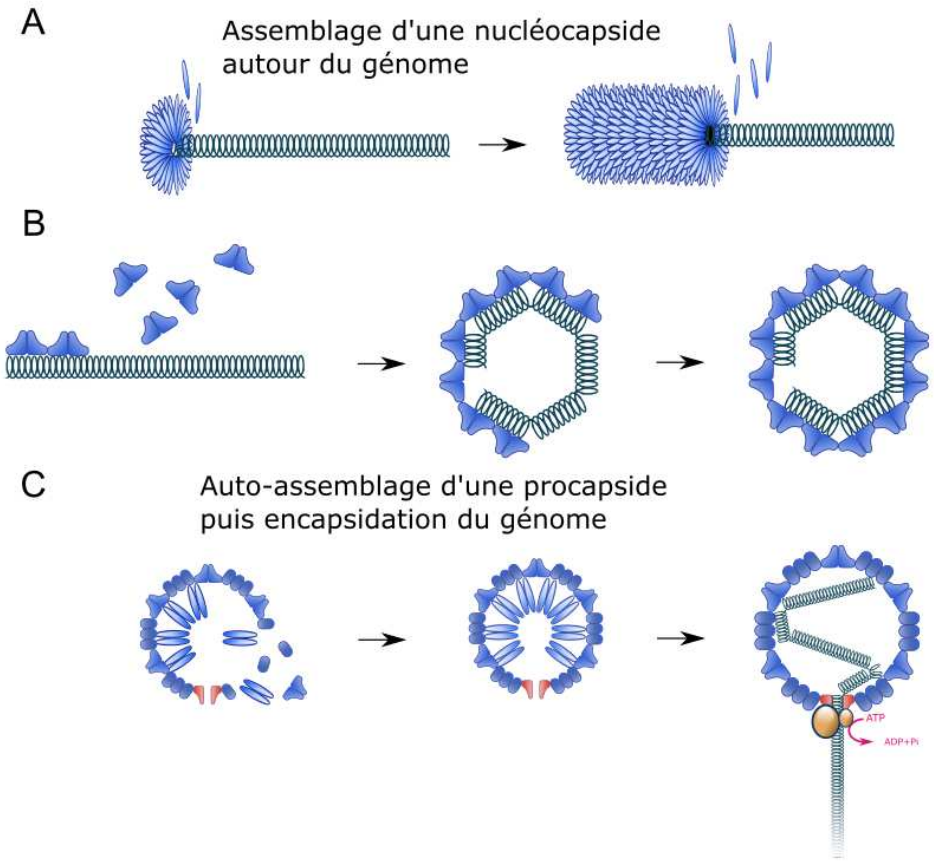


Figure 2. Différents modes d'encapsidation du génome viral. Chez la plupart des virus à capside hélicoïdale (A) et chez certains virus à capside icosaédrique (B), la nucléocapside s'assemble autour du génome. Chez d'autres virus à capside icosaédrique, une procapside s'auto-assemble à l'aide d'une protéine d'échafaudage et le génome est encapsidé à l'intérieur de cette structure par un complexe protéique spécialisé au niveau d'un pore dans la procapside (C).

I.II. Les bactériophages

Les bactériophages sont des virus qui infectent les bactéries. Tous les milieux où se développent les bactéries abritent des bactériophages. On les retrouve aussi bien dans le système digestif des animaux que dans l'océan ou le sol. On considère qu'il s'agit des entités biologiques les plus abondantes du globe (Coyette, 2013). Ils jouent ainsi un rôle important

dans la régulation des populations bactériennes au sein des écosystèmes et influent par conséquent sur les grands cycles biogéochimiques de la planète (Abedon et al, 2009).

Ces virus se répartissent en 14 ordres et 49 familles (ICTV, 2021). On y retrouve des virus à ADN simple et double brin, mais aussi, moins fréquemment, des virus à ARN simple brin. Ces virus arborent des formes très diversifiées comme l'illustre la figure 3. Notons que certaines familles regroupent à la fois des bactériophages et des virus d'archées. Parmi ces différentes familles de bactériophages, les 14 familles de l'ordre des *Caudovirales* sont les plus représentées, puisque 95% des bactériophages seraient des *Caudovirales* (voir partie I.II.II ; Pietilä et al, 2014).

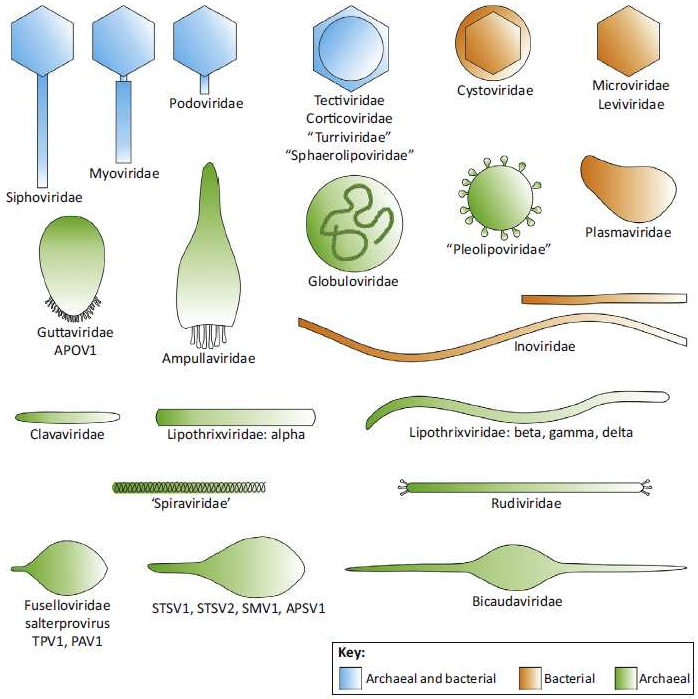


Figure 3. Morphologies des bactériophages. (Pietila et al, 2014). Aperçu non exhaustif des virus qui infectent les bactéries. Dans certaines familles, l'on retrouve des virus qui infectent aussi des archées. Les virus les plus communs dans la biosphère sont les *Caudovirales* (bactériophages qui possèdent une queue, ici *Siphoviridae*, *Myoviridae* et *Podoviridae*). Les virus d'archées ont des formes très diversifiées. Bleu : virus d'archées et de bactéries, Marron : uniquement virus de bactéries, Vert : uniquement virus d'archée.

I.II.I. Cycles lytique, lysogénique et chronique

Les bactériophages ont mis en place différentes stratégies d'infection et de dissémination. On les oppose souvent en deux grands types appelés cycle lytique et cycle lysogénique. Lors du cycle lytique, le bactériophage se fixe sur le récepteur de la cellule hôte et éjecte son génome dans le cytoplasme de la bactérie. Le métabolisme cellulaire de l'hôte est ensuite détourné

afin de générer de nouveaux virions, la lyse de la cellule infectée est ensuite induite par le bactériophage. Les virions sont libérés et peuvent infecter de nouvelles cellules.

Lors du cycle lysogénique, le phage éjecte son génome dans la cellule hôte et va soit s'intégrer au génome bactérien, on parle alors de prophage, soit se maintenir sous forme d'un épisome autonome dans la cellule hôte. Le virus se maintient ainsi sous cette forme et se perpétue dans la descendance de la cellule infectée au cours de divisions cellulaires successives. Lors d'un stress environnemental (activation de la réponse SOS, stress UV par exemple) le virus entre en cycle lytique. S'il s'est maintenu sous forme de prophage, ce-dernier est excisé du génome de l'hôte. Ensuite, le génome se réplique, est transcrit et traduit et de nouveaux virions sont générés (Wilson et Mann, 1997). L'entrée en lysogénie ou en cycle lytique est déterminée par de nombreux facteurs environnementaux, lors d'une infection un phage qui peut être lysogénique n'entre pas forcément en lysogénie si les conditions ne le permettent pas, il favorise alors un cycle lytique (Payet et al, 2013).

On distingue donc les phages strictement lytiques dits virulents des phages dits tempérés qui sont à la fois lysogéniques et lytiques. L'un des exemples le plus connu et le mieux caractérisé de phage tempéré est celui de lambda qui infecte la bactérie *Escherichia coli*. Nous verrons par la suite que le phage SPP1 est strictement lytique, donc virulent.

Ce modèle à deux cycles ne rend cependant pas compte de toutes les stratégies qui existent dans la nature (Hobbs et Abedon. 2016). Par exemple, certains phages infectent leur hôte, se multiplient et s'échappent de la cellule sans la lyser. Il s'agira alors d'une infection chronique. Le cycle aboutit à la production de virions mais pas à la lyse de la cellule, on ne peut alors plus raisonnablement parler de cycle lytique. Le cycle du phage M13 illustre bien ce type de stratégie. De même, il existe des phages qui se maintiennent sous forme de prophage mais qui conduisent à une infection chronique lorsqu'ils sont induits, c'est notamment le cas du phage CTXphi (voir figure 4 ; Howard-Varona et al, 2017).

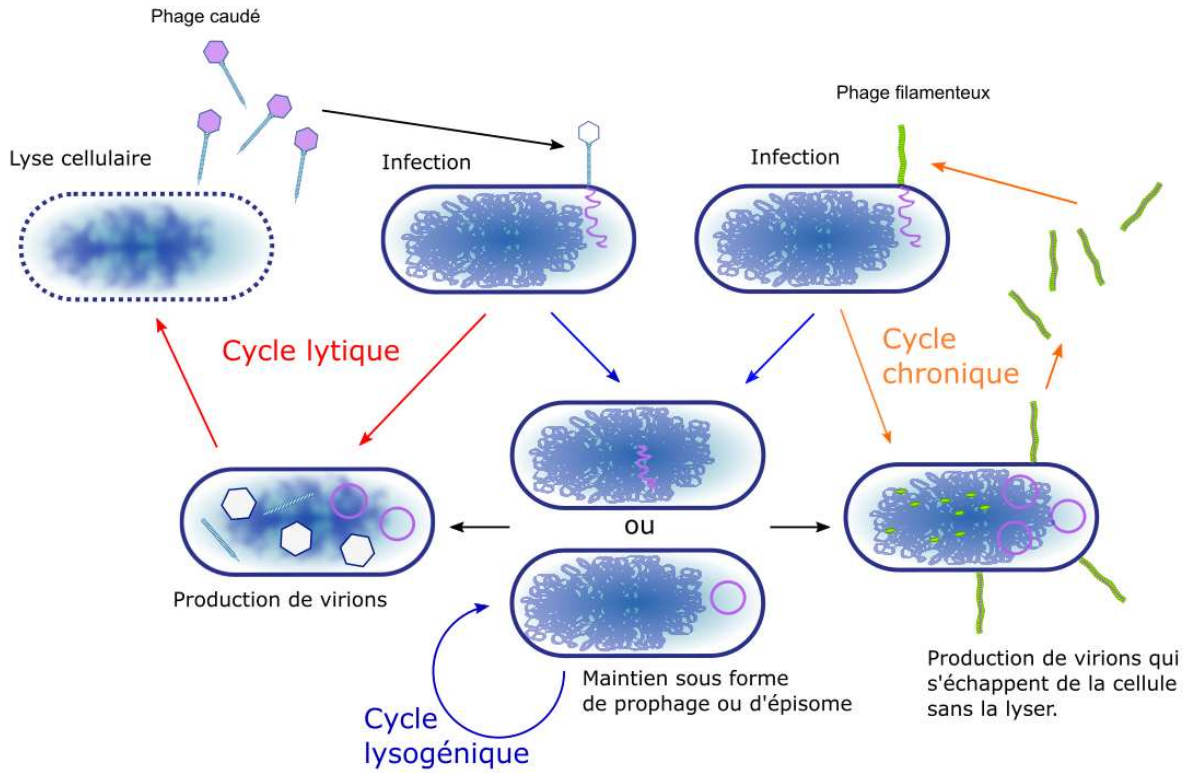


Figure 4. Cycles lysogénique, lytique et chronique. Les bactériophages ont des interactions complexes avec leur hôte. Ils peuvent se multiplier en lysant les cellules (cycle lytique ; rouge) ou s'échapper sans lyser les bactéries (cycle chronique ; orange) ou encore rester quiescent dans les bactéries (cycle lysogénique ; bleu). Certains virus combinent ces caractéristiques. Le cycle chronique est souvent possible chez les phages filamenteux. L'ADN bactérien est représenté en bleu et l'ADN viral en violet. Des hexagones blancs aux contours bleus représentent des capsides vides, et les petits ovales verts des protéines de nucléocapside.

La plupart des bactéries possèdent plusieurs prophages dans leur génome. Certains d'entre eux dégénèrent au fil de l'évolution et ne sont plus capables de donner des particules virales viables lorsqu'ils sont induits. De nombreux prophages offrent également certains avantages à leur hôte. Un des cas le plus connu est celui de la toxine cholérique (RTX) chez *Vibrio cholerae* qui est codée par un gène de prophage (Waldor et al, 1996). De même, certains systèmes de sécrétions utilisés par des bactéries pathogènes dériveraient de prophages (Tobe et al, 2006).

Certains prophages dégénérés deviennent des GTA pour *Gene Transfert Agent*. Lorsqu'ils sont induits, ils génèrent des particules proches de virions qui ne contiennent que de l'ADN bactérien. Ils permettent ainsi l'échange de matériel génétique au sein des populations de bactéries par transfert horizontal (Lang et al, 2012). Nous verrons plus en détails en quoi la

nature du cycle d'un phage considéré influe sur la manière dont celui-ci sert de vecteur au transfert horizontal de gènes.
Chez les bactériophages caudés, sont retrouvés à la fois des phages virulents et des phages tempérés.

I.II.II. Les bactériophages caudés

I.II.II.I. Taxonomie

De tous les bactériophages présents dans l'environnement, les *Caudovirales* sont les plus nombreux (Ackermann et al, 2007 ; Dion et al, 2020).
Les *Caudovirales*, également appelés phages caudés sont des bactériophages à ADN double brin. Leur génome est contenu dans une capside de forme icosaédrique à laquelle est reliée une queue de longueur variable. Ces virus possèdent souvent des fibres au bout de celle-ci. En fonction de leur morphologie et de leur séquence génomique, les *Caudovirales* sont répartis en 14 familles. Les bactériophages caudés possèdent une queue très courte et non contractile ou une queue longue et contractile ou encore une queue longue et non contractile.

I.II.II.II. La morphogénèse virale chez les bactériophages caudés

Les bactériophages caudés partagent de nombreuses similitudes quant aux voies d'assemblage qui conduisent à la formation de particules virales infectieuses. Les mécanismes moléculaires mis en jeu sont très semblables chez tous les bactériophages caudés. De même, une bonne partie des étapes clés de cette morphogénèse virale est partagée avec les virus eucaryotes de type Herpèsvirus, ce qui suggère qu'ils ont une origine évolutive commune (Abrescia et al, 2012).
Dans la cellule hôte, le métabolisme cellulaire est détourné afin de répliquer le génome et synthétiser les protéines virales. Parmi elles, il y a celles impliquées dans la réplication du génome viral, des protéines structurales, d'autres nécessaires à l'encapsidation du génome ou encore des protéines qui assurent la lyse cellulaire. Au sein de ces protéines de structure sont retrouvées celles qui forment la capside. Elles s'auto-assemblent en une structure appelée procapside (figure 5). Elles forment un pseudo-icosaèdre de morphologie et de taille variable en fonction des phages. Celui-ci est constitué de plusieurs centaines de sous-unités d'une protéine majeure de capside qui sont maintenues par une protéine dite d'échafaudage située à

l'intérieur de la structure (Ignatiou et al, 2019). À l'un des sommets de la procapside, on retrouve une protéine appelée portal qui forme un pore dans la procapside au travers duquel l'ADN viral va être encapsidé (Oliveira et al, 2013).

Les sous-unités de la protéine majeure de capside et de la protéine d'échafaudage s'agencent très précisément. Dans le cas le plus simple, la procapside est composée de 12 pentamères, soit 60 sous-unités de la protéine majoritaire de capside dont les interactions sont identiques. Mais il existe des capsides plus complexes dont les sous-unités établissent des contacts quasi-équivalents conduisant à la formation d'hexamères au niveau des surfaces plates de l'icosaèdre et de pentamères aux sommets de la structure, exception faite du sommet qui accueille la protéine portal. Les hexamères et pentamères sont appelés capsomères. À titre d'exemple, capside du phage SPP1 (voir partie I.V.IV) compte 415 sous-unités au total qui forment 60 hexamères et 11 pentamères retrouvés à chaque sommet de l'icosaèdre. L'assemblage de la capside est un processus complexe qui requière la présence d'une protéine échafaudage dont le rôle est de faciliter l'agencement correct des sous-unités de la protéine majeure de capside entre elles.

Bien que dans la majorité des cas une procapside soit composée d'un seul type de protéine majeure de capside et de protéine d'échafaudage, certains bactériophages caudés dérogent à cette règle. Le phage T4 par exemple, possède 2 protéines majeures de capside et plusieurs protéines d'échafaudage (Black. 2015). Certains phages comme T5, codent une protéine majeure de capside qui possède son propre domaine d'échafaudage. Dans ce cas précis, une seule protéine assure les deux fonctions (Huet et al, 2016).

Au cours de l'infection, le génome viral est synthétisé en de nombreuses copies, souvent sous forme de concatémères. Cela signifie que plusieurs copies du génome se retrouvent au sein de la même molécule les unes à la suite des autres. Ce type de molécule est généré lorsque le phage se réplique par réplication sigma qui est un mode de synthèse continu de l'ADN ou lors d'événements de recombinaison entre plusieurs copies du génome.

Ce concatémère d'ADN double brin linéaire est reconnu par un complexe protéique appelé terminase qui se charge de l'encapsidation. Les bactériophages caudés ont développé 2 grandes stratégies de reconnaissance. Soit cette reconnaissance est aspécifique, le génome de l'hôte est dégradé au cours de l'infection, ainsi seul le génome viral est encapsidé car il se retrouve seul dans la cellule. Plusieurs bactériophages, dont notamment T5, utilisent cette

stratégie (Gao et al, 2016). Soit le bactériophage reconnaît une séquence spécifique sur son propre génome pour initier le processus d'encapsidation (figure 5 ; Oliveira et al, 2013).

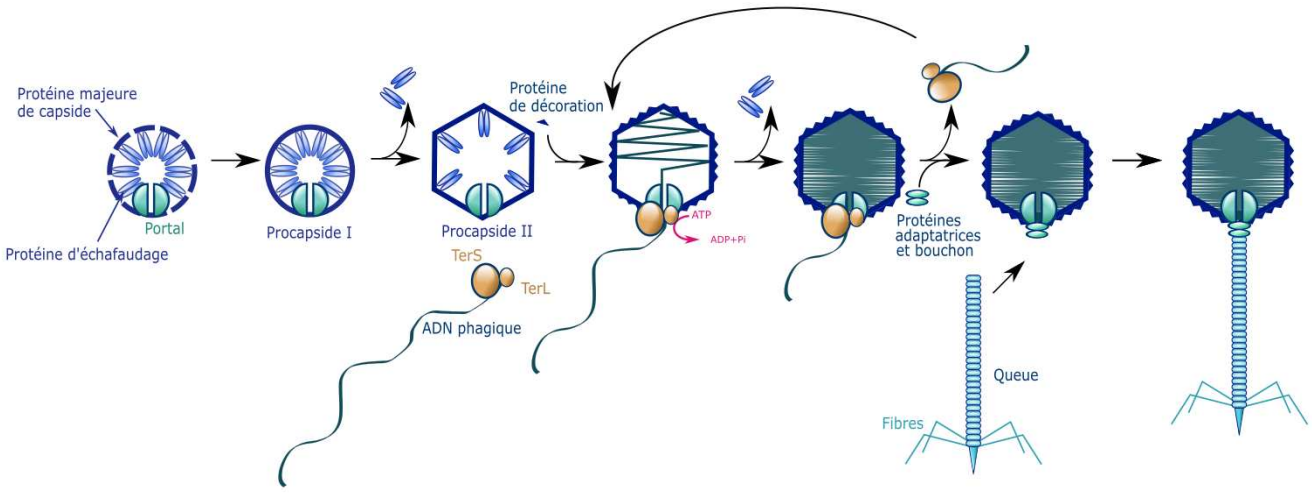


Figure 5. Morphogénèse virale chez les bactériophages caudés. Les protéines majeures de capside, d'échafaudage et la protéine portal s'assemblent pour former une procapside. L'ADN phagique, sous forme de concatémère, est reconnu et clivé par la terminase composée de TerS et TerL. La sortie de la protéine échafaudage de la procapside I permet sa maturation en procapside II, qui acquière un volume plus important et une forme plus angulaire, ainsi que l'encapsidation de l'ADN au cours de laquelle la procapside II va continuer sa maturation en capside. Une protéine de décoration se fixe parfois sur la capside mature. TerL et TerS se dissocient de la capside pour encapsider le reste du concatémère dans de nouvelles procapsides. Les proteines adaptatrices interagissent séquentiellement avec le sommet portal créant l'interface pour la fixation de la queue qui possède parfois des fibres.

Le complexe terminase est, chez la grande majorité des phages, constitué de deux protéines, une petite sous-unité TerS et une grande sous-unité appelée TerL. La protéine TerS est multimérique (souvent un nonamère), lorsqu'un signal d'encapsidation est présent sur le génome du phage, elle le reconnaît spécifiquement. Le complexe formé par TerS avec l'ADN est reconnu par TerL qui est une protéine à deux activités. En effet, elle va d'abord cliver l'ADN grâce à une activité endonucléase, puis l'ensemble du complexe va interagir avec une procapside vide au niveau du pore formé par la protéine portal, enfin TerL assure l'encapsidation de l'ADN grâce à son activité ATPase. Elle va agir comme un moteur moléculaire qui encapside l'ADN à l'intérieur de la procapside en utilisant l'ATP comme source d'énergie (Catalano. 2005). La sortie de la protéine d'échafaudage à l'extérieur de la procapside par des trous retrouvés au centre des hexamères créée l'espace nécessaire à l'entrée de l'ADN viral et la maturation de la procapside en un intermédiaire appelé procapside II puis en capside. Celle-ci devient plus volumineuse et plus anguleuse. A la fin de

l'encapsidation TerL clive de nouveau l'ADN. Ce clivage est induit, soit par une séquence spécifique, soit lorsque la capside est à son encombrement maximal (encapsidation par tête pleine). Le complexe terminase, toujours lié au concatémère de génomes se détache de la capside et rejoint une nouvelle procapside vide pour procéder à une nouvelle encapsidation (Fujisawa et Morita. 2003). Lors du départ du complexe terminase lié à l'ADN, des protéines se fixent sur la capside mature. Elles préviennent la sortie du génome de la capside et permettent la ligation de la queue.

La queue des bactériophages caudés suit une voie d'assemblage indépendante de celle de la capside à l'exception des *Podoviridae*, où la queue est assemblée par addition séquentielle de protéines au sommet de la portal fermée (Cuervo et al, 2013) Les longues queues des *Siphoviridae* et des *Myoviridae* sont généralement composées de plusieurs sous-unités en nombre variable qui forment un tube autour d'une protéine de mesure. Pour un même phage, cette protéine de mesure étant toujours identique, les queues font toujours la même taille (Seul et al, 2021). . Elles sont également souvent munies d'une pointe et de fibres à son extrémité.

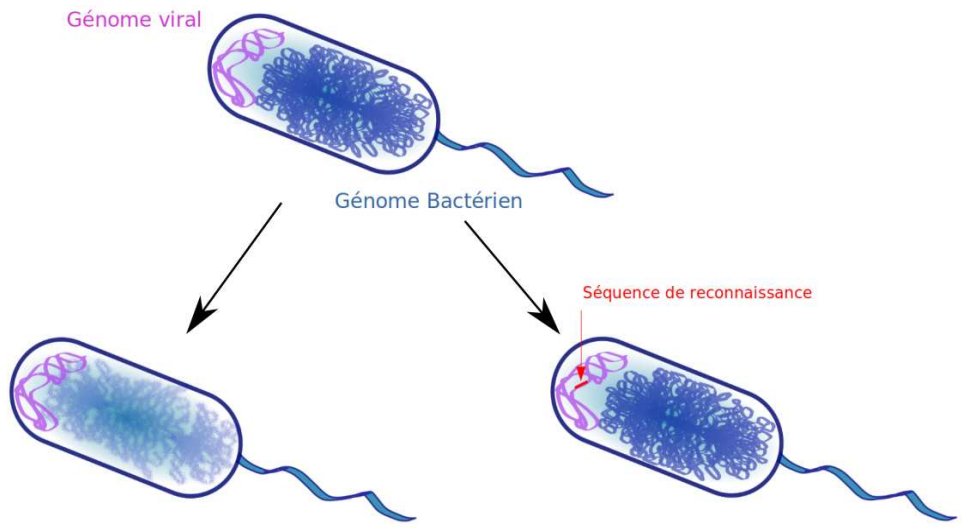


Figure 6. Deux stratégies mises en œuvre pour la reconnaissance du génome viral. (1) Dégradation du génome de l'hôte (gauche) (2) reconnaissance d'une séquence spécifique (droite).

I.III. Reconnaissance du génome viral et encapsidation chez les bactériophages et Herpèsvirus

I.III.I. Généralités

Lorsque le génome des phages caudés est spécifiquement reconnu, deux grandes stratégies de reconnaissance sont employées. La terminase reconnaît soit une séquence *cos* soit une séquence *pac*, en fonction de la nature de la séquence reconnue, les mécanismes d'encapsidation sont différents. L'ADN est clivé une première fois par la terminase au niveau de ces séquences de reconnaissance. L'ADN est encapsidé puis clivé une deuxième fois soit au niveau d'une séquence spécifique *cos*, soit, dans le cas des phages *pac*, de manière séquence indépendante lorsque la capside est à son encombrement maximal. Le complexe terminase/ADN se détache ensuite de la capside puis interagit avec une autre procapside dans laquelle elle va continuer d'encapsider le concatémère par un mécanisme processif. Chez les phages *cos*, chaque clivage s'effectue spécifiquement au niveau de la séquence de reconnaissance (figure 7.A). Chez les phages *pac*, la séquence *pac* n'est clivée qu'une seule fois par série de cycles d'encapsidation du concatémère, tous les autres clivages ne dépendant que de l'encombrement de la capside. Dans le ce second cas, on parlera d'encapsidation par tête pleine (figure 7.B). Lors de l'encapsidation par tête pleine plus d'une copie du génome est encapsidée ce qui génère des molécules partiellement permutées avec de la redondance terminale (figure 7.B). Chaque molécule possède donc des extrémités dont les séquences sont identiques, mais ces séquences varient au cours des différents cycles d'encapsidation à cause du décalage induit par le fait qu'à chaque cycle plus d'un génome est encapsidé. Chez les phages *cos*, cette permutation partielle et cette redondance de séquence ne sont pas retrouvées car ces phages encapsident strictement un génome (Fujisawa et Morita. 2003).

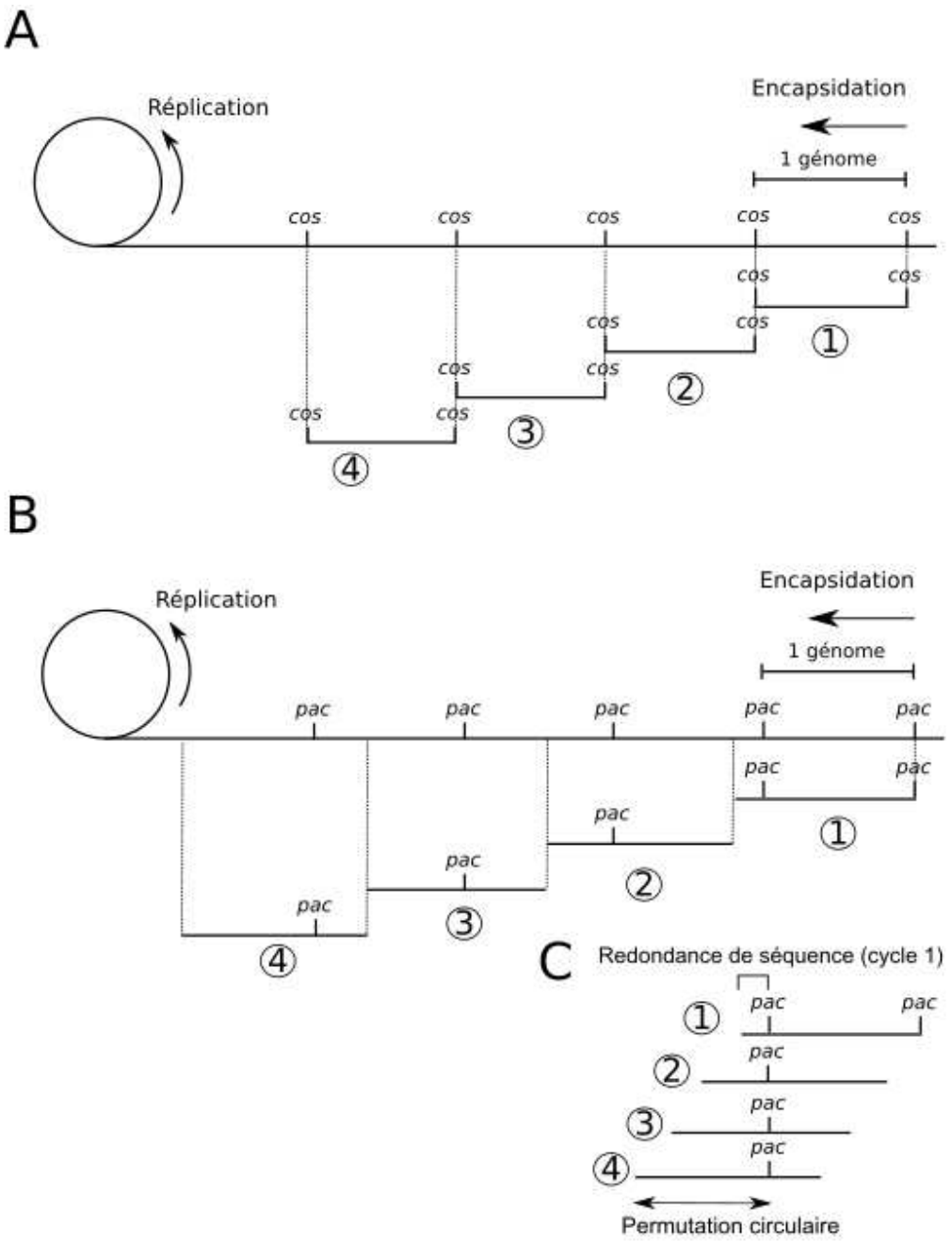


Figure 7. Mécanismes d'encapsidation les plus fréquents retrouvés chez les bactériophages caudés. (A) Encapsidation chez les phages *cos*. Une séquence *cos* est reconnue en début et en fin d'encapsidation sur le concatémère. Des molécules contenant strictement un seul génome sont donc encapsidées. (B) Encapsidation par tête pleine. La séquence *pac* est reconnue sur le concatémère et clivée une seule fois lors de l'initiation du processus d'encapsidation, puis l'ADN est de nouveau clivé à la fin de chaque cycle d'encapsidation lorsque la capside est pleine. La capside contient une molécule plus longue que le génome viral. (C) Les molécules d'ADN encapsidées ont une redondance terminale et une permutation circulaire partielle du génome.

I.III.II. Le phage lambda, un phage *cos*

L'un des phages *cos* les mieux caractérisés est le phage lambda. Le complexe terminase de lambda est composé de gpNuI et gpA qui correspondent respectivement à la petite et à la grande sous-unité. La séquence *cos* de lambda s'étend sur 200 pb et se divise en trois sites : *cos*N, *cos*B et *cos*Q. La petite sous-unité de la terminase gpNuI reconnait la séquence *cos*B, notamment 3 segments appelés *R elements* de 16pb chacun. La séquence *cos*B possède également un site I1 qui permet le recrutement d'un facteur de son hôte *E. coli*, la protéine IHF. Cette protéine appartient à la famille des NAP (nucleoid associated proteins) qui induisent des contraintes structurales à l'ADN et agissent souvent comme facteurs de transcription. Dans ce cas précis, IHF induirait une courbure de l'ADN qui favoriserait l'interaction de gpNuI avec son site de fixation. La grande sous-unité gpA sous forme dimérique interagit avec gpNuI liée à l'ADN et clive le concatémère au niveau de *cos*N. Le clivage génère une extrémité 5' protubérante de 12 nucléotides. Le complexe interagit ensuite avec une procapside et l'ADN est encapsidé. L'encapsidation se termine au niveau du site *cos* suivant sur le concatémère et est de nouveau clivé par gpA au niveau de *cos*N. Mais à la différence de l'initiation de l'encapsidation, le clivage est induit par une interaction de gpNuI sur *cos*Q et non sur *cos*B. Cette nouvelle coupure génère elle aussi une extrémité cohésive avec une extrémité protubérante de 12 nucléotides complémentaires à celle générée par le clivage ayant lieu au début de l'encapsidation. Cette propriété du génome encapsidé permet sa circularisation lors de l'infection (Fujisawa et Morita. 2003 ; Murialdo. 1991).

I.III.III. Le phage T4

Le phage T4 est un bactériophage chez lequel les mécanismes impliqués dans l'encapsidation ont été beaucoup étudiés. On a longtemps pensé que l'initiation de l'encapsidation avait lieu de manière aléatoire sur le génome de ce phage. Cependant, des expériences ont montré que l'initiation de l'encapsidation commençait préférentiellement au niveau du gène *16* qui code pour la petite sous-unité de la terminase. Lorsque ce gène est cloné dans un plasmide, ce-dernier est transduit plus efficacement dans des particules virales, suggérant là aussi une affinité de la petite sous-unité de la terminase pour cette séquence codante. Un site *pac* alternatif aurait aussi été identifié au niveau du gène *19*. Chose étonnante, malgré la présence d'une séquence d'initiation pour l'encapsidation, il a été démontré qu'il était possible d'encapsider de l'ADN linéaire *in vitro* avec seulement la grande sous-unité de la terminase, ce qui n'est pas possible chez d'autres phages comme SPP1. Cependant, la petite sous-unité

de la terminase reste nécessaire pour encapsider des concatémères *in-vivo*. Ceci semble indiquer que l'encapsidation commence sûrement à partir d'une séquence spécifique, mais qu'il doit également être possible, dans de rares cas, d'initier l'encapsidation ailleurs sur le génome du phage (Fujisawa et Morita, 2003 ; Gao et al, 2016).

Le processus d'encapsidation de T4 semble également être couplé à d'autres activités cellulaires comme la réplication et la transcription. En effet, il a été démontré que la protéine gp49, une résolvase de jonctions de Holliday qui résout les structures induites dans l'ADN au cours de la réplication, était nécessaire à l'encapsidation de concatémères *in vivo*. Le facteur sigma tardif gp55 est aussi nécessaire pour initier l'encapsidation en présence de la terminase (Golz et al, 1999; Black, 2015).

I.III.IV. Exemples de phages *pac* : SPP1, P1, P22, Sf6 et Mu

Les stratégies de reconnaissance et de clivage de *pac* mises en œuvre diffèrent beaucoup d'un phage à l'autre. Chez les phages SPP1, P1, P22, Sf6 ou Mu par exemple, la séquence *pac* possède une organisation distincte et les mécanismes de clivage diffèrent également (figure 8).

Chez le phage P22 de *Salmonella*, une séquence consensus de 22 nucléotides est reconnue par TerS sur son propre gène mais le clivage de *pac* peut avoir lieu à différents endroits de part et d'autre du site *pac* (Casjens et al, 1992a ; Wu et al, 2002).

Le mécanisme de reconnaissance du phage Sf6 de *Shigella flexneri* est proche de celui de P22. Une séquence *pac* est reconnue par TerS sur son gène mais le clivage peut avoir lieu 1600 pb de part et d'autre du site contre 120 pb pour P22 (Leavitt et al, 2013).

Chez P1 une séquence de 161 nucléotides est reconnue et le clivage a lieu sur cette séquence toujours au même endroit (Sternberg et al, 1987).

La TerS du phage Mu reconnaît un site *pac* de 23pb et le clivage a lieu de manière très imprécise 10 à 200bp en amont (Harel et al, 1990).

Finalement, chez SPP1 la séquence *pac* est divisée en 2 sites dont l'un chevauche le gène codant TerS et le clivage a lieu à une position précise (voir section I.V.IV.V).

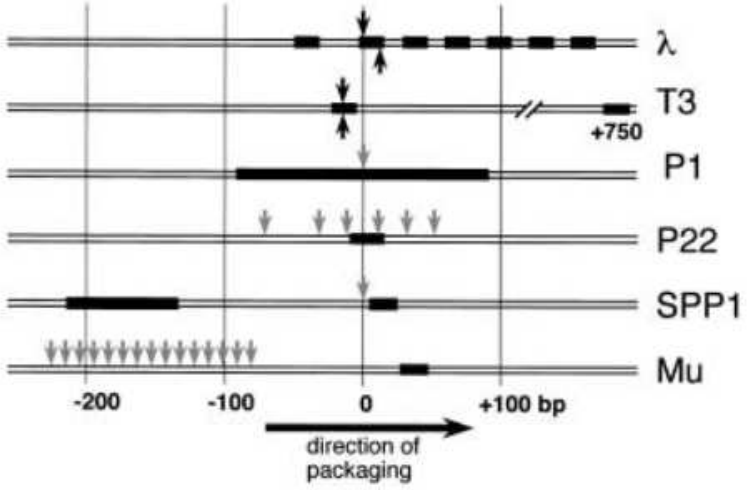


Figure 8. Organisation de la séquence de reconnaissance du génome viral chez des bactériophages à ADNdb. Chaque double ligne horizontale correspond à la région où le génome est reconnu. Les régions en noir sont les sites reconnus par la terminase. Les flèches verticales noires indiquent les sites de coupure de TerL et les flèches grises des sites de coupure plus hétérogènes. La flèche noire horizontale indique le sens de l'encapsidation. Reprise de Wu et al, 2002.

I.III.V. Les phages T3 et T7

T3 et T7 sont des phages très proches qui partagent de nombreuses caractéristiques communes et des mécanismes d'encapsidation similaires. Comme chez la majorité des phages, un ADN double brin concatémérique est synthétisé au cours de l'infection. Mais contrairement aux autres phages décrits, les molécules d'ADN encapsidées ont des séquences répétées identiques de 230 pb chez T3 et de 160 pb chez T7. Ces séquences auraient un rôle dans l'initiation de l'encapsidation. En effet, au sein de ces répétitions, est retrouvée une séquence *pac*B qui est le site reconnu par la terminase qui clive un site *pac*C en aval de *pac*B. On distingue ainsi les séquences *pac*CL et *pac*CR qui correspondent aux séquences *pac*C retrouvées au niveau des extrémités répétées gauche et droite d'un même génome. Comme tous les phages *pac*, seule la première coupure au niveau de cette séquence de reconnaissance est spécifique, les autres clivages ayant lieu sur le concatémère en fin d'encapsidation ne dépendent pas d'une séquence particulière (Catalano. 2005 ; Hashimoto &Fujisawa 1992).

La reconnaissance de *pac* et l'initiation de l'encapsidation, chez T3 comme T7, seraient liées au démarrage de la transcription. En effet, chez le phage T7 qui code sa propre ARN polymérase, des mutations sur cette protéine qui inhibent la pause transcriptionnelle (l'ARN polymérase reconnaît un signal qui arrête temporairement la transcription peu après l'initiation de la synthèse d'un ARN) empêchent également l'initiation de l'encapsidation. De même, chez T3 des mutations dans un signal de pause transcriptionnelle en aval de *pac*

inhibent l'initiation de l'encapsidation. Donc, il est fort probable que la petite sous-unité de la terminase reconnaisse le site *pac* seulement quand la transcription des régions répétées a commencé et que l'ARN polymérase est en attente au niveau d'un signal de pause transcriptionnelle.

Fujisawa et al ont présenté un modèle pour expliquer l'initiation de l'encapsidation (figure 9). Selon eux, la terminase reconnaîtrait l'ADN et induirait une coupure sur l'ADN simple brin généré par le déroulement de la double hélice induit par l'ARN polymérase. Ensuite, de l'ADN double brin serait re-synthétisé par l'ADN polymérase à partir de l'extrémité cohésive générée par la première coupure. À la suite de quoi, une nouvelle coupure sur l'ADN double brin serait induite sur le concatémère et l'encapsidation initiée à partir de ce point. La fin de l'encapsidation, quant à elle, ferait intervenir une exonucléase pour désolidariser le génome en cours d'encapsidation du reste du concatémère et à nouveau une ADN polymérase pour synthétiser de l'ADN double brin. Ce modèle permet ainsi d'expliquer comment les séquences répétées de part et d'autre de chaque génome sur le concatémère se répliquent pour que tous les génomes encapsidés possèdent des séquences répétées à leurs 2 extrémités. Il ne s'agit cependant que d'un modèle qui demande à être vérifié (Fujisawa & Morita. 1997).

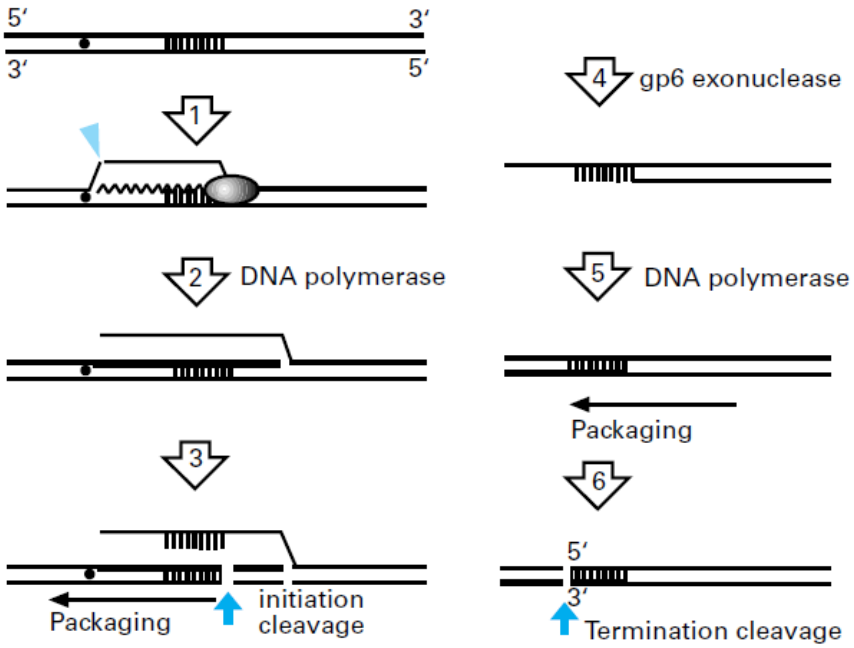


Figure 9. Modèle d'encapsidation chez les phages T3 et T7. La terminase induit une coupure sur l'ADN simple brin généré par la transcription. L'ADN polymérase synthétise un nouveau brin à partir de l'extrémité cohésive générée par la première coupure. Après un nouveau clivage, l'encapsidation démarre. Un clivage sur l'ADN double brin intervient en fin d'encapsidation. Figure adaptée de Fujisawa & Morita. 1997.

I.III.VI. Reconnaissance et encapsidation du génome chez les Herpèsvirus, exemple de HSV1

HSV1 pour Herpes Simplex Virus 1 (HSV1) est un virus de la famille des *Herpesviridae*. Les virus de cette famille forment des particules virales enveloppées. Le génome à ADN double brin est contenu dans une capside de forme icosaédrique. Ces virus à ADN infectent les humains et d'autres animaux, malgré tout, ils partagent de nombreuses similitudes avec les *Caudovirales*, notamment dans leur voie de morphogénèse. Comme les phages caudés, une procapside s'auto-assemble à l'aide de protéines échafaudage qui permettent le recrutement de la protéine majeure de capside. Ces virus possèdent eux-aussi un pore formé par une protéine portal à l'un des sommets de l'icosaèdre. Cependant, à la différence des phages caudés, les herpesvirus n'assemblent pas de queue. Les herpèsvirus possèdent également un tégument autour de leur capside ainsi qu'une enveloppe lipidique provenant de la membrane de la cellule hôte.

Les mécanismes de reconnaissance et d'encapsidation de l'ADN viral sont extrêmement proches de ceux mis en œuvre par les bactériophages caudés. L'ADN viral synthétisé au cours de l'infection forme des concatémères composés de plusieurs génomes les uns à la suite des autres selon une organisation tête-queue. Là encore, ces structures sont formées lorsque le virus réplique son ADN par cercle roulant (réplication sigma). Chez HSV1, la structure du génome possède certaines particularités. Celui-ci est composé de deux régions non-répétées contenant des séquences codantes nommées UL et US (*small* et *large*) en fonction de leur taille. Ces régions sont entourées de séquences répétées et inversées. La région UL, d'environ 128Kpb, est entourée de séquences répétées nommées ab et b'a' et la région US, de 25Kpb de long, des séquences a'c' et ca. La séquence a est répétée plusieurs fois au début du génome. Comme chez les phages caudés, ce génome est reconnu par un complexe spécialisé appelé terminase qui reconnait des sites spécifiques sur le génome. Chez HSVI, ce complexe est composé de 3 protéines UL15, UL28 et UL33. UL28 reconnaîtrait l'ADN alors qu'UL15 assurerait le clivage et la translocation ATP dépendante de l'ADN. À ce jour, le rôle d'UL33 reste moins bien caractérisé. La séquence de reconnaissance du génome viral se situe au niveau de la séquence a. Elle contient un site *Pac2* suivi d'un site *Pac1* qui sont tous deux impliqués dans la reconnaissance et le clivage de l'ADN viral. La séquence a se situant aux 2 extrémités du génome, *Pac1* et *Pac2* se retrouvent au début et à la fin de chaque génome (figure 10). La terminase reconnait d'abord *Pac2*. Elle clive l'ADN en amont de *Pac2* au niveau d'une séquence répétée appelée DR1. Ensuite, elle interagit avec une procapside et le

concatémère est encapsidé de façon processive. En fin d'encapsidation, le site *Pac1* est reconnu à l'extrémité du génome et l'ADN clivé au début de DR1 en aval de *Pac1* à la jonction entre deux génomes. Ensuite, *Pac2* est reconnu sur le génome suivant et l'encapsidation se poursuit dans une autre procapside. Les clivages générés par UL15 produisent des extrémités cohésives qui permettent la circularisation du génome lors de l'infection (Catalano. 2005 ; Adelman et al, 2001).

Les mécanismes mis en place par ces virus ressemblent à ceux des phages *cos*, mais ils restent différents à quelques égards. Chez les phages *cos*, le clivage à lieu au niveau d'un site *cos* qui est présent en une seule copie par génome, alors que chez HSVI l'information nécessaire aux clivages de début et de fin d'encapsidation est présente en plusieurs copies au sein d'un même génome. Cependant, comme chez le phage lambda par exemple, les signaux marquant le début et la fin de l'encapsidation constituent des séquences différentes. Une problématique similaire à celle des phages *pac* demeure : les séquences *Pac1* et *Pac2* sont retrouvées en de deux copies au sein d'un même génome mais elles ne sont pas toutes clivées.

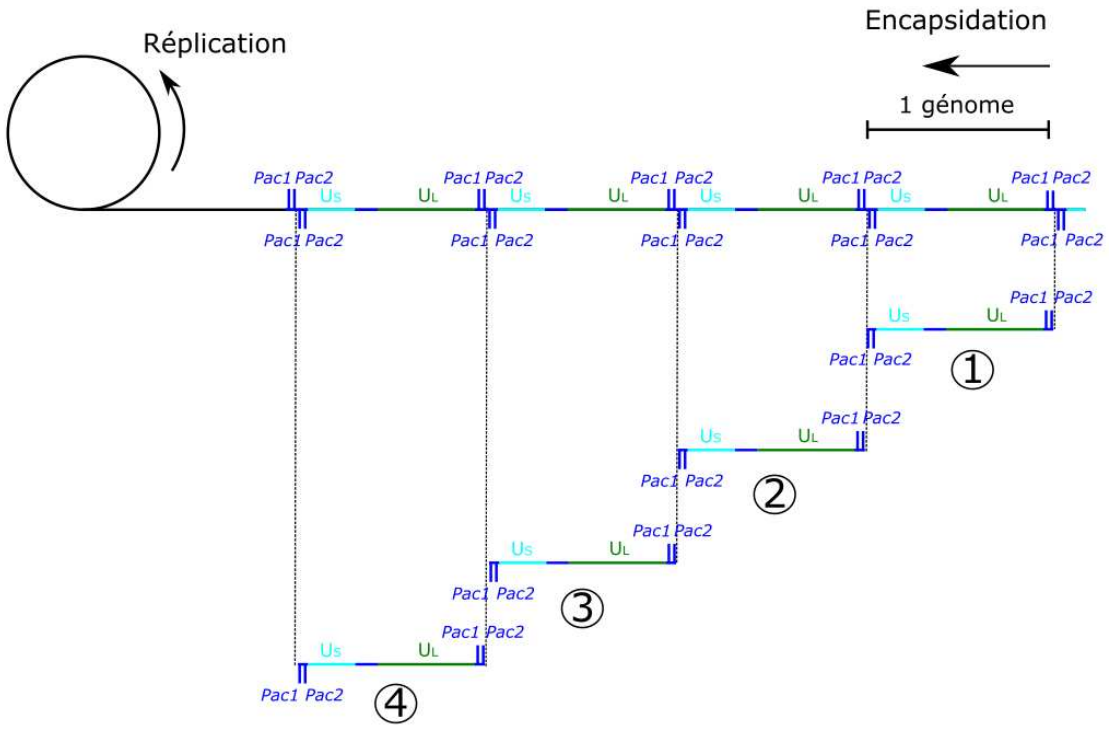


Figure 10. Encapsidation de l'ADN viral chez HSV1. Chez HSVI le substrat pour l'encasidation est également un concatémère. La terminase clive *Pac2* en début d'encapsidation et *Pac1* en fin d'encapsidation. Chaque ADN encapsidé contient donc la même séquence. Les régions US et UL sont représentées en turquoise et en vert respectivement.

I.IV. Défauts de reconnaissance du génome viral et transfert horizontal de gènes

Au cours de l'infection virale, il arrive parfois que le phage encapside de l'ADN de son hôte. Le bactériophage sert alors de vecteur pour le transfert horizontal de matériel génétique. Ce pocessus appelé transduction est un mécanisme qui contribue fortement à l'évolution des bactéries et des archées dans les différents écosystèmes où elles se développent. La transduction est notamment à l'origine de la diffusion de résistances aux antibiotiques au sein des populations bactériennes (Abedon et al, 2009;Volkova et al, 2014). Cependant, d'autres phénomènes comme la conjugaison et la transformation bactérienne contribuent eux aussi au transfert horizontal de gènes entre bactéries. Afin de mieux comprendre comment les bactéries évoluent dans leur milieu, il est donc important de prendre en considération la transduction dont le rôle semble être majeur.

La transduction de l'ADN de l'hôte dépend de différents mécanismes, et notamment du type de cycle lytique ou lysogénique mis en jeu lors de l'infection. À cet effet l'on différencie majoritairement deux types de transduction, la transduction généralisée et la transduction spécialisée.

I.IV.I. Transduction généralisée

Des événements de transduction généralisée se produisent lorsque le phage entre en cycle lytique. Le phage se trompe de génome et encapside de l'ADN bactérien. Se retrouvent alors dans la capside des virions formés au cours de l'infection des morceaux d'ADN bactérien de la taille d'une molécule d'ADN phagique encapsidée. Ce type de transduction est appelé généralisé car il permet le transfert de n'importe quelle région du génome de l'hôte. En plus, les molécules étant de grande taille, elles peuvent contenir plusieurs gènes bactériens. Ces évènements de transduction généralisée restent très rares mais ils sont suffisamment fréquents pour jouer un rôle dans l'échange et la diffusion de gènes dans les populations de bactéries (Adebon et al, 2009 ; Chiang et al, 2019).

Les mécanismes qui expliqueraient un tel processus sont encore mal connus. Le complexe terminase pourrait se fixer directement sur l'ADN de l'hôte et l'évènement de transduction s'expliquerait par un défaut de reconnaissance par le complexe terminase. Dans ce cas-ci, la terminase reconnaîtrait un site partageant une homologie plus ou moins forte avec la séquence qu'elle reconnaît sur le génome viral. Chez les phages *pac*, la terminase n'a besoin que d'un

site *pac* pour encapsider l'ADN issu d'un seul et même concatémère ensuite tous les clivages peuvent avoir lieu sur l'ADN bactérien indépendamment de la présence d'un site homologue à *pac* (figure 7). Chez les phages *cos,* la problématique est différente car une séquence spécifique est nécessaire en début et en fin d'encapsidation. Le phénomène décrit précédemment ne pourrait théoriquement pas se produire chez ces phages car la probabilité de retrouver deux sites homologues à *cos* distants d'une longueur équivalente à la taille d'une molécule encapsidée est quasiment nulle. On admet donc généralement que seuls les phages encapsidant leur ADN par tête pleine font de la transduction généralisée.

La transduction généralisée pourrait aussi être le résultat d'événements de recombinaison entre l'ADN du phage et l'ADN de l'hôte, on parlera alors de transduction facilitée. Cette seconde hypothèse implique qu'il y ait des séquences homologues entre l'hôte et le virus qui faciliteraient la recombinaison. Dans ce cas précis, l'encapsidation d'ADN bactérien n'impliquerait pas un défaut de reconnaissance d'un site *pac* ou *cos* par la terminase. Pour les phages encapsidant leur ADN par tête-pleine, la terminase reconnaitrait *pac*, encapsiderait de l'ADN phagique puis de l'ADN bactérien dans différentes procapsides à partir de la jonction produite par l'événement de recombinaison. Il serait alors possible de retrouver une molécule contenant à la fois de l'ADN de l'hôte et de l'ADN bactérien et des molécules contenant uniquement de l'ADN bactérien. Ce phénomène a déjà été observé chez les phages SPP1 et P22 qui peuvent transduire à haute fréquence des plasmides contenant des séquences homologues à leur génome (Canosi et al, 1982 ; Deichelbohrer et al, 1985).

Dans un autre cas, le phage encapsiderait directement de l'ADN bactérien sur lequel il n'y aurait ni de l'homologie avec une partie quelconque du génome du phage, ni aucune séquence proche d'une séquence *pac*. Dans ce contexte, le complexe terminase se fixerait aléatoirement sur le génome de l'hôte et tout le génome bactérien devrait être encapsidé à la même fréquence. Nous verrons dans la partie résultats de ce manuscrit que certaines expériences semblent indiquer que c'est effectivement ce qui se produit chez le bactériophage SPP1 (Yasbin et al, 1974 ; De Lencastre et al, 1980).

I.IV.II. Transduction spécialisée

La transduction spécialisée est retrouvée uniquement chez les phages tempérés. Lorsque ceux-ci entrent en cycle lysogénique, ils s'intègrent au niveau d'un site précis sur le chromosome bactérien à l'aide d'une intégrase généralement codée par le phage et parfois d'autres facteurs provenant de l'hôte (Fis et IHF dans le cas de lambda) ou du phage lui-même. Le site d'insertion est souvent unique. Par exemple, le phage lambda s'intègre systématiquement par recombinaison au niveau du site *attB* sur le chromosome de *E. coli*. Lors de l'induction du prophage, le génome est excisé par l'intégrase au cours d'évènements de recombinaison site-spécifique (Murialdo. 1991). Dans la grande majorité des cas, seul le génome du phage est excisé, répliqué et encapsidé. Cependant, il arrive que l'excision soit imparfaite et que de la recombinaison illégitime ait lieu sur le chromosome bactérien, ce qui conduit à la formation de molécules chimériques contenant à la fois de l'ADN viral et de l'ADN bactérien. Le génome est répliqué, puis si la molécule contient un site d'initiation de l'encapsidation, elle est encapsidée formant ainsi des particules capables de transférer de l'ADN bactérien. Cette transduction est dite spécialisée car seuls certains gènes en amont ou en aval du site d'insertion peuvent être transférés. Dans le cas du phage lambda, seuls les gènes *gal* et *bio* sont transduits. Cela signifie qu'il n'est pas possible pour la bactérie receveuse d'incorporer de nouveaux gènes mais seulement des nouveaux allèles des gènes en amont et en aval du site d'insertion (Chiang et al, 2019 ; Schneider, 2021).

I.IV.III. Transduction latérale

Chez certains phages tempérés, l'excision du prophage lors de l'induction du cycle lytique se produit tardivement. Il arrive donc parfois que le génome soit répliqué, reconnu et encapsidé avant l'excision du prophage. Si l'encapsidation s'effectue par tête pleine, plus d'un génome est encapsidé à partir d'une molécule. Autrement-dit, après avoir encapsidé son génome le phage continue d'encapsider l'ADN du chromosome bactérien. Ce type de transduction, appelé transduction latérale, a été étudié chez certains phages tempérés de *Staphylococcus aureus* qui transduisent l'ADN de la bactérie localisé d'un côté du prophage à très haute fréquence (Chiang et al, 2019 ; Chen et al, 2018). La principale différence avec la transduction spécialisée réside dans le fait que l'ADN est répliqué, reconnu et encapsidé avant l'excision du prophage et que de grandes portions du génome bactérien sont encapsidées par tête pleine.

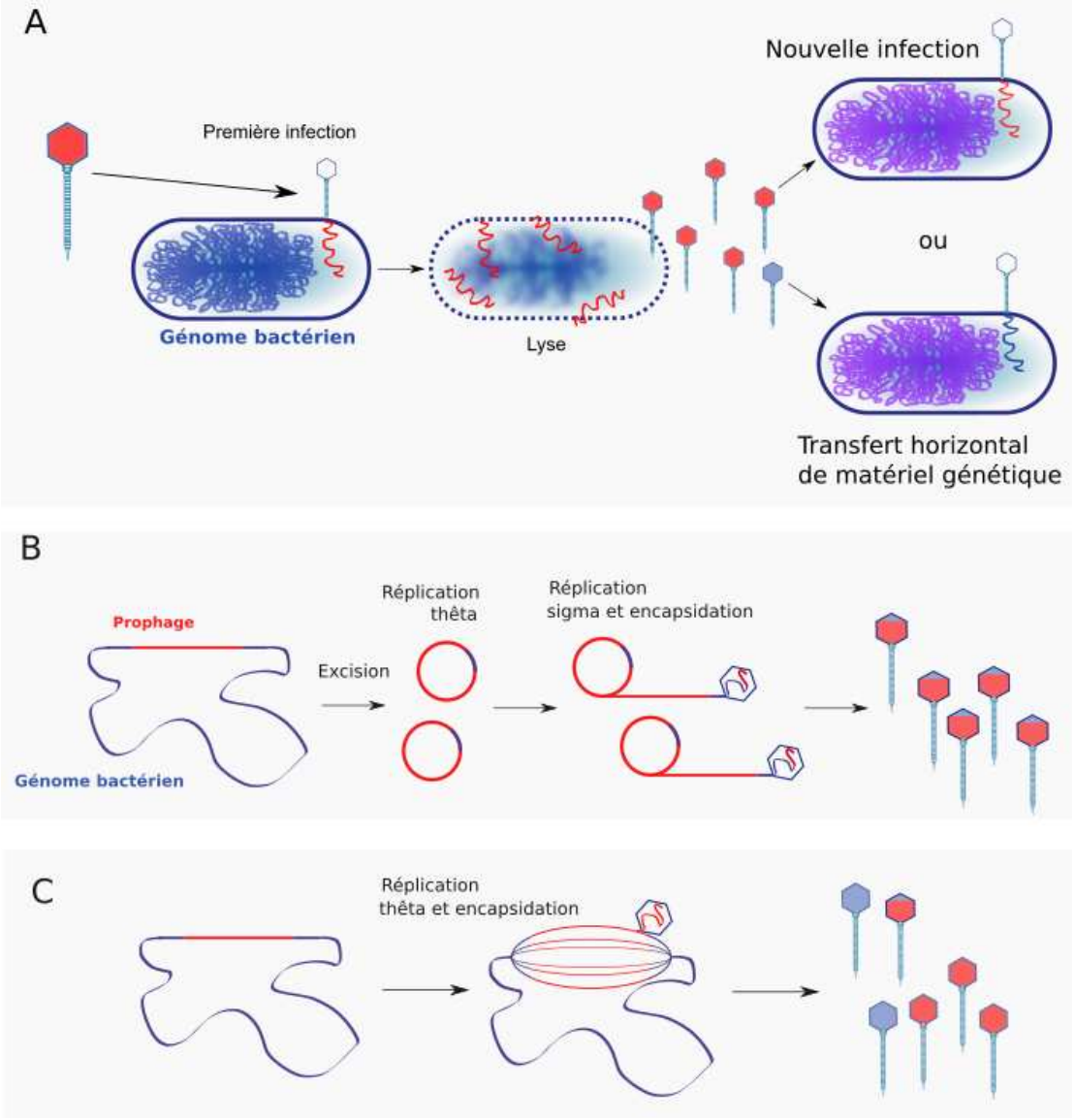


Figure 11. Transduction d'ADN bactérien. (A) Transduction généralisée : Une bactérie est infectée par SPP1 qui injecte son génome dans la cellule (rouge), à l'issue de l'infection des virions sont générés. Dans de rares cas, certaines particules contiennent de l'ADN bactérien (bleu) qui a été encapsidé par erreur au cours de l'infection. Un phage contenant de l'ADN de SPP1 (rouge) sera viable et pourra infecter une autre cellule. Les particules contenant de l'ADN bactérien vont éjecter cet ADN dans une nouvelle cellule qui pourra l'intégrer à son propre génome par recombinaison homologue (génome bleu dans génome violet). (B) Mécanismes impliqués dans la transduction spécialisée. Une excision anormale du prophage lors de son induction conduit à la formation d'une molécule hybride contenant de l'ADN phagique et bactérien. Cet hybride est ensuite répliqué selon un mode thêta puis sigma, cet ADN est encapsidé formant des virions contenant ces molécules hybrides. (C) Mécanismes impliqués dans la transduction latérale. Le prophage se réplique selon un mode thêta et est également encapsidé avant son excision, ce qui permet de transduire de l'ADN bactérien à haute fréquence.

I.V. Le Bactériophage SPP1

I.V.I. Généralités

La bactériophage SPP1 est un phage caudé de la famille des *Siphoviridae*, qui possède donc une longue queue non-contractile. Il infecte la bactérie à Gram positive *B. subtilis* qui vit dans le sol. SPP1 a été isolé par Riva et al dans le jardin botanique de Pavia en Italie en 1968, d'où son nom qui signifie Subtilis Pavia Phage 1 (SPP1). La photo ci-dessous est la première image de SPP1 obtenue par microscopie électronique dans la publication de 1968 (Riva et al, 1968).

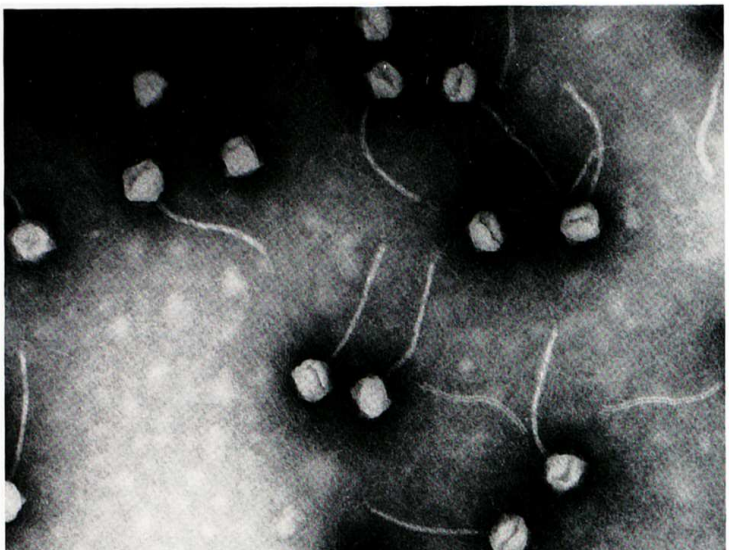


Figure 12. Image de particules virales du bactériophage SPP1 par microscopie électronique en transmission (coloration négative). Reproduite de Riva et al, 1968.

SPP1 est un phage strictement lytique, donc virulent, qui ne réalise pas de cycle lysogénique. La bactérie *B. subtilis* étant bien caractérisée et relativement simple à cultiver, SPP1 est un virus de choix pour l'étude des bactériophages. D'ailleurs, avec des phages de *E. coli* comme T4 ou lambda, il figure parmi les phages les plus étudiés. SPP1 est utilisé pour étudier la réplication et la morphogénèse virale. De manière plus détaillée, de nombreuses études sur la structure de la capside ou de la queue avec leurs étapes d'assemblage ou encore l'encapsidation qui va nous occuper dans la suite de ce manuscrit ont été entreprises.

Le cycle de SPP1 dure entre 30 et 60 minutes dans les conditions optimales de croissance exponentielle de *B. subtilis* à 37°C. Il suit un certain nombre d'étapes dont : la reconnaissance d'un récepteur membranaire (YueB) et l'éjection du génome dans le cytoplasme de l'hôte, une phase de réplication de l'ADN viral, l'assemblage de la procapside et de la queue,

l'encapsidation de l'ADN, l'assemblage final des virions avec la ligation de la queue et de la capside, et enfin la lyse de la cellule. Un cycle d'infection donne en moyenne plus de 100 virions (Labarde et al, 2021), c'est ce que l'on appelle *burst size* en anglais.

Comme il a été spécifié précédemment, SPP1 est un phage *pac* qui encapside son ADN par tête pleine (voir plus de détails dans les sections I.V.IV). Ces caractéristiques en font également un virus de choix pour l'étude des mécanismes de transduction généralisée qui occuperont une part conséquente de ce manuscrit. Cette propriété de SPP1 est d'ailleurs utilisée en biotechnologie notamment dans la création de mutants génétiques de *B. subtilis*. Elle a aussi été utilisée pour construire la carte génétique de *B. subtilis* (Lencastre et al, 1980). SPP1 est aussi un modèle important dans l'étude de la réplication du génome phagique qui fait intervenir 2 types de réplication différents dont les mécanismes ne sont pas encore très bien compris (voir sections suivantes).

I.V.II. Organisation du génome de SPP1

Comme tous les *Siphoviridae*, SPP1 possède un génome à ADN double brin. Sa longueur est de 44,016 kpb avec un contenu en GC de 43,7% (Godinho et al, 2018), ce qui est une taille moyenne pour un phage caudé. A titre de comparaison, le génome de phi29 ne fait que 19,3 kpb et celui de T4 168,9 kpb. Cependant, SPP1 étant un phage *pac*, sa capside contient un peu plus d'un génome soit environ 45.9 kpb. Le génome de SPP1 contient 80 cadres ouverts de lecture organisés en opéron. Des promoteurs spécifiques assurent la transcription séquentielle de ces gènes. Les promoteurs dits précoces sont reconnus par l'ARN polymérase de *B. subtilis* avec le facteur σ^A et sont transcrits en premier. Parmi ces gènes, l'on retrouve notamment tous ceux qui codent pour des protéines impliquées dans la réplication. Viennent ensuite les promoteurs tardifs qui requièrent un facteur provenant du phage qui n'a pas encore été identifié. Les gènes concernés codent pour des protéines structurales de la capside et de la queue, des protéines impliquées dans l'encapsidation ou encore dans la lyse cellulaire (voir figure 13). Dans la partie résultats, nous nous intéresserons en particulier au promoteur *P*L1 (promoteur tardif 1) qui contrôle la transcription des gènes *1* à *7*, ce qui inclut les sous-unités de la terminase et la protéine portal. Parmi tous les cadres de lecture identifiés, un grand nombre ne possède pas de fonction connue ou seulement une fonction théorique obtenue par analyse bioinformatique. Une partie du génome de SPP1 peut d'ailleurs être supprimée sans que cela affecte l'infectiosité des phages en conditions de laboratoire (partie hachurée figure 13 ; Godinho et al, 2018).

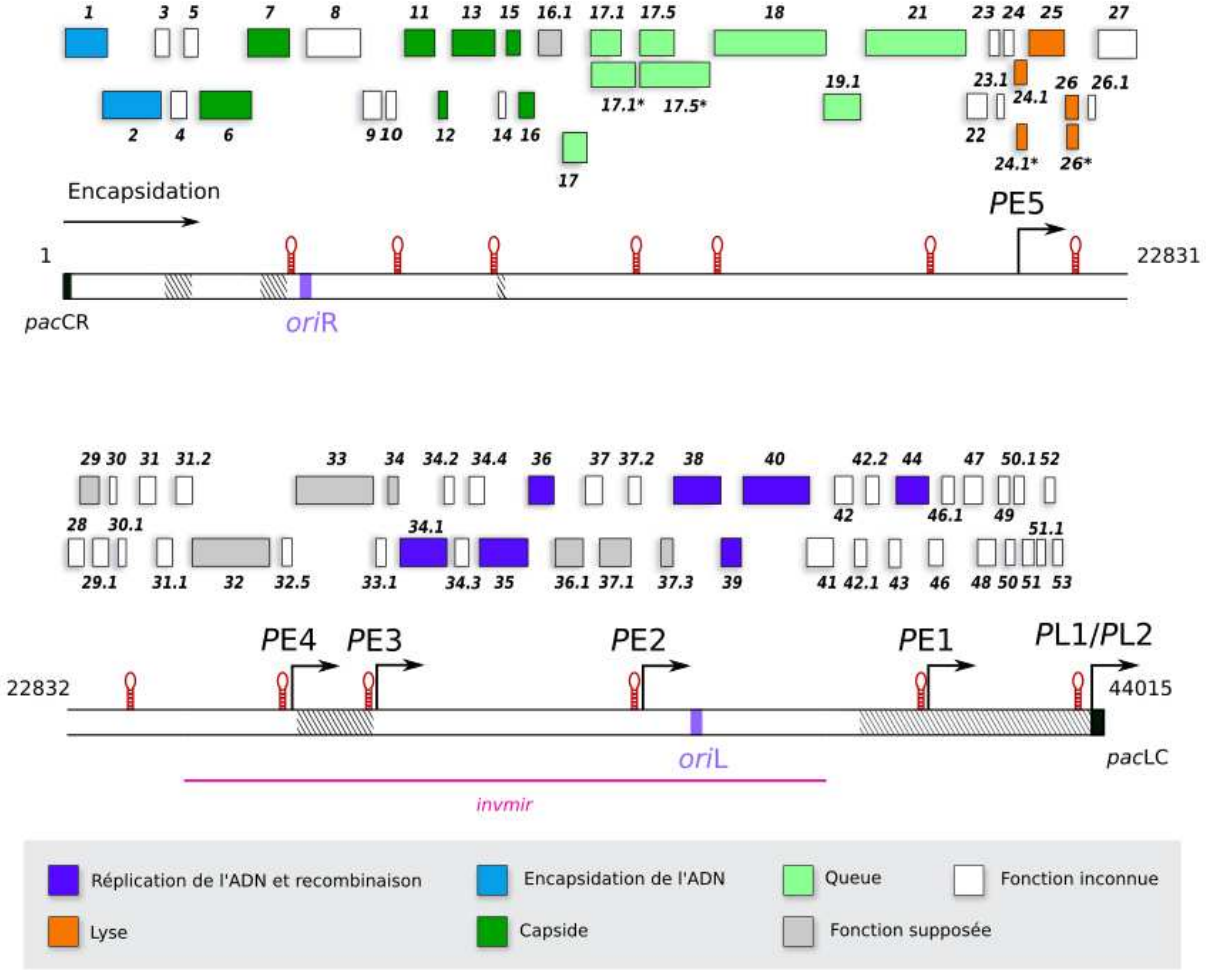


Figure 13. Organisation du génome de SPP1. La barre horizontale représente le génome de SPP1 d'une taille de 44016 pb. La coordonnée +1 correspond à la coupure *pac*. Les deux origines de réplication sont représentées en violet et les régions non essentielles en conditions de laboratoire par des hachures noires. La région invmir peut être inversée sans conséquences néfastes pour le phage. La position des promoteurs (flèches noires) et des terminateurs Rho-indépendants (en rouge) sont représentés le long du génome sur la ligne horizontale supérieure. Le sens d'encapsidation et de transcription sont identiques. Les cadres ouverts de lecture sont représentés par des rectangles numérotés et colorés en fonction du rôle de leur produit dans le cycle viral. Figure tirée de Godinho et al, 2018.

I.V.III. Réplication de l'ADN viral

Le génome phagique subit 2 modes de réplication au cours de l'infection. Ainsi, le génome de SPP1 dispose de 2 origines de réplication, *ori*R et *ori*L (pour right et left), dont l'intervention séquentielle dans chacun des modes de réplication, n'est, à ce jour, pas encore clairement élucidée. Lors du démarrage de l'infection, l'ADN de SPP1 est éjecté dans le cytoplasme de la bactérie. Comme nous l'avons vu précédemment, SPP1 encapside son ADN par tête pleine, ce qui signifie que plus d'un génome est contenu dans chaque particule virale, ce qui induit

des redondances de séquence aux 2 extrémités de la molécule encapsidée. Ainsi, le génome du phage peut être circularisé par recombinaison homologue une fois dans le cytoplasme bactérien.

La molécule circulaire possédant deux origines de réplication est ensuite répliquée selon un mode thêta. Ce type de réplication fait intervenir le réplisome de l'hôte qui forme une fourche de réplication et qui évolue ensuite de manière uni ou bi-directionnelle. À l'issue de la réplication, une nouvelle molécule d'ADN circulaire est formée. Certains facteurs du phage sont nécessaires à l'initiation de la réplication. La protéine gp38 se fixe au niveau d'*ori*L et ouvre la double hélice d'ADN, de multiples copies de la protéine gp36 interagissent ensuite avec l'ADN simple brin généré par l'ouverture de la double hélice ce qui permet de maintenir les 2 brins d'ADN séparés. La protéine gp39, couplée à l'hélicase du phage gp40, charge gp40 sur l'ADN simple brin, ce qui rompt l'interaction entre les 2 protéines. La libération de gp40 permet d'activer son activité hélicase et le recrutement du réplisome bactérien (Valero-Rello et al, 2017 ; Seco et al, 2017 ; Ayora et al, 2002).

Après plusieurs cycles de réplication thêta, les ADN circulaires générés sont ensuite répliqués selon un mode sigma, ce qui génère les concatémères qui vont servir de substrat pour l'encapsidation. Ce type de réplication mobilise l'origine de réplication *ori*R ou *ori*L. Les mécanismes impliqués dans la transition entre les 2 types de réplication ne sont pas encore bien compris.

Plusieurs modèles ont été proposés, dans l'un d'entre eux, la fourche de réplication est clivée puis chaque extrémité générée par la coupure est re-liguée à son brin adjacent. Zecchi et al, proposent que la protéine gp44 de SPP1 génère la coupure au niveau de la fourche de réplication. Ce qui est cohérent avec la fonction de gp44 qui est une résolvase de jonctions de Holliday. Ainsi, une extrémité 3' protubérante est générée (figure 14). Cette portion d'ADN simple brin est reconnue par la protéine phagique gp35 qui interagit avec une nouvelle fourche de réplication sur un autre génome en cours de réplication. Gp35 se dissocie ensuite de l'ADN.

A cette étape 2 hypothèses ont été formulées : (1) Une nouvelle coupure intervient au niveau de la deuxième fourche de réplication (peut-être médiée par gp44), et la réplication reprend au niveau des portions d'ADN simple brin générées par les réactions enzymatiques précédentes selon un mécanisme similaire à la réplication en cercle roulant. (2) La réplication continue au sein d'une fourche de réplication qui se déplace le long du génome circulaire (figure 14 .B). Dans les deux cas la réplication conduit à la formation de concatémères qui vont être reconnus

puis encapsidés dans des procapsides par des protéines spécifiques (Lo Piano et al, 2011 ; Ayora et al, 2002).

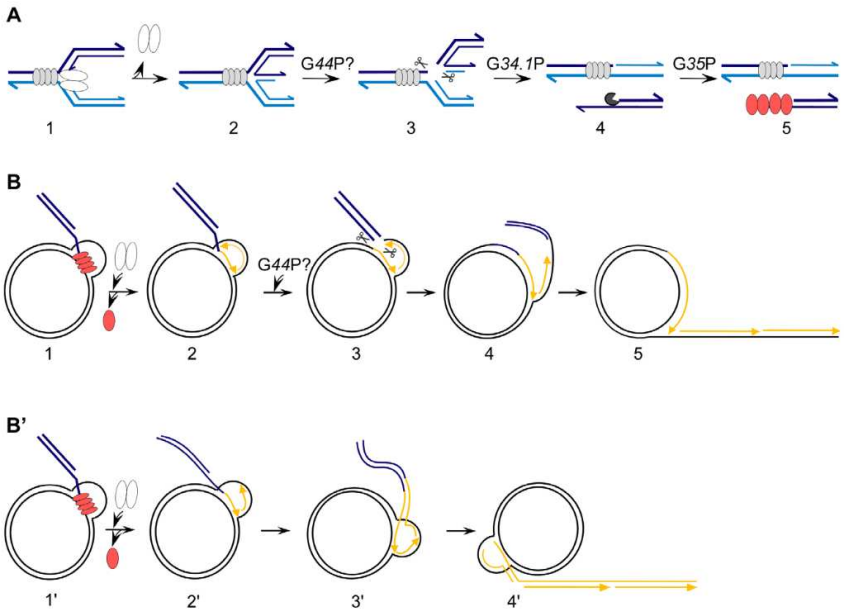


Figure 14. Transition entre la réplication thêta et sigma. (A) 1 :gp38 (ovales gris) se fixe sur la fourche de réplication (réplisomes en blanc). 2 : la réplication est stoppée et une jonction de Holliday se forme. 3 : gp44 induit une coupure double brin. 4 : une exonucléase (rond bleu marine) génère une portion d'ADN simple brin. 5 : gp35 (ovales rouges) se fixe sur la portion d'ADN simple brin. (B) Première possibilité : gp35 envahit une autre fourche de réplication qui est clivée par gp44 puis la réplication continue selon un mode par cercle roulant. (C) Seconde possibilité : gp35 liée à l'ADN simple brin envahit une nouvelle fourche de réplication, la synthèse d'ADN continue de mobiliser cette fourche de réplication, ce qui conduit à la synthèse d'un long ADN linéaire. Figure reproduite de Lo Piano et al, 2011.

I.V.IV. Morphogénèse virale

I.V.IV.I. Assemblage de la procapside

La capside du bactériophage SPP1 est un icosaèdre constitué de sous-unités de la protéine majoritaire de capside gp13. Les sous-unités de la protéine majeure de capside établissent des contacts quasi-équivalents formant des hexamères au niveau des surfaces plates de l'icosaèdre et de pentamères aux sommets. Les hexamères et pentamères sont appelés capsomères. La capside de SPP1 compte 415 sous-unités au total qui forment 60 hexamères et 11 pentamères retrouvés à chaque sommet de l'icosaèdre, à l'exception d'un seul sommet qui accueille la protéine portal gp6. Chaque face de l'icosaèdre est donc composée de 3 hexamères.

L'assemblage de la capside est un processus complexe qui requière la présence d'une protéine échafaudage dont le rôle est de faciliter l'agencement correct des sous-unités de la protéine majeure de capside entre elles. Chez SPP1, la protéine gp11 assure ce rôle. Cette protéine

interagit directement avec gp13 via des résidus spécifiques pour permettre à gp13 de s'agencer correctement. Gp11 et gp13 sont produites simultanément, et cette condition est nécessaire à la formation de la procapside (Dröge et al, 2000). Gp11 est une protéine chaperonne de gp13 assurant son repliement correct. Cette protéine de forme allongée est riche en structures secondaires de type hélice-alpha. Elle s'associe en dimères et peut également former des tétramères mais il a été proposé que ce sont les dimères qui participent à l'assemblage de gp13 (Poh et al, 2008).

Au cours de l'infection, les protéines gp6, gp11 et gp13 sont synthétisées simultanément. Ces protéines s'auto-assemblent en une structure appelée procapside I (figure 15, tableau 1). Cette structure est plus petite que la capside mature, elle ne mesure que 55nm de diamètre contre 61 nm pour la capside mature. Sa forme est légèrement plus arrondie. Dans cette structure, l'on retrouve la protéine portal, gp11 et gp13. La procapside I subit ensuite un processus de maturation en procapside II qui est concomitant avec le départ de gp11 (figure 15 .B). Ceci induit un changement de conformation des sous-unités de la protéine majeure de capside (élongation d'une partie N-terminale, et changement de structure d'une boucle appelée P). La structure formée par gp13 devient plus large et plus angulaire. La protéine gp11 est expulsée au niveau de trous présents au centre des hexamères. Seules les pentamères accueillent des sous-unités de gp11 dans la procapside II. C'est dans cette structure que l'ADN va être encapsidé (Ignatiou et al, 2019).

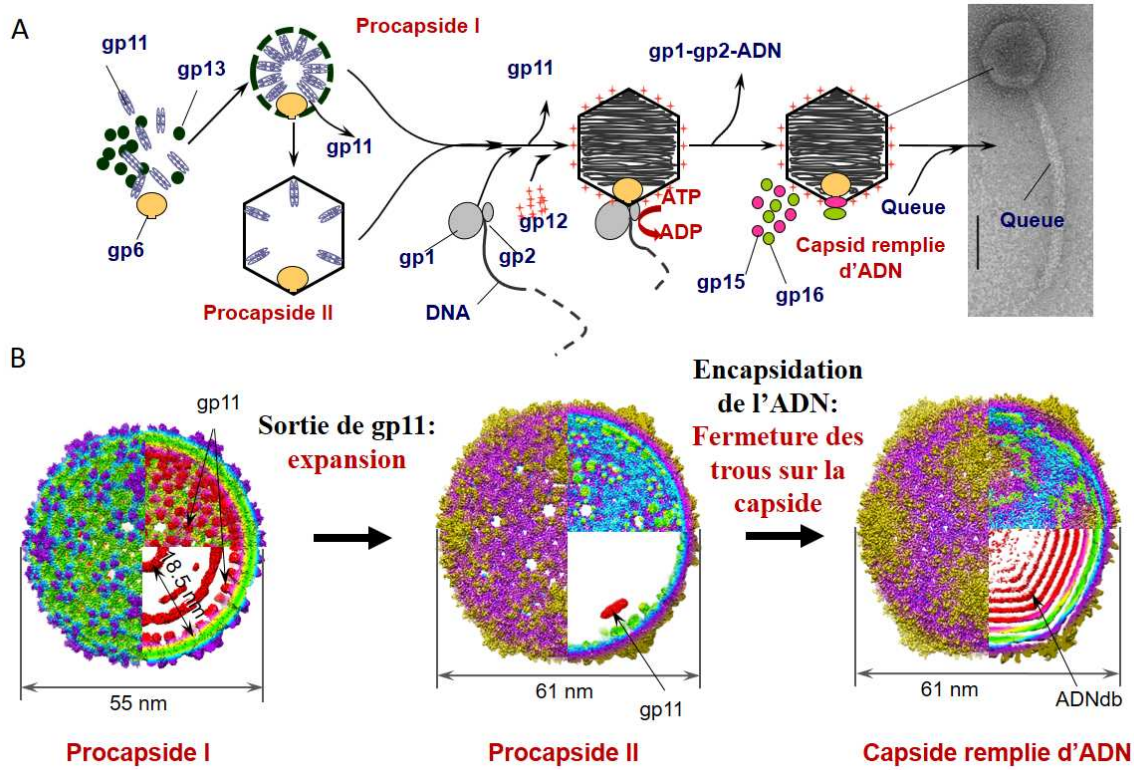


Figure 15. Voie d'assemblage des particules virales de SPP1 (A) et (B) structures obtenues par Cryo-EM de différents stades de maturation de la capside de SPP1. (A) Une procapside I se forme par auto-assemblage de gp11, gp13 et gp6, gp11 sort de la structure et la procapside I mature en procapside II. La terminase formée de gp1 et gp2 reconnait et clive un concatémère de génome qui est encapsidé dans la procapside II. Celle-ci mature en capside. À la fin de l'encapsidation gp15 et gp16 se fixent au niveau du sommet formé par la portal, et la queue qui suit sa propre voie d'assemblage, est liée à la capside. (B) Structure complète de la procapside I et II et de la capside mature. Une vue en coupe est réalisée afin de montrer les structures formées par gp11 ou l'ADN à l'intérieur des procapsides et de la capside. Adaptée de Ignatiou et al, 2019.

Tableau 1. Principales protéines impliquées dans la voie d'assemblage des particules virales de SPP1

Protéine	Fonction
gp1	Petite sous-unité de la terminase (TerS), reconnaît la séquence *pac* sur le génome de phage.
gp2	Grande sous-unité de la terminase (TerL), reconnaît la structure formée par gp1 en interaction avec *pac*, clive l'ADN viral et sert de moteur à l'encapsidation
gp6	Protéine portal, forme un pore dans la procapside du phage au travers duquel l'ADN est encapsidé.
gp11	Protéine d'échafaudage nécessaire à l'assemblage de la protéine majeure de capside.

gp12	Protéine de décoration,
gp13	Protéine majeure de capside.
gp15	Connecteur qui interagit avec gp6 en fin d'encapsidation
gp16	Protéine qui interagit avec gp15 et empêche la fuite de l'ADN viral hors de la capside

I.V.IV.IV. Reconnaissance de l'ADN viral

Chez SPP1, les protéines appelées gp1 et gp2 codées par les gènes *1* et *2* du phage forment le complexe terminase. Gp1, la petite sous-unité de la terminase (TerS), assure la reconnaissance de la séquence *pac*. La séquence *pac* de SPP1 est divisée en 2 sites, *pac*L et *pac*R qui ont été définis lors d'expériences d'empreintes à la DNAse avec gp1 *in-vitro* (figure 15 ; Chai et al, 1995). La structure formée par gp1 en interaction avec l'ADN est ensuite reconnue par gp2, la grande sous unité de la terminase (TerL), qui clive l'ADN viral. Le clivage, malgré sa précision quasi-nucléotidique, ne dépend pas d'une séquence particulière, mais seulement de contraintes structurales. Le complexe nucléoprotéique interagit ensuite avec une procapside virale au niveau de la protéine portal, gp6 chez SPP1, pour former le moteur de l'encapsidation. L'activité ATPase de gp2 apporte l'énergie nécessaire à la translocation du génome viral à l'intérieur de la procapside.

I.V.IV.V. Organisation de la séquence *pac*

La première étape nécessaire à l'encapsidation du génome de SPP1 est la reconnaissance du concatémère de génomes par la petite terminase gp1 au niveau d'une séquence spécifique, la séquence *pac* (figures 7.B, 16). D'une taille de 268 pb, la séquence *pac* se subdivise en deux régions, *pac*L et *pac*R. Le site *pac*L, constitué de 99 pb, est localisé en amont du gène *1* (codant lui-même pour la petite terminase). On y retrouve le terminateur de transcription en aval du gène *53*, les promoteurs de l'opéron des gènes *1* à *7* (opéron 1) et le Shine-Dalgarno du gène *1*. La séquence *pac*R, se situe dans la région codante du gène *1*, couvrant une région de 30 pb (figure 16.A). Les sites *pac*L et *pac*R ont initialement été identifiés par des expériences de protection à la DNAse qui ont démontré qu'il s'agissait des séquences d'interaction avec gp1. Gp1 interagirait notamment avec des segments de séquences répétées de 7 bp, dénommés *a1* et *a2* pour *pac*L et *c1* et *c2* pour *pac*R. La structure que forme gp1 liée à *pac*L et *pac*R est reconnue par gp2 qui clive l'ADN légèrement en amont de *pac*R. La

coupure médiée par gp2 ne requiert pas de séquence spécifique et ne dépend que de la structure formée par gp1 complexée à ADN (Djacem et al, 2017). Initialement un site *pac*C avait été identifié comme étant la séquence spécifique permettant la fixation de gp2. Cette fixation ne dépendant finalement pas d'une séquence spécifique, nous avons fait le choix de ne plus la considérer (Djacem et al, 2017). Cependant il est important de prendre en compte qu'il existe deux séquences répétées sur ce site initialement identifiées comme étant les boîtes *b1* et *b2* (Chai et al, 1992). Gp2 est une enzyme bifonctionnelle. Elle possède une activité endonucléase qui effectue d'une part une coupure double brin de l'ADN et d'autre part, un nouveau clivage de la molécule d'ADN à la fin de chaque cycle d'encapsidation. Gp2 dispose également d'une activité ATPase qui est nécessaire à la translocation du génome phagique à l'intérieur de la capside. Gp1 et gp2 sont nécessaires à l'encapsidation du génome dans la procapside et donc à la formation de particules virales viables.

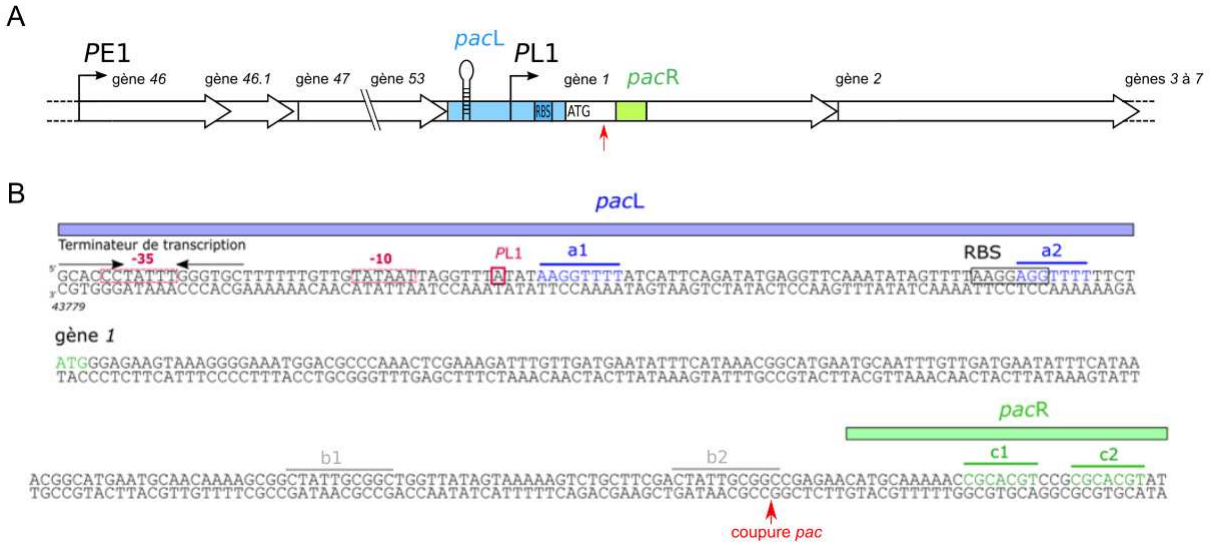


Figure 16. Contexte génomique de la séquence *pac*. (A) Celle-ci est composée de deux sites, *pac*L et *pac*R, qui sont respectivement en amont et sur le gène *1* (Chai et al, 1995). La séquence *pac*L contient les éléments de régulation de la transcription: le terminateur de l'opéron précoce en amont et le promoteur de l'opéron tardif en aval (Chai et al, 1992). Les flèches au-dessous de la carte génétique montrent la position de coupure sur *pac*. ATG indique le codon d'initiation du gène *1* et RBS le site de liaison au ribosome. (B) Sur la séquence, les boîtes -35 et -10 de *P*L1 sont représentées par des rectangles pointillés rouges et le site de démarrage par un rectangle gras rouge. Les boîtes *a1*, *a2*, *b1, b2, c1* et *c2* identifiées lors d'expériences d'empreintes à la DNAse sont identifiées en bleu et en vert. Le gène *1* commence au niveau de la deuxième ligne.

I.V.IV.VI. La petite sous-unité de la terminase gp1

Gp1 est la petite sous-unité de la terminase (TerS), elle assure la reconnaissance spécifique de la séquence *pac* et le recrutement de gp2 la grande sous-unité de la terminase. Chez SPP1 cette protéine est essentielle à l'encapsidation de l'ADN *in vivo* et *in vitro* (Oliveira et al, 2005). Un phage ne produisant pas gp1 ne génère pas de particules virales viables au cours de l'infection de l'hôte. Cependant, chez d'autres phages comme T4 et T3 cette protéine n'est pas essentielle à l'encapsidation *in vitro*, indiquant que, chez ces phages, la reconnaissance implique des mécanismes différents (Hamada et al, 1986). Bien que la structure de la petite terminase de SPP1 n'ait pas été résolue, la structure de la petite terminase du phage SF6 permet d'avoir une idée assez claire de ce à quoi devrait ressembler gp1. En effet, la petite sous-unité de la terminase de SF6 partage 86 % d'identité avec celle de SPP1. Elles ont des tailles très similaires avec 147 résidus pour celle de SPP1 et 145 pour celle de SF6 (Buttner et al, 2012). Gp1 possède un domaine de liaison à l'ADN N-terminal de 53 résidus, suivi d'un domaine d'oligomérisation de 67 résidus, et enfin d'un domaine C-terminal impliqué dans la formation d'un tonneau bêta de 25 résidus (figure 17). Tout comme celle de SF6, la petite sous-unité de la terminase de SPP1 forme des nonamères. L'oligomère présente un corps central formé par les domaines d'oligomérisation à sa base puis d'un tonneau bêta. Le domaine d'oligomérisation est formé d'une partie supérieure composée d'hélices alpha et d'une partie inférieure constitué de feuillets bêta. Cette structure, le domaine d'oligomérisation et le tonneau beta, est creuse en son centre. Son diamètre interne varie entre 11 et 29 Angströms. Les domaines de liaison à l'ADN forment des structures de type hélice-boucle-hélice qui sont reliées au corps central par une région flexible (Oliveira et al, 2013, Buttner et al, 2012). Certains modèles d'interaction de gp1 avec l'ADN suggéraient que l'acide nucléique passait au centre de la structure. Cependant, le diamètre du canal formé par gp1 semble légèrement trop petit. Le modèle privilégié, plus en phase avec les résultats obtenus par les expériences d'empreinte à la DNAse, suggèrent un enroulement de l'ADN au niveau du corps central garanti par une interaction avec les domaines de liaison à l'ADN. La présence de résidus chargés positivement sur la face externe du domaine d'oligomérisation supportent cette hypothèse (Oliveira et al, 2013).

La structure de gp1 en interaction avec l'ADN n'ayant jamais été déterminée, il n'y a toujours pas de consensus sur la manière dont gp1 interagirait avec l'ADN.

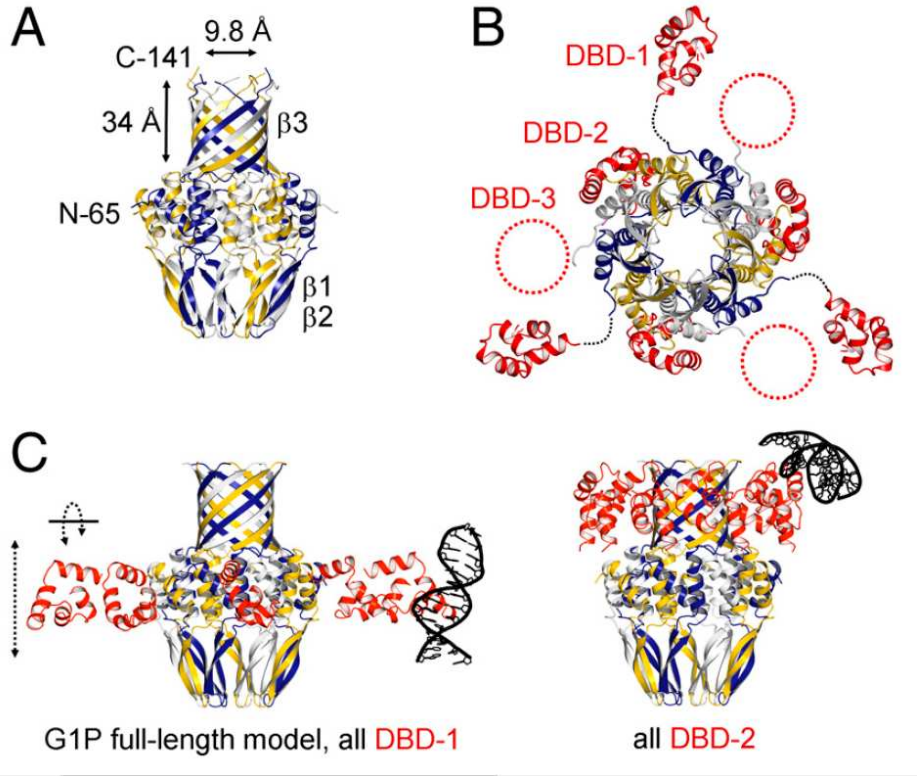


Figure 17. Structure cristallographique d'un nonamère de gp1 du bactériophage SF6. (A) Vue de côté de la structure du corps central sans les domaines de liaison à l'ADN. (B) Structure du nonamère vue du dessus. Les domaines de liaison à l'ADN occupent trois positions différentes (DBD-1, DBD-2 et DBD-3), DBD-3 n'est pas visible dans la structure (cercles rouges en pointillé. (C) Vue de côté de la structure d'un nonamère, modèles avec différents positionnements des domaines de liaison à l'ADN par rapport au corps central et à la double hélice d'ADN (en noir). Figure reproduite de Buttner et al, 2012.

I.V.III.VII. La grande sous-unité de la terminase gp2

La grande sous-unité de la terminase est une enzyme bi-fonctionnelle qui interagit avec gp1 liée à l'ADN. Contrairement à gp1, cette protéine est monomérique en solution. Elle possède 2 domaines séparés par une région flexible (figure 18). Au total, cette protéine compte 422 acides aminés. Chaque domaine assure une fonction différente, le domaine situé au niveau de la partie N-terminale de la protéine possède un site de liaison de l'ATP et assure la translocation de l'ADN viral à l'intérieur de la procapside en hydrolysant cette molécule. Le domaine occupant la partie C-terminale de la protéine est un domaine endonucléase qui assure quant à lui le clivage de l'ADN viral en début et en fin d'encapsidation. Ces deux activités, endonucléase et ATPase, interviennent séquentiellement à des moments différents de l'encapsidation, il y a donc un mécanisme, encore inconnu, qui doit permettre de favoriser une activité et de réprimer l'autre en fonction de la situation (Gual et al, 2000). Chez SPP1, seule

la structure du domaine nucléase a pu être déterminée (figure 18) par cristallographie (Smits et al, 2009). Sa structure est similaire à celle de la grande sous-unité de la terminase d'autres phages comme T4 et P22. HSV1 possède aussi une grande sous-unité de la terminase proche de celle de SPP1. Ceci suggère que les mécanismes impliqués dans le clivage du génome viral et dans le changement d'activité sont conservés chez ces virus et peut-être plus largement chez les bactériophages caudés et les *Herpesvirus*. Cependant, les *Herpesvirus* possèdent des terminases à 3 sous-unités (voir partie I.III.IV), on peut donc s'attendre à ce que les mécanismes de régulation des activités de la grande sous-unité de la terminase soient légèrement différents. Chez les bactériophages caudés, ces mécanismes de régulation restent inconnus, mais semblent impliquer à la fois la petite sous-unité de la terminase et la protéine portal. La protéine gp1 interviendrait notamment en inhibant l'activité nucléase de gp2 au profit de son activité ATPase, favorisant ainsi le changement d'une activité à l'autre (Camacho et al, 2003). Mais, dans ce cas, il faudrait que gp2 ait le temps de cliver l'ADN avant que son activité nucléase soit réprimée. Quant à la portal, on pourra supposer qu'un changement de la structure globale de la capside lors de l'encapsidation entraîne également un changement de conformation de la portal qui à son tour influe sur le complexe terminase en stimulant l'activité nucléase de gp2.

La structure du domaine nucléase de gp2 a été déterminée par cristallographie. Il se compose de 7 brins beta organisé en un feuillet entouré de chaque côté de 2 paires d'hélices alpha (Oliveira et al, 2013;Butner et al, 2012).

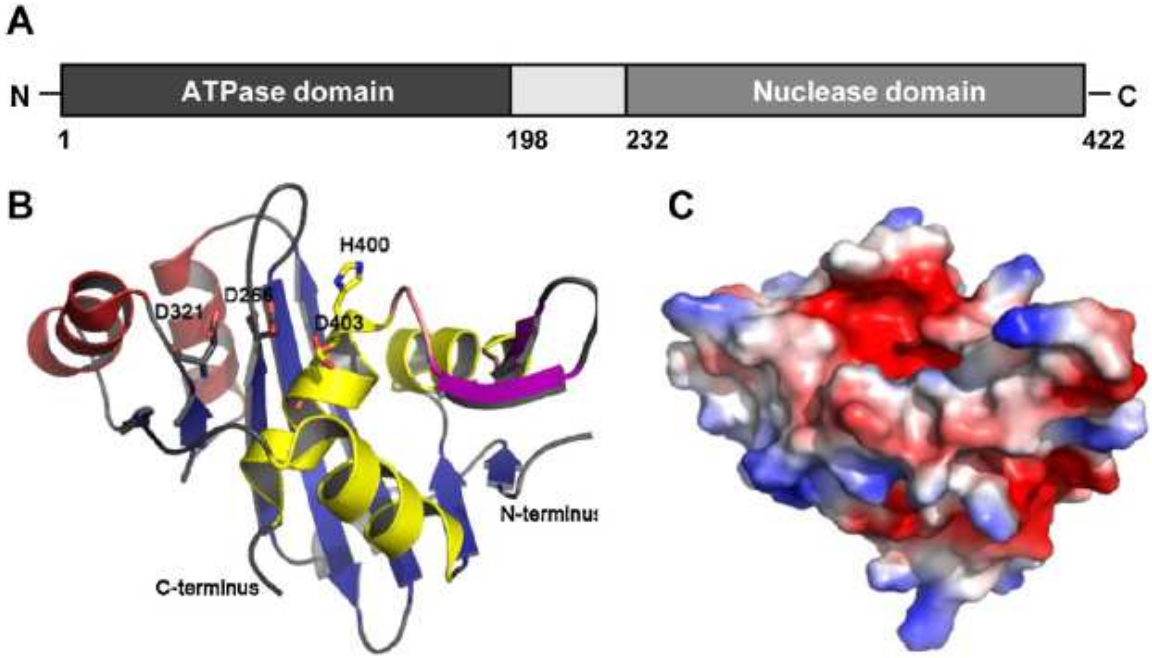


Figure 18. Structure de gp2. (A) Organisation de la protéine gp2 avec son domaine ATPase N-terminal et son domaine nucléase C-terminal. (B) Structure cristallographique du domaine nucléase de gp2. (C) Surface du domaine nucléase de gp2 avec un code couleur pour les acides aminés chargés : bleu pour les charges positives et rouge pour les négatives.

I.V.III.VIII. La protéine portal

La protéine portal est un oligomère de 12 sous-unités dans la capside virale en forme d'anneau créant un pore dans la capside de SPP1 se substituant à un des pentamères de la protéine majeure de capside. La structure de cette protéine, présentée figure 19, est assez conservée chez les bactériophages caudés et les herpèsvirus. Son rôle reste le même, permettre l'entrée de l'ADN viral à l'intérieur de la capside lors de l'encapsidation grâce au pore qu'elle forme dans la capside. Cette protéine jouerait aussi un rôle dans la régulation de l'activité des terminases. La figure 19 présente les portals des bactériophages phi29 et SPP1, même si leur forme est légèrement différente leur organisation globale reste la même. Notons qu'il est possible d'isoler des protéines portal avec un nombre de sous-unités variables lorsqu'elles sont produites hors du contexte de l'assemblage de la particule virale (13 sous-unités dans la figure 19 par exemple ; Rao et al, 2015).

Chez SPP1, chaque sous-unité de la portal possède 4 domaines. Un premier domaine, situé à l'extrémité la plus extérieure de la capside est impliqué dans les interactions entre les sous-unités et sert à stabiliser l'oligomère (trombone). De par sa localisation, c'est aussi une région qui interagit avec la terminase. Un deuxième domaine composé d'hélices alpha (tige) relie le

premier domaine au troisième et quatrième domaines en forme de couronne. Le troisième domaine, est une région globulaire qui va donner à l'oligomère sa forme caractéristique d'anneau. Notons que le dernier domaine (couronne) est quasiment absent chez phi29 (Rao et al, 2015).

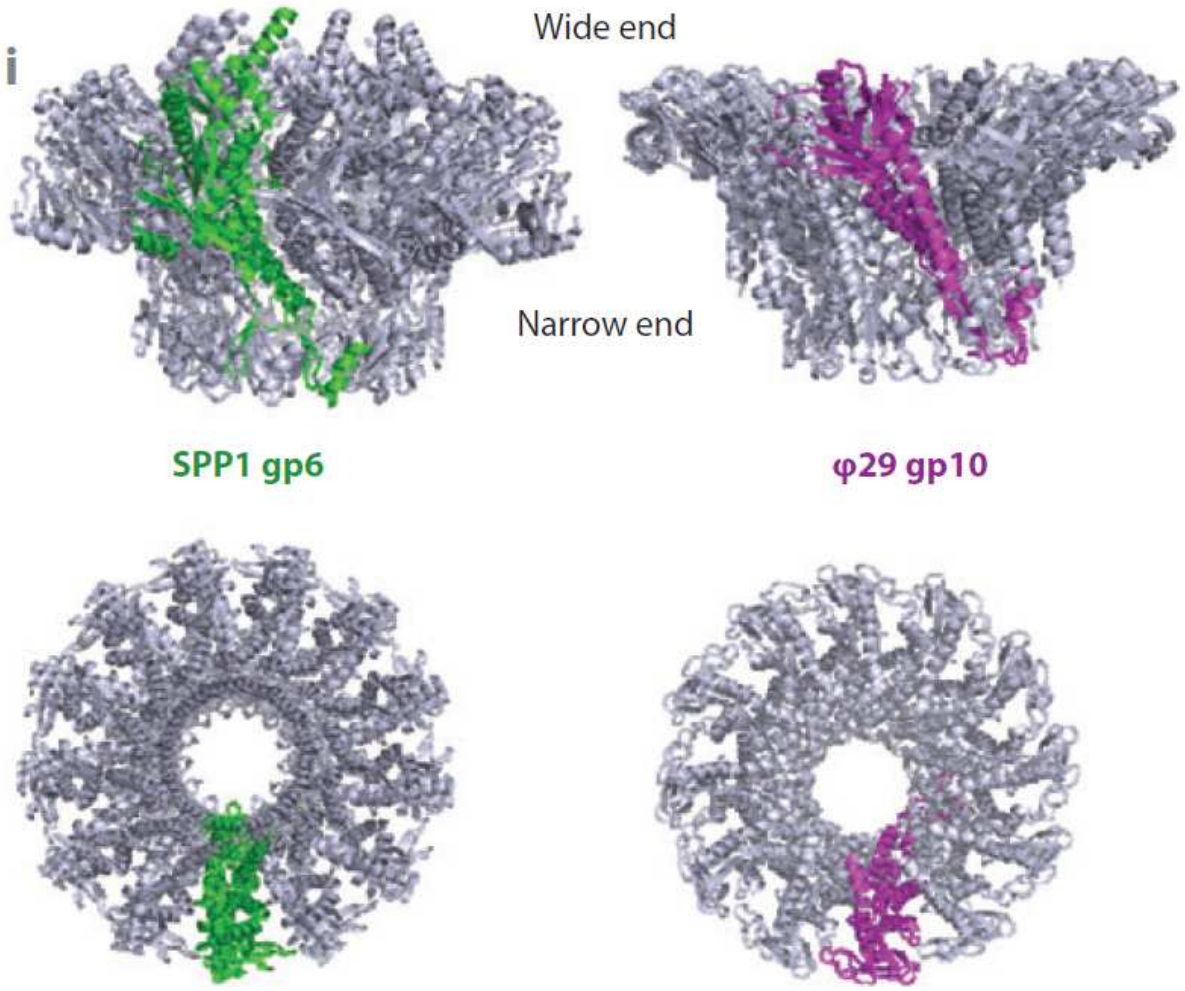


Figure 19. Structure de la portal. Comparaison des structures cristallographiques de la portal gp6 de SPP1 (à gauche) et de la gp10 de phi29 (à droite), vues de côté et du dessous. Une sous-unité est présentée en couleur. Figure reproduite de Rao et al, 2008.

I.V.III.IX. Encapsidation de l'ADN viral

L'étape d'encapsidation de l'ADN viral est relativement mal caractérisée par rapport aux étapes de reconnaissance de l'ADN viral, beaucoup de questions restent sans réponse. Nous avons vu que gp1 reconnaît *pac,* et gp2 la structure formée par gp1 en interaction avec l'ADN de manière séquence indépendante. La manière dont gp2 interagit avec gp1 reste floue et la structure d'un complexe gp1/gp2 n'a jamais pu être résolue. Gp1 semble protéger l'ADN viral au niveau de *pac*R alors que *pac*L est dégradé (Chai et al, 1995). Comme vu

précédemment, l'activité nucléase de gp2 est réprimée par gp1 puis gp2 interagit avec gp6. Là encore, l'interaction gp2/gp6 est très mal caractérisée. Il semblerait que gp2 interagit avec gp6 au niveau de son domaine C-terminal. Il a été démontré *in vitro* que gp2 avait une forte affinité pour gp6 et que celle-ci augmentait quand gp6 était insérée dans une procapside. Il semble évident que l'énergie nécessaire à l'encapsidation est fournie par gp2 qui hydrolyse l'ATP, en revanche, il est plus ardu de s'avancer sur les mécanismes précis mis en œuvre au cours de la translocation de l'ADN, notamment pour ce qui concerne l'évolution de l'interaction gp1/gp2/gp6. Certaines mutations ponctuelles d'acides aminés sur gp6 conduisent à l'encapsidation de molécules d'ADN plus courte que la normale (Tavares et al, 1992), ce qui suggère que gp6 agit comme senseur de l'encapsidation. Il semblerait que gp6 stimule l'activité nucléase en fin d'encapsidation à la suite d'un changement de conformation(Fujisawa & Morita, 1997). On pourra supposer que des mutations ciblées induisent ce changement de conformation plus tôt au cours de l'encapsidation. Cet aspect est important car ces gp6 mutées vont être utilisées pour étudier précisément les coupures induites par gp2 dans la partie résultats.

Plusieurs modèles ont été proposés pour expliquer comment s'effectuerait la translocation de l'ADN. Dans un premier modèle, gp2 ferait tourner la double hélice d'ADN sur elle-même ce qui entraînerait la rotation synchronisée de l'oligomère de gp6 sur lui-même permettant ainsi de faciliter l'entrée de l'ADN (Cuervo et al, 2007). Cependant, il est difficile d'imaginer comment la portal pourrait tourner sur elle-même alors qu'elle est en interaction avec la procapside. Un autre modèle plus pertinent suggère que chaque sous-unité de la portal est légèrement mobile. Elles bougeraient individuellement en formant une structure qui épouse la forme de la double hélice d'ADN lors de son passage dans le canal. Ce mouvement serait induit par l'hydrolyse de l'ATP de gp2 (Fujisawa & Morita, 1997 ; Oliveira et al, 2013). À ce jour aucun de ces modèles n'a pu être démontré expérimentalement.

Un réarrangement structural de la procapside permet de créer l'espace nécessaire à l'entrée de l'ADN viral, elle mature ensuite en capside (voir figure 14). Une fois la capside remplie, un signal, sûrement un changement de conformation de gp6, induit un changement d'activité de gp2 qui clive l'ADN viral. Il s'agit de la coupure par tête pleine qui a lieu à la fin de chaque encapsidation. Cette coupure est relativement imprécise puisqu'elle admet une incertitude de 2 kpb. On a longtemps pensé que cette coupure était non-spécifique, qu'elle ne dépendait pas

de séquences particulières. Cette assertion sera remise en cause dans la suite du manuscrit (voir partie résultats).

I.VI. Problématique de la thèse

Chez le bactériophage SPP1, la spécificité de la reconnaissance du génome viral est définie par l'interaction entre *pac* et gp1. La compréhension de la nature précise de cette interaction protéine-ADN est nécessaire pour éclairer la manière dont le virus discrimine son ADN de celui de l'hôte et aussi la nature des erreurs qui conduisent au transfert horizontal de gènes bactériens par transduction. Nous nous sommes donc focalisés sur l'étude de l'interaction entre *pac*, gp1 et gp2.

Des études précédentes *in vitro* ont conduit à un premier modèle de l'interaction dans lequel deux nonamères de gp1 se fixeraient respectivement sur *pac*L et *pac*R (Zecchi et al, 2012 ; Oliveira et al, 2013). Le profil de protection à la DNAse montre clairement des zones protégées espacées régulièrement sur *pac*L suggérant un enroulement de l'ADN par gp1. Dans ce système *in vitro*, il a été possible de remplacer une partie de *pac*L, mais pas sa totalité, par une autre séquence et maintenir le même espacement de protection à la DNAse dans la nouvelle séquence (Chai et al, 1995). Cette observation suggère qu'après une interaction spécifique avec une région de *pac*L, gp1 s'étale conduisant à l'enroulement de séquences adjacentes. Des approches de mutagénèse de *pac*L sur un plasmide où une séquence *pac* a été clonée avec les gènes de gp1 et gp2 semblent converger vers les mêmes conclusions. Les études sur plasmide, toujours par des approches de mutagénèse, ont aussi permis de déterminer qu'une séquence poly-A sur *pac*R était importante pour la reconnaissance de *pac* par gp1 (Djacem et al, 2017).

L'objectif de mon projet de thèse est de caractériser l'interaction terminase-*pac* dans le contexte de l'infection virale. Un premier volet a été l'identification de la séquence *pac* minimale nécessaire à la reconnaissance et l'encapsidation du génome du phage et ainsi d'apporter des éléments de réponse sur la spécificité de l'interaction *pac*/gp1. Le site *pac*L se situe juste en amont du gène *1*. Il possède les éléments nécessaires à la transcription de l'opéron codant les terminases et à la traduction de gp1. Ceci laisse penser à un rôle important de *pac*L dans la régulation de l'expression des gènes codant la terminase. Le rôle de *pac*L pourrait donc être double: réguler la quantité de terminase produite au cours de l'infection et/ou servir de site de reconnaissance pour gp1. La deuxième région de *pac* reconnue par gp1

est *pac*R qui est localisée sur le gène *1*. Elle n'inclut probablement pas d'élément de régulation de la transcription ou de la traduction. Le site de coupure de gp2 est très proche de cette séquence, ce qui laisse penser que *pac*R est uniquement impliqué dans la reconnaissance du génome viral. Nous tenterons de comprendre quels éléments de séquence sont nécessaires pour l'interaction de gp1-*pac*R.

Nous nous sommes d'abord focalisés sur une approche de mutagénèse de la séquence *pac* du génome du phage pour identifier les séquences de *pac* requises pour l'interaction avec gp1, la précision de clivage sur *pac* et la régulation de l'expression des gènes codant la terminase de SPP1. Ce travail sera poursuivi par l'analyse phénotypique et biochimique des mutants affectant ces activités. La nouveauté de ce travail est l'étude dans un contexte infectieux de l'interaction complexe entre la terminase et *pac* qui permet la reconnaissance et coupure d'un nombre réduit de séquences *pac* ainsi que la régulation de l'expression génique. Ces fonctions ont fort probablement un rôle coordonné essentiel pour l'efficacité du mécanisme d'encapsidation par tête pleine dans les phages caudés. Mes recherches apportent aussi des éléments permettant de mieux comprendre les mécanismes impliqués dans les erreurs de reconnaissance de l'ADN de SPP1 qui sont à l'origine de la transduction de gènes bactériens.

Matériel et méthodes

II. Matériel et méthodes

II.I. Phages, souches bactériennes et plasmides

II.I.I. Souches bactériennes

La souche de *B. subtilis* YB886, est semblable à la souche de référence AL009126.3, à l'exception que l'YB886 ne possède plus aucun prophage, ceci pour éviter leur induction lors des expériences d'infections avec SPP1. Elle est utilisée dans la majorité des expériences comme souche de référence de *B. subtilis*.

La souche de *B. subtilis* HA101B code un ARN de transfert suppresseur capable d'ajouter une lysine à une protéine en cours de synthèse lorsque le ribosome rencontre un codon TAA qui est habituellement reconnu comme un codon STOP. Cette souche est notamment utilisée pour les expériences de complémentation.

La souche *E. coli* DH5alpha est très fréquemment utilisée comme vecteur de clonage. Elle permet d'obtenir des transformants facilement.

La souche *E. coli* BL21 est optimisée pour la production de protéines, elle sera utilisée pour produire gp1.

II.I.II. Phages

Au cours de ce travail, de nombreux phages aux caractéristiques différentes ont été utilisés ou construits. Le tableau 2 présente la liste de ces phages avec une description de leur génotype et les travaux dans lesquels ils ont été décrits.

Tableau 2. Liste des bactériophages utilisés pour ce travail.

Phage	**Description**
SPP1 sauvage	Phage de référence (Riva et al, 1968)
SPP1*pac*L-99	Contrôle del41413-43779
SPP1*pac*L-93	Délétion dans *pac*L del41413-43785
SPP1*pac*L-76	Délétion dans *pac*L del41413-43802
SPP1*pac*L-68	Délétion dans *pac*L del41413-43810
SPP1*pac*L-58	Délétion dans *pac*L del41413-43820
SPP1*pac*L-54	Délétion dans *pac*L del41413-43824
SPP1*pac*L-36	Délétion dans *pac*L del41413-43842

SPP1*pac*L-27	Délétion dans *pac*L del41413-43851
SPP1*pac*L-15	Délétion dans *pac*L del41413-43863
SPP1*pac*L-99.1	dégénération de séquence dans *pac*L entre le promoteur et le RBS
SPP1*pac*L-99.2	dégénération de séquence dans *pac*L entre le promoteur et l'ATG
SPP1*pac*L-36REV1	Délétion dans *pac*L del41413-43842 et mutation compensatrice 40563A>G
SPP1*pac*L-27REV1	Délétion dans *pac*L del41413-43851 et mutation compensatrice 40563A>G
SPP1*pac*L-15REV1	Délétion dans *pac*L del41413-43863 et mutation compensatrice 40538T>C
SPP1*pac*L-15REV2	Délétion dans *pac*L del41413-43863 et mutations compensatrices 40499insT ; 23270G>T
SPP1*pac*L-15REV3	Délétion dans *pac*L del41413-43863 et mutations compensatrices 40495T>C ; 40566A>G
SPP1*pac*L-15REV4	Délétion dans *pac*L del41413-43863 et mutation compensatrice 40566A>G
SPP1*pac*L-0REV	Délétion dans *pac*L del41413-43863 et mutation 40538T>C Dégénération de séquence entre le RBS et l'ATG
SPP1*pac*R-0	Délétion del41413-43779 Dégénération du Poly-A et des boîtes c1 et c2 de *pac*R
SPP1*pac*R-0REV1	Délétion del41413-43779 Dégénération du Poly-A et des boîtes c1 et c2 de *pac*R ; mutation compensatrice 17G>A
SPP1*pac*R-0REV2	Délétion del41413-43779 Dégénération du Poly-A et des boîtes c1 et c2 de *pac*R ; mutation compensatrice 147G>A gp1:G94R
SPP1*pac*R-0REV3	Délétion del41413-43779 Dégénération du Poly-A et des boîtes c1 et c2 de *pac*R ; mutation compensatrice 165G>A gp1:E100K
SPP1gp1:G94R	Mutation 147G>A gp1:G94R
SPP1gp1:E100K	Mutation 165G>A gp1:E100K
SPP1*delX110*	Phage issu de SPP1 sauvage avec la délétion delX (Tavares et al, 1992)
SPP1*sus115delX110*	Phage SPP1*delX110* avec un codon STOP prématuré dans le gène *6* (Tavares et al, 1992)
SPP1*sus70sus115*	Phage SPP1 avec deux codons STOP prématurés d

	ans les gènes *1* et *6* (Tavares et al, 1992)

II.I.III. Plasmides

Le tableau suivant récapitule l'ensemble des plasmides utilisés pour les expériences avec des précisions pertinentes sur leurs caractéristiques.

Tableau 3. Plasmides utilisés

Plasmide	**Description**
pBT115	Code pour un gène de résistance à l'ampicilline et dispose de la séquence du gène *1* sous le contrôle d'un promoteur T7 (Oliveira et al, 2005).
pMS32	pBT115 avec une mutation sur le gène *1* qui conduit à la substitution gp1:G94R.
pMS33	pBT115 avec une mutation sur le gène *1* qui conduit à la substitution gp1:E100K.
pLysS	Plasmide contenant la lysozyme du phage T7 et un gène de résistance au chloramphénicol.
pBT233Neo	Plasmide contenant une résistance à la néomycine, mode de réplication thêta. Dérive de pBT233 (Ceglowski et al, 1993) avec le gène de résistance à la néomycine de pUB110 (Valero-Rello et al, 2017).
pUB110	Plasmide contenant une résistance à la néomycine, mode de réplication sigma (Leonhardt, 1990).
pUB110cop1	Dérivé de pUB110 présent en nombre de copies plus faible que pUB110.
pHP13	Plasmide dérivé de pTA1060 (Haima et al, 1987)
pMS34	Dérivé de pHP13 avec l'insertion d'un fragment du génome de SPP1 (gènes 1 à 6) et la mutation qui conduit à la substitution gp1:G94R
pMS35	Dérivé de pHP13 avec l'insertion d'un fragment du génome de SPP1 (gènes 1 à 6) et la mutation qui conduit à la substitution gp1:E100K
pSEA	Dérivé de pHP13 contenant le gène 6 avec la mutation *sizA*.

II.I.IV. Oligonucléotides utilisés

Cette rubrique regroupe l'ensemble des séquences des oligonucléotides utilisés pour ce travail.

Tableau 4. Oligonucléotides utilisés pour le séquençage Sanger

Nom	Séquence
Couple 1 NCS	TCCTTTGCTGATACTTCTACATGC
Couple 2 NCS	AGCTCCTCAATCCACATGCC
Eco9-1NCS	CGCATTATGCGTACCCCC
Eco12-2NCS	CTGCCCGCTAACACGTTCG
1057CS	AGGCTCACGCAGTGTTTA
3845CS	GATGAGCTACAGGAAGTC
PC1CS	GGAATTTTACCTACCTGACG
PC2CS	CCAAGGAAGGTTATTGTGG
PC3CS	GGAGTCCGCTATTGAGGC

Tableau 5. Oligonucléotides utilisés pour les clonages

Nom	Séquence
SPP1ΔpacL-99Fw	CGGACTAGTCCCCGATCGCGGAATTCCCGCACCCTATTTGGGTGCTTTTTTGTTG
SPP1ΔpacL-93Fw	CCCGATCGCGGAATTCGGGACTAGTCTAATATTTGGGTGCTTTTTTGTTGTATAATTAGG
SPP1ΔpacL-76Fw	CCCGATCGCGGAATTCGGGACTAGTCTAAGTTGTATAATTAGGTTTATATAAGG
SPP1ΔpacL-68Fw	CCCGATCGCGGAATTCGGGACTAGTCTAAATTAGGTTTATATAAGGTTTTATCATTC
SPP1ΔpacL-58Fw	CCCGATCGCGGAATTCGGGACTAGTCTAATATAAGGTTTTATCATTCAGATATGAG
SPP1ΔpacL-54Fw:	CCCGATCGCGGAATTCGGGACTAGTCTAAAGGTTTTATCATTCAGATATGAGG
SPP1ΔpacL-36Fw	CCCGATCGCGGAATTCGGGACTAGTCTAAATGAGGTTCAAATATAGTTTTAAGG
SPP1ΔpacL-27Fw	CCCGATCGCGGAATTCGGGACTAGTCTAAAAATATAGTTTTAAGGAGGTTTTTTCTATGG
SPP1ΔpacL-15Fw	CCCGATCGCGGAATTCGGGACTAGTCTAAAAGGAGGTTTTTTCTATGGGAGAAG
SalI-PstIgene6CS	TTCCGCGGCCGCTATGGCCGACGTCGACTGGTTCTGCAGGCCGAGTAGCGCATAAAGG
BamHI-pacRCS	CGGGATCCCGCGACTATTGCGGCCGAGAACATGCAAAAACCGCACGTCCGCGCACGTATCGAGG
G94RCS	GGAACAGGTGCTCATGAGAATTGGTAAAGGTGCGG
G94RNCS	CCGCACCTTTACCAATTCTCATGAGCACCTGTTCC
E100KCS	GGGAATTGGTAAAGGTGCGAAGACAAAAACGCATGTAG
E100KNCS	CTACATGCGTTTTTGTCTTCGCACCTTTACCAATTCCC

Tableau 6. Oligonucléotides utilisés pour les qRT-PCR

Nom	Séquence
SPP1-G1-F1	GACTATTGCGGCCGAGAACAT
SPP1-G1-R1	CGAGTGCAATGCGAGTCAAA
SPP1-G2-F1	CATCAACCCGCCGAAAAGA
SPP1-G2-R1	ATTGCGCCGTTTAACCTCCT
SPP1-G6-F1	CGGGXTGAAATACCTGTGGA
SPP1-G6-R1	CGGGCTGAAATACCTGTGGA
SPP1-G11-F1	ACGAGATAGCGGTGCGAAGA
SPP1-G11-R1	TCCTTGTACGCCTCCGTTTG
SPP1-G46-F1	GACCCAGAGGCTAACCCTTATCA
SPP1-G46-R1	GGCGTTCTCCATAACCCAATG
Bsera-F1	GTAGCGGCAACGATTGTGGT
Bsera-R1	GTAGACACGGGAGCCCAACA
BsdnaG-F1	GCCGGTTATGAAGCCACCTT
BsdnaG-R1	TAAATTTTTCCCCGCCGAAT
BssecA-F1	GGCGACGATTACGTTCCAAA
BssecA-R1	GGGATCGTGACAACCTGCAT
Bsgmk-F1	TCCCGGAAGGCCTGTTTATT
Bsgmk-R1	TCAGCTTTTGCGGCTTTCAT
BsGyrA-F1	GAATACGGCAGAACGGCAAA
BsGyrA-R1	TTCGTTTTGAAACCCCATGC

II.II. Mutagénèse de *pac*L

Un fragment du génome sauvage de SPP1 a été amplifié par PCR avec une amorce Eco9-1 NCS s'hybridant en aval du gène *3* et une autre complémentaire à la région *pac*L contenant un site *Spe*I et un codon STOP en 5' (voir séquences des amorces dans la partie Oligonucléotides). La région sur laquelle s'hybride cette amorce varie en fonction de la taille de la délétion considérée. Chez le mutant SPP1*pac*L-0, une partie de la séquence de l'amorce contient une séquence dégénérée entre le RBS et le codon d'initiation du gène *1*. La PCR a été réalisée dans 100 µl de réaction avec la Pfu polymérase. En fin de PCR, un fragment ne contenant que la région à conserver de *pac*L a été généré du fait que l'amorce utilisée ne permet pas l'amplification de tout le site *pac*L. Ce fragment possède également un codon STOP et un site *Spe*I à l'une de ces extrémités. Le fragment généré par PCR a été digéré par deux enzymes de restriction. SpeI clive le produit de PCR au niveau de l'une de ses extrémités et PspGI au niveau d'un site situé sur le gène *1*. Le génome de SPP1 sauvage,

contenant des sites *Spe*I et *Psp*GI uniques, a été digéré par les mêmes enzymes. Dans le cas de la construction du mutant SPP1*pac*L-0, c'est l'ADN génomique de SPP1*pac*L-15REV2 qui a été digéré (voir Résultats section III.I.). SpeI clive l'ADN dans le gène *47* et PspGI dans le gène *1*. Les fragments de PCR ont été purifiés avec le kit Macherey Nagel PCR Clean Up. Les échantillons d'ADN génomiques digérés ont été dilués dans du Tris 10 mM pH 7.5 jusqu'à un volume final de 200 µL puis extraits 2 fois au phénol et une fois au chloroforme. On a ensuite ajouté 20 µL de 3 M acétate de sodium pH 5.2 et 500 µL d'éthanol 100 %. Les échantillons ont été incubés 1 heure à -20°C et centrifugés 30 minutes à 14000 xg, à 4°C. Le surnageant a été éliminé, 500 µL d'éthanol 70 % ont été ajoutés et les échantillons ont été de nouveau centrifugés dans les mêmes conditions. L'éthanol 70 % a été retiré et le culot d'ADN séché 30 minutes à 37°C. Le culot d'ADN a ensuite été resuspendu dans 40 µL de 10 mM Tris pH 7.5. La T4 DNA ligase a été utilisée suivant les conditions du fournisseur d'enzyme (New England Biolab) pour lier les fragments de PCR à l'ADN de SPP1 sauvage ou du révertant SPP1pacL-15REV2 qui avaient été digérés. Après ligation, la séquence allant des gènes *47* à *53* ainsi qu'une partie de *pac*L ont été supprimées. Le codon STOP a servi à prévenir une traduction non désirée initiée à partir du gène *47*. Après ligation l'ADN a été digéré par SwaI afin de prévenir une religation du génome sauvage avec lui-même car seul SPP1 sauvage possède le site *Swa*I situé dans le gène *52*.

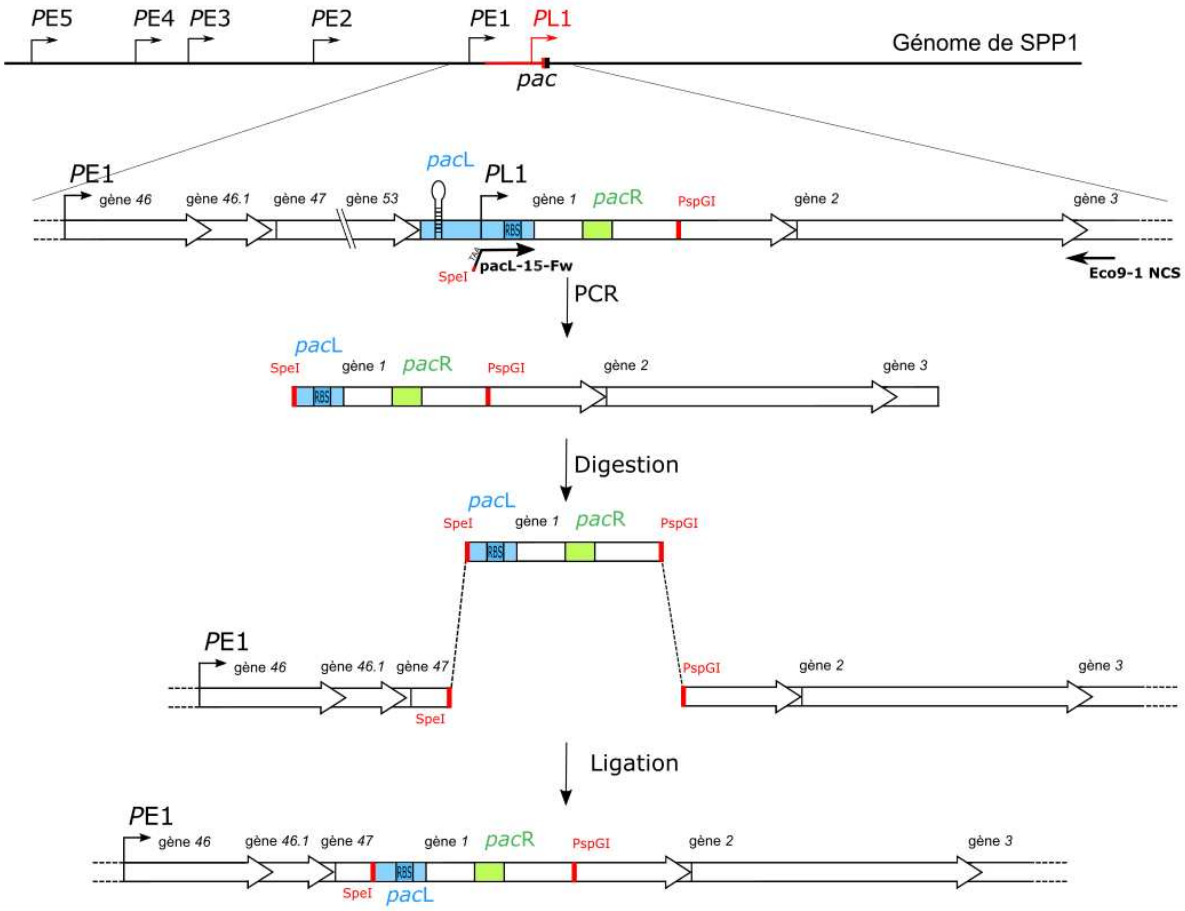


Figure 20. Construction des mutants avec une délétion dans *pac*L. Un fragment du génome de SPP1 est amplifié par PCR puis digéré avec SpeI (site ajouté avec un oligonucléotide) et PspGI. Le génome du phage sauvage est lui aussi digéré par les mêmes enzymes. Les deux digestions sont ensuite liguées entre-elles.

II.III. Mutagénèse de *pac*R

La technique de mutagénèse de *pac*R est en de nombreux points similaire à celle de *pac*L. L'amplification s'effectue cependant sur une matrice différente qui est le plasmide pBT163Mut16 qui contient une séquence *pac*R dégénérée (voir figure 21) ainsi que les gènes *1* à *6*. La séquence *pac*R est dégénérée au niveau des 2 boîtes c1 et c2 et d'une séquence poly-A. Les mutations ne changent pas la séquence en acides aminés de gp1. Les amorces Eco9-1NCS et Δ*pac*L-99 sont utilisées pour l'amplification. Le fragment de PCR ainsi que de l'ADN génomique de SPP1 sauvage sont digérés par SpeI et PspGI. Les fragments générés à l'issue des digestions sont purifiés puis ligués comme précédemment. *B. subtilis* YB886 est ensuite transfectée avec les ligations.

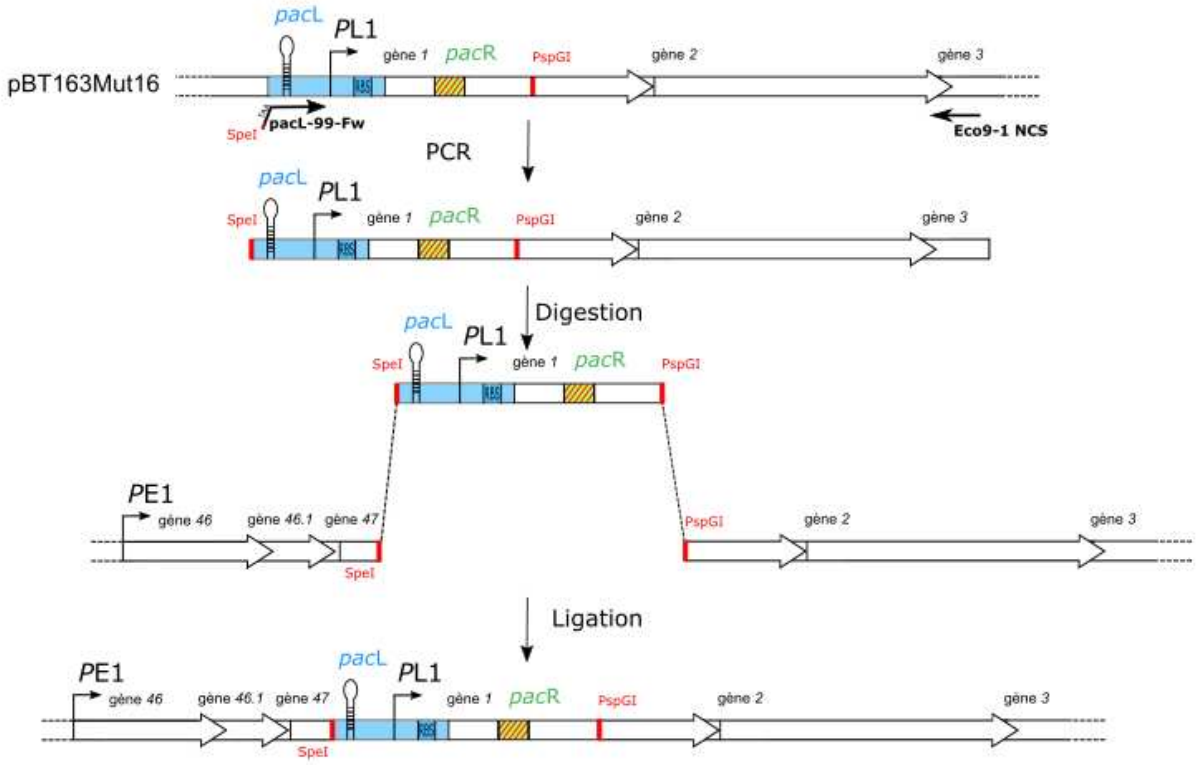


Figure 21. Construction des mutants avec une séquence *pac*R dégénérée. Un fragment du plasmide pBT163Mut16 est amplifié par PCR puis digéré avec SpeI (site ajouté avec un oligonucléotide) et PspGI. Le génome du phage sauvage est lui aussi digéré par les mêmes enzymes. Les deux digestions sont ensuite liguées entre-elles. La dégénération de séquence induite sur *pac*R est matérialisée par des hachures rouges.

II.IV. Transfections et transformations de *B. subtilis*

B. subtilis YB886 est transfectée par les génomes reconstitués à l'aide des méthodes décrites dans les sections II.II et II.III. Une culture de nuit à 30°C avec une agitation orbitale 150 rpm a été diluée au 50 ième dans du milieu SP1(2,6 g.l^{-1} K_2HPO_4, 1 g.l^{-1} KH_2PO_4, 2 g.l^{-1} $Na_3C_6H_5O_7$, 20 g.l^{-1} glucose, 1 g.l^{-1} hydrolysats de caséine, 1 g.l^{-1} L-Tryptophane, 22 g.l^{-1} Citrate d'ammonium ferrique, 40 g.l^{-1} Potassium Glutamate, 0,01 g.l^{-1} $MgSO_4$). Une courbe de croissance est réalisée. Elle permet de déterminer à quelle densité optique (600 nm) la transition entre la phase exponentielle et la phase stationnaire se produit. Au début de la transition, il faut compter une heure supplémentaire dans les mêmes conditions de culture avant de prélever 500 µl et de les mélanger avec l'ADN à transfecter ou transformer. Pour une transfection ou une transformation avec un plasmide, moins d'1 µg suffit. Les bactéries mélangées à l'ADN sont ensuite incubées 30 minutes à 37°C avec une agitation orbitale 150 rpm.

Ensuite, si les bactéries sont transfectées, elles sont mélangées avec 100 µl d'une culture de *B. subtilis* YB886(DO 600nm 0,8) dans 10 ml de LB agar 0.7% supplémenté avec 10 mM $CaCl_2$ sur boîte de Petri. Les boîtes ont été incubées toute la nuit à 37°C. Un contrôle négatif sans ADN et un contrôle positif avec de l'ADN sauvage ont été réalisés afin de vérifier le bon déroulement des transfections. Si la transfection a marché, des plages de lyse sont visibles après l'incubation sur la nuit. Il est alors possible de prélever des phages avec un cône de P200 et de le tremper dans 500 µl de TBT. Cette solution pourra alors servir à amplifier le mutant (voir section amplification des phages).

Dans le cas des transformations, après l'incubation 30 minutes à 37°C, les bactéries sont centrifugées 5 minutes à 13000xg, re-suspendues dans 100 µl de LB puis étalées sur boîte de LB agar 2 % contenant l'antibiotique correspondant à la résistance conférée par l'ADN incorporé. Les boîtes sont incubées sur la nuit à 37°C.

II.V. Création des mutants SPP1gp1:G94R et SPP1gp1:E100K

Les mutants SPP1gp1:G94R et SPP1gp1:E100K sont créés avec une approche différente de celle des mutants *pac*L et *pac*R. Un amplicon couvrant une région du génome de SPP1 allant de *pac*R aux 2/3 du gène *6* avec les mutations d'intérêt est généré par PCR. Pour cela, cette région est amplifiée à partir des ADN génomiques des révertants SPP1*pac*R-0REV2 et SPP1*pac*R-0REV3. Une amorce avec un site *Bam*HI et la séquence sauvage de *pac*R est utilisée ainsi qu'une amorce avec un site *Pst*I s'hybridant au milieu du gène *6*. L'amplicon généré possède ainsi la séquence *pac*R sauvage (celle de l'amorce) et les mutations d'intérêt. Les fragments de PCR ainsi que le plasmide pHP13 sont digérés par PstI et BamHI, puis la T4 DNA ligase est utilisée pour liguer les produits de digestion entre eux. Ensuite, *B. subtilis* YB886 est transformée (voir partie II.IV.) avec le nouveau plasmide généré par la ligation des fragments.

Afin d'obtenir un phage avec la séquence clonée dans le plasmide pBT165, YB886 contenant le plasmide (cultivée dans du LB+ 30 $\mu g.ml^{-1}$ Chloramphénicol + 10 $\mu g.ml^{-1}$ Erythromycine) est infectée par le phage SPP1*sus70sus115*. Ce phage possède deux mutations non-sens sur les gènes *1* et *6* qui induisent une fin précoce de la traduction. Les protéines gp1 et gp6 étant essentielles à la multiplication du phage, il ne peut pas se multiplier dans une souche comme YB886 dans laquelle il ne pourra pas produire gp1 et gp6, cette souche reconnaissant les codons mutés comme des codons STOP. En revanche, il existe une souche HA101B qui code

un ARNt qui reconnaît les codons STOP prématurés et qui, au lieu d'arrêter la traduction, intègre une lysine à la protéine en cours de synthèse.

Afin de générer les phages avec les mutations d'intérêt, une titration est réalisée où le phage SPP1*sus70sus115* est mis en contact avec l'YB886 possédant le plasmide pHP13 avec l'insert. 100 µl de culture à une DO 600 nm sont mélangés avec 100 µl de phages dilués dans 10 ml de LB Agar 0,7% supplémentés par 10 mM de $CaCl_2$. Cette méthode permet de sélectionner des phages issus d'un double événement de recombinaison entre la séquence de SPP1 sur le plasmide et le génome de SPP1*sus70sus115* qui supprime les 2 mutations présentes sur le génome de SPP1*sus70sus115*. Les phages issus de ce double événement de recombinaison produisent gp1 et gp6, et donc, peuvent se multiplier. SPP1*sus70sus115* est contre-sélectionné car il ne produit ni gp1 ni gp6. Lors de cette expérience, un contrôle avec la HA101B est réalisé pour avoir une idée du titre de SPP1*sus70sus115*. Un autre contrôle avec YB886 sans plasmide est réalisé pour voir à quelle fréquence apparaissent naturellement des réversions génétiques sur le génome de SPP1*sus70sus115*.

Les titrations sont incubées une nuit à 37°C. Le lendemain, si des plages de lyse sont apparues, elles sont prélevées avec un cône de P200 qui est ensuite trempé dans 500 µl de TBT.

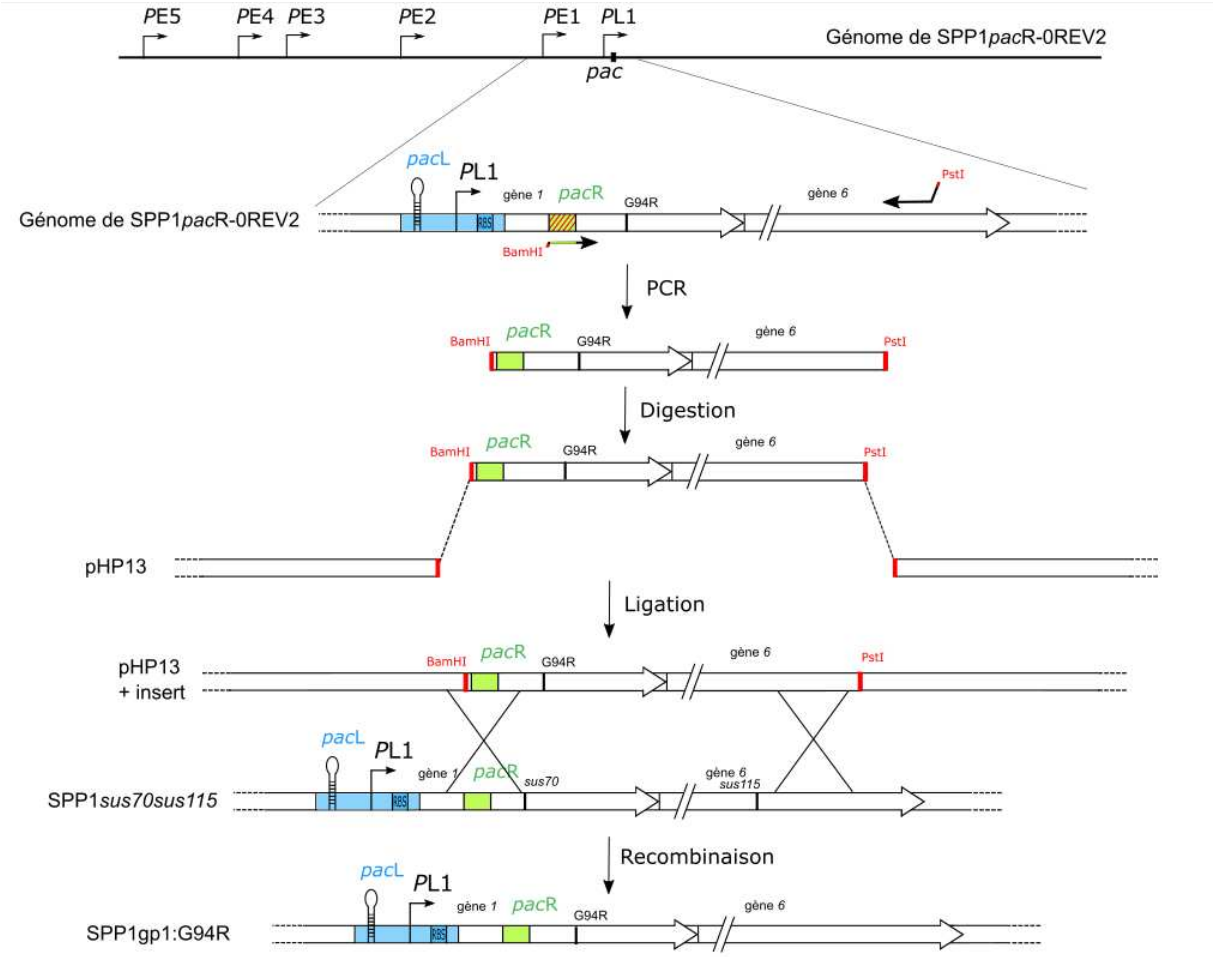


Figure 22. Construction de plasmides contenant un fragment des gènes *1* à *6* de SPP1 avec une mutation dans le gène *1*. Un fragment du génome de SPP1pacR-0REV2 ou SPP1pacR-0REV3 est amplifié par PCR puis digéré avec BamHI et PstI(sites ajoutés avec les oligonucléotides). Le plasmide pHP13 est lui aussi digéré par les mêmes enzymes. Les deux digestions sont ensuite liguées entre-elles. Une fois le plasmide construit, il est introduit dans *B. subtilis* YB886. Les cellules sont transfectées par le phage sauvage SPP1*sus70sus115*, les seuls phages viables dans ces conditions sont ceux qui ont recombiné avec le plasmide.

II.VI. Amplification des phages

II.VI.I. Plages de lyse isolées (SP, *single plaque*)

Les phages prélevés à partir des titrations (SP1 pour *single plaque* 1) re-suspendus dans du TBT sont de nouveau titrés. Ils sont dilués puis mélangés à une culture d'YB886 à une DO 600 nm de 0.9 dans 10 ml de LB+0.7% agar dans une boîte de Petri. Le facteur de dilution dépend du titre du SP1 qui varie selon les mutants. La figure 23 donne un aperçu du titre approximatif de phages à chaque étape d'amplification. Les boîtes sont incubées une nuit à 37°C, puis une plage de lyse isolée est prélevée avec un cône qui est ensuite trempé dans 500 µl de TBT (SP2 pour *single plaque* 2).

II.VI.II. Lysat sur boîte (PL, plate lysate)

Un volume de 100 µl de SP2 est mélangé à une culture d'YB886 à une DO 600 nm de 0.9 dans 10 ml de LB+0.7% agar dans une boîte de petri. Après solidification de l'agar, les boîtes sont incubées une nuit à 37°C. Le lendemain, les cellules sont complètement lysées. Un volume de 3 ml de LB est ajouté dans la boîte directement sur l'agar, puis la boîte est agitée. Après 2 heures à température ambiante, le LB est prélevé et filtré avec un filtre de 0.45 µm. Le lysat est conservé dans un tube à 4°C. Les lysats sont titrés pour déterminer leur concentration en phages.

II.VI.III. Grand lysat sur boîte (LPL large plate lysate)

Lorsque le titre de phage n'est pas suffisant pour réaliser des lysats en milieu liquide après un lysat sur boîte, un nouveau lysat sur boîte est effectué à partir du premier. Ce-dernier est réalisé dans une boîte de Petri carrée plus grande. On procède de la même manière que pour un PL, mais avec plus de phages et de milieu. 400 µl de PL sont mélangés à 30 ml de LB 10mM $CaCl_2$, avec 100 µl de culture d'YB886 dans du LB à une DO600nm de 0.9. L'ensemble est incubé une nuit à 37°C. Ensuite, 5 ml de LB sont déposés sur l'agar, la boîte est agitée et le LB récupéré après 2 heures à température ambiante. Le lysat est conservé à 4°C.

II.VI.IV. Lysats en milieu liquide (LL, liquid lysate)

Les lysats sur boîte sont utilisés pour réaliser les lysats en milieu liquide. Une culture de nuit de YB886 dans du LB est diluée au centième dans 3 ml de LB puis incubée à 37°C, 150 rpm. Lorsque sa densité optique 600 nm atteint 0.8, la culture est supplémentée avec 10 mM de $CaCl_2$ puis infectée par les phages issus du lysat sur boîte à une multiplicité d'infection de 3 cfu/pfu. Les cellules sont cultivées pendant deux heures à 37°C,150 rpm puis la culture est centrifugée à 4°C, 15 minutes, 10000 rpm dans un rotor JA-17 (Beckman). Le surnageant de phages est prélevé puis filtré avec un filtre de 0.45 µm. Le lysat (SSLL pour *Small Scale Liquid Lysate*) est conservé à 4°C. Les lysats sont titrés pour déterminer leur concentration en phages.

Les SSLL sont utilisés pour faire de nouveaux lysats plus concentrés en phages. Une culture de nuit d'YB886 dans du LB est diluée au centième dans 300 ml de LB à 37°C, 150 rpm. À

une OD600nm de 0.8, les cultures sont supplémentées par 10mM $CaCl_2$ puis infectées à une multiplicité d'infection de 3 pfu/cfu. Après 2 heures, les débris et les bactéries sont culottés par centrifugation à 4°C pendant 20 minutes à 8000 rpm dans un rotor JA-10. Le surnageant est prélevé et les phages culottés par centrifugation pendant 16 heures, à 4°C, 8000 rpm dans un rotor JA-10. Le surnageant est éliminé et le culot resuspendu dans 6 mL de TBT. Le lysat est conservé à 4°C.

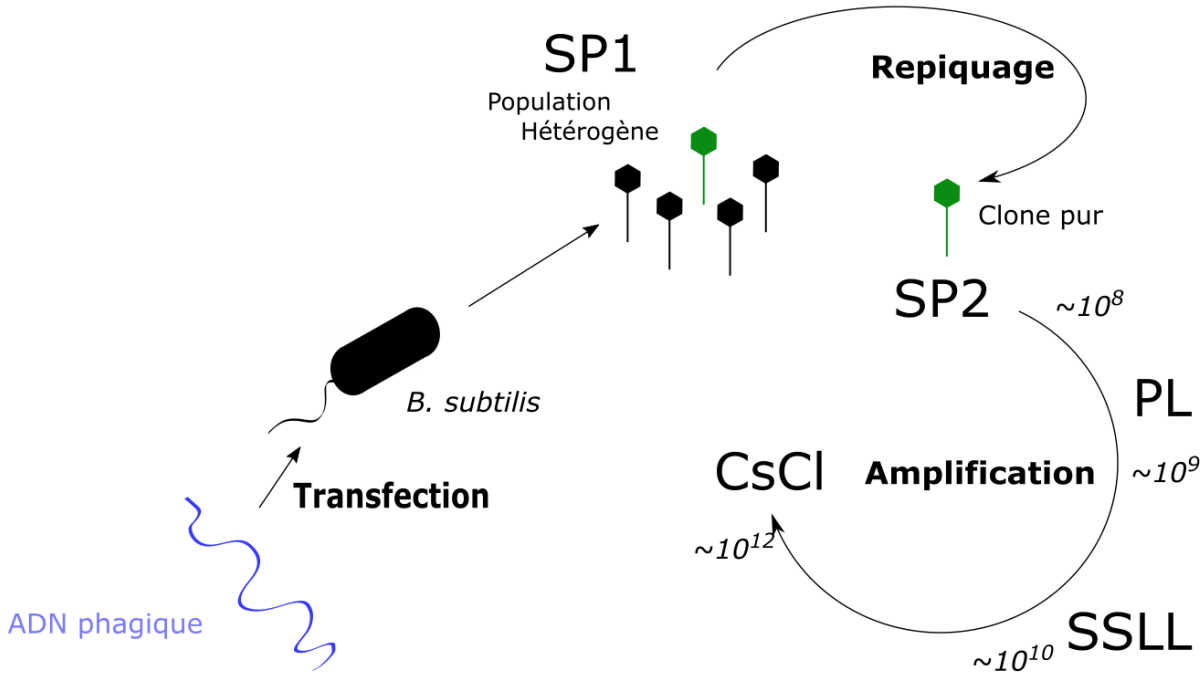


Figure 23. Méthode utilisée pour générer et amplifier les phages mutants. *B. subtilis* est transfectée avec la ligation d'ADN phagique avec le fragment de PCR contenant la délétion dans *pac*L. Plusieurs plages de lyse issues de la transfection (SP1) sont reprises et titrées pour obtenir des clones purs SP2. Les phages de chaque plage de lyse isolée SP2 sont amplifiés comme décrit ci-dessus. Le titre approximatif des phages à chaque étape d'amplification est présenté en italique en $pfu.mL^{-1}$.

II.VII. Purification de phages sur gradient de chlorure de césium

Le chlorure de césium est dilué dans du TBT afin d'atteindre des densités de 1.45, 1.50 et 1.70 respectivement. Dans chaque tube d'ultracentrifugation, sont séquentiellement déposés 750 µl de CsCl à une densité de 1.70, 2.5 ml de CsCl à 1.50, 3 ml de CsCl à 1.45 et 6 ml de LSLL concentré. L'ultracentrifugation a été réalisée dans un rotor SW41 à 20°C, 24000 rpm

pendant 5 heures. Les phages sont ensuite prélevés à l'aide d'une seringue au travers du tube entre les couches d'une densité de 1.50 et 1.45. Ils forment une bande bleue à l'interface entre ces couches de densités différentes. Les échantillons sont dialysés une nuit contre du TBT à 4°C puis conservés à la même température.

II.VIII. Extraction d'ADN génomique de SPP1

50 mM EDTA pH 8.0 sont ajoutés à 500 µL de particules virales issues des purifications sur gradient de chlorure de césium. Les échantillons sont incubés pendant 30 minutes à 55°C afin de détruire les capsides et de libérer l'ADN génomique du phage. Ils sont ensuite extraits 2 fois avec du phénol pour éliminer les protéines et une fois avec du chloroforme. L'ADN est ensuite dialysé une nuit contre 0.1 mM EDTA, 10 mM Tris HCl pH7.5, puis conservés à 4°C.

II.IX. Courbes de lyse

Une culture de nuit d'YB886 est diluée au centième dans 5 ml de LB puis incubée à 37°C, 150 rpm. À une densité optique (600nm) de 0.8, la culture est infectée par le phage d'intérêt à une multiplicité d'infection de 3. La densité optique est ensuite mesurée toutes les 15 minutes pendant 2 heures. Une condition contrôle sans infection est réalisée ainsi que deux autres contrôles avec une infection par SPP1 sauvage et une infection par SPP1*pac*L-99. Les valeurs de densité optique sont ensuite reportées dans un fichier informatique et traitées dans R pour réaliser des courbes avec ggplot2.

II.X. Séquençage des mutants, Illumina et Sanger

Le phage SPP1 sauvage ainsi qu'un clone des mutants SPP1*pac*L-15REV1, SPP1*pac*L-15REV2, SPP1*pac*R-0REV1, SPP1*pac*R-0REV2, SPP1*delX110* pUB110 sont séquencés avec la technologie Illumina. Les banques sont réalisées avec le kit « TrueSeq DNA Library Preps Kit » : L'ADN est fragmenté par sonication. Puis les extrémités cohésives d'ADN présentes sur les fragments sont réparées par la T4 polymérase, le fragment Klenow et la polynucléotide kinase. Ensuite, une adénine est ajoutée en 3' au niveau des extrémités franches générées par les traitements enzymatiques précédents. Cet ajout permet la ligation d'adaptateurs avec une extrémité cohésive d'un T en 5'. Les échantillons sont ensuite séquencés à l'aide d'un séquenceur NextSeq en *paired-end* 50-34 ou *paired-end* 80.

A l'issue du séquençage, la séquence des adaptateurs est retirée des fichiers .fastq en utilisant le programme cutadapt. Ensuite, l'alignement est réalisé à partir des fichiers .fastq avec bowtie2. Le génome de référence de SPP1 X97918.3 est utilisé pour l'alignement. Le fichier .sam généré par bowtie2 est converti en fichier .bam puis .sorted.bam en utilisant samtools. Les mutations sur le génome de SPP1 sont détectées avec freebayes. Un fichier d'index .bai est généré avec samtools à partir du fichier .sorted.bam, ce qui permet de lire le fichier .bam avec le logiciel IGV qui permet de visualiser l'alignement. La présence des mutations est également vérifiée avec ce logiciel.

II.XI. qRT-PCR des mutants *pac*L

B. subtilis YB886 est cultivée dans 5 ml de LB à 37°C avec une rotation orbitale de 150 rpm jusqu'à une densité optique (600 nm) de 0.8. La culture est ensuite supplémentée avec 10 mM de $CaCl_2$. Les cellules sont ensuite infectées à une multiplicité d'infection de 5 pfu/cfu avec des SSLL. A 15 minutes post-infection, 1 ml de culture est prélevé puis mélangé à du TBT avec 10 mM NaN_3 à 4°C. Les cellules sont immédiatement centrifugées 2 minutes à 13000xg. Le surnageant est éliminé et le culot congelé dans l'azote liquide puis conservé à -80°C. La même opération est répétée à 25 minutes post-infection. Une extraction d'ARN est réalisée à partir des culots en utilisant le kit Agilent « Total RNA Isolation Mini kit ». En fin de purification, après élution, les ARN sont digérés avec une DNAse I (NEB) pour éviter de fausser les résultats des qRT-PCR avec de l'ADN. Ensuite, les ARN sont de nouveau purifiés avec le kit à partir de l'étape où l'éthanol est ajouté (voir protocole du fournisseur du kit).
La reverse transcription est réalisée avec le kit « High Capacity cDNA Reverse Transcription Kits » de Applied Biosystems à partir des ARN purifiés et digérés à la DNAse. La PCR quantitative est ensuite réalisée à partir des ADNc dilués 300 fois. Un système « QuantStudio 12K Flex Real-Time PCR » avec une détection au SYBR green (Life technologies) est utilisé en suivant les instructions du fournisseur. 3 µl d'échantillon dilué sont mélangés avec un « Fast SYBR Green Master Mix » et avec 500 nM d'oligonucléotides dans un volume final de 10 µl. Les séquence des oligonucléotides utilisés est présentée dans la première rubrique de cette partie matériel et méthodes (section II.I.). Pour chacun des gènes d'intérêt, un couple d'oligonucléotides du tableau 6 est utilisé. Les gènes de *era*, *sgmK*, *secA*, *gyrA* et *dnaG* de *B. subtilis* servent de contrôle pour normaliser les données. Les gènes de SPP1 dont on souhaite connaître les niveaux d'expression sont les gènes *1*, *2*, *6*, *11* et *46*.

Le mélange de réaction est déposé sur une plaque de 384 puits puis soumis à 40 cycles de PCR comme suit : 95°C ;20 secondes, [95°C ;1 seconde, 60°C ;20 seccondes] X40.

Un contrôle négatif sans ADN est réalisé lors de chaque expérience. Les expériences sont réalisées 3 fois, des infections jusqu'à l'étape de PCR (triplicats biologiques), et les qPCR sont répétées 3 fois par réplicat biologique (réplicats techniques). La séquence des oligonucléotides utilisés est déterminée en utilisant l'outil Primer-BLAST. L'absence de potentielles structures secondaires ou de régions homologues non-désirées sur le génome d'intérêt a été vérifiée à l'aide des outils BLAST et oligocalc.

Les valeurs de Ct obtenues pour SPP1 sont normalisées avec la moyenne géométrique des Ct des gènes de référence de *B. subtilis* (1).

(1) $\Delta Ct = (Ct[era]*Ct[sgmK]*Ct[secA]*Ct[gyrA]*Ct[dnaG])^{1/2} - Ct[\text{gène de SPP1}]$

Ces valeurs de ΔCt sont importées dans R et représentées sous forme de boîtes à moustaches à l'aide de ggplot2.

Afin de déterminer le niveau d'expression des gènes d'intérêt chez les différents mutants par rapport au contrôle SPP1*pac*L-99, des taux d'expression relatifs notés ΔΔCt ont été calculés (2).

(2) $\Delta\Delta Ct = \Delta Ct[\text{mutant}] - \Delta Ct[\text{SPP1}pac\text{L-99}]$

Les valeurs de ΔΔCt sont ensuite transformées en valeurs logarithmiques (3).

(3) Niveau d'expression relatif $= 2^{-\Delta\Delta Ct}$

Les données sont ensuite importées dans R et représentées sous forme de *heatmaps* à l'aide du module ggplot2. Afin de déterminer la significativité des différences observées entre les mutants à différents temps, un test ANOVA est réalisé en utilisant la fonction aov(Valeurs de deltaCt~Temps post-infection+Clone). Les mutants sont ensuite comparés deux à deux à l'aide de la fonction glht() du paquet multcomp.

II.XIV. Détermination de la précision de la coupure de *pac* par séquençage Illumina

Afin de déterminer la précision de la coupure *pac*, 8 µg d'ADN génomique de SPP1 purifiés à partir de particules isolées sur gradient de chlorure de césium sont digérés par l'enzyme ApaLI selon les conditions définies par fournisseur de l'enzyme (NEB). Cette digestion génère un fragment 4600 pb qui est issu de la coupure du site *pac* et de celle de l'enzyme ApaLI. L'échantillon digéré est déposé sur un gel d'agarose 0.8% puis une migration à 100V

est réalisée. À l'issue de la migration, le fragment d'intérêt est purifié à partir du gel avec le kit Macherey Nagel « Gel and PCR clean up ». Le fragment purifié est ensuite séquencé par Illumina, selon les conditions décrites ci-dessus.

Après séquençage, la séquence des adaptateurs est retirée des fichiers .fastq à l'aide de cutadapt. L'alignement se fait sur le génome de référence de SPP1 X97918.3 avec bowtie2. Le fichier .sam généré par bowtie2 est converti en fichier .bam puis .sorted.bam en utilisant samtools. Un tableau avec la couverture à chaque position sur le génome est généré avec bedtools genomcov. Ce fichier est exporté dans R. Une séquence de 21 nucléotides correspondant à région où la coupure *pac* a lieu est sélectionnée puis des diagrammes de densité sont réalisés à partir de ces données.

II.XV. Détermination de la conservation la coupure de *pac* avec un profil de restriction par NcoI

1.8 µg d'ADN génomique purifié de chaque mutant de SPP1 a été digéré par l'enzyme de restriction NcoI dans du tampon NEB CutSmart Buffer dans un volume de 40 µL pendant la nuit. Les ADN digérés ont été ensuite déposés sur gel TAE 1X (40 mM Tris, 20mM acide acétique, 1 mM EDTA) agarose 0.8% et mis à migrer à 60 V pendant 6 heures. Le gel a ensuite été incubé dans un bain de bromure d'éthydium pendant 1 heure avant d'être photographié sous Ultraviolets.

II.XVI. Transfert d'un marqueur de résistance à un antibiotique

Des bactéries donneuses possédant un marqueur de résistance à un antibiotique (section II.I tableau 3) sur un locus du chromosome bactérien ou sur un plasmide sont incubées à 37°C, 150 rpm jusqu'à une DO_{600nm} de 0.8. Ensuite, 3 ml sont supplémentés par 10 mM $CaCl_2$ puis infectés par les phages d'intérêt à une multiplicité d'infection de 5 pfu/cfu. Après 2 heures d'infection, les cellules sont culottées par centrifugation 5 minute à 13000xg puis le surnageant de phages est prélevé et filtré avec un filtre de 0.45 µm. Il est ensuite conservé à 4°C. Les échantillons sont traités à la DNAse I à une concentration de 10 $\mu g.ml^{-1}$ pendant 1 heure à 37°C pour éviter que les bactéries réceptrices puissent incorporer de l'ADN libre présent dans le lysat.

Le lysat de phages issu de cette première infection a été utilisé pour infecter des bactéries réceptrices sensibles aux antibiotiques. Des cultures d'YB886 sont cultivées à 37°C, 150 rpm jusqu'à une DO600nm de 1.5. 300µL de cultures sont ensuite infectées à une multiplicité d'infection de 1 puis incubés 10 minutes à 37°C, 150 rpm. 1 ml de LB préchauffé à 37°C est ensuite ajouté aux cultures infectées qui sont de nouveau incubées à 37°C, 150 rpm pendant 10 minutes. Les cellules sont culottées par centrifugation 5 minute à 13000xg puis re-suspendues dans 300 µl de LB auxquels sont ajoutés 40 µl d'anticorps anti-SPP1 issus de sérum de lapin.

Les cellules sont incubées à 37°C, 150 rpm pendant 30 minutes avant d'être à nouveau culottées par centrifugation 5 minute à 13000xg. Les culots sont re-suspendus dans 120 µl de LB. Les cultures sont ensuite diluées à 10^{-5} et 10^{-6} ainsi qu'entre 10^{-1} et 10^{-4} et 100 µl des dilutions sont étalés sur des boîtes LB 2% agar avec et sans antibiotique respectivement. Des boîtes avec 5 $\mu g.ml^{-1}$ de chloramphénicol sont utilisées pour le marqueur sur le chromosome et des boîtes avec 5 $\mu g.ml^{-1}$ de néomycine pour les plasmides. Des cultures non-diluées sont étalées sur des boîtes avec 5 $\mu g.ml^{-1}$ de chloramphénicol dans le cas des expériences de transfert du marqueur sur le chromosome. Les boîtes sont incubées toute la nuit à 37°C puis les colonies dénombrées. Pour calculer les fréquences de transduction les boîtes contenant au minimum 29 et au maximum 330 colonies sont sélectionnées. Pour le marqueur de résistance sur le chromosome, les événements de transduction étant plus rares, des boîtes avec moins de 29 colonies sont parfois utilisées pour calculer des fréquences de transduction. Il n'a pas été possible d'augmenter le nombre de colonies sur ces boîtes puisque c'est l'échantillon non-dilué qui a été étalé.

La significativité des différences entre les mutants est déterminée grâce à un modèle linéaire généralisé avec la fonction glm(Fréquence de transduction~Clone, family=Gamma(link=log)) et les comparaisons des mutants deux à deux sont effectuées à l'aide de la fonction glht() du paquet multcomp.

II.XVII. Quantification de la transduction à partir des séquençages Illumina

Afin de vérifier les résultats des expériences de transduction avec une méthode complémentaire, les données Illumina initialement utilisées pour détecter des mutations sur le génome des phages ont été analysées différemment pour dénombrer les reads provenant de *B.*

subtilis. Les fichiers .fastq sans les adaptateurs sont alignés contre une référence qui contient le génome de SPP1 X97918.3, celui de *B. subtilis* YB886 et éventuellement la séquence d'un plasmide si nécessaire. Bowtie2 est utilisé pour l'alignement. Ensuite, les fichiers .sam générés par Bowtie2 sont analysés avec un script python3 qui permet d'extraire les reads qui se sont alignés sur le génome de *B. subtilis* ou sur la séquence d'un plasmide, mais également les paires de reads qui s'alignent à la fois sur *B. subtilis* et SPP1. En plus de compter le nombre de reads qui s'alignent sur différentes séquences, le script génère de nouveaux fichiers .sam qui peuvent être convertis en .bam puis sorted.bam par samtools. L'outil bedtools genomecov permet de générer un tableau avec la couverture obtenue pour chacune des références. Ensuite, le tableau est importé dans R, et la couverture des références est représentée sous forme d'histogrammes.

II.XVIII. Construction de plasmides produisant la gp1 avec les mutations G94R et E100K dans *E. coli*

Afin de produire les gp1:G94R et gp1:E100K, une mutagénèse dirigée est réalisée sur le plasmide pBT115 qui possède la séquence du gène *1* qui est transcrit à partir d'un promoteur T7 fort. Afin de modifier la séquence du gène *1*, le kit « Quickchange Site-directed Mutagenesis » d'Agilent est utilisé. Une PCR est réalisée avec pBT115 comme matrice. Des oligonucléotides, présentés tableau 5 possédant la modification à insérer dans le gène *1* sont utilisés pour la réaction. L'ensemble de la réaction est réalisée en suivant les recommandations du fournisseur. Après la PCR, les échantillons sont traités avec l'enzyme DpnI (NEB) pendant une heure pour dégrader l'ADN parental ayant servi de matrice (pBT115 sans mutation sur le gène *1*). La réaction est inactivée à 80°C, puis les plasmides sont directement transformés dans *E. coli* DH5alpha (voir protocole ci-dessous). Les colonies obtenues à l'issue de la transformation sont prélevées et mises en culture liquide dans 10 ml de LB+100 $\mu g.ml^{-1}$ ampiciline sur la nuit à 37°C, 150 rpm. Le lendemain, les plasmides sont purifiés en utilisant le kit Macheray Nagel « Nucleospin plasmid ». La présence de la mutation est vérifiée par séquençage Sanger en utilisant les oligonucléotides présentés dans la première partie et *E. coli* BL21 pLysS transformée avec les plasmides dont la séquence est correcte.

II.XIX. Transformations de *E. coli*

Une culture de nuit de *E. coli* dans du LB à 37°C, 150 rpm est diluée au cinquantième dans 300 ml de LB. La culture est incubée à 37°C, 150 rpm jusqu'à une DO600nm de 0.6. Les cellules sont centrifugées à 8000xg pendant 5 minutes à 4°C. Le culot est re-suspendu dans 15 ml de 100 mM $MgCl_2$ puis de nouveau centrifugé dans les mêmes conditions. Toutes les manipulations se font à 4°C et les solutions sont conservées à la même température. Le culot est re-suspendu dans 150 ml de 100 mM $CaCl_2$, puis l'ensemble est incubé 30 minutes dans la glace. Les bactéries sont de nouveau culottées comme précédemment et re-suspendues dans 20 ml d'une solution de 100 mM $CaCl_2$ et 15% glycérol. La culture est divisée en alicots conservés à -20°C ou utilisés directement pour des transformations.

Une quantité de 50 ng de plasmide est mélangée avec 120 µl de cellules compétentes, le tout est incubé au moins 40 minutes dans la glace. Les échantillons sont placés 40 secondes à 42°C puis 1 ml de LB préchauffé à 37°C est ajouté dans chaque tube. Ils sont ensuite incubés à 37°C, 150 rpm pendant 1 heure. Les cellules sont ensuite culottées par centrifugation 5 minutes à 13000xg puis re-suspendues dans 100 µl de LB. Elles sont ensuite étalées sur boîte LB agar 2 % contenant l'antibiotique désiré puis incubées une nuit à 37°C.

II.XX. Purification des protéines gp1 sauvage et avec la mutation gp1:G94R

La souche *E. coli* BL21 avec les plasmides pLysS et pBT115 avec la mutation d'intérêt est cultivée dans 1 litre de LB avec 30 $\mu g.ml^{-1}$ de chloramphénicol et 100 $\mu g.ml^{-1}$ d'ampicilline à 37°C, 150 rpm. A une DO_{600nm} de 0.6 la production de gp1 est induite par l'ajout de 10mM IPTG. Après 3 heures les cellules sont culottées par centrifugation 20 minutes à 8000xg. Le culot est re-suspendu dans 30 ml de 50mM Tris pH7.5, 1M NaCl, 5% glycérol, 1 $mg.ml^{-1}$ lysozyme et 1X inhibiteurs de protéases. Les cellules sont lysées au sonicateur puis centrifugées 45 minutes à 12000xg. Le surnageant est prélevé et précipité avec une solution contenant 70% de sulfate d'ammonium qui est ajoutée progressivement à une température ambiante de 4°C. L'échantillon est de nouveau centrifugé 30 minutes à 12000xg. Le surnageant est éliminé puis le culot re-suspendu dans 10 ml de 50mM Tris pH7.5, 5% glycérol et 300 mM NaCl. L'échantillon est dialysé une nuit à 4°C dans un tampon de même composition.

La protéine est purifiée une première fois par chromatographie échangeuse d'ions dans un purificateur AKTA (Amersham Parmacia Biotech) avec une colonne ressourceS. Un gradient linéaire progressif allant de 300 mM à 1 M NaCl est appliqué dans la colonne pour éluer la protéine gp1 (le tampon contient toujours 50mM de Tris pH7.5 et 5% glycérol, seule la concentration de NaCl change). La purification s'effectue à 4°C. Ensuite, les fractions collectées sont déposées sur un gel de polyacrylamide 15 % et une migration SDS-PAGE est réalisée. Le gel est ensuite coloré au bleu de coomassie. Les fractions contenant le plus de protéines sont concentrées en utilisant des colonnes Vivaspin 20 (MWCO 10,000, Sartorius).
Les fractions concentrées sont de nouveau purifiées par filtration en gel dans un purificateur AKTA à l'aide d'une colonne Superose6. Du dextran blue permet de déterminer le volume mort de la colonne et un marqueur moléculaire Biorad est utilisé pour pouvoir calculer la masse moléculaire des protéines gp1.
Ensuite, les fractions sont de nouveau déposées sur gel et colorées au bleu de coomassie. Les données sont extraites du logiciel Licorn (fourni avec la machine AKTA) et les chromatogrammes représentés sous forme de graphiques avec R et ggplot2. Le coefficient de partage (Kav) est calculé à partir des volumes d'élution des protéines d'intérêt et de ceux du marqueur avec la formule suivante : (Volume d'élution-Volume mort)/(Volume total-Volume mort). Une droite représentant la corrélation entre le Kav et la masse moléculaire permet de déduire celle des gp1.
Les fractions correspondant aux différents pics d'élutions sont utilisées pour la microscopie électronique. Un Bradford est réalisé pour estimer la concentration des protéines (kit « Biorad Protein Assay »). Les échantillons sont dilués de 2 à 10 fois en fonction de leur concentration puis marqués négativement avec 2 % d'acétate d'uranyle. Ils sont ensuite visualisés par microscopie électronique en transmission.

II.XXI. Séquençage Nanopore

Des ADN génomiques de SPP1 sont entièrement séquencés grâce à la technologie Nanopore. Les ADN génomiques sont obtenus à partir de particules purifiées sur gradient de chlorure de césium. L'intégrité des ADN est vérifiée par électrophorèse sur gel d'agarose 0,8 %. En cas de fragmentation, un *smear* est observé, si l'échantillon est de bonne qualité une bande nette apparaît sur le gel. Les banques sont réalisées à partir des ADN génomiques avec le kit SQK-LSK109 de Oxford Nanopore Technologies. Ce kit permet de générer des extrémités franches et de lier une adénine en 3' sur chaque ADN génomique. Il permet également de liguer les

adaptateurs nécessaires au séquençage. Le séquençage est réalisé avec un séquenceur GridION d'Oxford Nanopore technologies. La séquence des adaptateurs est retirée grâce au programme Porechop. Les reads de 40 à 45 kpb contenant un adaptateur à chaque extrémité sont sélectionnés par comparaison des fichiers .fastq avant et après le retrait des adaptateurs. Ils sont alignés avec des références contenant la séquence de *B. subtilis* YB886, celle du phage séquencé avec deux génomes de SPP1 à la suite l'un de l'autre (à cause de la redondance de séquence) et éventuellement celle d'un plasmide (pSEA ou pUB110). L'alignement est effectué avec minimap2. Il a ensuite été possible, à partir d'un fichier .sam, de comparer les alignements pour détecter les reads s'alignant sur deux séquences différentes (plasmide, génome de *B. subtilis* et génome de SPP1). Les coordonnées des jonctions sur les 2 références sont récupérées à partir du fichier .sam puis exportées dans R pour faire des graphiques avec les coordonnées respectives de la jonction sur SPP1 et sur le plasmide pSEA. Les coordonnées des jonctions, de début et de fin des reads sont représentées sur 2 séquences de génome de SPP1.Les données sont exportées dans R pour être présentées sous forme d'histogrammes avec la distribution de taille des différents types d'ADN. Les manipulations de fichiers .fastq et .sam sont effectuées avec Python3.

Résultats

III. Résultats

III.I. Rôle de la séquence *pac*L dans la reconnaissance du génome de SPP1

III.I.I. Mutagénèse de *pac*L

Des mutants du phage SPP1 avec des délétions graduelles de *pac*L ont été générés en utilisant la méthode de clonage décrite dans la rubrique matériel et méthodes de ce manuscrit (section II.II). Chaque mutant possède une délétion différente, nous avons ainsi réussi à obtenir les mutants suivants : SPP1*pac*L-99, SPP1*pac*L-93, SPP1*pac*L-76, SPP1*pac*L-68, SPP1*pac*L-58, SPP1*pac*L-54, SPP1*pac*L-36, SPP1*pac*L-27 et SPP1*pac*L-15 dont le nombre dans leur nom indique la longueur de *pac*L qui est graduellement réduite à partir de l'extrémité gauche de la séquence (voir figure 24). Ces mutations ont été choisies en fonction des zones couvertes par gp1 dans les expériences d'empreintes à la DNAse de Chai et al, 1995, des expériences réalisées dans un vecteur plasmidique (Djacem et al, 2017) et également des positions du promoteur et des régions répétées sur *pac*L (figure 24.A). Ainsi le mutant SPP1*pac*L-99 possède toute la séquence *pac*L, mais porte également une délétion en amont rendue inévitable par la technique utilisée pour le clonage. Dans la suite des expériences, il servira de contrôle pour confirmer que les phénotypes observés ne sont pas dus à la délétion en amont de *pac*L mais bien aux délétions sur *pac*L. Chez les autres mutants, les motifs importants pour la transcription et supposés essentiels pour la reconnaissance de *pac* par la terminase ont été peu à peu supprimés. Ainsi, le mutant SPP1*pac*L-93 ne dispose plus du terminateur de transcription (voir figure 24), car la formation d'une structure tige-boucle est rendue impossible par une délétion de 6 nucléotides. Chez le mutant SPP1*pac*L-76, le terminateur de transcription ainsi que la boîte -35 du promoteur *P*L1 ont été supprimés. Le mutant SPP1*pac*L-68 possède la séquence minimale à partir de laquelle il n'est plus possible de couper la séquence *pac* dans un système plasmidique (voir Djacem et al, 2017). Cette délétion inclue également la boîte TATA du promoteur *P*L1. Le motif *a1,* site de fixation supposé de la terminase sur *pac*L, n'est plus présent à partir du mutant SPP1*pac*L-36. Ensuite, la séquence *pac* est réduite au minimum jusqu'au RBS du gène *1* (figure 24.C).

Afin de déterminer l'éventuelle influence de l'absence du promoteur *P*L1 sur la viabilité des phages mutants décrits dans le paragraphe précédent, nous avons créé 2 nouveaux phages SPP1*pac*L-99.1 et SPP1*pac*L-99.2 dont la séquence de *pac*L a été remplacée par une

séquence aléatoire entre *P*L1 et le RBS du gène *1,* et également entre le RBS et le début du gène *1* chez le mutant SPP1*pac*L-99.2.

III.I.II. Phénotypes de plages de lyse et courbes de lyse

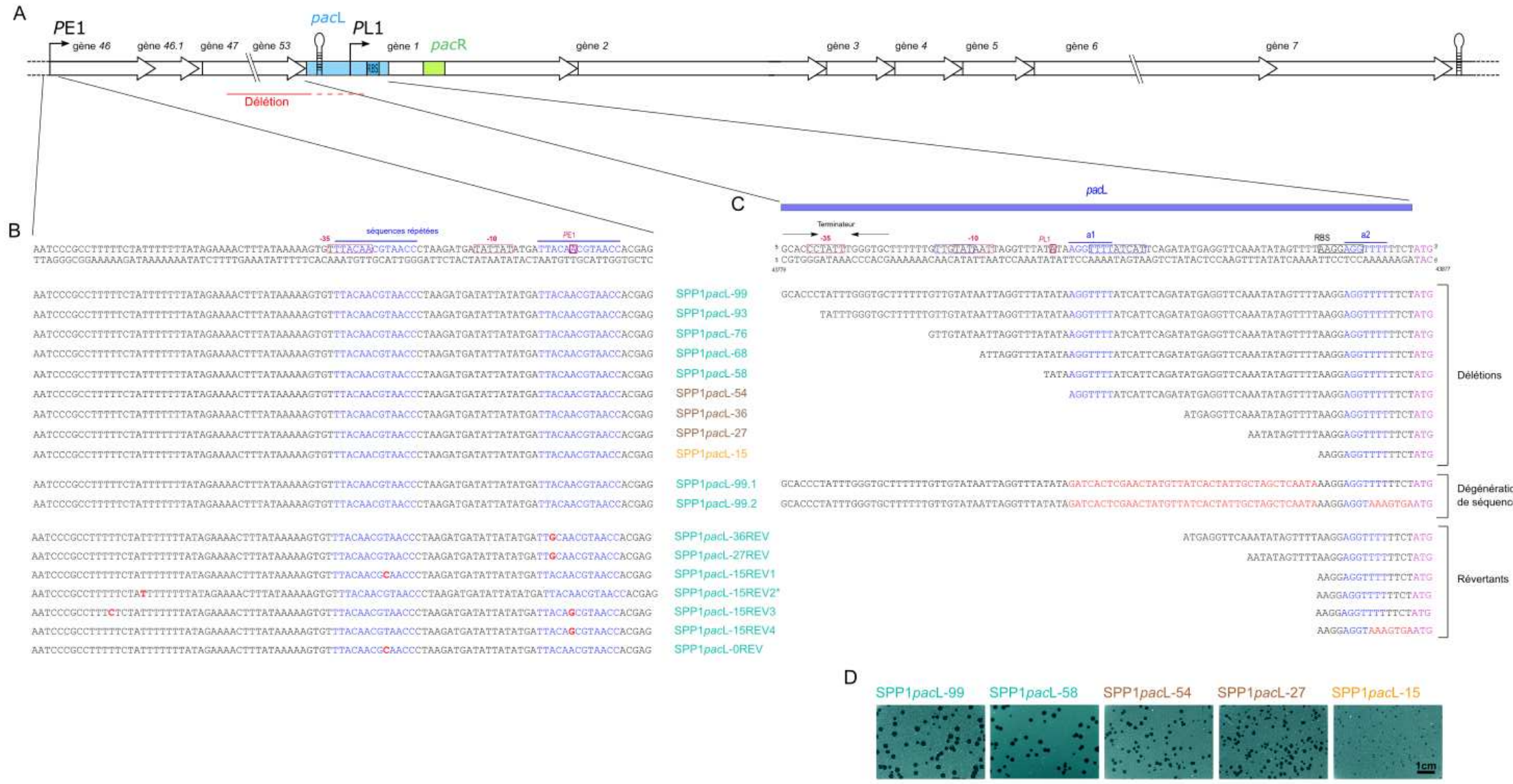
L'ensemble de ces mutants a été titré afin de déterminer des phénotypes de plages de lyse. À l'issue de cette expérience, il est possible d'évaluer la capacité de multiplication des phages en observant la taille des plages de lyse (figure 24.D). Chacune d'entre elles correspond à un phage qui a infecté une cellule puis qui s'est multiplié. Au cours d'infections successives, une plage de lyse se forme sur un tapis confluent de bactéries. Plus le phage infecte efficacement les bactéries, plus les plages de lyse qu'il génère sont larges. Tous les mutants mentionnés ci-dessus forment des plages de lyse, ce qui signifie que toutes les mutations permettent la formation d'une particule virale capable de se multiplier. Autrement dit, il est possible de supprimer la quasi-totalité de *pac*L et d'obtenir des phages viables. En revanche, différents phénotypes de plages de lyse sont observés, ce qui indique que certaines mutations semblent affecter l'infectiosité des phages. Nous avons ainsi classé les mutants en trois groupes. Le premier groupe se compose des mutants qui ne présentent pas de phénotype discernable de SPP1 sauvage et du mutant SPP1*pac*L-99. Il regroupe ainsi les mutants SPP1*pac*L-99, SPP1*pac*L-99.1, SPP1*pac*L-99.2, SPP1*pac*L-93, SPP1*pac*L-76, SPP1*pac*L-68 et SPP1*pac*L-58. Viennent ensuite, dans un second groupe, les mutants SPP1*pac*L-54, SPP1*pac*L-36 et SPP1*pac*L-27 qui semblent moyennement affectés par les mutations. Le mutant SPP1*pac*L-15 a été classé à part dans un troisième groupe, car il arbore des plages de lyse bien plus petites que l'ensemble des autres mutants (voir figure 24.D). Ces résultats indiquent que les phages semblent être affectés par la délétion lorsque celle-ci entame les 54 paires des bases en amont du codon d'initiation du gène *1*. Nous pourrions être tentés de conclure que l'élimination des séquences répétées *a1* et *a2* de *pac*L impacterait l'interaction de gp1 avec *pac*L affectant ainsi l'infectiosité. En vérité, il n'en est rien, puisque les mutants SPP1*pac*L-99.1 et SPP1*pac*L-99.2 ne semblent pas affectés par la dégénération de cette séquence. Un rôle redondant de ces séquences, l'une devenant importante seulement en l'absence de l'autre, ne peut être exclu.

Une fois obtenus, ces phages ont été amplifiés en suivant la procédure décrite dans la rubrique matériel et méthodes (section II). En titrant les phages à différentes étapes d'amplification, il a été possible d'observer un phénomène assez singulier chez les mutants les plus affectés par

les délétions. Une nouvelle population de plages de lyse plus grosses apparaît en coexistence avec de petites plages (voir section III.I.III). Lors des cycles d'amplification suivants, cette nouvelle population devient dominante. En isolant des phages à partir de ces grosses plages de lyse et en les titrant à nouveau, il a été possible d'obtenir des plages identiques à celles obtenues en titrant SPP1 sauvage ou SPP1*pac*L-99. Ceci indique l'apparition de révertants génétiques. Ce phénomène est observable chez le mutant SPP1*pac*L-15 à partir de lysats sur boîte, et plus tardivement au cours de l'amplification des mutants SPP1*pac*L-36 et SPP1*pac*L-27 lors de la réalisation de lysats en culture liquide. Une fois isolé, le phage SPP1*pac*L-15 révertant, a permis la construction d'un mutant SPP1*pac*L-0REV dont la séquence entre le RBS et le début du gène *1* (codant pour gp1 la petite sous-unité de la terminase, voir tableau 1) a été dégénérée, ce phage possède un phénotype de plages de lyse sauvage. La caractérisation génétique de ces révertants est présentée dans la section suivante.

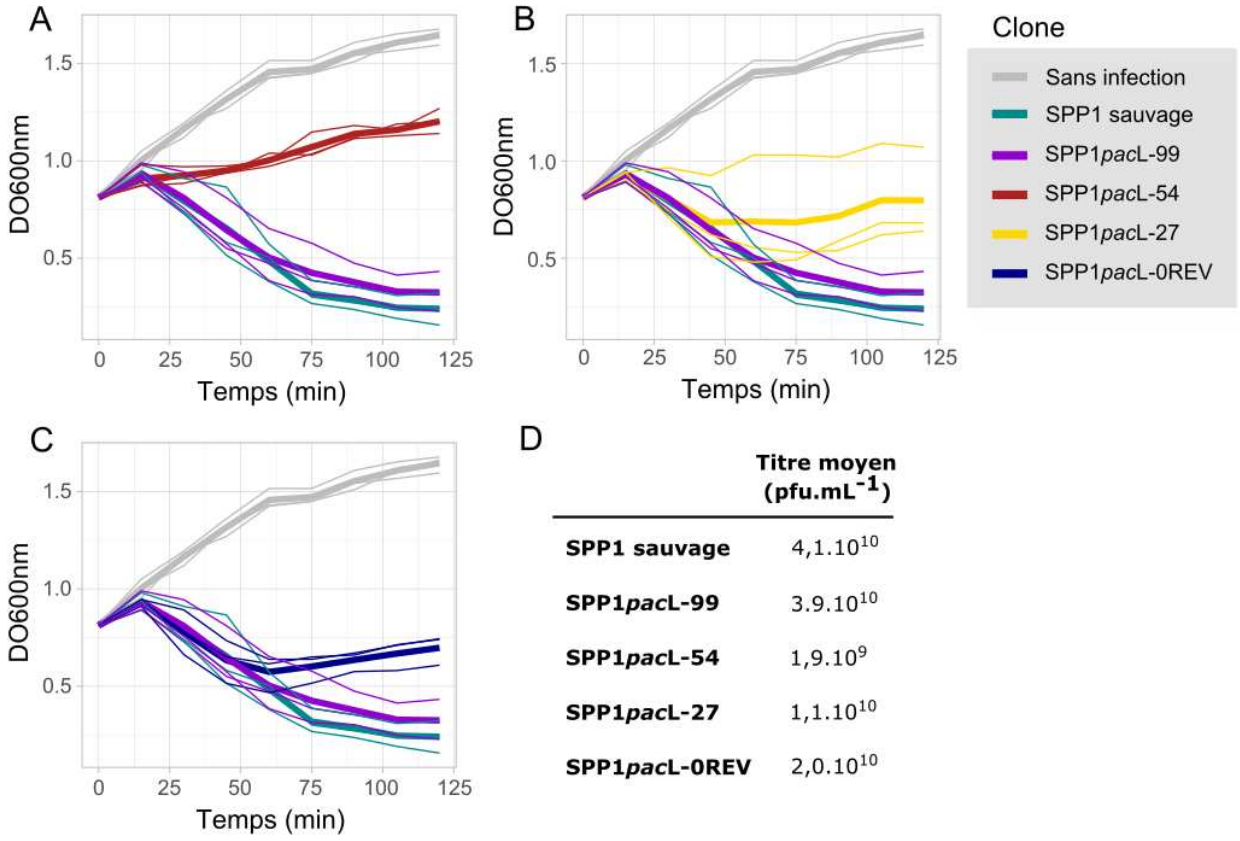
Figure 24. Mutagénèse de *pac*L. (A) Contexte génomique de *pac*. *pac*L est représentée en bleu avec le terminateur et le promoteur *P*L1 et *pac*R en vert. Les gènes en amont et en aval de ces séquences figurent en blanc, PE1 est en amont du gène 47. Représentation des génotypes de chacun des révertants et mutants obtenus. La séquence de la région en amont et en aval de *P*E1 (B) et celle de *pac*L (C) sont représentées. Les substitutions de nucléotides sont représentées en rouge et les séquences répétées en bleu. Le code couleur associé au nom de chaque mutant de SPP1 correspond aux phénotypes observés: turquoise - sauvage; marron - moyennement affecté; orange - très affecté. Les phénotypes de plages de lyse sont représentés en bas à droite de la figure (D).

***** Ce révertant possède une mutation additionnelle dans le gène *29* (codon STOP prématuré en position 68).


A
PE1
pacL
PL1
gène 46
gène 46.1
gène 47
gène 53
gène 1
pacR
gène 2
gène 3
gène 4
gène 5
gène 6
gène 7
Délétion
B
séquences répétées
-35
-10
PE1
C
pacL
Terminateur
-35
-10
PL1
a1
RBS
a2
SPP1pacL-99
SPP1pacL-93
SPP1pacL-76
SPP1pacL-68
SPP1pacL-58
SPP1pacL-54
SPP1pacL-36
SPP1pacL-27
SPP1pacL-15
Délétions
SPP1pacL-99.1
SPP1pacL-99.2
Dégénérations de séquence
SPP1pacL-36REV
SPP1pacL-27REV
SPP1pacL-15REV1
SPP1pacL-15REV2*
SPP1pacL-15REV3
SPP1pacL-15REV4
SPP1pacL-0REV
Révertants
D
SPP1pacL-99
SPP1pacL-58
SPP1pacL-54
SPP1pacL-27
SPP1pacL-15
1cm

Afin de caractériser plus précisément les mutants obtenus, des courbes de lyse ont été réalisées (figure 25). Des cellules sont cultivées en milieu liquide, puis infectées par des lysats des phages d'intérêt issus de cultures liquides, une mesure de la turbidité (densité optique à 600nm) est ensuite effectuée toutes les 15 minutes. Nous avons choisi des phages représentatifs de chacun des groupes déterminés précédemment : SPP1 sauvage, SPP1*pac*L-99, SPP1*pac*L-93, SPP1*pac*L-54, SPP1*pac*L-27, SPP1*pac*L-15REV et SPP1*pac*L-0REV. Une diminution de la valeur de la densité optique atteste de la lyse des cellules par le phage. Si elles lysent rapidement, la densité optique décroît elle-aussi plus rapidement, ce qui indique que l'infection est efficace. Le seuil minimal de densité optique atteint au cours de l'expérience peut aussi renseigner sur l'efficacité de l'infection, plus le seuil est bas, plus un nombre important de cellules a été lysé. De ces expériences, se dégagent 3 types de comportement détaillés figure 25. Dans le premier cas, la densité optique évolue de manière similaire au sauvage et au mutant SPP1*pac*L-99. Ce comportement est observé chez le mutant SPP1*pac*L-93 et les révertants SPP1*pac*L-15REV et SPP1*pac*L-0REV. Dans un deuxième cas, la densité optique continue d'augmenter lentement sans pour autant atteindre les valeurs de densité optique d'une culture sans infection. Seul le mutant SPP1*pac*L-54 présente cette tendance. Enfin, dans un dernier cas, différents réplicats avec le même mutant donnent des résultats qui s'approchent soit de la tendance observée chez SPP1*pac*L-54, soit de celle observée chez SPP1*pac*L-99 (figure 25). Ces résultats sont cohérents avec les phénotypes de plages de lyse ; Chez le mutant SPP1*pac*L-54, l'infectiosité des phages est fortement affectée, ce qui est visible à la fois au niveau des plages de lyse qui sont plus petites que chez le contrôle, et aussi dans les expériences de cinétique où la densité optique ne décroît pas. Chez le mutant SPP1*pac*L-27, de petites plages de lyse indiquent que les phages sont affectés par les mutations, mais elles sont rapidement supplantées par une population de plages de lyse plus grosses qui indiquent l'apparition de révertants. Dans les expériences de courbes de lyse, la situation évolue différemment à chaque fois, parfois la lyse des cellules suit le même tracé que le contrôle, ce qui doit signifier que des révertants apparaissent au cours de l'expérience, parfois l'allure de la courbe est proche celle de SPP1*pac*L-54, ce qui indique que le phage est toujours affecté par la mutation et que des révertants n'apparaissent pas ou tardivement dans la population. Ces résultats montrent que les lysats de SPP1*pac*L-27 présentent des populations hétérogènes rendant la caractérisation phénotypique de ce phage très difficile.

Ce résultat est intéressant, car il indique que des révertants n'apparaissent pas systématiquement au cours de l'expérience. Il sera à prendre en considération lorsqu'il sera question de faire des expériences où des lysats en culture liquide seront utilisés.



	Titre moyen (pfu.mL^{-1})
SPP1 sauvage	$4{,}1.10^{10}$
SPP1*pac*L-99	$3.9.10^{10}$
SPP1*pac*L-54	$1{,}9.10^{9}$
SPP1*pac*L-27	$1{,}1.10^{10}$
SPP1*pac*L-0REV	$2{,}0.10^{10}$

Figure 25. Courbes de lyse de cultures infectées par des mutants sur *pac*L. Les bactéries sont infectées à une OD600nm de 0.8 au temps 0 par SPP1*pac*L-54 (A) SPP1*pac*L-27 (B) et SPP1*pac*L-0REV (C), des infections par SPP1 sauvage et SPP1*pac*L-99 servent de contrôle avec une culture sans infection. (D) Titre de phages obtenu à l'issue de l'expérience (surnageant). Les expériences sont répétées 3 fois, chaque réplicat est représenté par une ligne fine et leur moyenne avec un trait plus épais.

III.I.III. Caractérisation de révertants génétiques des délétions sur *pac*L et analyse des mutations compensatrices par séquençage

Nous avons vu dans la section précédente que des suppressions de phénotype se produisaient chez les mutants SPP1*pac*L-36, SPP1*pac*L-27 et SPP1*pac*L-15, semblant indiquer l'apparition de révertants génétiques. Afin de le prouver et de déterminer les éventuels mécanismes impliqués dans la réversion, nous avons d'abord séquencé entièrement le génome de 2 phages révertants SPP1*pac*L-15 (SPP1*pac*L-15REV1 et SPP1*pac*L-15REV2)

précédemment isolés en utilisant la technologie Illumina. Après séquençage et analyse des données (voir matériel et méthodes), il a été possible de déterminer la position des mutations (figure 24). Nous avons localisé une mutation ponctuelle sur deux séquences répétées qui chevauchent le promoteur *P*E1 en amont du promoteur *P*L1 absent chez les mutants séquencés (séquences en bleu dans la figure 24). Chacun d'eux possède une mutation ponctuelle différente. En séquençant des lysats à différentes étapes d'amplification, il a été possible de suivre les changements génétiques qui se sont opérés. Comme le montre la figure 26, dans le lysat initial, la réversion n'est pas présente, elle apparaît ensuite dans les lysats sur boîte (voir matériel et méthodes) où elle forme une population minoritaire de plages de lyse plus grosses par rapport aux phages non-révertants. Le repiquage de ces grosses plages montre que les phages héritent de ce phénotype, démontrant ainsi un changement d'origine génétique.

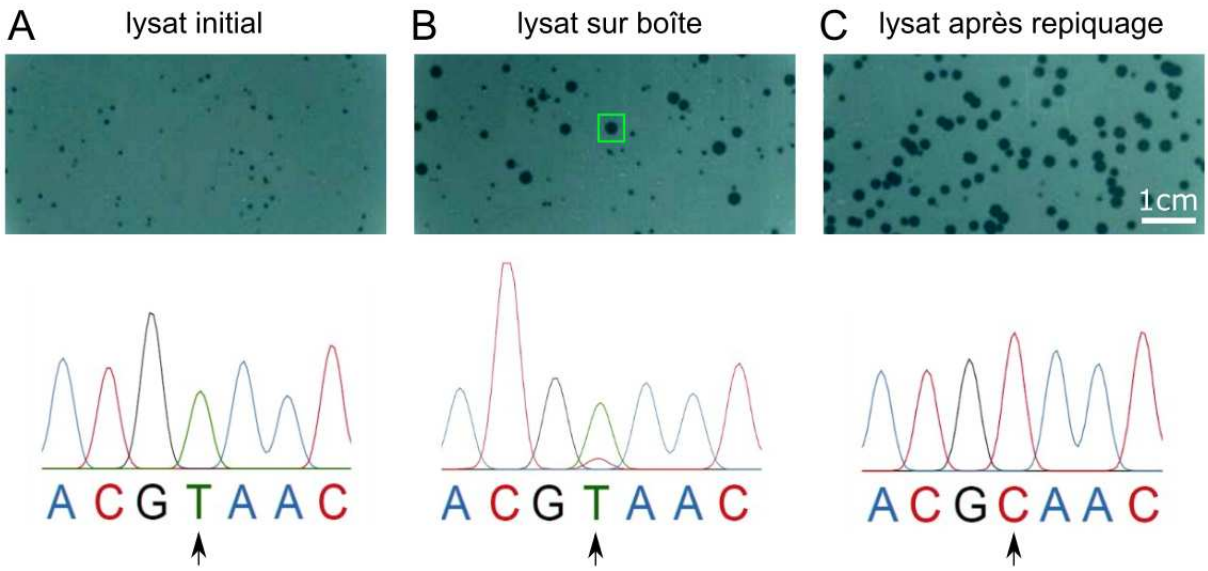


Figure 26. Apparition de révertants chez SPP1*pac*L-15. Titration de SPP1*pac*L-15 avec des plages de lyses très petites (A), réversion du phénotype chez un certain nombre de phages après amplification (B), puis repiquage d'une grosse plage de lyse et titration (C), les plages de lyse sont globalement plus grosses. Corrélation entre le phénotype et le génotype avec apparition d'une mutation proche du promoteur précoce 1 en position 40538, au centre des chromatogrammes. Au milieu, le chromatogramme indique l'existence d'une sous-population ou le T est remplacé par un C.

Nous avons voulu savoir si les autres révertants identifiés présentaient le même type de mutation sur ces mêmes séquences. Nous avons donc séquencé une région de quelques centaines de paires de bases de part et d'autre de *P*E1 chez un révertant SPP1*pac*L-36, un révertants SPP1-*pac*L-27 et 6 révertants SPP1*pac*L-15. Nous avons repéré des mutations sur

les séquences répétées chez les révertants SPP1*pac*L-36, SPP1-*pac*L-27 et 4 révertants SPP1*pac*L-15, et une mutation ponctuelle sur une séquence poly-T en amont du promoteur *P*E1 chez 1 révertant SPP1*pac*L-15. Ces mutations sont décrites dans la figure 24, notons que les révertants possédant les mêmes mutations ont hérité du même nom. Précisons également que le révertant *pac*L-15REV2 possède une seconde mutation dans le gène *27* dont la fonction est encore inconnue. On ne peut donc pas exclure un rôle de cette mutation dans la réversion dans ce cas précis. L'ensemble de ces résultats suggère que les mutations servent à compenser un défaut de transcription induit par l'absence de *P*L1. On supposera que la séquence répétée 2 fois qui chevauche partiellement le promoteur *P*E1 est le site de fixation d'un régulateur de transcription encore inconnu chargé de réprimer le promoteur *P*E1. En l'absence de *P*L1, le promoteur *P*E1 doit assurer la transcription de l'opéron en aval de *pac*L. Lorsque le répresseur est présent, un bas niveau de transcription doit affecter le phage dans la production de l'ARNm puis de protéines essentielles à sa multiplication (gp1, gp2, gp6 et gp7), ce qui génère, du moins en partie, le phénotype observé précédemment. Les mutations ponctuelles sur la séquence consensus du répresseur doivent empêcher sa fixation et rétablir un niveau suffisant de transcription de l'opéron en aval de *pac*L. Afin de vérifier cette hypothèse, des expériences de qRT-PCR ont été réalisées.

III.I.IV. Analyse du niveau de transcription chez des mutants *pac*L par qRT-PCR

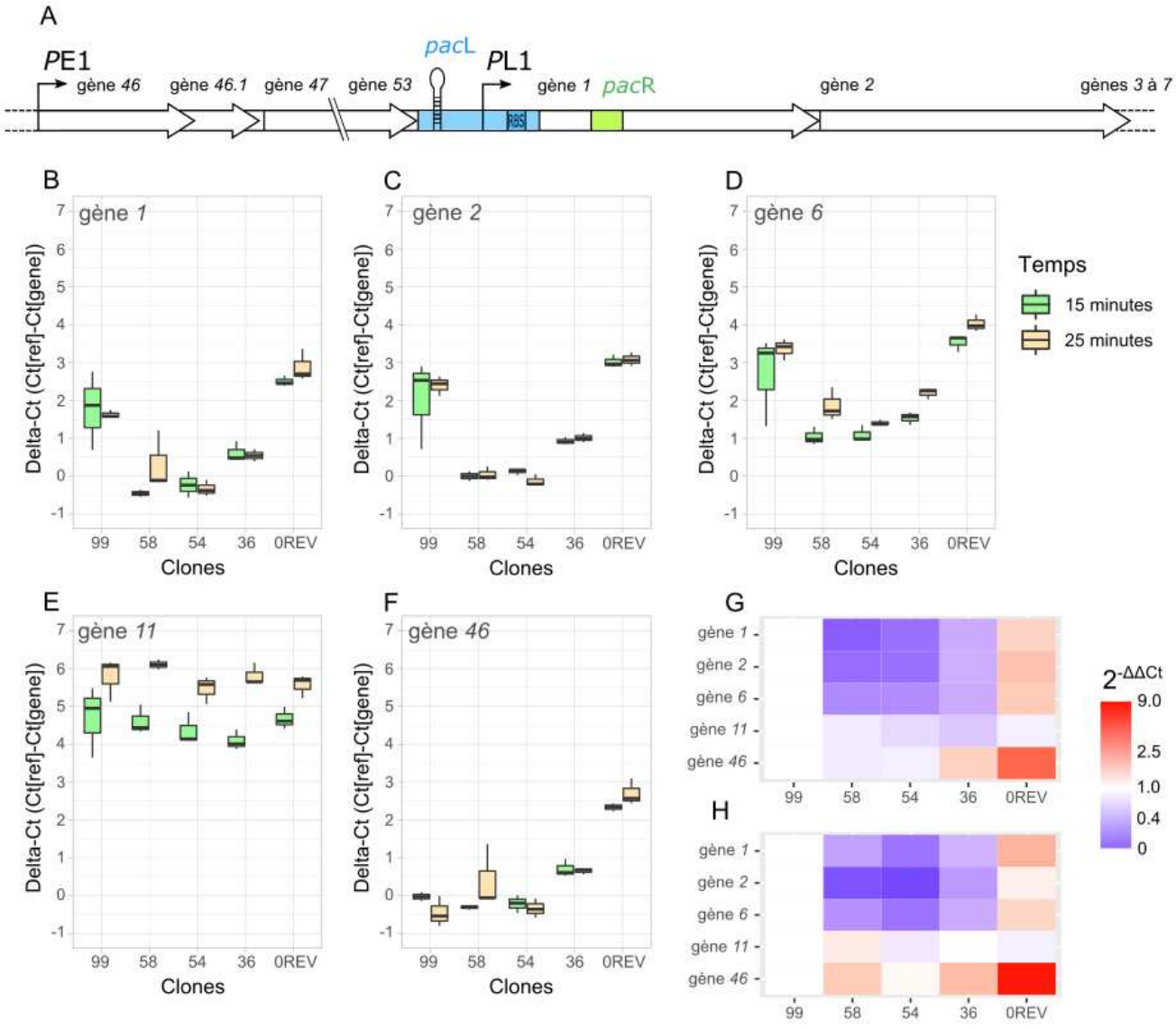


Figure 27. qRT-PCR de mutants sur *pac*L pour quantifier l'influence des mutations sur le niveau de transcription.Rappel du contexte transcriptionnel de la région étudiée. Boîtes à moustaches représentant les ΔCt des gènes *1* (B), *2* (C), *6* (D), *11* (E) et *46* (F) chez SPP1*pac*L-99, SPP1*pac*L-58, SPP1*pac*L-54, SPP1*pac*L-36 et SPP1*pac*L-0REV à 15(vert) et 25 minutes (beige) post-infection. Chaque expérience est réalisée en 3 fois. Heatmap représentant le niveau d'expression des gènes par rapport à celui de SPP1*pac*L-99 à 15 (G) et 25 minutes (H) post-infection. Statistiques présentées en annexe.

Afin de déterminer comment est affectée la transcription chez les mutants ayant une délétion dans *pac*L et de comprendre comment la réversion influe sur la régulation de la transcription, des qRT-PCR des gènes *46*, *1*, *2*, *6*, et *11* de SPP1 ont été réalisées à 15 et 25 minutes post-infection chez SPP1*pac*L-99, SPP1*pac*L-58, SPP1*pac*L-54, SPP1*pac*L-36 et SPP1*pac*L-

0REV. La transcription du gène *46* est analysée car ce gène est situé juste en aval de *P*E1 et fourni des renseignements précis sur l'activité de *P*E1 dans ces différentes conditions. Les gènes *1* et *2* sont ceux de la terminase (tableau 1), leur analyse permet de déterminer comment le niveau de transcription des terminases est modifié par les différentes mutations. Quantifier la transcription du gène *6*, codant pour gp6 (tableau 1), est utile pour savoir si le niveau de transcription est identique à celui des gènes *1* et *2* en fin d'opéron. Enfin, le gène *11*, qui code pour gp11 la protéine d'échafaudage (tableau 1), servira de contrôle car il n'est ni sous le contrôle de *P*E1, ni sous le contrôle de *P*L1.

Les résultats en delta-Ct, correspondant aux valeurs de Ct des gènes d'intérêt normalisées par les Ct des gènes de référence, (figure 27.A à E) montrent clairement une diminution de la transcription des gènes *1*, *2* et *6* chez les mutants SPP1*pac*L-58, SPP1*pac*L-54 et SPP1*pac*L-36 par rapport au contrôle SPP1*pac*L-99 à 15 et 25 minutes post-infection. Cette diminution peut s'interpréter comme étant la conséquence de la délétion du promoteur *P*L1 et de l'absence, du moins partielle, de dé-répression du promoteur *P*E1. La transcription du gène *46* semble inchangée chez ces mutants, ce qui est cohérent avec l'idée que le contexte transcriptionnel reste identique. Cependant, la transcription des gènes *1*, *2* et *6* parait un peu plus élevée chez SPP1*pac*L-36 comparé aux mutants SPP1*pac*L-58 et SPP1*pac*L-54. Si l'on considère ce qui a été discuté dans la partie précédente, ce fait est aisément explicable ; le lysat SPP1*pac*L-36 doit contenir une population mixte où des révertants ont déjà commencé à apparaître. Ainsi l'augmentation de la transcription serait due à une dé-répression de *P*E1 dans une population minoritaire de révertants. Malheureusement, chez SPP1*pac*L-36, il n'est pas possible d'obtenir un lysat qui soit suffisamment concentré en phages pour infecter les cellules qui ne contiennent pas déjà des révertants. Le phage SPP1*pac*L-0REV a un profil différent des autres mutants, les gènes *1*, *2*, *6* et *46* sont plus exprimés chez ce clone par rapport à l'ensemble des autres mutants. Ceci s'explique par la présence de la mutation compensatrice sur *P*E1 qui n'est plus réprimé par le répresseur de transcription. Cette dé-répression est très visible dans les qRT-PCR du gène *46* (figure 27.F) où le mutant SPP1*pac*L-0REV affiche un niveau de transcription plus élevé que les autres clones. Ainsi, l'ensemble de ces gènes étant sous le contrôle de *P*E1 dans cette situation, ils sont fortement exprimés. Cette surexpression est visible aussi bien à 15 qu'à 25 minutes post-infection. Lorsque l'expression de ces gènes est normalisée avec celle de SPP1*pac*L-99, il est possible de savoir combien de fois un gène d'intérêt est plus ou moins transcrit par rapport à SPP1*pac*L-99. Ces résultats sont présentés sous forme de *heatmap* figure 27.F-G. Il est très

clair que les gènes *1*, *2*, *6* et *46* sont plus exprimés chez SPP1*pac*L-0REV que chez SPP1*pac*L-99. Cette différence est surtout visible pour le gène *46* qui est jusqu'à 9 fois plus exprimé que chez SPP1*pac*L-99 à 25 minutes post-infection. Chez SPP1*pac*L-36, le gène *46* est quasiment 2 fois plus transcrit que chez SPP1*pac*L-99, ce qui est cohérent avec les delta-Ct déterminés précédemment. Les résultats obtenus pour la transcription du gène *11* montrent une expression plutôt homogène de ce gène à 15 et 25 minutes post-infection respectivement. Son expression est globalement plus haute à 25 minutes post-infection qu'à 15 minutes post-infection. Ceci est cohérent avec le fait que gp11 est une protéine structurale (protéine échafaudage, voir introduction) qui intervient vers la fin du cycle viral et son gène est transcrit tardivement.

III.I.V. Étude de la régulation de la transcription de l'opéron des terminases par gp1 et gp2.

Pour mieux comprendre comment la transcription de l'opéron de la terminase est régulée par l'activité de la terminase elle-même, des qRT-PCR ont été réalisées avec les phages SPP1*sus19* et SPP1*sus70*. Le phage SPP1*sus70* ne produit pas la gp1, il permet donc de voir comment la transcription de l'opéron évolue en absence de gp1 et également lorsque la coupure *pac* n'a pas lieu, gp2 seule ne pouvant pas cliver l'ADN au niveau de *pac*. Le mutant SPP1*sus19* ne produit pas gp2, il est ainsi possible de déterminer comment la transcription est modifiée lorsque gp1 peut se fixer sur *pac* mais que cette séquence n'est pas clivée. Des qRT-PCR sur les gènes *1*, *2*, *6*, *11* et *46* ont été réalisées à 15 et 25 minutes post-infection. Chez les gènes de l'opéron de la terminase (*1*, *2* et *6*), une légère augmentation du niveau de transcription est visible dès 15 minutes post-infection chez les mutants SPP1*sus70* et SPP1*sus19* par rapport à SPP1 sauvage (figure 28). Cette différence devient bien plus importante à 25 minutes post-infection. En revanche, les niveaux de transcription observés chez les deux mutants semblent comparativement assez similaires à 15 comme à 25 minutes post-infection respectivement. Ceci indique que gp1 et gp2 sont bien impliquées dans la régulation de leur opéron, puisqu'en leur absence, la transcription augmente. Ces protéines auraient donc tendance à diminuer le niveau de transcription de leurs propres gènes. Deux mécanismes peuvent expliquer cette baisse. Dans un premier cas, la fixation de gp1 sur la séquence *pac* bloquerait l'ARN polymérase, ce qui aurait pour conséquence un arrêt de la transcription. Dans un second cas, la coupure de l'ADN par gp2 entre *pac*L et *pac*R pourrait

simplement empêcher la synthèse d'ARN, le promoteur et la séquence à transcrire se trouvant dorénavant sur 2 molécules d'ADN séparées.

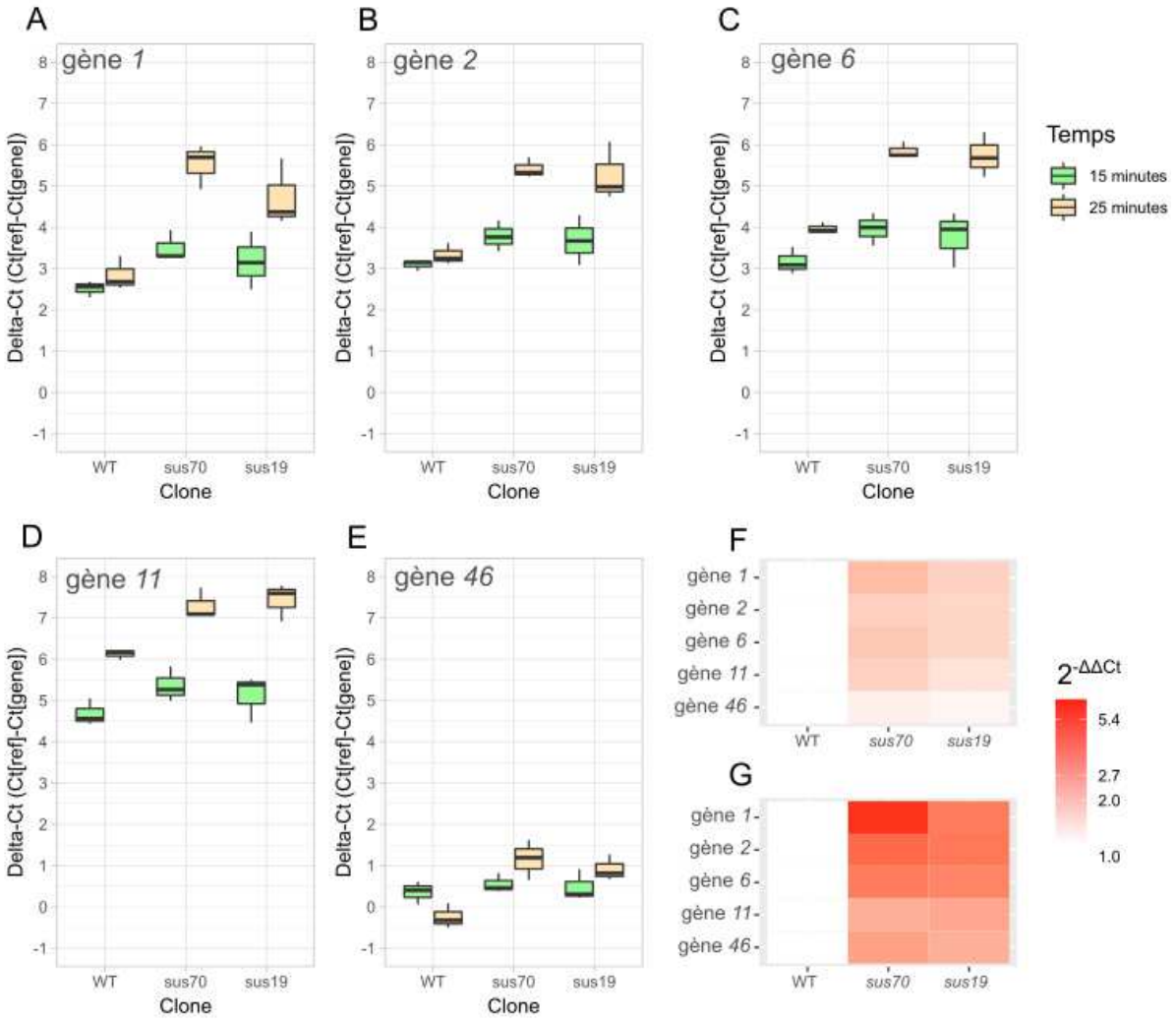


Figure 28. qRT-PCR de SPP1 sauvage et des mutants SPP1*sus19* et SPP1*sus70* pour quantifier l'influence des terminases sur le niveau de transcription de leur opéron. Boîtes à moustaches représentant les ΔCt des gènes *1* (A), *2* (B), *6* (C), *11* (D) et *46* (E) à 15(vert) et 25 minutes (beige) post-infection. Chaque expérience est réalisée en 3 fois. Heatmap représentant le niveau d'expression des gènes par rapport à celui de SPP1 sauvage à 15 (F) et 25 minutes (G) post-infection. Statistiques présentées en annexe.

Les niveaux d'augmentation étant quasiment similaires chez les deux mutants, on déduit que c'est la coupure de l'ADN par gp2 qui semble avoir le plus grand impact. En effet, si la fixation de gp1 sur *pac* était la cause principale d'une diminution de la transcription, le mutant SPP1*sus19*, qui possède une gp1 viable, devrait avoir un niveau plus bas, or ce n'est pas le

cas. La hausse du niveau de transcription chez les 2 mutants serait donc essentiellement due à l'absence de coupure de *pac*. En regardant attentivement les niveaux d'expression à 25 minutes post-infection, une légère différence est cependant visible, les niveaux de transcription de SPP1*sus70* sont légèrement supérieurs à ceux de SPP1*sus19*, ce qui semble indiquer que la fixation de gp1 sur *pac* a quand même un léger impact sur la transcription.

III.I.VI. Détermination de la précision de la coupure de *pac* par Illumina

Il a précédemment été montré qu'il était possible de supprimer une grande partie de *pac*L et de générer des phages qui soient viables. Les délétions les plus importantes impactent négativement les phages qui voient leur infectiosité diminuée. Celle-ci est rétablie à un niveau proche de la normale grâce à l'apparition de révertants dont les mutations compensatrices présentes sur *P*E1 provoquent une surexpression de l'opéron des terminases. Afin de voir si la gp1 se fixe toujours sur *pac* et si gp2 est capable de cliver cette séquence, nous avons cherché à déterminer si la précision de coupure de *pac* était conservée. Il a été possible d'obtenir cette information par des approches de séquençage Illumina qui sont décrites dans la rubrique matériel et méthodes. Trois mutants représentatifs ont été choisis. SPP1*pac*L-99 sert de contrôle pour déterminer si la délétion induite par la méthode utilisée pour le clonage n'a pas d'impact sur la précision de coupure de *pac* et permet de comparer les résultats obtenus pour les autres mutants ; SPP1*pac*L-54 a été également choisi, car il s'agit du seul mutant qui soit fortement affecté par la délétion dans *pac*L qui ne produise pas de révertants lors des cinétiques d'infections (voir figure 25). Enfin, SPP1*pac*L-0REV a été sélectionné car il s'agit d'un révertant et qu'il possède la séquence *pac*L la plus courte.

Des ADN génomiques de phages purifiés ont été digérés par l'enzyme de restriction ApaLI et un fragment d'ADN de 4200 pb a été purifié. Il est issu d'une part du clivage de *pac* lors de la reconnaissance de l'ADN viral par la terminase et du clivage de l'enzyme de restriction d'autre part, le site de restriction se situant à 4200 pb du site *pac* (Djacem et al 2017 ; voir matériel et méthodes). Ces fragments ont ensuite été séquencés avec la technologie Illumina.

À l'issue d'un séquençage Illumina, le nucléotide présent aux extrémités des fragments séquencés a été chaque fois cartographié, ce qui a permis de réaliser des diagrammes de

densité qui rendent compte de la précision de la coupure *pac* figure 29. Ils représentent le nombre de reads qui commencent à chaque position en abscisse de l'histogramme.

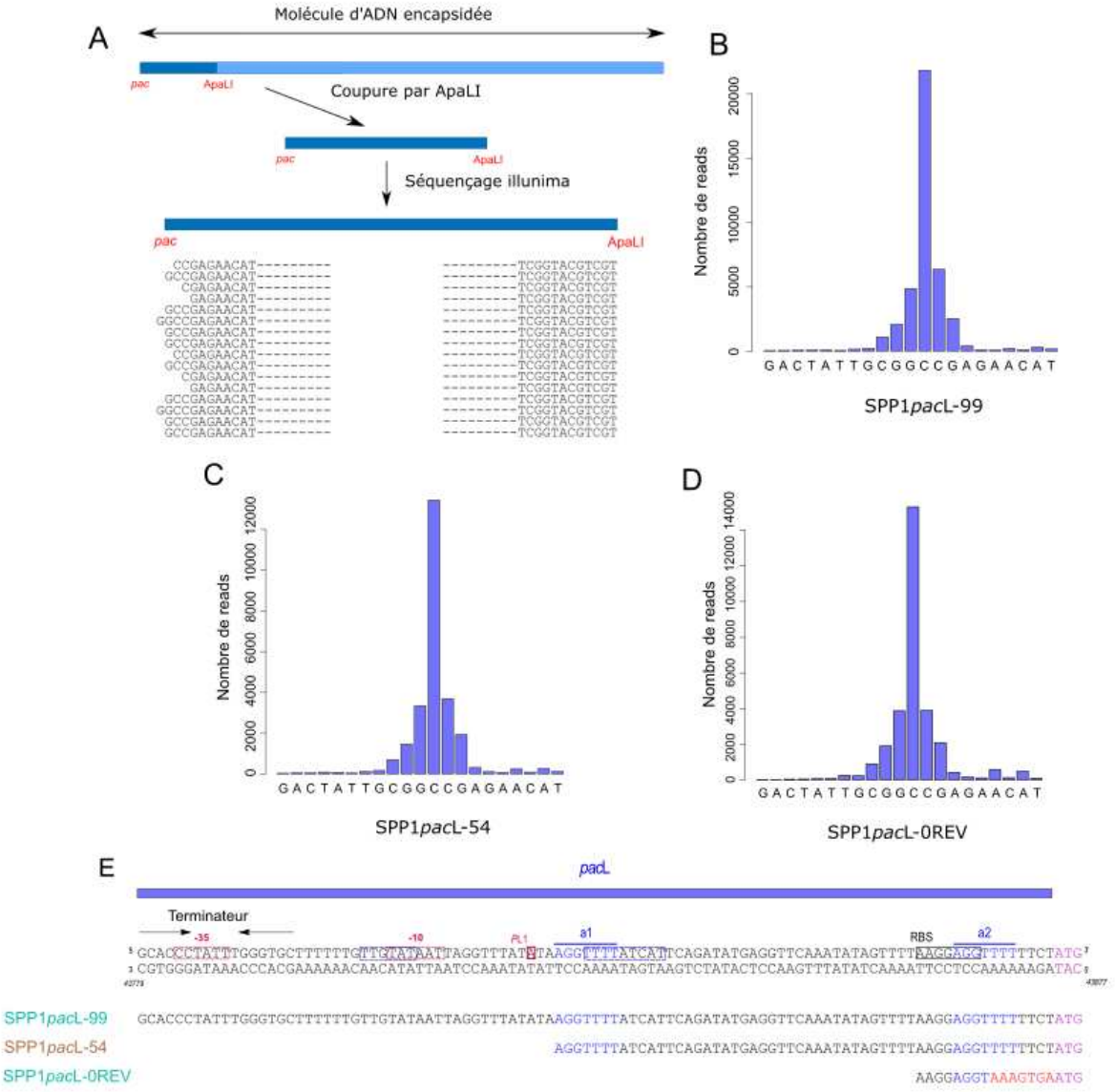


Figure 29. Détermination de la précision de la coupure *pac* par Illumina. (A) Stratégie utilisée pour déterminer la position de coupure sur *pac* : une molécule d'ADN encapsidée est purifiée puis digérée avec ApaLI, un fragment *pac-Apa*LI est isolé puis séquencé par Illumina, les reads obtenus sont alignés au niveau des extrémités du fragment. Diagrammes de densité représentant la position de la coupure sur *pac* chez SPP1*pac*L-99 (B), SPP1*pac*L-54 (C) et SPP1*pac*L-0REV (D). Chaque barre de l'histogramme correspond au nombre de reads finissant par le nucléotide en abscisse. La séquence des mutants dans la région *pac*L est rappelée en bas de la figure (E), les couleurs employées sont les mêmes que dans la figure 24.

Dans les trois cas, il est clair que le même profil est observé, ce qui indique que la précision de coupure de *pac* est conservée chez l'ensemble des mutants. On remarque que la coupure

n'est pas vraiment précise au nucléotide près, mais qu'elle peut s'étaler de quelques nucléotides de part et d'autre du site majoritaire de coupure (la position où le nombre d'extrémités de fragments est le plus élevé). Ce motif est similaire à celui obtenu au cours des travaux de Djacem et al, (2017), où le phage sauvage avait été analysé de la même manière. Nous pouvons donc en conclure que les mutations présentes sur le génome de ces phages n'affectent pas la précision de coupure de *pac* par gp2. Cependant, il est important de faire remarquer que les résultats n'indiquent pas nécessairement que l'efficacité de coupure est conservée. En effet, nous ne séquençons que de l'ADN purifié à partir de particules virales, nous n'analysons donc que l'ADN qui a été encapsidé. Un problème dans la coupure de la séquence *pac* pourrait altérer le processus d'encapsidation et l'ADN où la coupure ne s'est pas effectuée correctement ne serait alors jamais encapsidé. Avec la méthode employée ci-dessus, il n'est pas possible de détecter ce type de problème. Ce que nous pouvons conclure, c'est que la séquence *pac* est clivée et au bon endroit lorsque les molécules d'ADN sont encapsidées. Ces résultats ne sont pas en adéquation avec ce qui a été observé dans un système plasmidique, quand *pac* est cloné dans un plasmide, il faut que *pac*L fasse au moins 68 nucléotides pour qu'elle soit clivée de façon détectable (Djacem et al, 2017), alors que dans le système phagique, cette séquence peut être supprimée en grande partie et le clivage conservé.

III.I.VII. Détermination de la conservation la coupure de *pac* avec un profil de restriction par NcoI

Afin de déterminer si le clivage de *pac* était conservé chez les mutants dont l'extrémité *pac* n'a pas été séquencée, un profil de restriction a été réalisé à partir d'ADN encapsidés digérés par NcoI. Cet enzyme coupe le génome de SPP1 en 5 positions. Si le clivage de *pac* s'effectue correctement chez les mutants, l'on s'attend à observer un fragment de 4256 pb issu des clivages respectifs de *pac* et du site *Nco*I adjacent. Après digestion, quatre fragments de 15962, 351, 2495 et 9747 pb et ainsi qu'un fragment d'une longueur variable en fin de génome (à cause de la permutation partielle de séquence qui a lieu lors de l'encapsidation) sont générés. Sur les gels de la figure 30, la bande attendue de 4256 pb est observée chez tous les mutants indiquant une conservation du clivage de *pac*. L'intensité de ce signal est plus faible de celle des autres bandes du profil car *pac* n'est clivé qu'une seule fois dans la série de cycles d'encapsidation. Chez le sauvage, un *smear* est observé entre ~13 kpb et ~15 kpb. Ceci est attendu, il s'agit de l'ensemble des fragments générés par clivage du dernier site *Nco*I (figure 30) et par le clivage par tête pleine qui génère des molécules de tailles variables à

cause de son imprécision. Chez les mutants, le *smear* est absent, car la délétion induite par la technique de clonage permet l'encapsidation du site *Nco*I situé juste en aval de *pac*, et donc la génération d'un fragment issu du double clivage des sites *Nco*I en amont et en aval de *pac*. Des différences d'intensités entre les bandes issues du clivage *pac*/*Apa*LI sont visibles entre les mutants, mais elles peuvent simplement être dues à des différences dans la quantité d'ADN déposé. Donc, ces résultats sont plutôt qualitatifs que quantitatifs, et une normalisation des données n'a pas vraiment été possible.

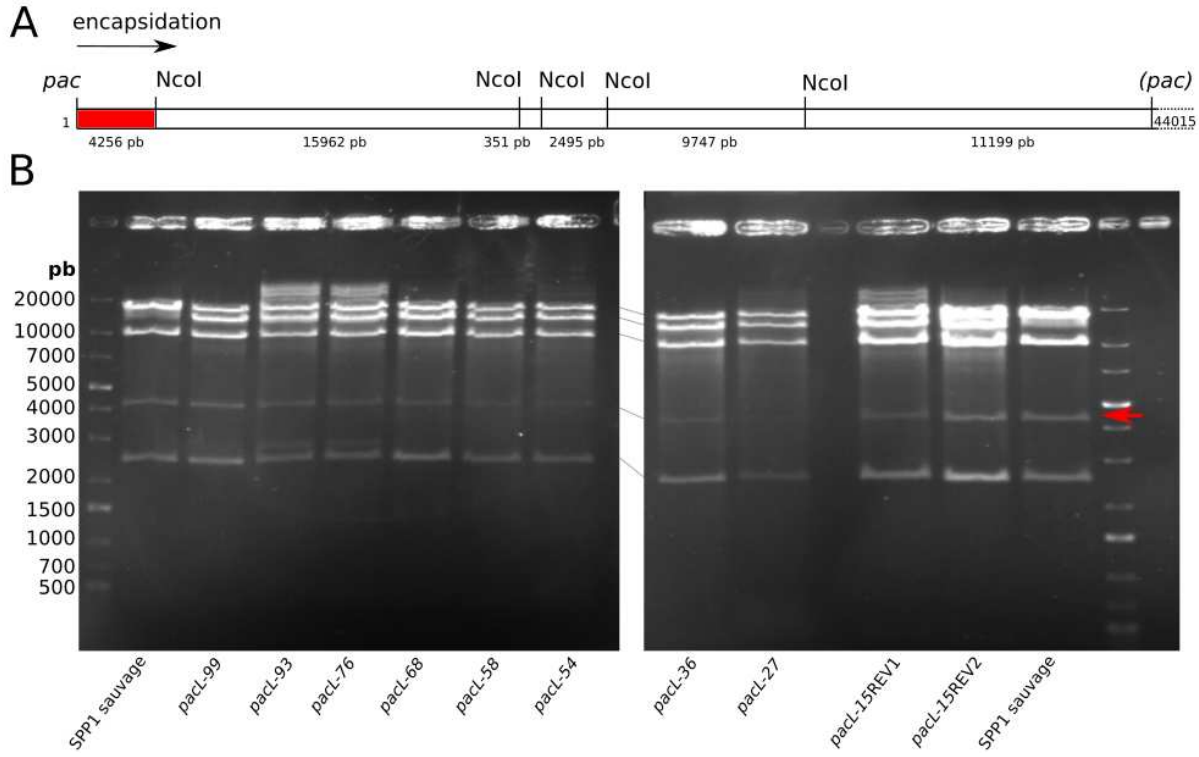


Figure 30. Analyse du clivage sur *pac* par profil de restriction avec NcoI. (A) Carte de restriction du génome de SPP1 clivé par NcoI. Le 1 représente la position de *pac*. Le fragment *pac*-*Nco*I de 4256 pb est présenté en rouge. (B) Profil de restriction NcoI de l'ADN de SPP1 sauvage, SPP1*pac*L-99 et de différents mutants de délétion de *pac*L. Le fragment *pac*-*Nco*I identifié par la flèche rouge est sub-stoichiometrique par rapport aux autres fragments de restriction car il n'est généré que dans le premier cycle d'encapsidation de l'ADN de SPP1 lors d'une série de cycles d'encapsidation par tête pleine (figure 7).

III.II. Rôle de la séquence *pac*R dans la reconnaissance du génome de SPP1

III.II.I. Mutagénèse de *pac*R

Après s'être intéressé à la séquence *pac*L, notamment à son rôle dans la reconnaissance du génome viral et dans la régulation de la transcription de l'opéron des gènes *1* à *7*, nous aborderons dans cette partie le rôle de la séquence *pac*R qui est l'autre région de fixation de gp1 (figure 24). Une approche de mutagénèse dirigée a là aussi été utilisée.

Des mutants possédant des mutations sur *pac*R ont été construits selon la procédure décrite dans la section matériel et méthodes. La séquence *pac*R est plus petite que *pac*L, elle ne couvre que 30 nucléotides, elle ne possède pas d'élément régulateur de la transcription et se situe sur la séquence codante du gène *1* (figure 31.A.B). Comme *pac*L, elle comporte 2 boîtes de séquences identiques (*c1* et *c2* ; figure 31 .C) qui seraient des sites de fixation de la gp1. En plus de ces séquences remarquables, elle possède également un motif poly-A qui a été identifié comme étant important pour la fonction de la terminase dans des expériences en système plasmidique (Djacem et al, 2017). L'ensemble de ces caractéristiques de la séquence *pac*R, nous a permis de mieux déterminer quelles mutations il était possible d'y introduire. Des substitutions ont été réalisées sur les 2 boîtes *c1* et *c2* de *pac*R ainsi que sur la séquence poly-A, en s'assurant de conserver la séquence en acides aminés de gp1 (Djacem et al, 2017 ; ce travail). Il a été possible de construire un tel mutant, les mutations insérées dans son génome sont présentées figure 31.C. Dans la suite du manuscrit, ce mutant sera nommé SPP1*pac*R-0.

III.II.II. Phénotypes de plages de lyse et caractérisation de révertants

Comme précédemment, des titrations ont été réalisées afin de déterminer les capacités infectieuses du phage à partir des phénotypes de plages de lyse. Ce mutant donne de très petites plages de lyse, et c'est un mutant très difficile à amplifier (voir matériel et méthodes section II.VI), ce qui indique que son cycle infectieux est très peu efficace. Une seule étape

d’amplification suffit à la survenue d’une population de phages qui produisent de plus grosses plages de lyse. Comme pour les mutants *pac*L, des phages ont été prélevés à partir de plages de lyse uniques, puis re-titrés pour obtenir des clones purs. Les plages de lyse obtenues à l’issue de ces titrations présentent la même apparence que celles d’un phage sauvage. On en déduit que des révertants sont de nouveau apparus.

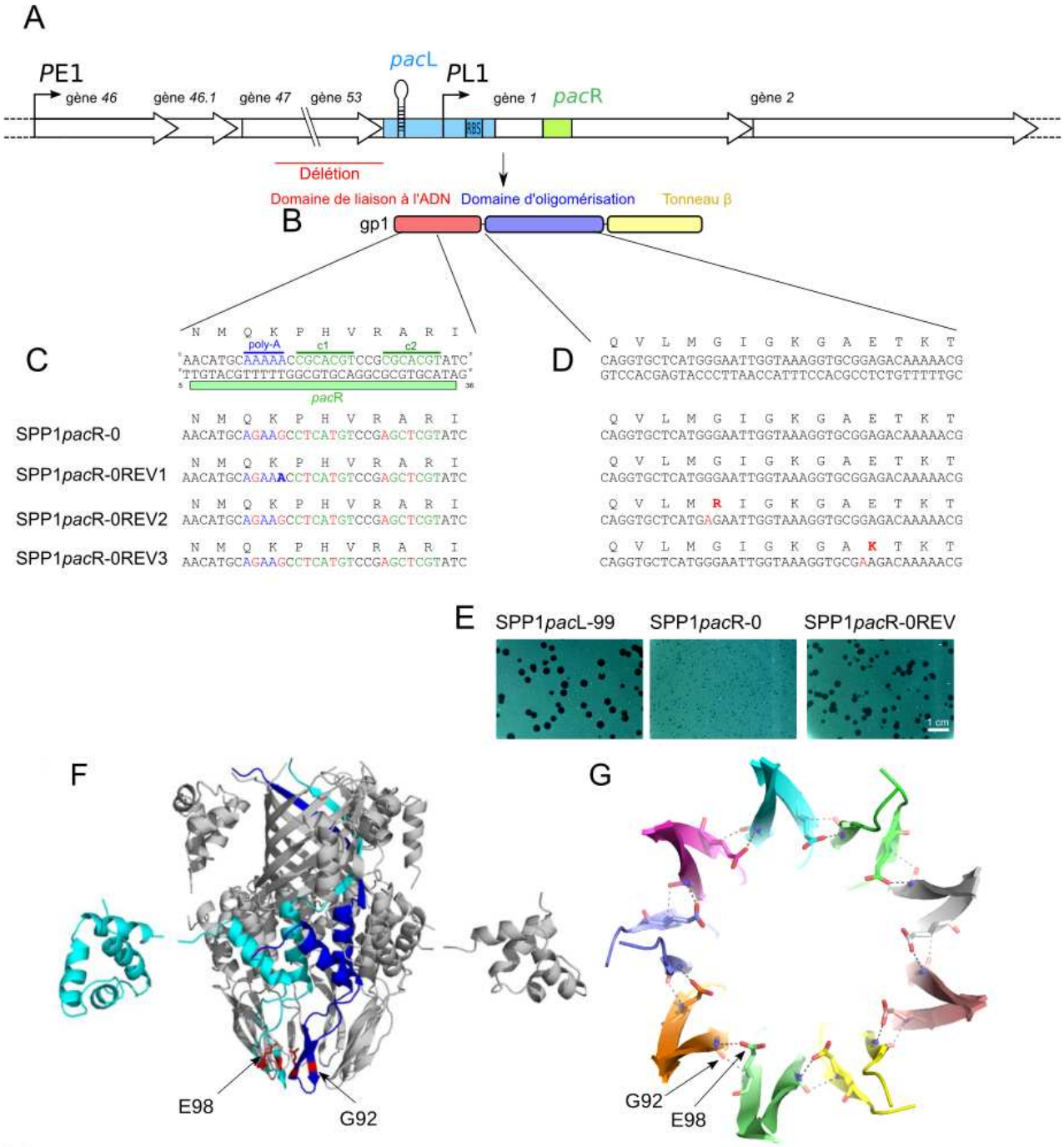

A
PE1
gène 46
gène 46.1
gène 47
gène 53
pacL
PL1
gène 1
pacR
gène 2
RBS
Délétion
Domaine de liaison à l'ADN
Domaine d'oligomérisation
Tonneau β
B
gp1
C
N M Q K P H V R A R I
poly-A
c1
c2
AACATGCAAAAACCGCACGTCCGCGCACGTATC
TTGTACGTTTTTGGCGTGCAGGCGCGTGCATAG
pacR
D
Q V L M G I G K G A E T K T
CAGGTGCTCATGGGAATTGGTAAAGGTGCGGAGACAAAAACG
GTCCACGAGTACCCTTAACCATTTCCACGCCTCTGTTTTTGC
SPP1pacR-0
AACATGCAGAAGCCTCATGTCCGAGCTCGTATC
CAGGTGCTCATGGGAATTGGTAAAGGTGCGGAGACAAAAACG
SPP1pacR-0REV1
AACATGCAGAAACCTCATGTCCGAGCTCGTATC
CAGGTGCTCATGGGAATTGGTAAAGGTGCGGAGACAAAAACG
SPP1pacR-0REV2
AACATGCAGAAGCCTCATGTCCGAGCTCGTATC
CAGGTGCTCATGAGAATTGGTAAAGGTGCGGAGACAAAAACG
SPP1pacR-0REV3
AACATGCAGAAGCCTCATGTCCGAGCTCGTATC
CAGGTGCTCATGGGAATTGGTAAAGGTGCGAAGACAAAAACG
E
SPP1pacL-99
SPP1pacR-0
SPP1pacR-0REV
1 cm
F
E98
G92
G
G92
E98
H
SF6 --MKEPKLSPKQERFIEEYFINDMNATKAAIAAGYSKNSASAIGAENLQKPAIRARIDAR
SPP1 MGEVKGKWTPKLERFVDEYFINGMNATKAAIAAGYSKKSASTIAAENMQKPHVRARIEER
: * :** ***::*****.**************:***:*.***:*** :****: *
SF6 LKEINEKKILQANEVLEHLTRIALGQEKEQVLMGIGKGAETKTHVEVSAKDRIKALELLG
SPP1 LAQMDKKRIMQAEEVLEHLTRIALGQEKEQVLMGIGKGAETKTHVEVSAKDRIKALELLG
* ::::*:*:**:**
SF6 KAHAVFTDKQKVETNQVISLTIAAMQNEENKAF
SPP1 KAHAVFTDKQKVETNQVIIVDDSGDAE------
****************** : :. :

Figure 31. Génotype et phénotype du mutant SPP1*pac*R0 et ses révertants. (A) Environnement génomique de la séquence *pac* (B). Représentation schématique des domaines de gp1 basée sur la structure cristallographique de gp1 du phage SF6 (figure 17). (C). Position des mutations sur *pac*R et leur génotype. Le mutant initial SPP1*pac*R-0 est muté sur une séquence poly-A et sur 2 sites avec une séquence répétée en vert, identifiés comme étant les sites de fixation de la gp1 lors des expériences d'empreinte à la DNAse (Chai et al 1995). Des mutations compensatrices apparaissent sur la séquence poly-A et dans le domaine d'oligomérisation de gp1 (D). (E) Phénotypes de plages de lyse. Les plages de lyse de SPP1*pac*R0 sont très petites. Après une étape d'amplification, des plages de lyse plus grosses apparaissent. L'analyse de clones individuels montre qu'il s'agit de révertants (C,D). (F,G) Structure de la TerS de SF6 (nonamère) vue de côté (F) et du dessous (G). Les substitutions d'acide aminé sur gp1 compensatrices des mutations sur *pac*R sont indiquées pas des flèches. La numérotation des résidus correspond à celle de la gp1 phage SF6. (H) Alignement des séquences protéiques de gp1 de SF6 et SPP1. * Résidu identique; : substitution conservative; . Substitution semi-conservative. Les résidus changés chez les révertants sont colorés en vert.

III.II.III. Analyse des révertants par séquençage

Afin d'identifier les mutations impliquées dans la réversion du phénotype du mutant SPP1*pac*R-0, 8 révertants indépendants ont été isolés. Trois d'entre eux ont été entièrement séquencés par Illumina. Les mutations retrouvées étant toutes dans le gène *1*, nous avons choisi de séquencer la région qui couvre *P*E1, *pac*L et les gènes *1* à *6* dans les 5 autres mutants. Trois types de révertants ont été identifiés : 4 dont un A qui avait été transformé en G dans le poly-A est redevenu un A comme chez SPP1 sauvage (figure 31 C.D) et des révertants avec des mutations ponctuelles dans le gène *1* qui conduisent à des substitutions d'acides aminés dans la protéine gp1. Parmi ceux-ci 2 possèdent une mutation E100K et 3 une mutation G94R (figure 31 .D). La réversion sur la séquence poly-A indique qu'elle est importante pour la reconnaissance du génome viral par gp1. Ceci vient confirmer les résultats obtenus par Djacem et al (2017) où la séquence poly-A était nécessaire au clivage de *pac* dans un plasmide où cette séquence avait été clonée. En revanche, chez les autres mutants une autre stratégie a été adoptée, ce n'est pas la séquence *pac* qui a été de nouveau mutée mais la protéine elle-même. Dans le premier cas, il semble assez évident que la nouvelle mutation sur *pac*R doit restaurer l'interaction entre gp1 et l'ADN. En revanche, rien n'est moins sûr dans le second cas.

Une analyse de la structure de la TerS du phage SF6 dont la séquence protéique est très similaire (figure 31.H), montre que ces deux résidus, G92 et E98, se situent à la base du

domaine d'oligomérisation de la protéine (figure 31.F.G). La séquence de ce domaine est 100 % identique entre la TerS de SF6 et SPP1. Ils interagissent ensemble en formant des liaisons hydrogène, une entre la chaîne latérale du glutamate et la chaîne principale de la glycine, une autre entre les deux chaînes principales. Ces liaisons font environ 2.7 Angström. Le remplacement de la glycine par une arginine ajoute des charges positives à l'oligomère. De plus, cet acide aminé possède une chaîne latérale beaucoup plus volumineuse qui devrait repousser les résidus présents à proximité et potentiellement altérer la structure de la protéine. Le remplacement E100K conduit aussi à un changement par un résidu de charge positive et élimine le pont hydrogène inter-sous-unité entre la chaîne latérale de E100 et la chaîne principale de G94. Peut-être que ces mutations, en modifiant la structure de gp1, changent la spécificité de gp1 pour l'ADN viral. On pourrait effectivement imaginer que, comme dans le cas des révertants mutés sur le poly-A, l'interaction est rétablie, mais aussi l'opposé. Une gp1 moins spécifique aurait tendance à se fixer sur différentes régions du génome et initier l'encapsidation à partir de celles-ci. Si cette hypothèse est vraie, il devrait en être de même pour le génome de l'hôte, une gp1 moins spécifique devrait reconnaître plus fréquemment le génome de *B. subtilis*. Afin de tester cette hypothèse, nous avons (1) cherché à déterminer si la coupure *pac* était toujours aussi précise et si elle avait lieu au même endroit (2) expérimentalement mesuré la transduction de marqueurs de résistance à un antibiotique lors d'infections par les phages révertants.

III.II.IV. Détermination de la précision de la coupure de *pac* par Nanopore

La précision de la coupure *pac* du mutant SPP1*pac*R-0REV3 a été déterminée en utilisant une méthode sensiblement différente de celle utilisée pour les mutants *pac*L. Ce mutant ayant été séquencé par Nanopore pour les besoins des expérimentations sur la transduction dans les sections suivantes, nous avons utilisé ces données de séquençage pour déterminer la précision de coupure de *pac*. En effet, le séquençage Nanopore permet de générer de très longs reads. Il est donc possible de sélectionner les reads suffisamment longs pour couvrir toute une molécule encapsidée de SPP1, et de retenir parmi eux seulement ceux qui commencent par *pac*. Il est ainsi possible de procéder comme précédemment en cartographiant la position de début des reads d'intérêts. Le diagramme de densité obtenu avec cette méthode est présenté figure 32. Il est important de noter la différence d'échelle entre les deux histogrammes où le

nombre de molécules commençant sur la région *pac* du génome de SPP1*pac*R-0REV3 (à droite dans la figure 32B) est ~175x inférieur à celui du phage contrôle (à gauche dans la figure 32B). On ne retrouve pas la distribution observée pour les mutants *pac*L. Ceci indique que la séquence *pac* n'est plus reconnue correctement fort probablement parce que l'interaction entre *pac* et gp1 n'est pas restaurée par la présence de la mutation E100K. Ce résultat ne peut pas être dû à un biais méthodologique, puisque nous l'avons testé sur le phage SPP1*sus115delX110*, dont on sait qu'il reconnaît *pac* et un motif similaire à celui des mutants *pac*L. Afin de tester d'une autre manière l'hypothèse d'une perte de spécificité entre *pac* et gp1, des expériences de transfert d'un marqueur de résistance à un antibiotique par transduction lors d'infections par les mutants SPP1*pac*R-0REV ont été menées.

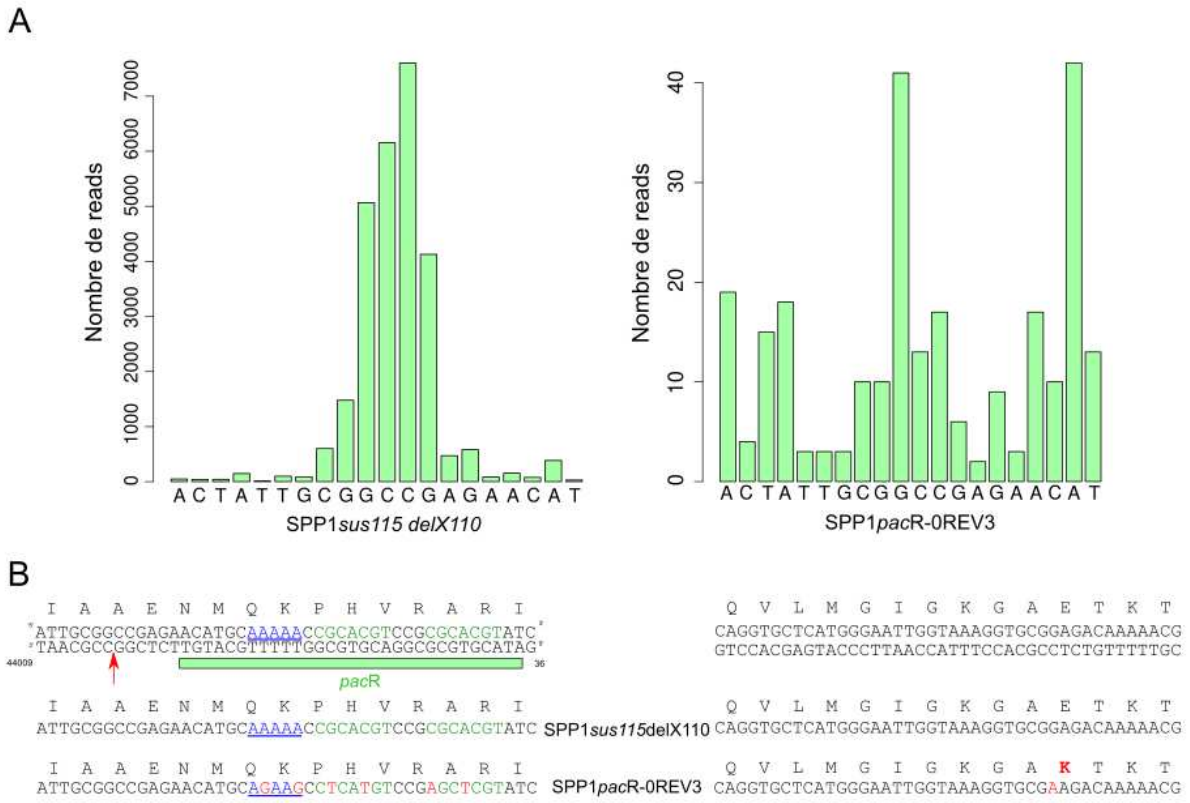


Figure 32. Diagrammes de densité de la position de la coupure sur *pac* chez SPP1*sus115delX110* et SPP1*pac*R-0REV3. Chaque barre de l'histogramme correspond au nombre de reads obtenus par séquençage Nanopore finissant par le nucléotide en abscisse. La séquence des mutants dans la région *pac*R et celle avec la substitution sur gp1 sont rappelées en bas de la figure.

III.III. Analyse de la transduction chez les mutants *pac*L et *pac*R

III.III.I. Transfert de marqueurs de résistance à un antibiotique

Nous avons vu que les mutants sur *pac*L continuaient à couper correctement la séquence *pac.* On en a déduit que la reconnaissance de *pac* avait lieu correctement malgré les délétions de *pac*L. Les expériences menées n'ont cependant pas permis de déterminer le niveau d'efficacité de la coupure puisque seules les molécules encapsidées ont été analysées (voir partie I.VI). Il est donc fort possible que la spécificité du complexe terminase pour l'ADN de SPP1 soit quand même impactée. Chez les mutants *pac*R, il est clair que la spécificité pour *pac* est perdue puisque les diagrammes de densité obtenus à partir de données de séquençage Nanopore montrent clairement que la séquence *pac* n'est plus clivée au même endroit (figure 32). Une autre approche pour tester la spécificité de la terminase pour l'ensemble de ces mutants, serait de voir s'ils sont capables de reconnaître de l'ADN étranger ne comportant pas la séquence *pac*. Nous avons donc choisi de mesurer la transduction d'ADN bactérien par des expériences de transfert d'un marqueur de résistance à un antibiotique. Une souche contenant une cassette de résistance est infectée par les phages d'intérêt. À l'issue de cette première infection, le lysat de phages est récupéré puis utilisé pour infecter des cellules ne possédant pas le marqueur de résistance à l'antibiotique. Si le gène contenant la cassette a été encapsidé durant la première infection, il aura pu être transmis aux bactéries sensibles à l'antibiotique lors de la seconde. Il suffit donc de compter les bactéries devenues résistantes à un antibiotique afin de déterminer indirectement la fréquence à laquelle est encapsidé ce marqueur (voir matériel et méthodes section II.XVI. pour plus de détails). Cette méthode atteste de la survenue de tout le processus de transduction conduisant au transfert du marqueur de résistance dans la souche receveuse et non uniquement des erreurs d'encapsidation de la terminase. Il permet cependant indirectement de rendre compte de la fréquence des erreurs. Nous avons choisi de mesurer la transduction de marqueurs provenant du chromosome bactérien et de différents plasmides. Des plasmides se répliquant différemment ont été choisis pour tester si la fréquence d'encapsidation de l'ADN bactérien est influencée par le type de réplication.

Un marqueur de résistance au chloramphénicol a été choisi sur le chromosome de *B. subtilis*, et un marqueur de résistance à la néomycine pour les vecteurs plasmidiques. Le plasmide

pUB110 a été sélectionné car il se réplique par cercle roulant et qu'il est présent en de nombreuses copies dans la bactérie (à peu près 40, Valero-Rello et al, 2017). Le plasmide pBT233 a été utilisé car il se réplique selon un mode thêta. Et enfin, pUB110cop1 a aussi été testé car il se réplique par cercle roulant comme pUB110 mais est présent en nombre de copies similaire à pBT233 dans la cellule (environ 9 copies). La comparaison de leurs niveaux de transduction permettra d'estimer l'influence du type de réplication et du nombre de copies du réplicon dans le processus de transduction.

Le mutant SPP1*pac*L-99 et SPP1 sauvage sont choisis comme contrôles. Le mutant SPP1*pac*L-0REV est sélectionné car il s'agit du mutant ayant la séquence *pac*L la plus courte et qu'il possède une mutation compensatrice. Le clone *pac*L-99.2 est choisi car toute sa séquence *pac*L après *P*L1 est dégénérée, ce qui permettra de voir si le niveau de transduction change lorsque l'on modifie *pac*L indépendamment de l'influence que pourrait avoir le promoteur *P*L1. Il est ainsi possible de déterminer l'éventuel rôle des mutations sur la séquence *pac*L avec et sans le promoteur *P*L1. Ensuite, ont été testés les trois révertants SPP1*pac*R-0 pour déterminer l'influence de leurs mutations respectives.

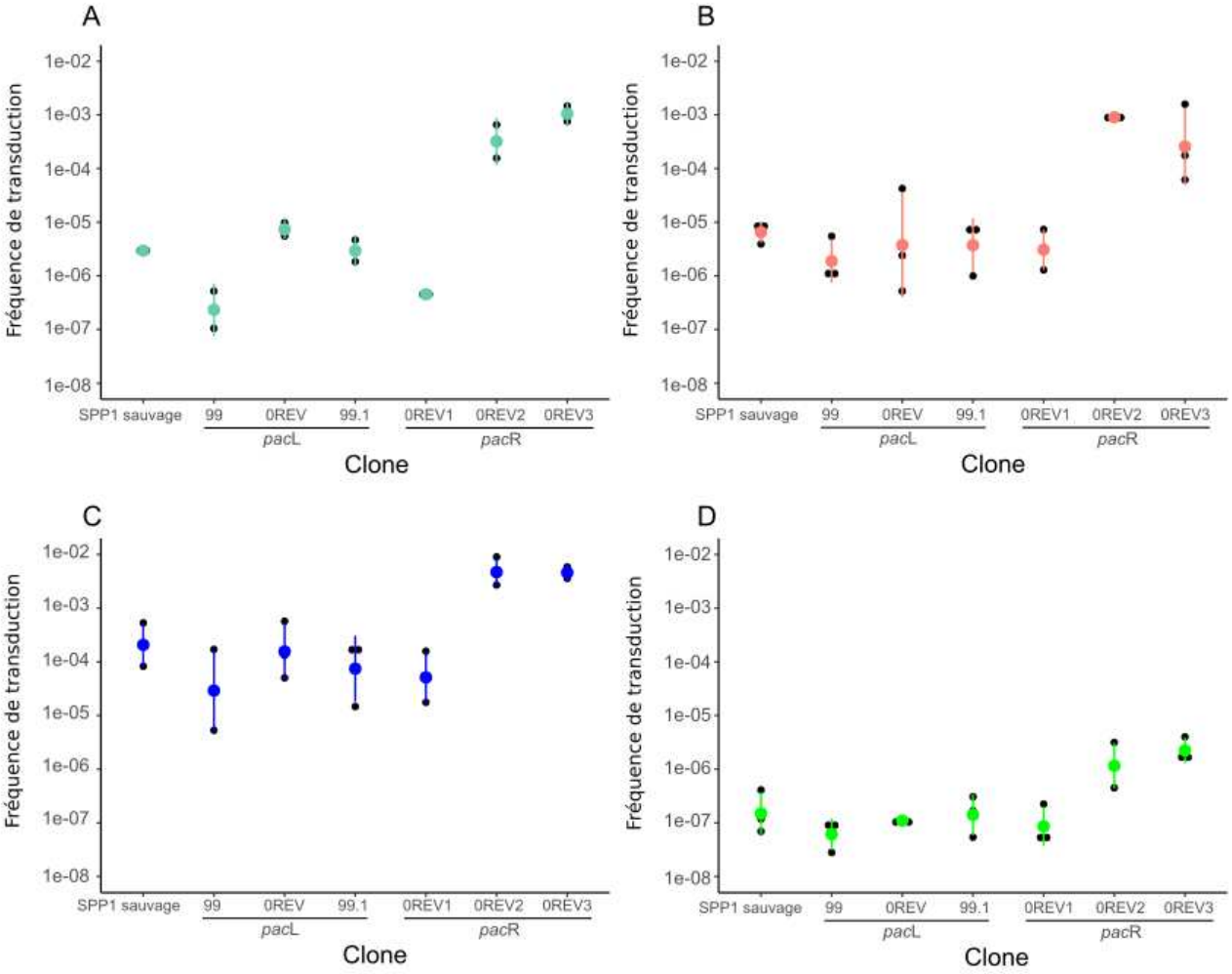


Figure 33. Fréquences de transduction d'un marqueur de résistance à la néomycine dans les plasmides pBT233Neo (A), pUB110 (B) et pUB110cop1 (C) et au chloramphénicol dans le chromosome (D) par différents mutants sur *pac*R et *pac*L. Les expériences sont réalisées en triplicats. Les deux révertants de SPP1*pac*R-0 mutés sur gp1 transduisent les deux marqueurs à une fréquence significativement plus élevée que l'ensemble des autres mutants. Un modèle linéaire généralisé indique que les niveaux de transduction obtenus pour les mutants SPP1*pac*R-0REV2 et SPP1*pac*R-0REV3 sont significativement différents de ceux de l'ensemble des autres mutants avec une p-valeur au moins inférieure à 0.01. Statistiques en annexe.

Les graphiques présentés dans la figure 33 résument l'ensemble des expériences qui ont été effectuées au moins 3 fois. Chez tous les mutants testés les fréquences de transduction obtenues pour les plasmides pUB110cop1 et pBT233Neo sont similaires, ce qui indique que le type de réplication, thêta ou sigma, du vecteur n'a pas d'influence sur le niveau de transduction. La fréquence de transduction du marqueur issu de pUB110 est globalement plus élevée chez l'ensemble des mutants par rapport à celles obtenues pour les plasmides pUB110cop1 et pBT233, ce qui est attendu puisque ce plasmide est présent en nombre plus important dans les cellules bactériennes. La probabilité d'encapsider le marqueur est donc

plus importante car il est en nombre de copies plus grand dans la bactérie. En ce qui concerne le marqueur de résistance au chloramphénicol présent sur le chromosome, c'est l'inverse qui est observé. Ceci s'explique par les mêmes raisons que précédemment, le chromosome étant unique ou en 2 copies si l'ADN chromosomal de *B. subtilis* se réplique, le marqueur est moins souvent encapsidé que dans les autres conditions. De plus, il faut également qu'il puisse être intégré dans le chromosome de la bactérie receveuse, ce qui requière un double événement de recombinaison. Cela pose d'ailleurs un problème d'ordre statistique, puisque dans certaines conditions, il n'est pas possible de compter plus de 20 colonies résistantes au chloramphénicol (voir matériel et méthodes section II.XVI.). Ensemble, ces résultats démontrent sans ambiguïté que les 2 mutants SPP1*pac*R-0 ayant des substitutions d'acides aminés dans gp1 transduisent bien plus d'ADN bactérien que l'ensemble des autres mutants, et ce quel que soit le type de marqueur considéré. La fréquence est souvent de 50 à 100 fois supérieure chez ces derniers par rapport à l'ensemble des autres clones. Les statistiques réalisées démontrent que cet écart est statistiquement très significatif (voir figure 33 et annexe).Ce phénomène a l'air indépendant du type de réplication de l'ADN transduit. Dans la partie suivante, seront présentées des analyses de séquençage à haut débit permettant de déterminer les fréquences auxquelles sont encapsidés différents types d'ADN bactériens.

III.III.II. Mesure de fréquences de transduction du génome bactérien à partir de données de séquençage à haut débit

Afin de confirmer les résultats précédents avec une approche différente, des séquençages Illumina de différents mutants et SPP1 sauvage ont été analysés. Les 2 clones SPP1*pac*L-0REV1 et SPP1*pac*L-0REV2 séquencés précédemment dans l'objectif de déceler des mutations compensatrices des délétions de *pac*L (voir partie IV.I) ont été intégrés dans ces nouvelles analyses. Le mutant SPP1*pac*R-0REV3 a aussi été séquencé avec la technologie Nanopore qui génère des reads de la taille d'une molécule encapsidée (voir parties IV.II). Ainsi, il est possible de déterminer la séquence entière de molécules contenues chacune dans une capside. Enfin, SPP1*pac*L-99 a été utilisé comme contrôle. L'objectif a été de détecter l'ADN de *B. subtilis* dans des échantillons issus d'une extraction d'ADN génomique de SPP1 à partir de particules virales purifiées (voir matériel et méthodes section II.VIII.) et ainsi de dénombrer l'ensemble des paires de reads qui s'alignent sur le génome de *B. subtilis* et celles

qui s'alignent à la fois sur le génome de *B. subtilis* et celui de SPP1. Ensuite, afin de voir si certaines régions du génome de *B. subtilis* sont encapsidées plus fréquemment que d'autres, la couverture du génome de *B. subtilis* a été déterminée. La couverture correspond au nombre de fois que chaque position sur le génome a été séquencée.

La figure 34 représente les proportions de reads s'alignant sur le génome de SPP1, celui de *B. subtilis* ou les deux, on parle alors de jonctions. Chez les révertants SPP1*pac*L-15 et SPP1*pac*L-99, le niveau d'ADN de *B. subtilis* encapsidé est très faible, puisque l'on ne retrouve qu'au maximum 4 % des reads qui s'alignent sur le chromosome de *B. subtilis*. Les révertants semblent encapsider sensiblement plus d'ADN de *B. subtilis* que le clone SPP1*pac*L-99, le pourcentage de reads de *B. subtilis* allant jusqu'à 4% chez les clones SPP1*pac*L-15REV contre moins de 1% pour SPP1*pac*L-99. Difficile de dire si cette différence est significative puisque chaque génome n'a été séquencé qu'une seule fois. Cependant, si l'on se fie aux résultats des expériences de transfert de marqueurs de résistance à un antibiotique, aucune différence significative n'a été décelée. En revanche, l'analyse des séquençages des mutants SPP1*pac*R-0REV2 et SPP1*pac*R-0REV3 offre un contraste plus saisissant puisque 20% des reads alignés correspondent à de l'ADN bactérien. Ce résultat est confirmé par Nanopore pour l'un des mutants SPP1*pac*R-0REV (voir partie III.IV.II.). Ceci est entièrement cohérent avec ce qui a été précédemment démontré. Le pourcentage de jonctions détectées chez l'ensemble des mutants est très faible au regard du nombre de reads correspondant à de l'ADN bactérien chez chacun d'entre eux, ce qui laisse suggérer que l'encapsidation démarre directement au niveau du chromosome bactérien dans la grande majorité des cas. Les erreurs n'impliqueraient que très rarement des événements de recombinaison mais plutôt une mauvaise fixation du complexe terminase qui reconnaîtrait directement l'ADN bactérien. Les résultats obtenus avec le séquençage Nanopore de SPP1*pac*R-0REV3 viennent confirmer cette hypothèse, puisque très peu de jonctions sont détectées et qu'elles correspondent pour la plupart à des artefacts de séquençage (nous avons détecté des molécules de plus de 80 kpb qui correspondent en réalité à des ADN qui se sont ligués entre eux lors de la réalisation des banques). Cet aspect du phénomène de transduction est abordé avec plus de précision dans la partie III.IV. de ce manuscrit.

La couverture du génome de *B. subtilis* par les reads issus du séquençage Illumina de chacun des phages dont la couverture est suffisamment importante est présentée figure 34.A.B.C, elle est globalement homogène sur l'ensemble du génome, et ce pour tous les mutants. La couverture du génome est bien sûr plus importante pour les mutants qui encapsident plus

souvent de l'ADN bactérien. Pour les phages avec une couverture très importante, on note une légère augmentation progressive de la couverture au niveau de l'origine de réplication qui se situe au début et à la fin de l'axe des abscisses sur les graphiques. Cette augmentation s'explique par le simple fait que lors de l'infection, le chromosome de certaines bactéries est en cours de réplication, et donc que les parties proches de l'origine de réplication, qui se répliquent les premières, ont plus de chance d'être encapsidées car elles sont présentes en plus grand nombre. L'ensemble de ces résultats suggère donc que l'encapsidation du génome bactérien s'effectue de manière aléatoire et que la probabilité d'encapsidation d'un locus donné dépend de son nombre de copies dans la cellule.

A

	Pourcentage de reads		
	SPP1	*B. subtilis*	Jonctions SPP1/ *B. subtilis*
SPP1 sauvage	99,265	0.731	0.004
SPP1*pac*L-99*	99,927	0.070	0.003
SPP1*pac*L-15REV1*	95,637	4,335	0.028
SPP1*pac*L-15REV2*	94,920	5,057	0.023
SPP1*pac*R-0REV2	75,605	24.259	0.135
SPP1*pac*R-0REV3	75,584	24,301	0.115

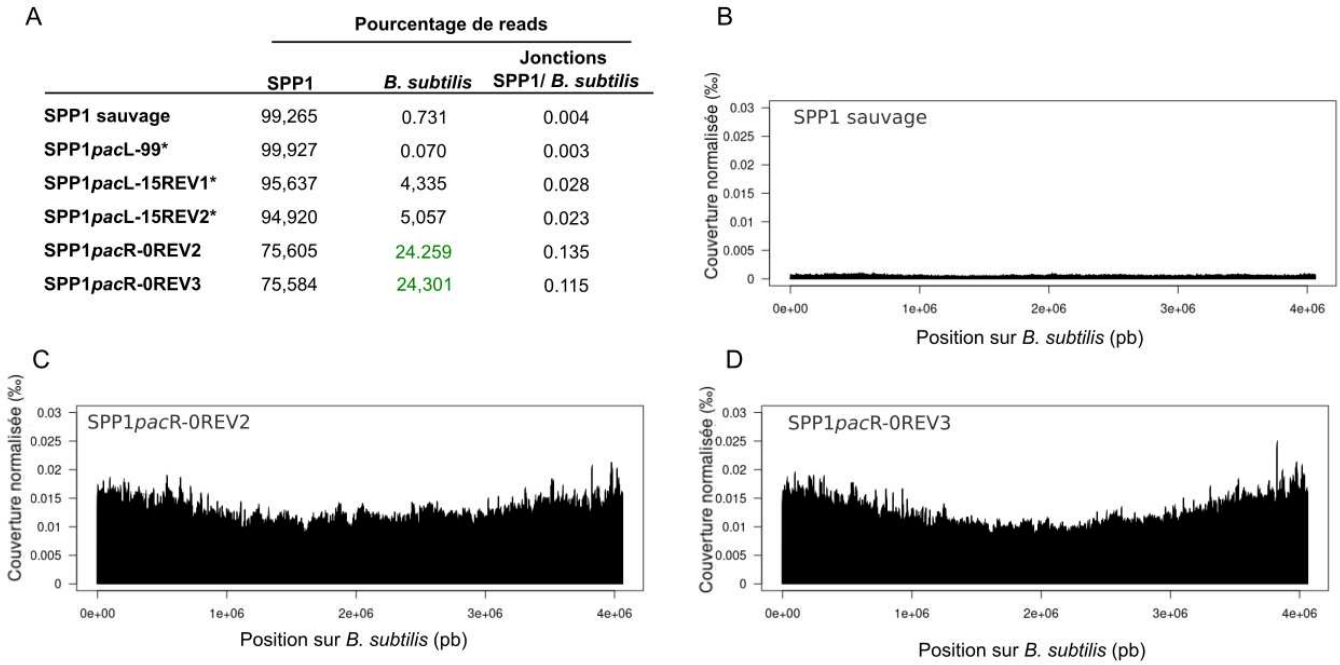


Figure 34. Analyse de la transduction des révertants SPP1*pac*R-0 à partir de données de séquençage Illumina. (A) Tableau représentant les proportions de paires de reads s'alignant sur le génome de SPP1, celui de *B. subtilis* ou les deux. *couverture sur *B. subtilis* trop faible pour réaliser des histogrammes. Histogrammes représentant la couverture du génome de *B. subtilis* en nombre de reads par position sur le génome chez SPP1 sauvage (B), SPP1*pac*R-0REV2 (C) et SPP1*pac*R-0REV3 (D).

L'ensemble de ces résultats suggèrent donc que les substitutions d'acides aminés G94R et E100K dans gp1 réduisent la spécificité de la terminase pour le génome de SPP1 et facilitent en conséquence l'encapsidation d'ADN bactérien. Une question demeure, est-ce seulement les substitutions d'acides aminés sur gp1 qui sont responsables de cette augmentation de transduction ? En effet, les mutations induites sur *pac*R pourraient tout autant être

responsables de cette augmentation. Étant donné qu'il n'a pas été possible d'amplifier un mutant possédant uniquement les mutations induites artificiellement sur *pac*R à cause de l'apparition de révertants, il n'est toujours pas possible de répondre à cette interrogation. Un moyen efficace pour déterminer la réelle influence des substitutions sur gp1 serait de créer des mutants à partir d'un génome sauvage n'ayant chacun qu'une substitution dans gp1. C'est ce qui sera présenté dans la partie suivante.

III.IV. Recherches globales sur les mécanismes impliqués dans la transduction généralisée

III.IV.I. Mutations sur gp1 et hyper-transduction

Après avoir artificiellement dégénéré la séquence *pac*R sur le génome de SPP1, nous avons pu isoler des mutants suppresseurs transduisant de l'ADN bactérien à des fréquences bien plus élevées que la normale. Ceci a permis d'identifier deux mutations entraînant des substitutions d'acides aminés sur gp1, G94R et E100K, qui semblaient être à l'origine du phénomène. Afin d'affirmer qu'il s'agit bien des mutations responsables de ce phénotype et d'exclure un effet des mutations sur *pac*R seules ou en synergie avec les mutations sur gp1, des mutants possédant les mutations sur gp1 sans les mutations sur *pac*R ont été créés à partir d'un génome de SPP1 sauvage (voir matériel et méthodes section II.V.).

III.IV.I.I. Transfert de marqueurs de résistance à un antibiotique

Le niveau de transduction de ces mutants a été mesuré en utilisant les mêmes méthodes que précédemment. Des expériences de transfert d'un marqueur de résistance à un antibiotique présent sur les plasmides pUB110, pUB110cop1 et pBT233Neo et le chromosome bactérien ont été réalisées (figure 35). Le niveau de transduction mesuré est presque 100 fois supérieur à celui d'un phage sauvage, ce qui suggère que ce sont bien les mutations sur gp1 qui sont à l'origine de la hausse du niveau de transduction.

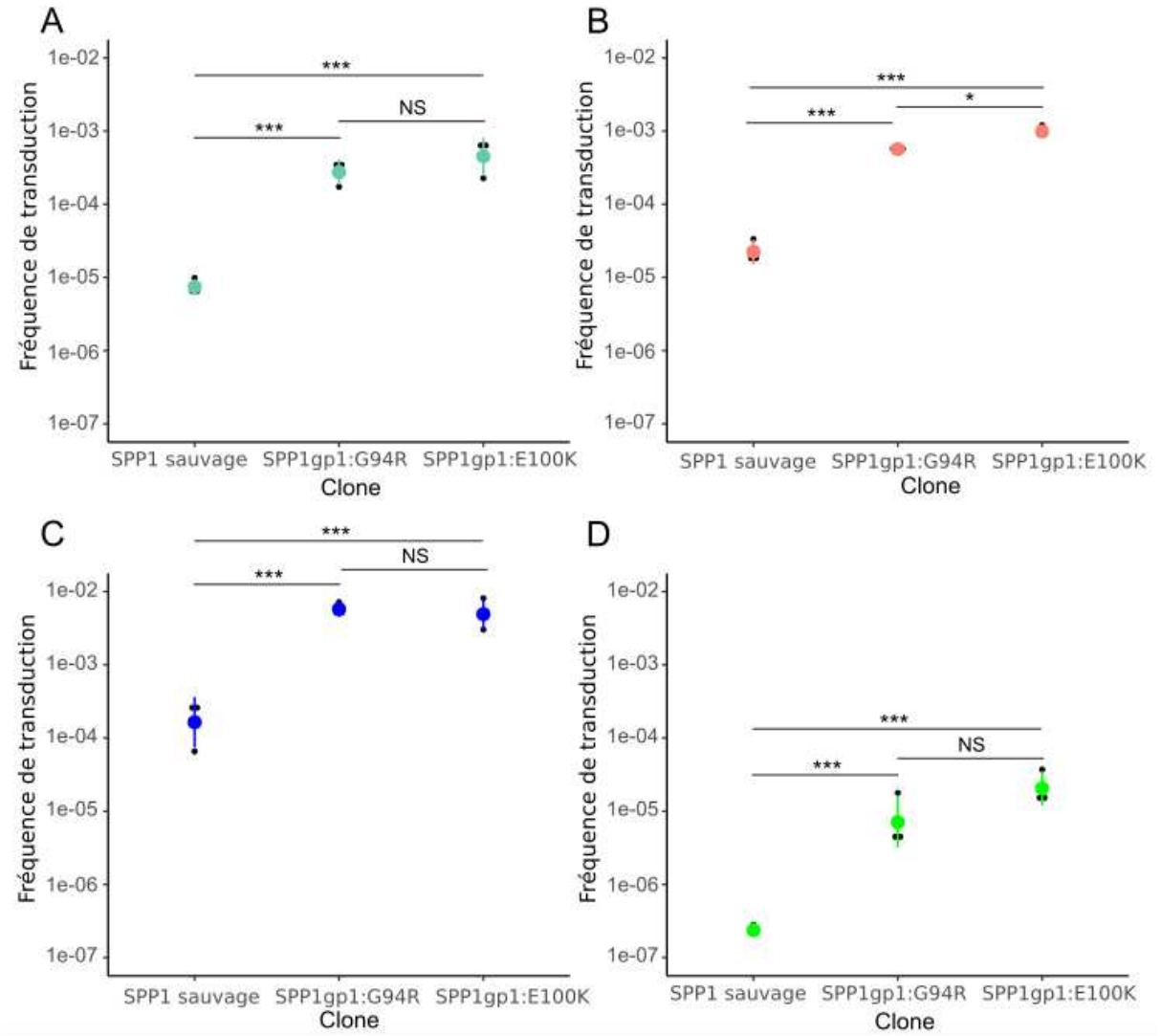


Figure 35. Fréquences de transduction d'un marqueur de résistance à la néomycine dans les plasmides pBT233Neo (A), pUB110 (B) et pUB110cop1 (C) et au chloramphénicol dans le chromosome (D) par les clones mutés sur gp1. Les expériences sont réalisées en triplicats. Un modèle linéaire généralisé est réalisé pour comparer les niveaux de transduction : *** p-valeur<0.001, * p-valeur<0.1, NS : non significatif.

III.IV.I.II. Mesure de fréquences de transduction à partir de données de séquençage à haut débit

Comme précédemment, les mutants SPP1gp1:G94R et SPP1gp1:E100K ont été séquencés par Illumina pour détecter d'éventuelles mutations complémentaires induites par les modifications opérées sur le génome de ces 2 phages. Les reads qui correspondent à de la transduction d'ADN génomique de *B. subtilis* ont aussi été recherchés.

Les résultats des expériences de transfert d'un marqueur de résistance à un antibiotique ont été confirmés avec l'analyse des données de séquençage à haut débit (figure 36). Chez les mutants SPP1gp1:G94R et SPP1gp1:E100K les reads s'alignant uniquement sur *B. subtilis* représentent à peu près 20% du nombre total de reads alignés. Ce qui est un niveau comparable à celui observé chez les révertants SPP1*pac*R-0. Comme précédemment, le nombre de jonctions est relativement faible puisqu'il ne représente qu'en moyenne 0.01 % des reads totaux. Ceci laisse également penser que la transduction ne fait pas intervenir de recombinaison ou de ligation entre l'ADN de SPP1 et celui de *B. subtilis* mais que le complexe terminase initie l'encapsidation directement à partir de l'ADN étranger.

A

	Pourcentage de reads		
	SPP1	*B. subtilis*	Jonctions SPP1/ *B. subtilis*
SPP1 sauvage	99,265	0,731	0,004
SPP1gp1:G94R	82,100	17,887	0,013
SPP1gp1:E100K	77,824	22,166	0,010

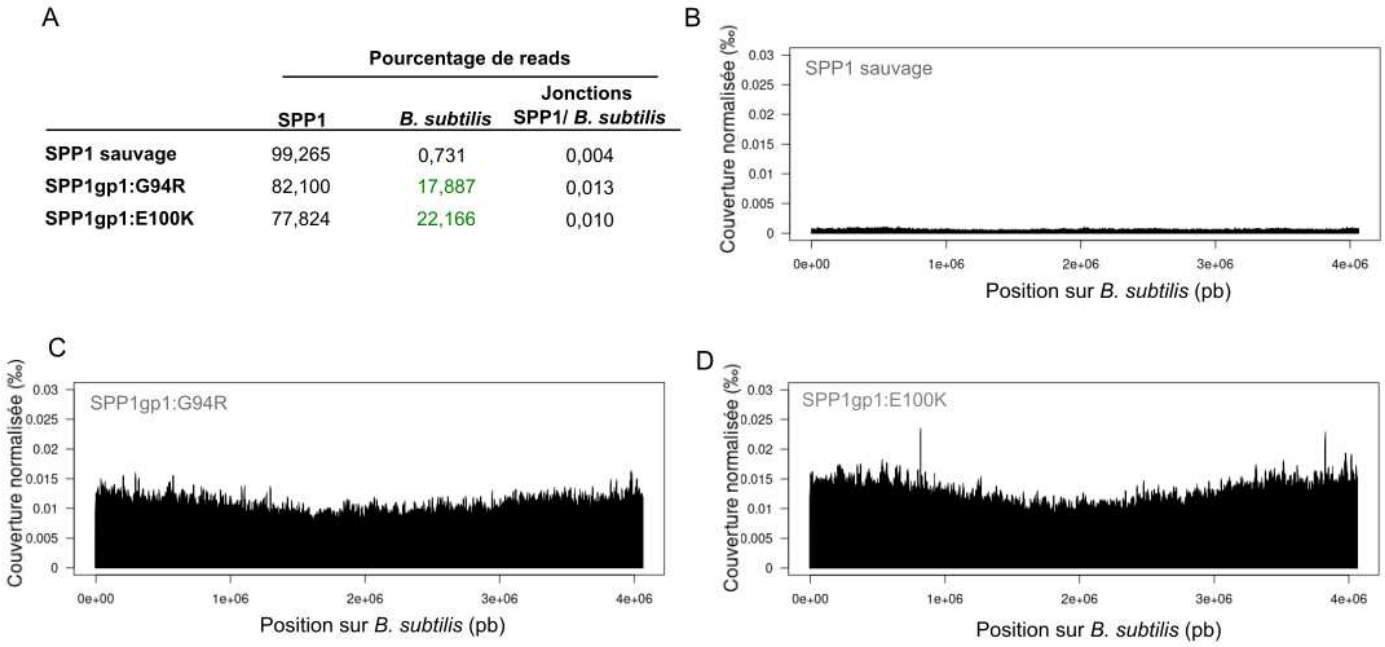
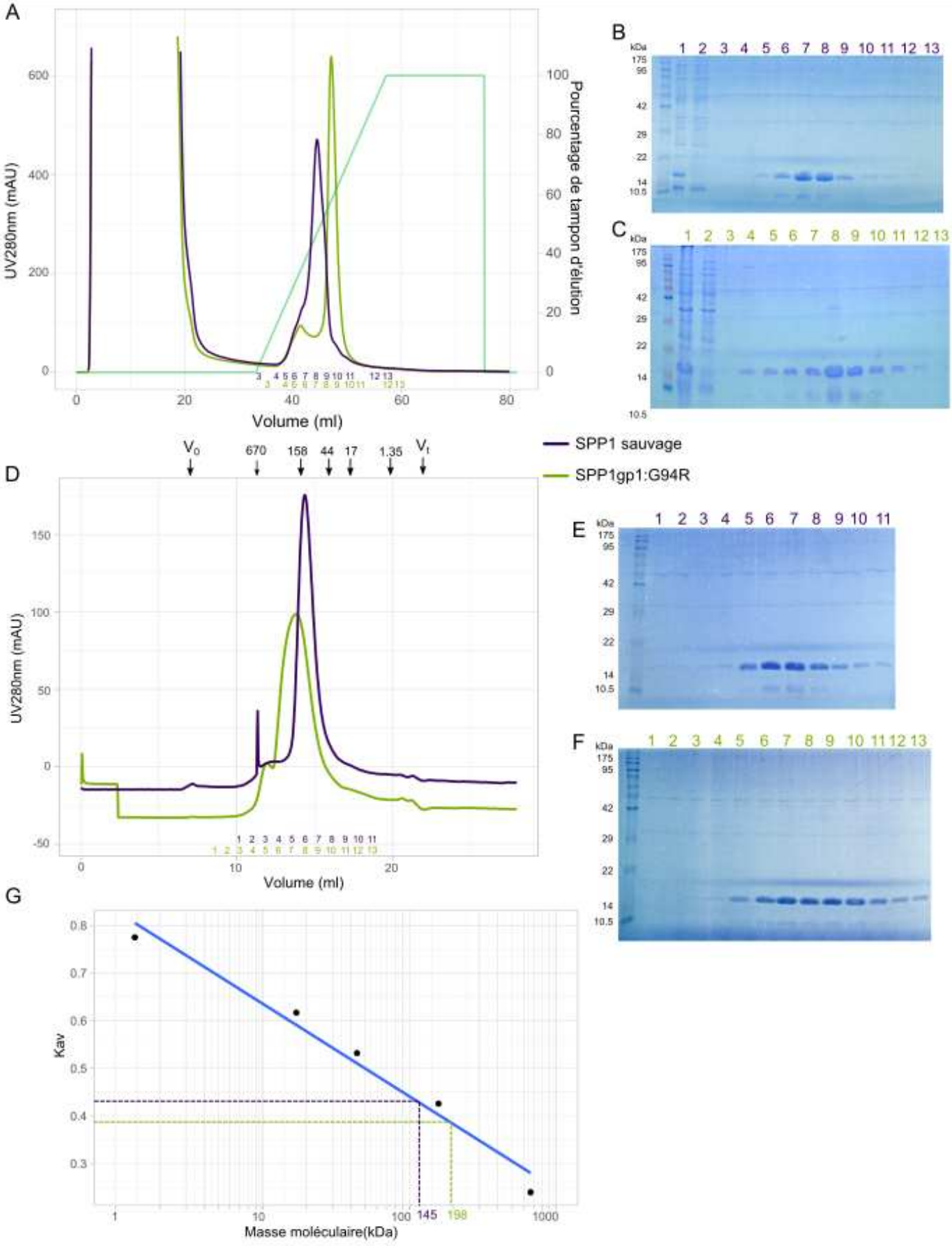


Figure 36. Analyse de la transduction des clones mutés sur gp1 à partir de données de séquençage Illumina. (A) Tableau représentant les proportions de paires de reads s'alignant sur le génome de SPP1, celui de *B. subtilis* ou les deux. Histogrammes représentant la couverture du génome de *B. subtilis* en nombre de reads par position sur le génome chez SPP1 sauvage (B), SPP1gp1 :G94R (C) et SPP1gp1 :E100K (D).

III.IV.I.III. Purification des protéines gp1 sauvage et avec la mutation gp1 :G94R

Afin de mieux comprendre comment les substitutions dans gp1 affectent éventuellement son oligomérisation et son activité, les protéines gp1 sauvage et mutée ont été produites chez *E. coli* à partir d'un plasmide (pBT115) puis purifiées (voir matériel et méthode section II.XX.). Il n'a été possible que de produire la gp1 sauvage et une gp1 avec la mutation G94R. En effet, le changement de la séquence nucléotidique du gène *1* qui génère la mutation E100K dans la protéine ne donne pas de gp1 viable. Ce qui laisse supposer une éventuelle toxicité ou une instabilité de la protéine chez *E. coli*.

La protéine gp1 ayant une forte charge positive, une colonne échangeuse d'ions a été utilisée comme première étape de purification. Les profils d'élution présentés figure 37.A, montrent que la protéine gp1:G94R est éluée un peu après la gp1 sauvage, ce qui peut simplement s'expliquer par le fait que la glycine 94 a été remplacée par une arginine qui est un acide aminé chargé positivement. La grande majorité des protéines cellulaires ne se fixe pas sur la colonne, comme l'atteste l'augmentation de l'absorbance à 280 nm entre 0 et 20 ml lorsque le lysat cellulaire est injecté sur la colonne. Sur le profil d'élution de la gp1, un petit infléchissement est visible avant le pic principal. Il correspondrait à une isoforme de gp1 de charge plus négative que la forme majoritaire, que l'on retrouve d'ailleurs dans les gels colorés au bleu de coomassie figure 37 .B et .C. Bien que son existence ait déjà été démontrée (Oliveira et al, 2005 ; Chai et al, 1994), on ne la retrouve pas chez *B. subtilis*, il s'agit donc probablement d'un artefact de la surproduction de gp1 dans *E. coli*. Après élution, les fractions correspondant au pic d'élution sur le profil figure 37.A sont collectées puis un SDS-PAGE en condition dénaturante est réalisé suivi d'une coloration au bleu de coomassie. Les puits 1 et 2 des gels figure 37.B.C correspondent respectivement à l'échantillon avant passage sur la colonne échangeuse d'ions et à l'échantillon des protéines ne se fixant pas sur la colonne, au moment où l'essentiel des protéines non-désirées est élué (à peu près à 7 ml sur la figure 37.A). On remarque sur les deux gels que gp1 est présente dans l'échantillon 1 mais pas dans le 2, ce qui indique que la protéine a bien été retenue par la colonne. Ensuite, les puits suivants correspondent aux fractions collectées au niveau du pic d'élution de gp1, dont les numéros sont reportés sur la figure 37.A. La coloration révèle une bande de la taille attendue correspondant à la gp1 dénaturée sous forme monomérique (figure 37.B.C).


A
UV280nm (mAU)
Pourcentage de tampon d'élution
Volume (ml)
B
C
kDa
175
95
42
29
22
14
10.5
1 2 3 4 5 6 7 8 9 10 11 12 13
SPP1 sauvage
SPP1gp1:G94R
D
V0
670
158
44
17
1.35
Vt
UV280nm (mAU)
Volume (ml)
E
1 2 3 4 5 6 7 8 9 10 11
F
1 2 3 4 5 6 7 8 9 10 11 12 13
G
Kav
Masse moléculaire(kDa)
145
198

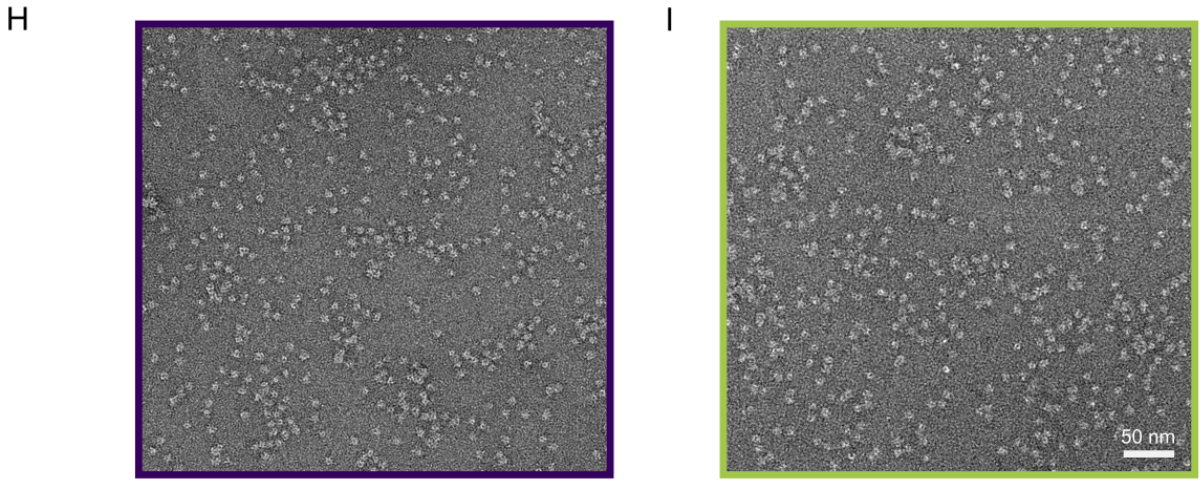


Figure 37. Purifications de gp1 sauvage et gp1 :G94R et observations au microscope électronique en transmission. (A) Chromatogrammes de la purification sur colonne échangeuse d'ions d'échantillons contenant gp1 sauvage (violet) ou gp1 :G94R (vert). La courbe turquoise correspond au pourcentage de tampon d'élution. Les fractions collectées pour le SDS-PAGE sont représentées par des numéros au-dessus de l'axe des abscisses. (B et C) Gels de polyacrylamide (15%) colorés au bleu de coomassie avec la gp1 sauvage et gp1:G94R respectivement : les puits 1 et 2 correspondant respectivement à l'échantillon avant passage sur la colonne échangeuse d'ions et à l'échantillon récolté au tout début de la purification (7 ml), les puits 3 à 13 correspondant aux fractions collectées. (D) Chromatogrammes issus de la purification par exclusion de taille d'échantillons contenant gp1 sauvage (violet) ou gp1 :G94R (vert). Les fractions collectées pour le SDS-PAGE sont représentées par des numéros au-dessus de l'axe des abscisses. Les nombres noirs associés à des flèches indiquent les poids moléculaires en kDa. (E et F) Gels de polyacrylamide (15%) colorés au bleu de coomassie avec les fractions correspondant à la gp1 sauvage et à gp1:G94R respectivement. (G) Relation linaire entre la masse moléculaire des protéines du marqueur de poids moléculaire et le coefficient de partage (Kav) permettant d'estimer la masse moléculaire de gp1 sauvage et gp1 G94R. (H et I) Observation de gp1 sauvage et gp1 :G94R au microscope électronique après coloration négative.

On note aussi que la gp1:G94R a un profil d'élution plus étalé, avec deux pics, que celui de la gp1 sauvage suggérant une certaine hétérogénéité de la protéine mutante. Un signal faible de plus petit poids moléculaire (~ 10 kDa) correspondant à l'isoforme de gp1 est aussi visible dans les 2 gels. Les fractions collectées au centre du pic d'élution sont celles qui contiennent le plus de protéines. Les fractions 6, 7, 8 et 9 de gp1 sauvage et 7, 8, 9 et 10 de gp1:G94R sont concentrées au vivaspin puis purifiées en utilisant une chromatographie par exclusion de taille (voir matériel et méthodes section II.XX.). Dans le chromatogramme présenté figure 37.D, le pic observé correspondant à l'élution de la gp1 possède également un léger infléchissement en début d'élution qui est plus marqué dans le cas de gp1:G94R. Celui-ci correspond possiblement à une forme minoritaire de gp1 de haute masse moléculaire. La gp1:G94R a un profil d'élution plus étalé que celui de la gp1 sauvage (Figure 37D-F).

La protéine gp1 sauvage nonamérique fait 144 kDa. Après calcul du coefficient de partage (Kav), on obtient une masse moléculaire de 145 kDa pour gp1 sauvage, ce qui correspond à la taille attendue (figure 37.G). La gp1:G94R est éluée un peu avant la gp1 sauvage. On pourra supposer que la mutation déstabilise la structure rendant ainsi la protéine plus volumineuse. Les résultats semblent potentiellement indiquer que gp1 :G94R forme des oligomères avec un nombre de sous-unités différent du sauvage, grâce au calcul du coefficient de partage, on peut déduire que la protéine ferait environ 198 kDa, ce qui est bien au-dessus de la masse moléculaire de la protéine sauvage (figure 37.G). Un monomère faisant à peu près 16 kDa, on peut déduire que la protéine forme potentiellement des dodecamères. Il n'a pas été possible de détecter des monomères dans les 2 conditions. Donc, gp1:G94R forme sûrement elle aussi exclusivement des oligomères.

Pour obtenir plus d'informations concernant une éventuelle différence dans la structure des 2 gp1, les fractions ayant le plus de protéines (voir gels figure 37.E.F) ont été observées au microscope électronique en transmission à l'aide d'une coloration négative. Les images sont présentées figure 37.G.H. À cette résolution, la différence entre les 2 protéines n'est pas flagrante, il faudra pousser l'analyse plus loin pour détecter d'éventuelles modifications de structure.

III.IV.II. Étude des mécanismes impliqués dans la transduction généralisée par séquençage Nanopore

Notre étude nous a permis d'identifier des mutations uniques sur gp1 qui rendent SPP1 hypertransducteur suggèrant que la transduction est le résultat d'une erreur qui conduit le complexe terminase à initier l'encapsidation directement sur l'ADN bactérien. Afin de comprendre plus en détails les mécanismes qui régissent les phénomènes de transduction chez SPP1, la technologie de séquençage Nanopore a été utilisée. En effet, ce type de séquençage permet de générer des reads de la longueur d'une molécule d'ADN encapsidée de SPP1, ce qui n'est pas possible avec un séquençage Illumina classique où l'ADN est fragmenté par sonication. Ceci permet non seulement d'identifier des événements de transduction, mais également d'analyser au nucléotide près la séquence des molécules contenant de l'ADN de l'hôte. Nous avons choisi d'analyser trois conditions différentes. Dans le premier cas, afin d'étudier les phénomènes de transduction facilitée, une bactérie possédant le plasmide pSEA contenant le gène *6* de SPP1 est infectée par un phage SPP1*sus115delX110* qui ne produit pas la gp6 et possède une délétion de 3419 nucléotides dans son génome. À noter que la mutation

sizA sur la gp6, qui complémente l'absence de gp6 chez SPP1*sus115delX110,* affecte le senseur de tête pleine lors de l'encapsidation du génome de SPP1 réduisant la taille des molécules encapsidées (Tavares et al 1992). Dans un deuxième cas, une bactérie possédant le plasmide pUB110 est infectée avec le phage SPP1*delX110* qui possède la même délétion que précédemment mais une gp6 fonctionnelle. Cette condition permettra d'étudier la transduction généralisée non-facilitée d'un plasmide qui ne possède pas d'homologie avec l'ADN de SPP1. Enfin, dans un dernier cas nous avons choisi d'infecter des cellules avec le mutant SPP1*pac*R-0REV3 pour vérifier si l'on retrouve le même niveau de transduction qu'avec le séquençage Illumina et identifier des jonctions SPP1/*B. subtilis* en plus grand nombre que dans les conditions précédentes, ce phage ayant un niveau de transduction plus élevé. Les histogrammes présentés figure 38 représentent le nombre de reads qui correspondent à une taille donnée dans les trois séquençages. La majorité des molécules se situe à la taille attendue pour une molécule encapsidée. Notons que dans les figures 38.A et C qui correspondent au séquençage effectué sur SPP1*delX110* et SPP1*pac*R-0REV3, la plupart des reads se situent à 45000 pb alors que dans la figure 38.B, qui correspond au séquençage de SPP1*sus115delX110*, la majorité des reads font 42500 pb. Ceci est dû à la présence de la mutation *sizA* sur la gp6 qui complémente l'absence de gp6 chez SPP1*sus115delX110* qui réduit la taille des molécules encapsidées. Les reads présents à 2500pb de part et d'autre de ces positions (voir figure 38 panneau de gauche) ont été sélectionnés pour les analyses dans les parties suivantes. Le panel du centre est un agrandissement de ces régions et celui de droite les mêmes graphiques après une transformation logarithmique de l'axe des ordonnées. Il est ainsi possible de voir la proportion d'ADN plasmidique en bleu et celle d'ADN de *B. subtilis* en jaune par rapport à celle de SPP1 en magenta. Des données chiffrées sur ces différentes proportions sont présentées dans le tableau 7. Il est très clair que le clone SPP1*pac*R-0REV3 transduit beaucoup plus d'ADN bactérien que les autres phages, ce qui est cohérent avec ce qui est discuté dans la partie III.II. Ce clone ne transduit pas de plasmide, car son hôte n'en possède pas. Le phage SPP1*sus115delX110* transduit bien plus de plasmide que SPP1*delX110*, ce qui s'explique par la présence d'une séquence homologue sur pSEA qui favorise la transduction facilitée. En revanche, le phage SPP1*sus115delX110* transduit un peu moins d'ADN de *B. subtilis* que SPP1*delX110*, mais l'impossibilité d'effectuer des analyses statistiques due à l'absence de réplicat ne permet pas de déterminer si cet écart est réellement significatif.

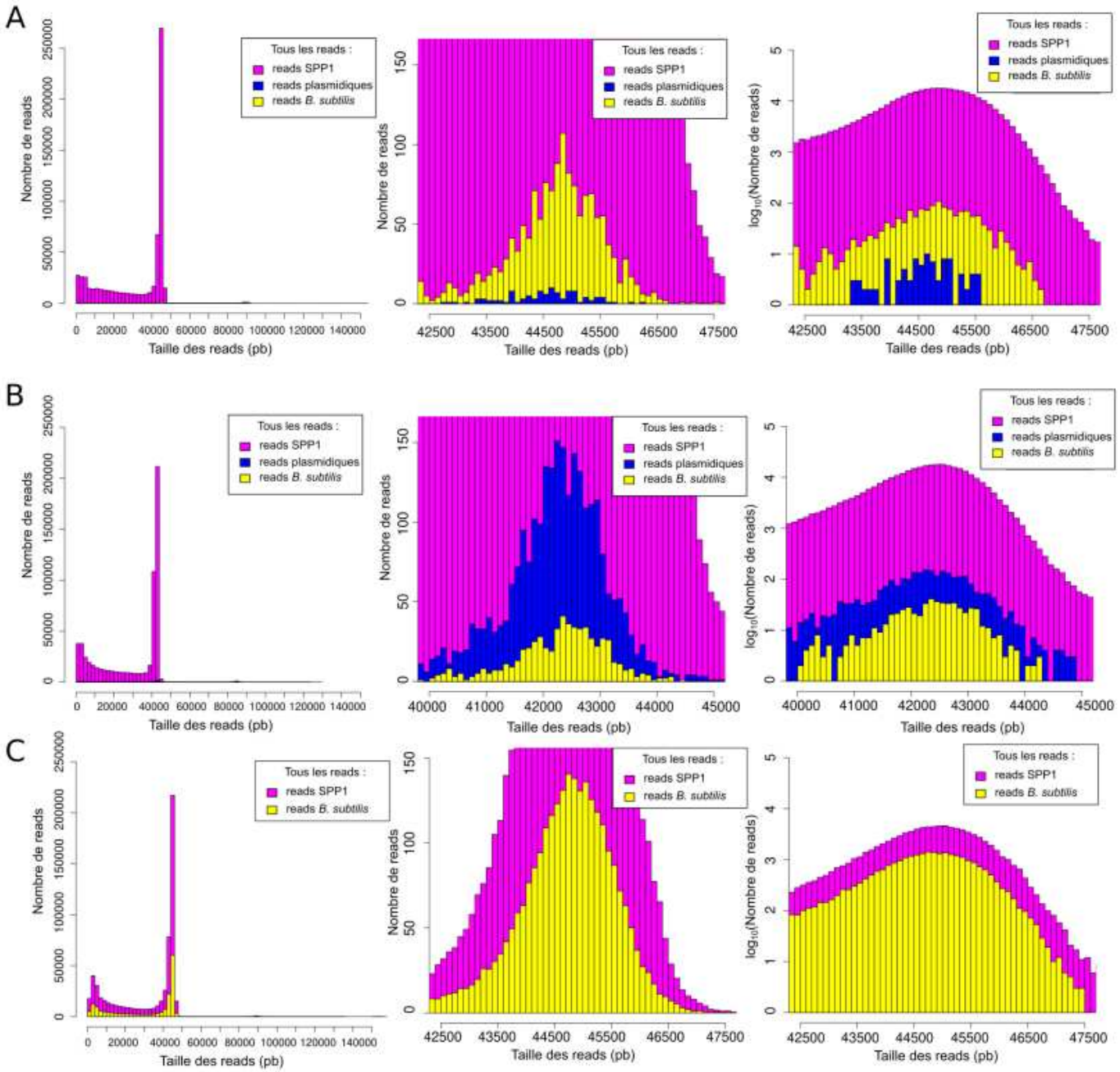


Figure 38. Répartition des tailles de reads de SPP1 et *B. subtilis* dans les séquençages Nanopore. (A) Séquençage SPP1*delX110* pUB110, (B) Séquençage SPP1*sus115delX110*, (C) Séquençage SPP1*pac*R-0REV3. Première colonne, répartition de l'ensemble des reads. Deuxième colonne, agrandissement sur une région de 5000 nucléotides autour de la taille la plus surreprésentée. Troisième colonne, graphiques de la deuxième colonne avec une échelle logarithmique. Le magenta représente les reads de SPP1, le bleu l'ADN plasmidique et le jaune l'ADN génomique de *B. subtilis*.

	Pourcentage de reads				
	SPP1	*B. subtilis*	Plasmide	Jonctions SPP1/ plasmide	Jonctions SPP1/ *B. subtilis*
SPP1*sus115delX110* + pSEA	98,971	0,196	0,833	0,668	0,000
SPP1*delX110* + pUB110	99,540	0,433	0,026	0,000	0,001
SPP1*pac*R-0REV3	76,800	23,387	-----	-----	0,009

Tableau 7. Pourcentage de reads correspondant à de l'ADN de SPP1, de l'ADN de *B. subtilis* ou des jonctions issues d'événements de transduction.

On remarque également, que chez l'ensemble des conditions, le type d'ADN, que cela soit du génome de phage ou de bactérie ou encore de plasmide, n'influe pas sur la taille des molécules encapsidées.

III.IV.III.I. Transduction facilitée de pHP13gp6 :*sizA*

Les capsides générées après infection de YB886+pSEA par SPP1*sus115delX110* sont purifiées et l'ADN qu'elles contiennent extrait. Le séquençage de molécules issues de cette expérience comporte deux avantages. Le premier, est que le plasmide possède de l'homologie avec le génome du phage, et qu'il sera donc possible de voir si de la recombinaison homologue entre de l'ADN plasmidique et de l'ADN de SPP1 est possible et induit de la transduction facilitée. Le second est que le phage possède une délétion *delX110* (voir figure 39) et que l'absence de gp6 provenant de son génome est complémentée par une gp6*sizA* mutante codée par pSEA. Dans ces conditions, la permutation entre les séquences issues de cycles différents est plus grande et la gp6*sizA* génère des molécules encapsidées plus courtes.

La figure 39.A représente le type de molécules transduites théoriquement possibles issues d'événements de recombinaison entre le gène *6* du génome de SPP1 et celui du plasmide pSEA au cours de différents cycles d'encapsidation. Si la recombinaison survient lors du premier cycle, une molécule chimérique est formée. Celle-ci se compose alors d'un fragment de génome de SPP1 allant de *pac* au gène *6* puis d'une suite de plasmides pSEA jusqu'à l'extrémité de la molécule formée par la coupure par tête pleine. En effet, le plasmide pSEA se réplique par cercle roulant (voir matériel et méthodes section III.I.), donc si la recombinaison à lieu lorsque le plasmide se réplique, une molécule contenant plusieurs séquences de plasmides les unes à la suite des autres devrait être encapsidée. Lors des cycles suivants, seul de l'ADN plasmidique est encapsidé. Si la recombinaison a lieu lors du deuxième cycle, le gène *6* étant à la fin des molécules générées, à cause de la permutation

circulaire, les molécules hybrides devraient essentiellement être constituées d'ADN de SPP1 avec une petite partie d'ADN plasmidique. Notons également que les molécules issues du cycle 2 commencent au niveau du gène *6*, de la recombinaison pourrait donc également avoir lieu à ce niveau, mais les molécules issues de tels événements ne sont pas identifiables car la séquence du gène *6* de SPP1 est quasiment identique à celle présente sur le plasmide. Donc, si de la recombinaison intervient à ce niveau, les molécules générées seront en tout point similaires à des molécules seulement composées de plasmide. Lors des cycles suivants, le gène *6* se retrouve également proche de l'extrémité des molécules encapsidées, mais avec la permutation circulaire, le gène *6* aura tendance à s'en éloigner. Au fur et à mesure des cycles d'encapsidation, les molécules hybrides observées auront donc une part de plus en plus importante de plasmide au détriment de l'ADN de SPP1.

Avant de vérifier la véracité du modèle théorique décrit ci-dessus, il a fallu localiser la position des jonctions sur les molécules hybrides SPP1/plasmide. En théorie, elles devraient toutes se situer au niveau du gène *6*. Les gènes *6* du plasmide et de SPP1 étant similaires, les jonctions devraient donc être détectées au début ou à la fin de ces-derniers. Le graphique figure 39.B représente la position des jonctions en fonction de leurs coordonnées respectives sur le génome de SPP1 en abscisse ou sur le plasmide en ordonnée. Le début et la fin du gène *6* sur les 2 séquences sont représentés par des lignes pointillées grises. On voit que la grande majorité des jonctions se situe au début ou à la fin du gène *6* de SPP1 et du plasmide comme attendu.

Ensuite, les séquences des molécules hybrides obtenues à l'issue du séquençage ont été alignées sur une référence contenant 2 séquences génomiques de SPP1 l'une à la suite de l'autre. En effet, les molécules encapsidées étant partiellement permutées et plus longues qu'un génome, la fin de certaines séquences peut s'aligner avec le début du génome du phage. À l'issue de l'alignement, nous avons vérifié si le modèle théorique figure 39.A était exact. La figure 39.C représente l'ensemble des reads alignés sur la référence. L'ADN de SPP1 est représenté en valeurs de magenta de plus en plus foncées au fur et à mesure des cycles et l'ADN plasmidique en bleu. Comme attendu, les molécules issues du cycle 1 sont composées d'une courte partie correspondant au génome de SPP1 allant de *pac* au gène 6 puis d'ADN plasmidique jusqu'à l'extrémité de la molécule formée par la coupure par tête pleine. Les molécules des cycles suivants sont essentiellement constituées d'ADN de SPP1 avec une petite partie d'ADN plasmidique, mais cette tendance tend à s'inverser au fur et à mesure des cycles.

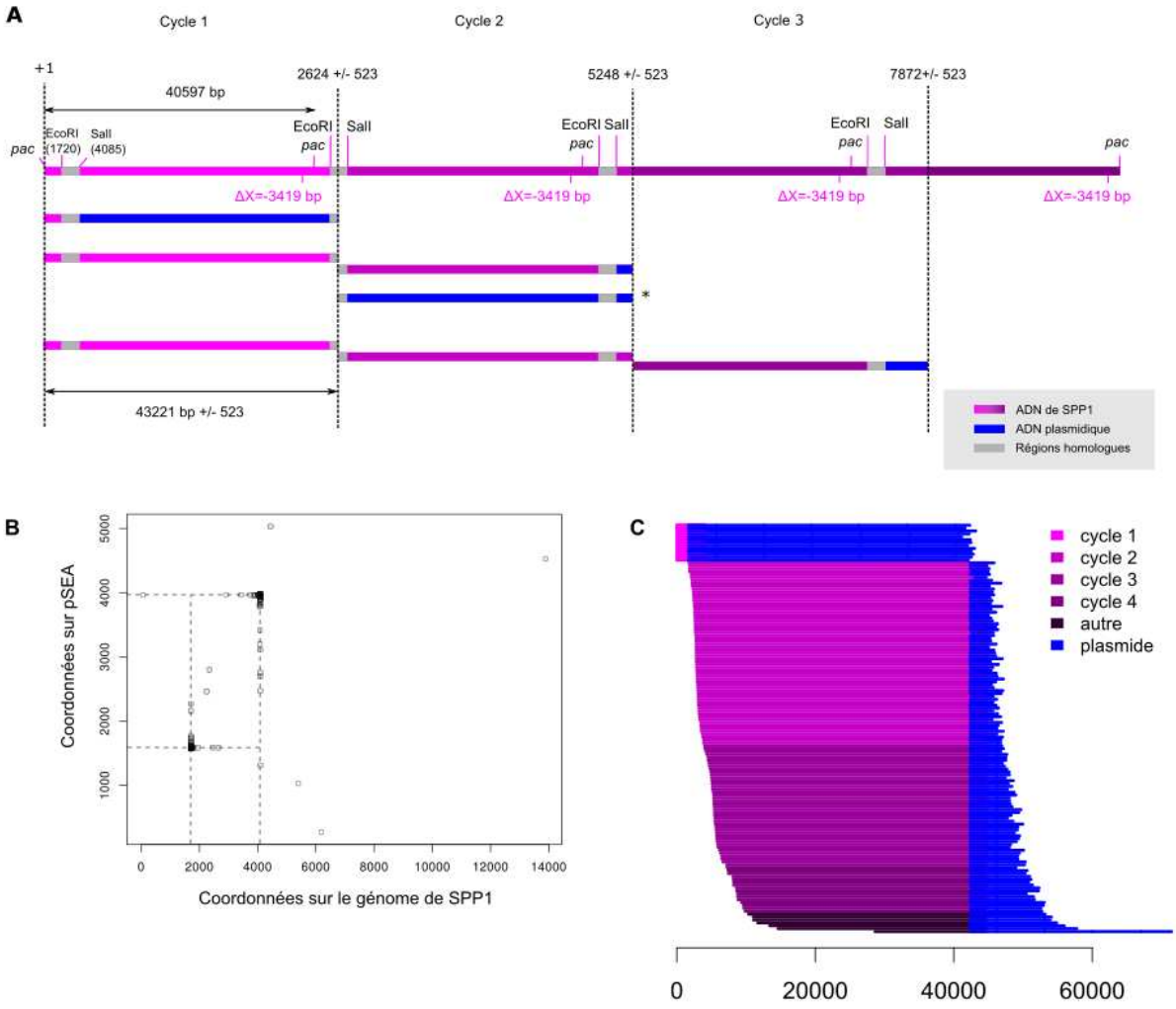


Figure 39. Transduction facilitée de pSEA chez SPP1*sus115delX110*. (A) Types de molécules obtenus par recombinaison homologue au niveau du gène 6 (gris) entre de l'ADN SPP1 (magenta) et le plasmide pSEA (bleu) au cours de différents cycles d'encapsidation (l'ADN de SPP1 est représenté de plus en plus foncé au fur et à mesure des cycles). (B) Coordonnées des jonctions SPP1/plasmide sur SPP1 et pSEA (C) Alignement des molécules hybrides issues du séquençage contre 2 séquences de génome de SPP1, code couleur identique à (A).

Conclusions et discussion

IV. Conclusions et discussion

IV.I. Rôle de la séquence *pac*L pendant l'infection par SPP1

L'organisation génomique de la région contrôlant l'expression de l'opéron terminase-portal (gènes *1* à *7*) est complexe. Celle-ci fait partie de la séquence *pac*L, une des régions qui interagit avec gp1 (Chai et al, 1995). Le terminateur de transcription qui définit la fin de l'unité transcriptionelle contrôlée par le promoteur précoce *P*E1 et le promoteur tardif *P*L1 responsable de la transcription de l'opéron terminase –portal sont localisés dans *pac*L (figure 40A). Au cours de ce travail, nous avons vu qu'il était possible de supprimer toute la séquence *pac*L, à l'exception du RBS du gène *1*, nécessaire pour la traduction de gp1. Les résultats démontrent clairement un effet négatif des mutations sur l'infectiosité des phages. Nous avons donc cherché à comprendre la cause des phénotypes que nous observions. À ce titre, nous avons formulé plusieurs hypothèses : (1) les phénotypes observés proviennent d'un défaut de reconnaissance de *pac*L par gp1, (2) la diminution de l'infectiosité est due à des problèmes de transcription conséquences de la délétion du promoteur *P*L1 et du terminateur en amont (figure 40.B.C), ou (3) les deux hypothèses n'étant pas exclusives elles pourraient être toutes les deux vraies. L'apparition de mutations compensatrices sur *P*E1 nous a plutôt orienté vers la seconde hypothèse, sans pour autant la confirmer totalement. Nous avons effectivement vu que l'absence de *P*L1 et de mutations compensatrices sur *P*E1 conduisaient à une baisse du niveau de transcription (section III.I.IV, figure 40.C). Sûrement que sans *P*L1 la transcription de l'opéron continue quand même à partir de *P*E1 car le terminateur de transcription normalement présent sur *pac*L est absent. *P*E1 étant réprimé en l'absence de mutations compensatrices, le niveau de transcription des gènes de l'opéron des terminases est plus faible que la normale. La figure ci-dessous résume comment la transcription pourrait se voir modifiée en fonction du contexte transcriptionnel. Dans la situation sauvage, quand *P*L1 est présent la transcription a lieu normalement à partir de ce promoteur et *P*E1 est réprimé par un facteur de transcription (figure 40.A). Dans le cas où le terminateur de transcription est éliminé par délétion, l'on s'attend à ce que la transcription de l'opéron terminase-portal soit plus élevée, résultant de la transcription précoce de *P*E1 et tardive de *P*L1 (Figure 40B). Lorsque *P*L1 est partiellement voire totalement supprimé et que *P*E1 est toujours réprimé, le niveau de transcription baisse (figure 40.C). Pour compenser ce défaut, des mutations apparaissent sur la séquence consensus du répresseur qui ne peut plus interagir avec l'ADN au niveau de *P*E1. La transcription peut dorénavant avoir lieu à partir de ce promoteur (figure

40.D). On note aussi que lorsque le terminateur de transcription est éliminé, l'expression de l'opéron terminase-portal est contrôlée par *P*E1, elle devient donc précoce en contraste avec l'expression tardive issue de *P*L1 dans un contexte sauvage.

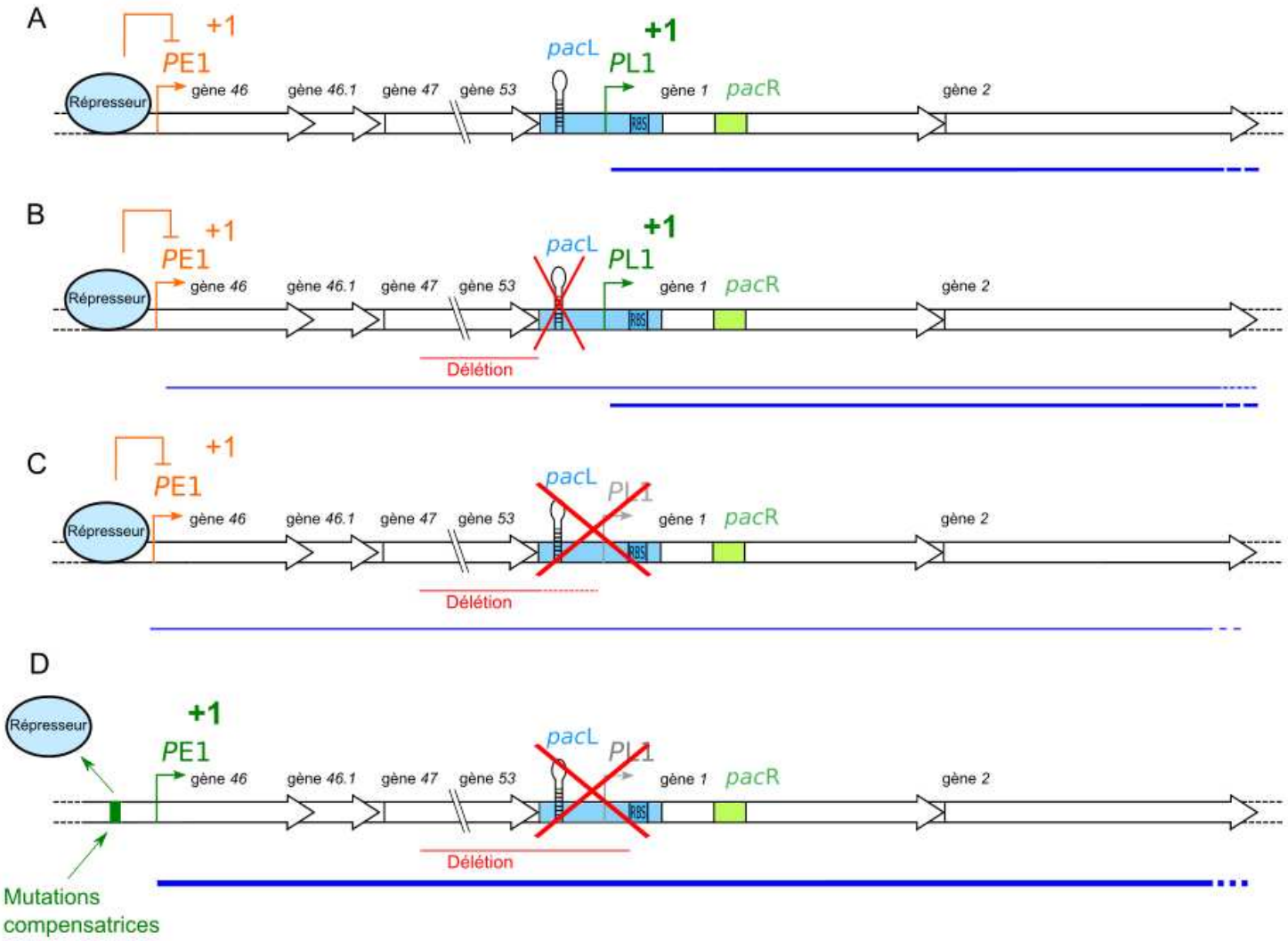


Figure 40. Mécanismes d'expression de l'opéron des gènes *1* à *7* chez les mutants ayant une délétion sur *pac*L et leurs révertants. (A) Sur les délétions plus courtes *P*E1 est réprimé et en présence de *P*L1 la transcription de l'opéron des terminases assure un phénotype sauvage. (B) Les délétions plus importantes réduisent l'activité de *P*L1 et *P*E1 est toujours réprimé conduisant à un phénotype de petite plage de lyse. (C) La délétion de *P*L1 fait que l'opéron des terminases n'est pas assez transcrit à partir de *P*E1 conduisant à un phénotype fort. (D) Des mécanismes compensatoires apparaissent conduisant à la levée de la répression sur *P*E1 et la transcription de l'opéron de la terminase augmente fortement.

En plus d'avoir mis en évidence ces défauts de transcription, nous avons également réussi à prouver que lorsque l'ADN de SPP1 était encapsidé, *pac* était correctement clivé, et ce, pour les mutants avec et sans réversion (section III.I.VI). L'ensemble de ces données suggère donc

que la cause majeure des phénotypes observés est bien un problème de transcription, la reconnaissance de *pac* s'effectuant correctement, il n'y a pas de raison de penser à un quelconque défaut de reconnaissance de cette séquence. Une analyse en profondeur des résultats démontre pourtant que la situation n'est pas si simple. Prenons par exemple les phénotypes de plages de lyse présentés dans la figure 24, le mutant SPP1*pac*L-54 est plus affecté par la délétion que le mutant SPP1*pac*L-58. Pourtant, le promoteur *P*L1 est absent chez les 2 mutants. Dans ce contexte, comment expliquer la différence de phénotype observée ? Il doit donc y avoir autre chose qui rend la délétion de SPP1*pac*L-54 plus problématique que celle de SPP1*pac*L-58.

De même, lorsque l'on s'attarde sur les résultats des qRT-PCR de la section III.I.IV, on se rend compte que la baisse du niveau de transcription des gènes de l'opéron de la terminase est du même ordre de grandeur chez SPP1*pac*L-58 et SPP1*pac*L-54, pourtant le phénotype de SPP1*pac*L-54 est plus fort que celui de SPP1*pac*L-58. Il y a donc autre chose qui affecte l'infectiosité des phages.

On pourra rétorquer que les expériences des sections III.I.VI et VII prouvent que *pac* est correctement reconnu et clivé et que, en conséquence, les phénotypes ne sont pas dus à des défauts dans l'interaction *pac*/terminase. Cependant, ces expériences sont effectuées à partir d'ADN qui ont été encapsidés, on ne regarde donc potentiellement qu'une partie de l'ensemble des ADN synthétisés. Il est fort possible que les ADN n'étant pas correctement reconnus ne soient même pas encapsidés, et donc invisibles dans les expériences réalisées. Ce que nous disent les données obtenues, c'est que lorsque l'ADN est encapsidé, *pac* est correctement reconnu et clivé. Si des défauts de reconnaissance se produisent, l'encapsidation ne doit pas être possible, ou du moins elle ne débute pas à *pac*. Il est en effet possible que l'encapsidation débute plus rarement à d'autres endroits sur le génome de SPP1 mais que ces sites d'initiation ne soient pas surreprésentés dans les données d'Illumina, et donc indétectables lors des analyses. Cependant, tout semble indiquer que la majorité des événements d'encapsidation commencent à *pac*.

Ces résultats suggèrent donc que *pac* est correctement clivé lorsque *pac*L est quasiment absente. Cette séquence n'est donc pas nécessaire pour qu'il y ait reconnaissance spécifique de *pac*. Cependant *pac*L pourrait avoir un rôle dans la fréquence de cette reconnaissance. Elle influencerait donc, non pas la spécificité de l'interaction, mais sa fréquence.

Des effets croisés entre la régulation de la transcription et la reconnaissance de l'ADN viral peuvent aussi être envisagés. Il a déjà été démontré, et les résultats des qRT-PCR sur les mutants *sus19* et *sus70* vont dans ce sens, que la terminase était capable d'affecter la transcription de son propre opéron (Chai et al, 1997). D'une part, la fixation de gp1 sur *P*L1 empêcherait l'ARN polymérase de synthétiser de l'ARNm, d'autre part la coupure de gp2 pourrait rendre impossible la transcription de l'opéron de la terminase en séparant physiquement *P*L1 des gènes qu'il contrôle. De ce fait, il n'y a pas de stricte séparation entre les 2 phénomènes discutés plus haut. Il se peut donc qu'un défaut de clivage de *pac* entraîne une modification de transcription et, qu'inversement, un défaut de transcription affecte l'efficacité de la coupure de *pac*. S'il est donc très difficile, voire impossible de découpler les deux phénomènes, c'est qu'ils sont totalement interdépendants. Il s'agit peut-être également de la raison pour laquelle nos résultats sont différents de ceux obtenus dans un système plasmidique minimal mimant la coupure sur *pac* (Djacem et al, 2017). En effet, dans un plasmide le contexte transcriptionnel est différent. Même si *P*L1 a été conservé, il n'est pas actif car sous le contrôle d'un facteur phagique, et *P*E1 est absent. La transcription est assurée par un promoteur plasmidique en amont des gènes *1* et *2* (Djacem et al, 2017). De plus, en dehors du contexte de l'infection, certains facteurs phagiques pourraient manquer. Nous avons d'ailleurs pu voir qu'un répresseur pouvait jouer un rôle dans la régulation de *P*E1, il n'est pas impossible qu'il soit codé par le phage (Godinho et al, 2018). Peut-être également que la quantité de terminase produite influe sur les niveaux de transcription observés.

Dans un plasmide, la topologie de l'ADN est potentiellement différente de celle adoptée par l'ADN viral qui est encapsidé à partir d'un concatémère linéaire (Figure 7B). Donc, si l'interaction nécessite certains types de structures secondaires, elles ne sont pas présentes dans un système plasmidique.

Dans le cas précis des mutations induites dans le phage sur *pac*L, il est clair que les réversions sur *P*E1 modifient totalement le contexte transcriptionnel de l'opéron de la terminase. D'ailleurs, les niveaux de transcription observés chez SPP1*pac*L-0REV (figure 27) pour les gènes *1*, *2* et *6* sont significativement plus élevés que dans les conditions normales. On aurait donc une augmentation globale du niveau de transcription de ces gènes qui serait suffisante pour compenser les délétions dans *pac*L. Cette augmentation conduit sûrement à l'augmentation de la quantité de protéines synthétisées *in-fine*, ce qui pourrait influencer les mécanismes mis en jeu lors de la reconnaissance et du clivage de *pac*.

IV.II. Rôle de la séquence *pac*R dans la reconnaissance du génome de SPP1

Les résultats obtenus pour la séquence *pac*R sont plus clairs et moins ambigus que ceux obtenus pour *pac*L. Le phénotype observé chez le mutant SPP1*pac*R-0 prouve que cette séquence est très importante et qu'un changement sur celle-ci diminue drastiquement l'infectiosité des phages. L'apparition de révertants avec une mutation compensatrice au niveau de la séquence ploy-A de *pac*R montre clairement que cette séquence doit être déterminante dans l'interaction terminase/*pac*. Cependant, les deux séquences répétées *c1* et *c2*, qui semblaient importantes dans les expériences d'empreintes à la DNAse *in vitro* (Chai et al, 1995), ne le sont pas dans le contexte d'une véritable infection. Aucun révertant ne présente de suppression sur ces séquences. Donc, s'il est clair que le poly-A joue un rôle clé dans la reconnaissance du génome viral, comment expliquer qu'une séquence aussi courte et aussi commune puisse être suffisante pour établir la spécificité de la terminase pour l'ADN viral ? Il y a sûrement d'autres mécanismes en jeu nécessaires à la reconnaissance spécifique de *pac* qui n'ont pas encore été identifiés.

Les autres révertants qui conservent la séquence poly-A de *pac*R dégénérée possèdent des substitutions de nucléotides qui conduisent à la synthèse de protéines avec un changement d'acide aminé sur gp1, G94R ou E100K. Les deux résidus se situent sur le domaine d'oligomérisation de la protéine et interagissent tous deux entre eux en établissant des liaisons hydrogène (figure 31). Les changements doivent donc rompre cette interaction et potentiellement changer la structure de la protéine.

Toujours d'après le modèle issu des expériences d'empreinte à la DNAse (Chai et al, 1995), le nonamère de gp1 enroulerait l'ADN. Il est donc possible que cet enroulement de l'ADN se fasse autour du domaine d'oligomérisation. Même si les domaines de liaison à l'ADN assurent un rôle critique, le domaine d'oligomérisation pourrait aussi être impliqué dans l'interaction avec l'ADN. Ceci est d'ailleurs cohérent avec la présence de nombreux résidus positifs sur sa surface extérieure (Büttner et al, 2012 ; Oliveira et al, 2013).

Nos résultats montrent clairement que la spécificité de l'interaction est perdue chez les révertants avec une substitution sur gp1, les expériences de clivage de *pac* et de transduction le démontrent. Il est tout à fait probable qu'une simple substitution change l'état oligomérique de la gp1, ce phénomène a notamment été observé chez P22 où des mutations dans la protéine

entraînaient la formation d'un decamère au lieu d'un nonamère (Němeček et al, 2007). De plus, il a été démontré que ces mutations dans la TerS de P22 conduisaient à une augmentation de la fréquence de transduction d'ADN bactérien (Cajens et al, 1992), exactement comme ce que nous observons pour la gp1 de SPP1. Les estimations de poids moléculaire obtenues à l'issue de la chromatographie par exclusion de taille semblent indiquer que la gp1 mutante s'assemble en dodecamère, au moins pour la substitution G94R. Ceci impliquerait donc un nombre plus important de domaines de liaison à l'ADN (3 de plus) et un corps central de diamètre plus grand, un peu comme dans le cas de la TerS mutante de P22 (Němeček et al, 2007). Comment expliquer que ce changement dans la structure de la TerS puisse affecter sa spécificité pour *pac* ? Le fait qu'il y ait un nombre plus important de domaines de liaison à l'ADN dans l'oligomère pourrait accroître l'affinité de la protéine pour l'ADN de manière générale, et elle aurait alors tendance à se fixer plus facilement sur un ADN étranger. L'interaction avec la séquence *pac* pourrait également nécessiter un enroulement de l'ADN autour du corps central de la protéine (Chai et al, 1995), il faudrait donc que ce dernier ait un diamètre bien défini. Une augmentation de diamètre pourrait donc perturber son interaction avec *pac*.

Maintenant, comment expliquer que cette perte de spécificité puisse compenser les mutations sur *pac*R. En effet, le phage adopte deux stratégies opposées pour contourner l'effet des mutations sur ce site, soit il restaure la spécificité de l'interaction, soit il la supprime complètement. Si le premier cas semble évident, le second l'est beaucoup moins. Les résultats du séquençage Illumina montrent que le phage avec la gp1 mutante encapside de l'ADN de *B. subtilis* à hauteur de 20% et de l'ADN de SPP1 à 80%, tandis que la quantité d'ADN phagique est seulement ~1,2 fois supérieure à celle d'ADN bactérien lors de l'infection par SPP1 (Labarde et al 2021). Donc, le phage encapside malgré tout beaucoup plus de son ADN, comment l'expliquer ? Soit la spécificité n'est pas totalement perdue, soit la terminase a accès plus facilement à l'ADN du phage. En effet, les mutations produisent peut-être un oligomère qui interagit plus facilement à l'ADN quelle que soit sa composition en nucléotides et qu'il se fixe préférentiellement à l'ADN de SPP1 car il est tout simplement plus accessible dans la cellule. De précédentes études (Labarde et al, 2021) ont démontré que la réplication de l'ADN viral et potentiellement l'encapsidation avaient lieu à des endroits très localisés dans la cellule. Donc, la terminase, qui est synthétisée à partir de l'ADN viral, doit se trouver plus à proximité de l'ADN phagique que de celui de *B. subtilis*, ce qui expliquerait les niveaux de transduction observés.

IV.III. Modèle d'interaction de gp1 avec *pac*L et *pac*R

Dans les deux premières parties, nous avons cherché à expliquer les résultats obtenus après mutagénèse de *pac*R et *pac*L, et d'en déduire certains mécanismes impliqués dans la reconnaissance de l'ADN viral par gp1. Dans cette troisième partie nous intégrerons plutôt ces résultats dans une vision plus globale pour définir un modèle général de l'interaction gp1/*pac*.

Il semble tout d'abord, que dans toutes les séquences modifiées ou supprimées au cours de ce travail, la séquence poly-A soit celle qui est la plus déterminante. Il est donc fort possible qu'il s'agisse de l'endroit où l'interaction de gp1 avec l'ADN débute. Cette séquence n'a rien de particulier et est retrouvée partout dans le génome du phage ou de la bactérie. Cependant, elle se localise précisément au début du gène *1*, ce qui n'est pas anodin. La protéine gp1 pourrait très bien se fixer à l'ADN au cours ou peu après sa propre synthèse au niveau du poly-A de *pac*R. Cela pourrait expliquer pourquoi cette protéine ne va pas se fixer sur d'autre séquences identiques mais spécifiquement sur celle-ci. Des expériences *in vitro* ont permis de démontrer que lorsque des extraits de cellules infectées sont mélangés, l'ADN de SPP1 est toujours encapsidé sélectivement à partir de l'extrait donneur de la terminase (Dröge et Tavares, 2002). Ces résultats vont clairement dans le sens d'une interaction entre gp1 et l'ADN qui a servi à sa propre synthèse.

Peut-être aussi que des séquences qui n'ont pas été mutées entre *pac*L et *pac*R sont reconnues. Toujours est-il que l'interaction débute très certainement au niveau de *pac*R.

Nous avons vu que *pac*L était importante pour la transcription de l'opéron de la terminase, mais qu'elle pourrait tout autant être impliquée dans la reconnaissance de l'ADN viral. Si c'est bien le cas, son implication est sûrement moins importante que *pac*R. Il a déjà été démontré que la gp1 nonamérique reconnaissait plus facilement l'ADN lorsque ce dernier était courbé (Chai et al, 1996). La prédominance de bases A et T sur *pac*L pourrait également être à l'origine d'une courbure de l'ADN qui faciliterait la fixation de gp1 (Chai et al, 1996). Peut-être qu'une délétion de cette séquence pourrait avoir un impact sur la topologie de l'ADN et entraîner des défauts de reconnaissance. On pourra donc imaginer que la fixation d'un premier oligomère sur *pac*R créé une courbure de l'ADN qui facilite la fixation d'un nouvel oligomère et ainsi de suite. Il s'agirait donc d'un mécanisme de fixation coopérative comme il en a déjà été caractérisé dans d'autres systèmes comme parABS ou encore le

répresseur H-NS de *E. coli* (Jalal et al, 2020 ; Ali et al, 2012) . Cependant, ces systèmes font intervenir des protéines dont les structures sont relativement lointaines de celle de la petite sous-unité de la terminase. La plupart d'entre elles forment des dimères dont chaque sous-unité possède un domaine de liaison à l'ADN de type hélice-boucle-hélice. Il n'y a à présent pas de preuve formelle que la petite terminase puisse former des dimères *in vivo*, elle s'assemblerait uniquement en multimères cycliques (Buttner et al, 2012; Oliveira et al, 2013). Les expériences d'empreinte à la DNAse montrent clairement une alternance de zones couvertes et de zones libres régulièrement espacées suggérant clairement un enroulement de l'ADN par gp1, du moins au niveau de *pac*L (Chai et al, 1995). Donc, si plusieurs oligomères interagissent sur l'ADN, ils doivent plutôt le faire un peu à la manière d'un nucléosome. Ce type de modèle pourrait expliquer pourquoi chez les phages Mu et P22 le clivage de l'ADN viral a lieu sur une région plus ou moins éloignée du site de fixation de la petite sous-unité de la terminase (Wu et al, 2002). En effet, si TerS se fixe de manière coopérative sur l'ADN, elle peut le couvrir sur une certaine distance, et alors l'interaction TerS/TerL ainsi que le clivage du génome viral peuvent avoir lieu sur des séquences en amont ou en aval de *pac*.

Une autre question importante se pose également : Est-ce qu'un ou plusieurs facteur(s) issus de l'hôte de SPP1 pourraient être impliqué(s) dans la reconnaissance du génome viral ? En effet, on sait aujourd'hui que c'est déjà le cas chez d'autres phages comme P1 où deux NAP (Nucleoid Associated Protein) IHF et HU sont nécessaires à la reconnaissance du génome viral (Skorupski et al, 1994). De plus, des expériences avec un système *in vitro* composé de gp1, gp2 et *pac* chez *E. coli* (Djacem, 2016) ont montré que *pac* n'était pas clivé dans ces conditions laissant pressentir la possible intervention d'un facteur de *B. subtilis*. Aussi, il n'a jamais été possible de reconstituer la coupure de *pac in vitro* avec seulement gp1, gp2, des procapsides purifiées et un ADN contenant *pac* (Oliveira et al, 2005), ce qui est un argument supplémentaire en faveur de l'intervention d'un facteur de l'hôte.

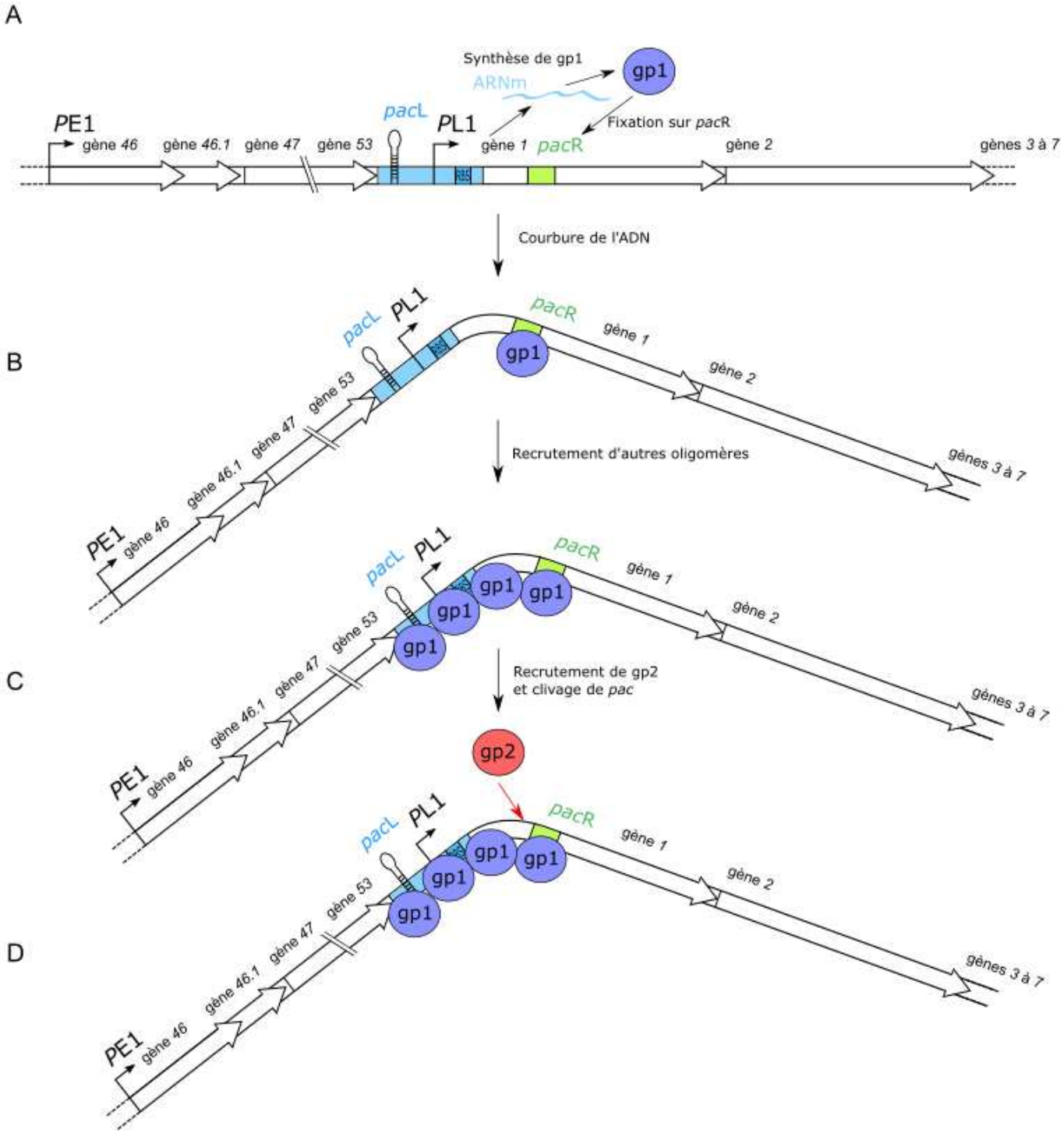


Figure 41. Modèle d'interaction de gp1 avec *pac*. Gp1 se fixe sur *pac*R juste après sa synthèse (A), les changements topologiques induits par la fixation de gp1 sur *pac*R (B) facilitent le recrutement d'autres oligomères de gp1 qui viennent eux aussi d'être synthétisés (C). Gp1 couvre ainsi une large région de l'ADN jusqu'à *pac*L réprimant ainsi sa propre synthèse. La structure du complexe nucléoprotéique est reconnue par gp2 qui clive l'ADN viral entre *pac*R et *pac*L (D).

IV.IV. Mécanismes impliqués dans la transduction d'ADN bactérien

Au cours de nos travaux, nous avons pu étudier deux types de transduction, facilitée et généralisée.

IV.IV.I. La transduction facilitée

La transduction facilitée fait intervenir des événements de recombinaison qui nécessitent la présence de séquences homologues au génome du phage dans la séquence transduite (Canosi et al, 1982 ; Deichelbohrer et al, 1985). Nous avons pu étudier ce type de mécanisme grâce au plasmide pSEA qui contient le gène *6* de SPP1. Dans ce type de transduction, la séquence *pac* semble bien reconnue par gp1 et l'ADN bactérien encapsidé à partir du gène *6*, suggérant la survenue d'événements de recombinaison avant l'encapsidation de l'ADN bactérien. La recombinaison est d'ailleurs sûrement assurée par des facteurs cellulaires de l'hôte. Grâce au séquençage Nanopore, nous avons pu déterminer que les événements de recombinaison pouvaient survenir à n'importe quel cycle d'encapsidation. Chose étrange, il semblerait que les événements de recombinaison lors du cycle 1 (figure 39) soient plus rares que pour les autres cycles. Pour l'instant, ce résultat est difficile à expliquer et devra être reproduit avec une autre méthode. Les résultats montrent très clairement qu'une séquence homologue sur le plasmide augmente très largement la fréquence de transduction, ce qui a déjà été démontré dans d'autres études (Valero-Rello et al, 2017, Deichelbohrer et al, 1985).

IV.IV.II. La transduction généralisée

La transduction généralisée fait sûrement intervenir d'autres mécanismes que la transduction facilitée. On remarque en effet qu'un très faible nombre de molécules hybrides SPP1/*B. subtilis* a été retrouvé dans les ADN encapsidés séquencés et qu'aucune région du génome de *B. subtilis* est transduite à une fréquence plus élevée que la moyenne. Ensemble, ces résultats indiquent que dans la très grande majorité des cas, il ne semble pas y avoir d'événements de recombinaison entre l'ADN de SPP1 et celui de son hôte. L'absence de régions surreprésentées sur le génome de *B. subtilis* (figures 34 et 36) semble indiquer qu'il ne présente pas d'homologie avec le génome du phage. Les résultats du Nanopore montrent également que la plupart des molécules transduites à partir du génome de *B. subtilis* ne présentent aucune trace de l'ADN de SPP1. Tout paraît donc indiquer que l'encapsidation

commence directement au niveau de l'ADN bactérien. Parmi les hypothèses que nous avons émises en introduction (section I.IV.I), la plus probable semble être celle selon laquelle la terminase se fixerait directement sur le génome de l'hôte puis commencerait l'encapsidation à partir de cet ADN. Selon cette hypothèse la transduction serait donc directement la cause d'erreurs de reconnaissance de la terminase. Ceci expliquerait pourquoi il s'agit d'événements très rares en conditions normales. En effet, selon notre modèle, la terminase interagirait directement avec l'ADN qui a servi à sa synthèse, cette interaction en *cis* doit d'une certaine manière séquestrer la terminase et réduire la possibilité qu'elle puisse se fixer à l'ADN de son hôte. De plus, nous avons vu que l'ADN de SPP1 était localisé en des endroits très précis dans la cellule (Labarde et al, 2021), donc la terminase devrait être spatialement plus proche de l'ADN du phage que de celui de la bactérie. Chez les mutants possédant des changements dans la séquence de gp1, le génome bactérien est reconnu bien plus souvent. Très certainement, qu'à cause des changements structuraux induits par les substitutions, l'interaction en *cis* avec l'ADN phagique se fait bien moins efficacement. Ce qui devrait permettre une plus grande diffusion de la protéine dans la cellule.

Ensemble, ces résultats démontrent que des changements minimes sur le génome du phage, (une simple substitution d'un nucléotide) peuvent grandement affecter sa capacité à reconnaître son propre génome et à transduire de l'ADN bactérien. Ces deux aspects sont d'ailleurs interdépendants dans le sens où une baisse de spécificité de SPP1 pour son propre génome conduit à une augmentation de la transduction. La capacité d'un phage à transduire plus ou moins fréquemment de l'ADN bactérien ne dépend donc finalement que de peu de facteurs. Il est donc tout à fait probable que ce type d'événement puisse survenir dans le milieu naturel. Des études montrent d'ailleurs que la fréquence de transduction au sein de certaines populations de bactéries d'eaux douces est assez élevée et qu'elle devrait jouer un rôle prépondérant dans la dynamique de ces populations (Kenzaka et al, 2010). Peut-être donc que ces phages hyper-transductants existent bien dans la nature et, qu'à l'instar des GTA (Gene Transfert Agent, voir section I.II.I et Lang et al, 2012), ils participent à l'évolution des populations qu'ils infectent en permettant l'échange de loci génétiques entre bactéries, ce type de phage posséderait cependant un mode de dissémination différent de celui des GTA. En effet, les GTA sont des prophages dégénérés qui n'ont aucune affinité pour leur propre ADN et qui encapsident systématiquement de l'ADN bactérien (Lang et al, 2012). Donc une fois que le GTA a éjecté l'ADN qu'il contenait dans une nouvelle bactérie, il ne peut pas former de nouveaux vecteurs. Le GTA étant intégré au génome de son hôte, il se multiplie donc

uniquement lorsque la bactérie se divise. Chez les phages hyper-transductant, l'infection produit toujours une majorité de virions viables. Donc, si de tels phages existent dans la nature, ils peuvent coexister avec leur hôte tout en étant virulent et en générant un nombre important de particules contenant de l'ADN bactérien. Ceci peut d'ailleurs être très avantageux pour la bactérie qui peut acquérir de nouveaux gènes utiles à sa survie dans l'environnement, et ce, à des fréquences élevées.

Perspectives

V. Perspectives

V.I. Rôle de *pac*L dans la régulation de la production des terminases

Afin d'avoir une vision plus claire des mécanismes mis en jeu dans les réversions des mutants *pac*L, il faudra également quantifier la quantité de terminases produite au cours de l'infection. Pour l'instant, nous avons uniquement estimé les niveaux de transcription de l'opéron des terminases, il nous manque donc des informations quantitatives sur les protéines produites *in-fine*. Peut-être que des mécanismes de régulation post-transcriptionnels sont également à l'oeuvre et qu'ils jouent un rôle déterminant dans la reconnaissance de *pac*, le clivage de cette séquence pouvant être affecté par les proportions stœchiométriques de chaque sous-unité de la terminase. Il faudrait également pouvoir mesurer l'influence de la synthèse d'ARN à partir du promoteur *P*E1 sur les niveaux de protéines obtenus après traduction. À cet effet, on pourra envisager de réaliser des western-blot à différents temps post-infection, de préférence après 15 ou 25 minutes pour que les résultats soient comparables avec les qRT-PCR. Ces expériences pourront aussi être effectuées avec les mutants SPP1*sus70* et SPP1*sus19* pour confirmer que gp1 et gp2 régulent leur propre synthèse.

V.II. Rôle de *pac*L et *pac*R dans la reconnaissance du génome viral

Nous avons vu que, des 2 sites *pac*L et *pac*R, la séquence *pac*R était la plus importante. Nous en avions déduit que l'interaction entre gp1 et l'ADN devait débuter à cet endroit sans pour autant exclure un rôle de *pac*L. Il est toujours impossible de conclure sur ce qui est concrètement reconnu sur ces sites ou sur les régions de part et d'autre de ces dernières. Les expériences d'empreintes à la DNAse (Chai et al, 1995) nous ont donné des indices sur les séquences reconnues, mais comme nous l'avons déjà vu, ce qui est reconnu *in-vitro* n'est pas vraiment comparable avec ce qui l'est *in-vivo* dans le contexte de l'infection. Nous proposons donc des expériences de ChIP ou ChAP seq réalisées lors d'infections de SPP1 sauvage et de mutants sur *pac*L pour déterminer à quelles séquences se lient précisément les nonamères de gp1. Cela permettra aussi de tester le modèle développé rubrique IV.III. où gp1 couvre de longues régions d'ADN. Des expériences de microscopie électronique avec gp1 lié à l'ADN permettraient aussi de tester le modèle, car si gp1 se fixe de manière coopérative sur l'ADN cela devrait être visible par des techniques d'imagerie de complexes protéine-ADN (Lurz et

Spiess, 1988). Une autre possibilité serait de purifier le complexe gp1/gp2 en interaction avec l'ADN et d'en déterminer la structure par Cryo-électromicroscopie. Cependant, cette approche a quelques limites, notamment parce qu'elle nécessite la formation du complexe *in-vitro* et que nous avons vu que ces expériences donnent des résultats différents de ceux *in-vivo*.

Pour tester l'hypothèse d'une interaction en *cis* de gp1 avec l'ADN, nous pourrons avoir recours à des tests de complémentation dans lesquels *B. subtilis* qui contient un gène qui code pour gp1 est infectée avec SPP1*sus70* qui ne produit pas cette protéine. Si, dans ces conditions, l'infection produit des phages viables, il faudra renoncer à notre hypothèse. En effet, cela démontera que la gp1 produite par la bactérie peut interagir en *trans* avec la séquence *pac* du génome de SPP1*sus70*. On prendra garde à utiliser un gène synthétique sans homologie avec le gène *1* de SPP1 afin d'éviter tout événement de recombinaison.

Pour mettre en évidence un éventuel rôle de la proximité spatiale entre l'ADN de SPP1 et gp1 dans la reconnaissance du génome viral, des expériences de microscopie pourront être réalisées. Nous construirions alors une gp1 couplée à une protéine fluorescente pour voir si elle colocalise avec l'ADN de SPP1 également marqué (voir Labarde et al, 2021 pour la méthode).

L'ensemble de ces études permettront d'apporter une vision intégrée des mécanismes moléculaires complexes qui régissent la reconnaissance spécifique de *pac* par la terminase.

V.III. Etude des mécanismes de la transduction généralisée

Nous avons vu que les événements de transduction impliquaient sûrement des erreurs de la terminase qui se fixerait directement sur l'ADN bactérien. Afin de confirmer cette hypothèse, nous proposons également de réaliser des expériences de ChIP ou ChAP -seq qui permettront de voir concrètement les ADN sur lesquels la gp1 se fixe. Il faudra procéder de la même manière que rubrique II.XVII et réaliser un profil de la couverture du génome de *B. sutilis* comme dans les figures 34 et 36. Si les résultats sont identiques, ils nous indiqueront que gp1 interagit sur tout le génome à des fréquences identiques, alors que les expériences précédentes permettaient seulement de démontrer que tout le génome de *B. subtilis* était encapsidé à la même fréquence.

V.IV. Perspectives générales

Ce travail apporte une compréhension nouvelle sur l'interaction de gp1 avec *pac*, mais les mécanismes qui sous-tendent cette reconnaissance de l'ADN viral restent pour la plupart mystérieux. Nos approches expérimentales ont permis de déterminer les séquences qui jouaient un rôle important, mais elles n'ont pas permis de caractériser directement l'interaction entre les différents facteurs en jeu. En effet, on ne sait toujours pas comment gp1 et gp2 interagissent ensemble, ni comment le complexe nucléoprotéique formé par la terminase lié à l'ADN se fixe au sommet portal de la procapside II, ni quelles différentes conformations adopte l'ensemble de ces facteurs lors de l'encapsidation. Les résultats obtenus au cours de cette thèse donnent cependant des indices sur la manière dont gp1 interagit avec l'ADN viral. Notre modèle de fixation coopérative devra être testé grâce aux différentes approches décrites plus haut, chez SPP1, voire chez d'autres phages *pac* qui pourraient partager les mêmes mécanismes. S'il s'avère juste, il faudra chercher à comprendre comment gp2 est recrutée et pourquoi le clivage a systématiquement lieu au même endroit chez SPP1.

L'étude de la transduction généralisée donne des résultats intéressants qui favorisent l'hypothèse selon laquelle les événements de transduction sont bien dus à des erreurs de gp1 qui se fixe sur l'ADN bactérien. Une fois cette hypothèse confirmée par les expériences décrites plus haut, il faudra aussi déterminer si ces mécanismes sont identiques chez les hypertransductants. Enfin, dans une optique plus large, il serait intéressant de voir si de tels phages peuvent exister dans le milieu naturel et quelle serait leur influence sur le transfert horizontal de gènes au sein des populations bactériennes.

Références

Abedon, S. T., Duffy, S. & Turner, P. E. Bacteriophage Ecology. in Encyclopedia of Microbiology 4257 (Elsevier, 2009). doi:10.1016/B978-012373944-5.00022-5.

Abrescia, N. G. A., Bamford, D. H., Grimes, J. M. & Stuart, D. I. Structure Unifies the Viral Universe. Annu. Rev. Biochem. 81, 795–822 (2012).

Ackermann, H.-W. 5500 Phages examined in the electron microscope. Arch. Virol. 152, 227–243 (2007).

Adelman, K., Salmon, B. & Baines, J. D. Herpes simplex virus DNA packaging sequences adopt novel structures that are specifically recognized by a component of the cleavage and packaging machinery. Proc. Natl. Acad. Sci. 98, 3086–3091 (2001).

Ahi, Y. S. & Mittal, S. K. Components of Adenovirus Genome Packaging. Front. Microbiol. 7, (2016).

Ali, S. S., Xia, B., Liu, J., & Navarre, W. W. (2012). Silencing of foreign DNA in bacteria. Curr. Opin. Microbiol., 15(2), 175-181.

Ansaldi, M. et al. A century of research on bacteriophages. Virologie 24, 9–22 (2018).

Ayora, S. et al. Homologous-pairing Activity of the Bacillus subtilis Bacteriophage SPP1 Replication Protein G35P. J. Biol. Chem. 277, 35969–35979 (2002).

Baines, J. D. Herpes simplex virus capsid assembly and DNA packaging: a present and future antiviral drug target. Trends Microbiol. 19, 606–613 (2011).

Beijerinck, M. W. CONCERNING A CONTAGIUM VIWM FLUIDUM AS CAUSE OF THE SPOT DISEASE OF TOBACCO LEAVES. (1898).

Black, L. W. Old, new, and widely true: The bacteriophage T4 DNA packaging mechanism. Virology 479–480, 650–656 (2015).

Bolte, H., Rosu, M. E., Hagelauer, E., García-Sastre, A. & Schwemmle, M. Packaging of the Influenza Virus Genome Is Governed by a Plastic Network of RNA- and Nucleoprotein-Mediated Interactions. J. Virol. 93, (2019).

Bos. L. Beijerinck's Work on Tobacco Mosaic Virus: Historical Context and Legacy. Philosophical Transactions: Biological Sciences vol. 354 (2014).

Buttner, C. R. et al. Structural basis for DNA recognition and loading into a viral packaging motor. Proc. Natl. Acad. Sci. 109, 811–816 (2012).

Camacho, A. G., Gual, A., Lurz, R., Tavares, P. & Alonso, J. C. Bacillus subtilis Bacteriophage SPP1 DNA Packaging Motor Requires Terminase and Portal Proteins. Journal of Biological Chemistry 278, 23251–23259 (2003).

Canosi, U., Lüder, G. & Trautner, T. A. SPP1-mediated plasmid transduction. J Virol 44, 431–436 (1982).

Casjens, S. et al. Bacteriophage P22 portal protein is part of the gauge that regulates packing density of intravirion DNA. J. Mol. Biol. 224, 1055–1074 (1992a).

Casjens, S. et al. Molecular genetic analysis of bacteriophage P22 gene 3 product, a protein involved in the initiation of headful DNA packaging. Journal of Molecular Biology 227, 1086–1099 (1992b).

Catalano, C. E. Viral genome packaging machines: genetics, structure, and mechanism. (Landes Bioscience/Eurekah.com Kluwer Academic/Plenum Publishers, 2005).

Comas-Garcia, M. Packaging of Genomic RNA in Positive-Sense Single-Stranded RNA Viruses: A Complex Story. Viruses 11, 253 (2019).

Chai, S. Distamycin-induced inhibition of formation of a nucleoprotein complex between the terminase small subunit G1P and the non-encapsidated end (pacL site) of Bacillus subtilis bacteriophage SPP1. Nucleic Acids Research 24, 282–288 (1996).

Chai, S., Kruft, V. & Alonso, J. C. Analysis of the Bacillus subtilis Bacteriophages SPP1 and SF6 Gene 1 Product: A Protein Involved in the Initiation of Headful Packaging. Virology 202, 930–939 (1994).

Chai, S., Lurz, R. & Alonso, J. C. The Small Subunit of the Terminase Enzyme ofBacillus subtilisBacteriophage SPP1 forms a Specialized Nucleoprotein Complex with the Packaging Initiation Region. J. Mol. Biol. 252, 386–398 (1995).

Chai, S., Szepan, U. & Alonso, J. C. Bacillus subtilis bacteriophage SPP1 terminase has a dual activity: it is required for the packaging initiation and represses its own synthesis. Gene 184, 251–256 (1997).

Chen, J. et al. Genome hypermobility by lateral transduction. Science 362, 207–212 (2018).

Ceglowski, P., Boitsov, A., Chai, S., and Alonso, J. C. (1993a). Analysis of the stabilization system of pSM19035-derived plasmid pBT233 in Bacillus subtilis. Gene 136, 1–12.

Chiang, Y. N., Penadés, J. R. & Chen, J. Genetic transduction by phages and chromosomal islands: The new and noncanonical. PLOS Pathog. 15, e1007878 (2019).

Coyette, J. & Mergeay, M. Microbiologie. (De Boeck, 2013).

Cuervo, A. et al. Structural Characterization of the Bacteriophage T7 Tail Machinery. Journal of Biological Chemistry 288, 26290–26299 (2013).

Cuervo, A., Vaney, M.-C., Antson, A. A., Tavares, P. & Oliveira, L. Structural Rearrangements between Portal Protein Subunits Are Essential for Viral DNA Translocation. J. Biol. Chem. 282, 18907–18913 (2007).

Deichelbohrer, I., Alonso, J. C., Lüder, G. & Trautner, T. A. Plasmid transduction by Bacillus subtilis bacteriophage SPP1: effects of DNA homology between plasmid and bacteriophage. J Bacteriol 162, 1238–1243 (1985).

De Lencastre, H. & Archer, L. J. Characterization of Bacteriophage SPP1 Transducing Particles. Microbiology 117, 347–355 (1980).

D'Hérelle, F. Sur le microbe bactériophage. Comptes rendus de la Société de Biologie vol. 82 1237 (1919).

Dion, M. B., Oechslin, F. & Moineau, S. Phage diversity, genomics and phylogeny. Nat Rev Microbiol 18, 125–138 (2020).

Djacem, K. (2016). Mécanisme moléculaire de reconnaissance et de clivage du génome chez le bactériophage SPP1, un virus à ADN double-brin (Doctoral dissertation, Université Paris-Saclay (ComUE)).

Djacem, K., Tavares, P. & Oliveira, L. Bacteriophage SPP1 pac Cleavage: A Precise Cut without Sequence Specificity Requirement. J. Mol. Biol. 429, 1381–1395 (2017).

Dröge, A. et al. Shape and DNA packaging activity of bacteriophage SPP1 procapsid: protein components and interactions during assembly 1 1Edited by J. Karn. Journal of Molecular Biology 296, 117–132 (2000).

Dröge, A. & Tavares, P. In vitro Packaging of DNA of the Bacillus subtilis bacteriophage SPP1 1 1Edited by J. Karn. Journal of Molecular Biology 296, 103–115 (2000).

D'Souza, V. & Summers, M. F. How retroviruses select their genomes. Nat. Rev. Microbiol. 3, 643–655 (2005).

Dublanchet, A. & Fruciano, E. Brève histoire de la phagothérapie. Médecine Mal. Infect. 38, 415–420 (2008).

Fujisawa, H. & Morita, M. Phage DNA packaging. Genes Cells 2, 537–545 (2003).

Gao, S., Zhang, L. & Rao, V. B. Exclusion of small terminase mediated DNA threading models for genome packaging in bacteriophage T4. Nucleic Acids Res. 44, 4425–4439 (2016).

Giese, S., Bolte, H. & Schwemmle, M. The Feat of Packaging Eight Unique Genome Segments. Viruses 8, 165 (2016).

Godinho, L. M. et al. The Revisited Genome of Bacillus subtilis Bacteriophage SPP1. Viruses 10, 705 (2018).

Golz, S. & Kemper, B. Association of Holliday-structure Resolving Endonuclease VII with gp20 from the Packaging Machine of Phage T4. J. Mol. Biol. 285, 1131–1144 (1999).

Green, T. J. et al. Common Mechanism for RNA Encapsidation by Negative-Strand RNA Viruses. J. Virol. 88, 3766–3775 (2014).

Grmek D. , M. Histoires de la virologie, des viroses et des virologues. History and Philosophy of the Life Sciences vol. 16 (1994).

Gual, A., Camacho, A. G. & Alonso, J. C. Functional Analysis of the Terminase Large Subunit, G2P, of Bacillus subtilis Bacteriophage SPP1. J. Biol. Chem. 275, 35311–35319 (2000).

Hashimoto, C. & Fujisawa, H. Transcription dependence of DNA packaging of bacteriophages T3 and T7. Virology 191, 246–250 (1992).

Hamada, K., Fujisawa, H. & Minagawa, T. A defined in vitro system for packaging of bacteriophage T3 DNA. Virology 151, 119–123 (1986).

Haima, P., Bron, S., and Venema, G. (1987). The effect of restriction on shotgun cloning and plasmid stability in Bacillus subtilis Marburg. Mol. Gen. Genet. 209, 335–342. doi: 10.1007/BF00329663

Harel, J., Duplessis, L., Kahn, JeffreyS. & DuBow, MichaelS. The cis-acting DNA sequences required in vivo for bacteriophage Mu helper-mediated transposition and packaging. Arch. Microbiol. 154, (1990).

Hobbs, Z. & Abedon, S. T. Diversity of phage infection types and associated terminology: the problem with 'Lytic or lysogenic'. FEMS Microbiol. Lett. 363, fnw047 (2016).

Howard-Varona, C., Hargreaves, K. R., Abedon, S. T. & Sullivan, M. B. Lysogeny in nature: mechanisms, impact and ecology of temperate phages. ISME J. 11, 1511–1520 (2017).

Huet, A., Duda, R. L., Hendrix, R. W., Boulanger, P. & Conway, J. F. Correct Assembly of the Bacteriophage T5 Procapsid Requires Both the Maturation Protease and the Portal Complex. Journal of Molecular Biology 428, 165–181 (2016).

Ignatiou, A. et al. Structural transitions during the scaffolding-driven assembly of a viral capsid. Nat. Commun. 10, 4840 (2019).

Jalal, A. S. B. & Le, T. B. K. Bacterial chromosome segregation by the ParABS system. *Open Biol.* **10**, 200097 (2020).

Kenzaka, T., Tani, K. & Nasu, M. High-frequency phage-mediated gene transfer in freshwater environments determined at single-cell level. ISME J 4, 648–659 (2010).

Kausche G.A. . Die Sichtbarmachung von pflanzlichem Virus im Übermikroskop. Naturwissenschaften 27, 292–299 (1939).

Labarde, A. et al. Temporal compartmentalization of viral infection in bacterial cells. Proc Natl Acad Sci USA 118, e2018297118 (2021).

Lang, A. S., Zhaxybayeva, O. & Beatty, J. T. Gene transfer agents: phage-like elements of genetic exchange. Nat. Rev. Microbiol. 10, 472–482 (2012).

Leavitt, J. C., Gilcrease, E. B., Wilson, K. & Casjens, S. R. Function and horizontal transfer of the small terminase subunit of the tailed bacteriophage Sf6 DNA packaging nanomotor. Virology 440, 117–133 (2013).

Leonhardt, H. (1990). Identification of a low-copy-number mutation within the pUB110 replicon and its effect on plasmid stability in Bacillus subtilis. Gene 94, 121–124. doi: 10.1016/0378-1119(90)90477-9

Lo Piano, A., Martínez-Jiménez, M. I., Zecchi, L. & Ayora, S. Recombination-dependent concatemeric viral DNA replication. Virus Res. 160, 1–14 (2011).

Murialdo, H. Bacteriophage λ DNA Maturation and Packaging. Annu. Rev. Biochem. 60, 125–153 (1991).

Němeček, D. et al. Subunit Conformations and Assembly States of a DNA-translocating Motor: The Terminase of Bacteriophage P22. Journal of Molecular Biology 374, 817–836 (2007).

Oliveira, L., Alonso, J. C. & Tavares, P. A Defined in Vitro System for DNA Packaging by the Bacteriophage SPP1: Insights into the Headful Packaging Mechanism. J. Mol. Biol. 353, 529–539 (2005).

Oliveira, L., Tavares, P. & Alonso, J. C. Headful DNA packaging: Bacteriophage SPP1 as a model system. Virus Res. 173, 247–259 (2013).

Payet, J. P. & Suttle, C. A. To kill or not to kill: The balance between lytic and lysogenic viral infection is driven by trophic status. Limnol. Oceanogr. 58, 465–474 (2013).

Pietilä, M. K., Demina, T. A., Atanasova, N. S., Oksanen, H. M. & Bamford, D. H. Archaeal viruses and bacteriophages: comparisons and contrasts. Trends Microbiol. 22, 334–344 (2014).

Poh, S. L. et al. Oligomerization of the SPP1 Scaffolding Protein. Journal of Molecular Biology 378, 551–564 (2008).

Rao, V. B. & Feiss, M. Mechanisms of DNA Packaging by Large Double-Stranded DNA Viruses. Annu. Rev. Virol. 2, 351–378 (2015).

Riva, S., Polsinelli, M. & Falaschi, A. A new phage of Bacillus subtilis with infectious DNA having separable strands. J. Mol. Biol. 35, 347–356 (1968).

Schneider, C. L. Bacteriophage-Mediated Horizontal Gene Transfer: Transduction. in Bacteriophages (eds. Harper, D. R., Abedon, S. T., Burrowes, B. H. & McConville, M. L.) 151–192 (Springer International Publishing, 2021). doi:10.1007/978-3-319-41986-2_4.

Schneemann, A. The Structural and Functional Role of RNA in Icosahedral Virus Assembly. Annu. Rev. Microbiol. 60, 51–67 (2006).

Seco, E. M. & Ayora, S. Bacillus subtilis DNA polymerases, PolC and DnaE, are required for both leading and lagging strand synthesis in SPP1 origin-dependent DNA replication. Nucleic Acids Res. 45, 8302–8313 (2017).

Seul, A. et al. Biogenesis of a Bacteriophage Long Non-Contractile Tail. J. Mol. Biol. 433, 167112 (2021).

Skorupski,K., Sauer,B., Sternberg,N. (1994) Faithful cleavage of the P1 packaging site (pac) requires two phage proteins, PacA and PacB, and two Escherichia coli proteins, IHF and HU. J Mol Biol. 243: 268-282.

Smits, C. et al. Structural basis for the nuclease activity of a bacteriophage large terminase. EMBO Rep. 10, 592–598 (2009).

Spiess, E. & Lurz, R. 13 Electron Microscopic Analysis of Nucleic Acids and Nucleic Acid Protein Complexes. in *Methods in Microbiology* vol. 20 293–323 (Elsevier, 1988).

Sternberg, N. & Coulby, J. Recognition and cleavage of the bacteriophage P1 packaging site (pac). J. Mol. Biol. 194, 469–479 (1987).

Tavares, P. et al. Identification of a gene in Bacillus subtilis bacteriophage SPP1 determining the amount of packaged DNA. J. Mol. Biol. 225, 81–92 (1992).

Tobe, T. et al. 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. 103, 14941–14946 (2006).

Volkova, V. V., Lu, Z., Besser, T. & Gröhn, Y. T. Modeling the Infection Dynamics of Bacteriophages in Enteric Escherichia coli: Estimating the Contribution of Transduction to Antimicrobial Gene Spread. Appl. Environ. Microbiol. 80, 4350–4362 (2014).

Waldor, M. K. & Mekalanos, J. J. Lysogenic Conversion by a Filamentous Phage Encoding Cholera Toxin. Science 272, 1910–1914 (1996).

Valero-Rello, A., López-Sanz, M., Quevedo-Olmos, A., Sorokin, A. & Ayora, S. Molecular Mechanisms That Contribute to Horizontal Transfer of Plasmids by the Bacteriophage SPP1. Front. Microbiol. 8, 1816 (2017).

Wilson, W. & Mann, N. Lysogenic and lytic viral production in marine microbial communities. Aquat. Microb. Ecol. 13, 95–100 (1997).

Wu, H., Sampson, L., Parr, R. & Casjens, S. The DNA site utilized by bacteriophage P22 for initiation of DNA packaging: Bacteriophage P22 pac site. Mol. Microbiol. 45, 1631–1646 (2002).

Yasbin, R. E. & Young, F. E. Transduction in Bacillus subtilis by Bacteriophage SPP1. J Virol 14, 1343–1348 (1974).

Zecchi, L. et al. Characterization of the Holliday Junction Resolving Enzyme Encoded by the Bacillus subtilis Bacteriophage SPP1. PLoS ONE 7, e48440 (2012).

Annexes

Résultats des tests statistiques

qRT-PCR

qRT-PCR des gènes *1*, *2*, *6*, *11* et *46* Comparaison des valeurs de deltaCT obtenues pour chaque mutant deux à deux, tableau avec les p-valeurs obtenues après le test ANOVA. *** p-valeur<0.001, ** p-valeur<0.01, * p-valeur<0.1.

Hypothèse	**Gène *1***	**Gène *2***	**Gène *6***	**Gène *11***	**Gène *46***
36 - 0REV == 0	< 0.001 ***	0.00776 **	0.0335 *	0.990	< 0.001 ***
54 - 0REV == 0	< 0.001 ***	< 1e-04 ***	<0.001 ***	0.805	0.00326 **
58 - 0REV == 0	< 0.001 ***	< 1e-04 ***	<0.001 ***	0.720	0.99978
99 - 0REV == 0	0.00976 **	< 1e-04 ***	<0.001 ***	0.990	0.69390
54 - 36 == 0	0.02970 *	< 1e-04 ***	<0.001 ***	0.967	< 0.001 ***
58 - 36 == 0	0.14753	< 1e-04 ***	<0.001 ***	0.930	< 0.001 ***
99 - 36 == 0	0.00316 **	< 1e-04 ***	<0.001 ***	0.883	< 0.001 ***
58 - 54 == 0	0.93510	0.00149 **	0.1091	1.000	0.00215 **
99 - 54 == 0	< 0.001 ***	0.00203 **	0.4346	0.532	0.06565 .
99 - 58 == 0	< 0.001 ***	0.99995	0.9172	0.440	0.58919

qRT-PCR des gènes *1*, *2*, *6*, *11* et *46* pour les mutants SPP1*pac*L-99, SPP1*pac*L-58, SPP1*pac*L-54, SPP1*pac*L-36 et SPP1*pac*L-0REV Comparaison des valeurs de deltaCT obtenues à 15 et 25 minutes post-infection, tableau avec les p-valeurs obtenues après le test ANOVA. *** p-valeur<0.001, ** p-valeur<0.01, * p-valeur<0.1.

Hypothèse	**p-valeur**
Gène 1 15 min - 25 min == 0	0.342
Gène 2 15 min - 25 min == 0	0.684
Gène 6 15 min - 25 min == 0	0.000661 ***
Gène 11 15 min - 25 min == 0	5.07e-08 ***
Gène 46 15 min - 25 min == 0	0.495

qRT-PCR des gènes *1*, *2*, *6*, *11* et *46* pour les mutants SPP1*sus70*, SPP1*sus115* et SPP1 sauvage. Comparaison des valeurs de deltaCT obtenues pour chaque mutant à 15 et 25 minutes post-infection, tableau avec les p-valeurs obtenues après le test ANOVA. *** p-valeur<0.001, ** p-valeur<0.01, * p-valeur<0.1.

Hypothèse	**gène *1***	**gène *2***	**gène *6***	**gène *11***	**gène *46***
sus19 25 min - *sus19* 15 min == 0	0.04063 *	0.00792 **	0.00105 **	<0.001 ***	0.00101 **
sus70 15 min - *sus19* 15 min == 0	0.97257	0.99971	0.99165	0.9713	0.99164
sus70 25 min - *sus19* 15 min == 0	0.00205 **	0.00412 **	< 0.001 ***	<0.001 ***	< 0.001 ***
WT 15 min- *sus19* 15 min== 0	0.67686	0.58122	0.52675	0.7744	0.52669
WT 25 min - *sus19* 15 min== 0	0.96853	0.91216	0.99113	0.0796 .	0.99113
sus70 15 min - *sus19* 25 min == 0	0.13690	0.01272 *	0.00253 **	<0.001 ***	0.00254 **
sus70 25 min - *sus19* 25 min == 0	0.51020	0.99799	0.99943	0.9982	0.99943
WT 15 min - *sus19* 25 min == 0	0.00335 **	< 0.001 ***	< 0.001 ***	<0.001 ***	< 0.001 ***
WT 25 min- *sus19* 25 min == 0	0.01125 *	0.00163 **	0.00261 **	0.0180 *	0.00269 **
sus70 25 min - *sus70* 15 min == 0	0.00700 **	0.00631 **	0.00156 **	<0.001 ***	0.00156 **
WT 15 min- *sus70* 15 min== 0	0.29284	0.42925	0.25924	0.3672	0.25928
WT 25 min - *sus70* 15 min == 0	0.66782	0.79594	1.00000	0.2542	1.00000
WT 15 min- *sus70* 25 min == 0	< 0.001 ***	< 0.001 ***	< 0.001 ***	<0.001 ***	< 0.001 ***
WT25 min- *sus70* 25 min== 0	< 0.001 ***	< 0.001 ***	0.00154 **	0.0357 *	0.00159 **
WT 25 min - WT 15 min == 0	0.97498	0.98272	0.25651	0.0089 **	0.25656

Expériences de transduction

Comparaison des fréquences de transduction des différents marqueurs mesurées pour chaque mutant deux à deux, tableaux avec les p-valeurs obtenues avec un modèle linéaire généralisé. *** p-valeur<0.001, ** p-valeur<0.01, * p-valeur<0.1.

Marqueur sur le chromosome	SPP1 sauvage	SPP1*pac*L-99	SPP1*pac*L-0REV	SPP1*pac*L-99.1	SPP1*pac*R-0REV1	SPP1*pac*R-0REV2	SPP1*pac*R-0REV3
SPP1 sauvage		0.00230 **	0.65892	1.00000	0.02256 *	< 0.001 ***	< 0.001 ***
SPP1*pac*L-99			< 0.001 ***	0.00118 **	0.99539	< 0.001 ***	< 0.001 ***
SPP1*pac*L-0REV				0.76614	< 0.001 ***	< 0.001 ***	< 0.001 ***
SPP1*pac*L-99.1					0.01265 *	< 0.001 ***	< 0.001 ***
SPP1*pac*R-0REV1						< 0.001 ***	< 0.001 ***
SPP1*pac*R-0REV2							0.59793
SPP1*pac*R-0REV3							

pBT233	SPP1 sauvage	SPP1*pac*L-99	SPP1*pac*L-0REV	SPP1*pac*L-99.1	SPP1*pac*R-0REV1	SPP1*pac*R-0REV2	SPP1*pac*R-0REV3
SPP1 sauvage		0.00230 **	0.65892	1.00000	0.02256 *	< 0.001 ***	< 0.001 ***
SPP1*pac*L-99			< 0.001 ***	0.00118 **	0.99539	< 0.001 ***	< 0.001 ***
SPP1*pac*L-0REV				0.76614	< 0.001 ***	< 0.001 ***	< 0.001 ***
SPP1*pac*L-99.1					0.01265 *	< 0.001 ***	< 0.001 ***
SPP1*pac*R-0REV1						< 0.001 ***	< 0.001 ***
SPP1*pac*R-0REV2							0.59793
SPP1*pac*R-0REV3							

pUB110cop1	SPP1 sauvage	SPP1*pac*L-99	SPP1*pac*L-0REV	SPP1*pac*L-99.1	SPP1*pac*R-0REV1	SPP1*pac*R-0REV2	SPP1*pac*R-0REV3
SPP1 sauvage		0.59431	1.00000	0.94486	0.67826	0.00393 **	< 0.001 ***
SPP1*pac*L-99			0.66111	0.99304	1.00000	< 0.001 ***	< 0.001 ***
SPP1*pac*L-0REV				0.96610	0.74105	0.00268 **	< 0.001 ***
SPP1*pac*L-99.1					0.99770	< 0.001 ***	< 0.001 ***
SPP1*pac*R-0REV1						< 0.001 ***	< 0.001 ***
SPP1*pac*R-0REV2							0.95271
SPP1*pac*R-0REV3							

pUB110	SPP1 sauvage	SPP1*pac*L-99	SPP1*pac*L-0REV	SPP1*pac*L-99.1	SPP1*pac*R-0REV1	SPP1*pac*R-0REV2	SPP1*pac*R-0REV3
SPP1 sauvage		0.876	0.962	1.000	0.991	< 0.001 ***	< 0.001 ***
SPP1*pac*L-99			0.290	0.973	0.999	< 0.001 ***	< 0.001 ***
SPP1*pac*L-0REV				0.848	0.624	< 0.001 ***	< 0.001 ***
SPP1*pac*L-99.1					1.000	< 0.001 ***	< 0.001 ***
SPP1*pac*R-0REV1						< 0.001 ***	< 0.001 ***
SPP1*pac*R-0REV2							0.999
SPP1*pac*R-0REV3							

Articles

Changements structuraux à l'œuvre durant l'assemblage de la capside de SPP1

Au cours de ce travail réalisé en collaboration avec l'institut de biologie structurale et moléculaire de Londres (Birkbeck College) lors de mes stages de fin de licence et de master 2, nous avons analysé les différentes étapes impliquées dans la formation de la capside de SPP1. Les structures de la procapside de SPP1 ainsi que celle de la capside mature ont pu être déterminées par cryo-électromicroscopie. Nous avons pu identifier deux intermédiaires de la procapside, les procapsides I et II, aux structures légèrement différentes. Elles correspondent à deux stades de maturation qui coexistent avant l'étape d'encapsidation de l'ADN. La protéine d'échafaudage gp11 interagit avec la protéine majeure de capside gp13 permettant ainsi l'assemblage de la procapside I. Lors d'un premier processus de maturation, une partie des protéines d'échafaudage est éjectée hors de la procapside I, ce qui conduit à des réarrangements structuraux des protéines majeures de capside entre elles et à la formation de la procapside II. Celle-ci est plus volumineuse et possède une forme plus angulaire que la procapside I. L'encapsidation de l'ADN et l'éjection de la totalité des protéines d'échafaudage génèrent de nouveaux changements structuraux qui aboutissent à la formation de la capside mature. Mon travail a consisté à identifier plus précisément les résidus de gp13 impliqués dans les différents changements de conformations induits lors de ces étapes de maturation. Nous avons ainsi pu confirmer l'importance de certaines régions de gp13 dans ces réarrangements structuraux. Nous avons identifié des résidus importants dans les interactions inter-capsomères au niveau d'une région de gp13 appelée boucle E, ainsi que des acides aminés au niveau de l'extrémité N-terminale flexible de gp13 qui sont notamment impliqués dans le processus de maturation de la procapside I en procapside II.

Ré-annotation du génome du phage SPP1

Dans ce papier, nous avons présenté une ré-annotation complète du génome de SPP1 qui pourra servir de référence à toute personne désireuse de travailler sur ce phage. En effet, la dernière annotation disponible datait de 1997, nous avons donc pensé qu'une mise à jour s'imposait. J'avais également besoin d'un génome de référence fiable pour mes expériences de thèse où j'ai beaucoup eu recours à des approches de séquençage à haut débit. Afin de mener à bien ce projet, le génome du phage a été séquencé en utilisant du séquençage Sanger et Illumina. Nous avons également utilisé un certain nombre d'outils bioinformatiques afin d'identifier la potentielle fonction de gènes dont le rôle n'a pas encore identifié ainsi que de nouveaux promoteurs. Ceci n'était pas possible en 1997, les outils nécessaires à cette analyse n'étant pas encore disponibles.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-12790-6 OPEN

Structural transitions during the scaffolding-driven assembly of a viral capsid

Athanasios Ignatiou[1], Sandrine Brasilès[2], Mehdi El Sadek Fadel[2], Jörg Bürger[3,4], Thorsten Mielke[3], Maya Topf[1], Paulo Tavares[2*] & Elena V. Orlova[1*]

Assembly of tailed bacteriophages and herpesviruses starts with formation of procapsids (virion precursors without DNA). Scaffolding proteins (SP) drive assembly by chaperoning the major capsid protein (MCP) to build an icosahedral lattice. Here we report near-atomic resolution cryo-EM structures of the bacteriophage SPP1 procapsid, the intermediate expanded procapsid with partially released SPs, and the mature capsid with DNA. In the intermediate state, SPs are bound only to MCP pentons and to adjacent subunits from hexons. SP departure results in the expanded state associated with unfolding of the MCP N-terminus and straightening of E-loops. The newly formed extensive inter-capsomere bonding appears to compensate for release of SPs that clasp MCP capsomeres together. Subsequent DNA packaging instigates bending of MCP A domain loops outwards, closing the hexons central opening and creating the capsid auxiliary protein binding interface. These findings provide a molecular basis for the sequential structural rearrangements during viral capsid maturation.

[1] Institute of Structural and Molecular Biology, Birkbeck College, Malet Street, London WC1E 7HX, UK. [2] Department of Virology, Institut de Biologie Intégrative de la Cellule (I2BC), CEA, CNRS, Université Paris-Sud, Université Paris-Saclay, 91198 Gif-sur-Yvette, France. [3] Max-Planck-Institut für Molekulare Genetik, Microscopy and Cryo-Electron Microscopy Group, Ihnestraße 63-73, 14195 Berlin, Germany. [4] Medizinische Physik und Biophysik, Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany. *email: paulo.tavares@i2bc.paris-saclay.fr; e.orlova@mail.cryst.bbk.ac.uk

Viruses use a limited number of structural arrangements to build stable infectious viral particles. These primordial assembly strategies were established during evolution, resulting in lineages of very distant viral species that infect different domains of Life. Tailed bacteriophages, which account for more than 10^{31} virions on Earth, are the most abundant viruses in the Biosphere[1]. In the tailed bacteriophages-herpesviruses lineage, a strict order of macromolecular interactions leads to assembly of icosahedral structures larger than 20 MDa, which protect the linear double−stranded DNA (dsDNA) viral genome.

A procapsid, also named prohead, (a virion precursor without DNA) is assembled first (Fig. 1a). Its capsid subunits establish quasi-equivalent interactions during formation of an icosahedral lattice[2]. Correct positioning of the procapsid subunits requires the internal scaffolding protein (SP), a chaperone, which can be an independent protein (e.g. phages P22, SPP1, phi29, herpesviruses)[3] or fused to the N-terminus of the major capsid protein (MCP) (phages HK97[4] and T5[5]). Maturation of the procapsid to the DNA-filled capsid state requires SP release and DNA packaging (Fig. 1a). These processes are accompanied by dramatic overall rearrangements of the MCP lattice that increases in size and becomes thinner. The resulting structure is highly resistant to environmental insult and withstands an internal pressure of ~60 atm applied by the viral dsDNA dense packing[6].

Structures of MCPs were determined using X-ray crystallography for the mature capsid and prohead II of bacteriophage HK97[7,8] and for the MCP pentamer of T4[9]. Cryo-electron microscopy (cryo-EM) sub-nanometre structures of procapsids and capsids have been obtained for several other tailed bacteriophages[10–14]. These studies revealed that tailed phages MCPs have a common fold, named after the prototype virus HK97. Comparison of procapsid and mature capsid structures showed that the MCP undergoes significant conformational changes during capsid maturation[10,11,14]. The MCP of herpesviruses capsids share a similar fold[15,16]. However, the structural events associated specifically with SP release and DNA packaging have remained unknown. These events are critical for the production of infectious virions. Thus, understanding the molecular mechanisms underpinning the sequential transformation of the viral capsid lattice may reveal drug targets to disrupt virus assembly in tailed phages and herpesviruses.

Here we report near-atomic models built into cryo-EM structures of procapsid I, intermediate procapsid II which has partially released the SP, and the mature DNA-filled capsid from bacteriophage SPP1. The structures are T = 7 *laevo* icosahedra composed of 415 copies of the MCP gp13 (35.4 kDa), several hundred copies of SP gp11 (23.5 kDa) inside procapsids, 180 copies of gp12 in DNA-filled capsids, and a gp6-gp7 complex which forms the specialised portal vertex[17,18]. The structures obtained reveal distinct steps of the MCP conformational transitions during capsid maturation that correlate with release of the SP and with DNA packaging that were uncoupled in this study.

Results

General organisation of SPP1 procapsids and mature capsid. Structures of SPP1 procapsids and capsids purified from *Bacillus subtilis* infected cells were determined by cryo-EM. Images were collected on a 300-keV electron cryo-microscope Polara (FEI, equipped with a K2 camera operated in the counting mode, see Methods). We found that the preparations of SPP1 procapsids, which are biologically active for DNA packaging[19], contain two major populations. Both structures are present in extracts of *B. subtilis* infected with a SPP1 mutant that does not package DNA (Supplementary Fig. 1a, b) indicating that they represent different states of the procapsid that co-exist in infected bacteria. One type of procapsids has a diameter of ~55 nm (named procapsid I) while the other one has a diameter of ~61 nm (procapsid II). Cryo-EM images of the two types of particles (Supplementary Fig. 1c, e) were separated by multivariate statistical analysis[20] and the subpopulations extracted were used for structure determination. The mature virion capsid (MVC) packed with DNA was reconstructed from images of the infectious SPP1 phage particles. The three structures (procapsid I, procapsid II, and the mature capsid) have icosahedral $T = 7$ *laevo* symmetry and were determined at a resolution of 5.2 Å, 4.7 Å and 4.5 Å at the threshold of 0.5 (4.5 Å, 4.3 Å and 4.1 Å at threshold of 0.143), respectively (Fig. 1b; Supplementary Fig. 2; Supplementary Table 1). High quality of the EM maps allowed tracing the polypeptide chains de novo (Supplementary Fig. 2). The asymmetric units (ASUs) comprise seven SPP1 gp13 subunits: six in a hexon (subunits A to F) and one from the capsid penton (subunit G) (Fig. 1c, d).

Procapsid I has spiked vertices, creased faces and its shell thickness varies between 3.5 and 5 nm. There are bulky areas of density attached to each gp13 subunit that project towards the capsid centre (shown in grey in Fig. 1d). These areas have an average length of ~3.5 and a width of 3.0 nm. The inner densities in procapsid I (Fig. 1b, top panel) were assigned to the SP gp11. Procapsid II is larger, more angular and has a much thinner shell than procapsid I, ranging between 2 and 2.7 nm. Procapsid II has bulky density for the SP attached to the capsid lattice interior only underneath the MCP five−fold vertices (Fig. 1b, d, middle panels). The MVC has the same size and protein shell thickness as procapsid II. Its size and organisation is consistent with the previous 8.8 Å resolution structure[21]. The capsid is filled with densely packed dsDNA appearing as concentric layers of density (Fig. 1b, bottom panel; Supplementary Fig. 1d, e, right). These densities do not have connections to the inner surface of the capsid lattice.

The shape of hexon capsomeres changes during assembly (Fig. 1b–d). They are skewed in procapsid I with a size of 14.5 nm × 11.5 nm and a central opening of ~3.0 nm × 2.2 nm (Fig. 1c, d, upper panels). The hexon becomes less skewed in procapsid II, more flat and expands to a size of 16.5 nm × 14.2 nm with the central opening getting slightly larger (~3.0 nm × 3.7 nm) (Fig. 1c, d, middle panels). The opening is closed in the MVC state whose hexon is nearly flat and only 1 nm longer than in procapsid II (Fig. 1c, d, bottom panels). The release of SP is coupled with transition from a curved shape to the flattened conformation in all MCP subunits.

Fold of the SPP1 gp13 protein in the MVC. The overall fold of the 324 amino acids-long SPP1 gp13 in the MVC is similar to gp5 in the mature empty capsid of phage HK97[7] in spite of only 12% sequence identity (Fig. 2a, b; Supplementary Movie 1). Gp13 has the characteristic L-shape of capsid proteins of other tailed bacteriophages, with well-defined A and P domains, an extended N-terminus arm, and the E-loop (Fig. 2a; Supplementary Fig. 2; Supplementary Movie 1)[7–16]. Superposition of the gp13 (SPP1) and gp5 (HK97, PDB 1OHG[7]) atomic models indicated clearly the location of 42 additional amino acids in gp13 (Fig. 2b). They form an additional β-hairpin (β9–β10, residues Val224-Phe234) in the A domain, and a small domain that comprises helix α7 together with the short loop Lys282-Gln305 positioned above the connection of the P domain to the E-loop (Fig. 2b, Supplementary Figs. 2 and 3). The SPP1 gp13 helix α1 (Asp39-Ala45) is a bit longer and shifted slightly away from spine helix α3 when compared to its location in gp5.

The gp13 A domain consists of two helices α5 and α6 (labelled as in HK97 gp5) that sandwich a six-stranded β-sheet (shown in yellow in Fig. 2a, Supplementary Fig. 3). Its A-loop (residues

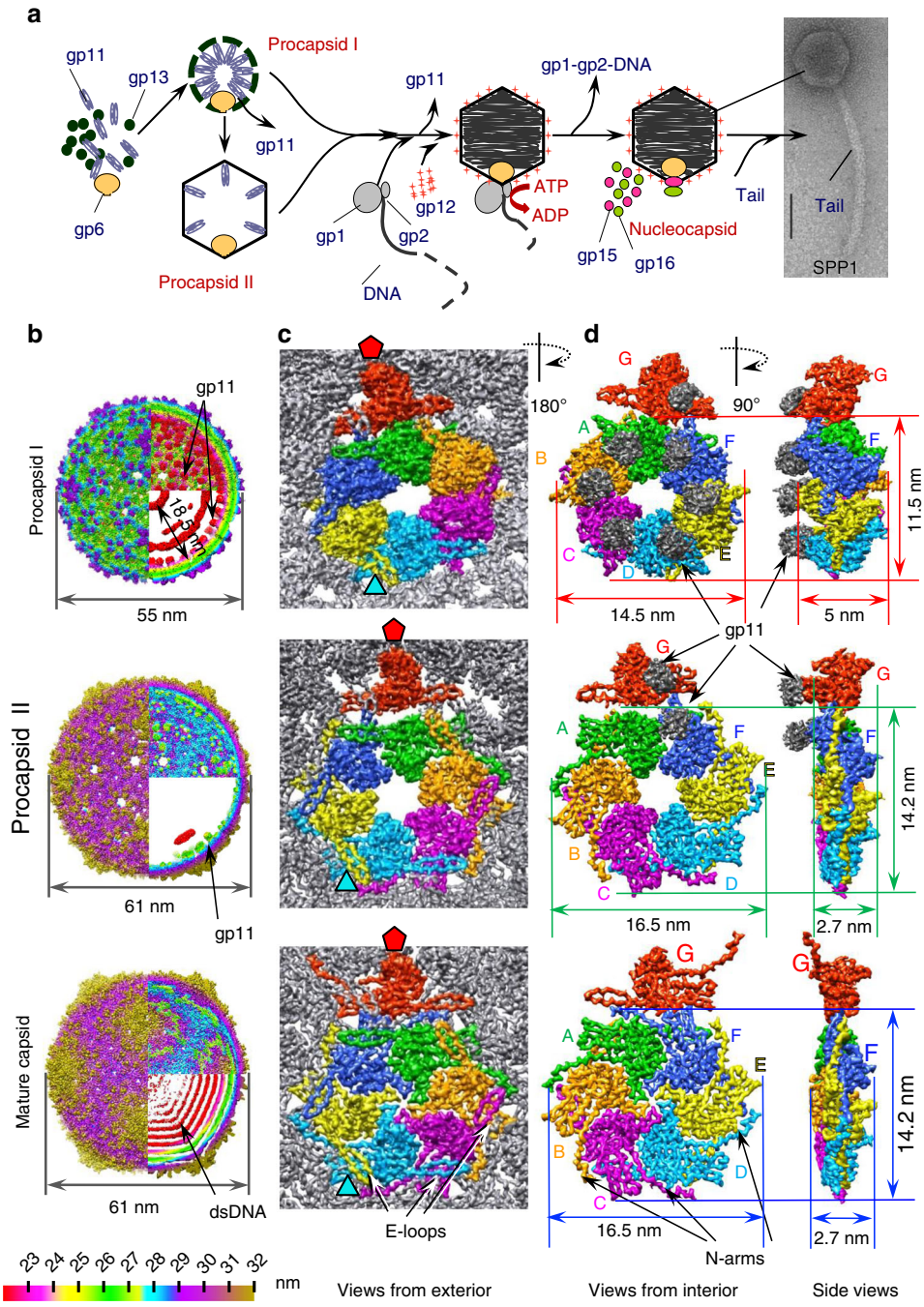


Fig. 1 3D structures of the SPP1 procapsids and mature capsid. **a** Schematic representation of bacteriophage SPP1 assembly pathway. **b** Structures of procapsid I, procapsid II, and mature virion capsid (MVC) (from the top to the bottom). The left part of the images shows the outer capsid surface, the top right quarter the inner side of the capsid shell; and the bottom quarter is a central slice of the structures. All structures are radially coloured from the origin to the outer surface. Procapsid I has three dense layers inside of the shell (shown in red) that were assigned to the gp11 SP. The outermost layer demonstrates well-defined bulks of densities attached to the capsid shell. Procapsid II has such inner bulks of density only underneath pentons. The mature capsid has densely packed layers of dsDNA. **c** Asymmetric unit (ASU) of each structure is viewed from outside. **d** ASUs viewed from inside the (pro) capsids. The densities of the gp11 SP are in grey at a lower threshold. The right panels show the asymmetric units side view. Each subunit of the ASU is colour coded and labelled with letters according to the nomenclature used throughout the manuscript

Asp194-Arg207) projects towards the centre of hexons closing their central openings in the MVC (Fig. 1c, d, bottom panels). The P domain comprises a 50 Å long helix α3 (the spine helix) and three antiparallel β-strands with β12 and β13 connected by the P-loop. The N-arm (Ala2-Ala28) protrudes outwards from the middle of spine helix α3 by 70 Å, passing underneath the E-loop of a neighbouring subunit located on the left side (when observed from the capsid exterior). The N-arm of the A subunit goes underneath of the E-loop of the B subunit while the N-arm of the B subunit is located underneath the E-loop of the C subunit and so on (Fig. 1c, d, bottom panel). The E-loop (Pro57-Gln84) extends by 42 Å on the other side of the P domain (Fig. 2a, Supplementary Figs. 2 and 3). The E-loop covers the N-arm and P domain of the neighbour MCP on the right side. Both N-arms and E-loops establish inter-capsomere contacts (see below).

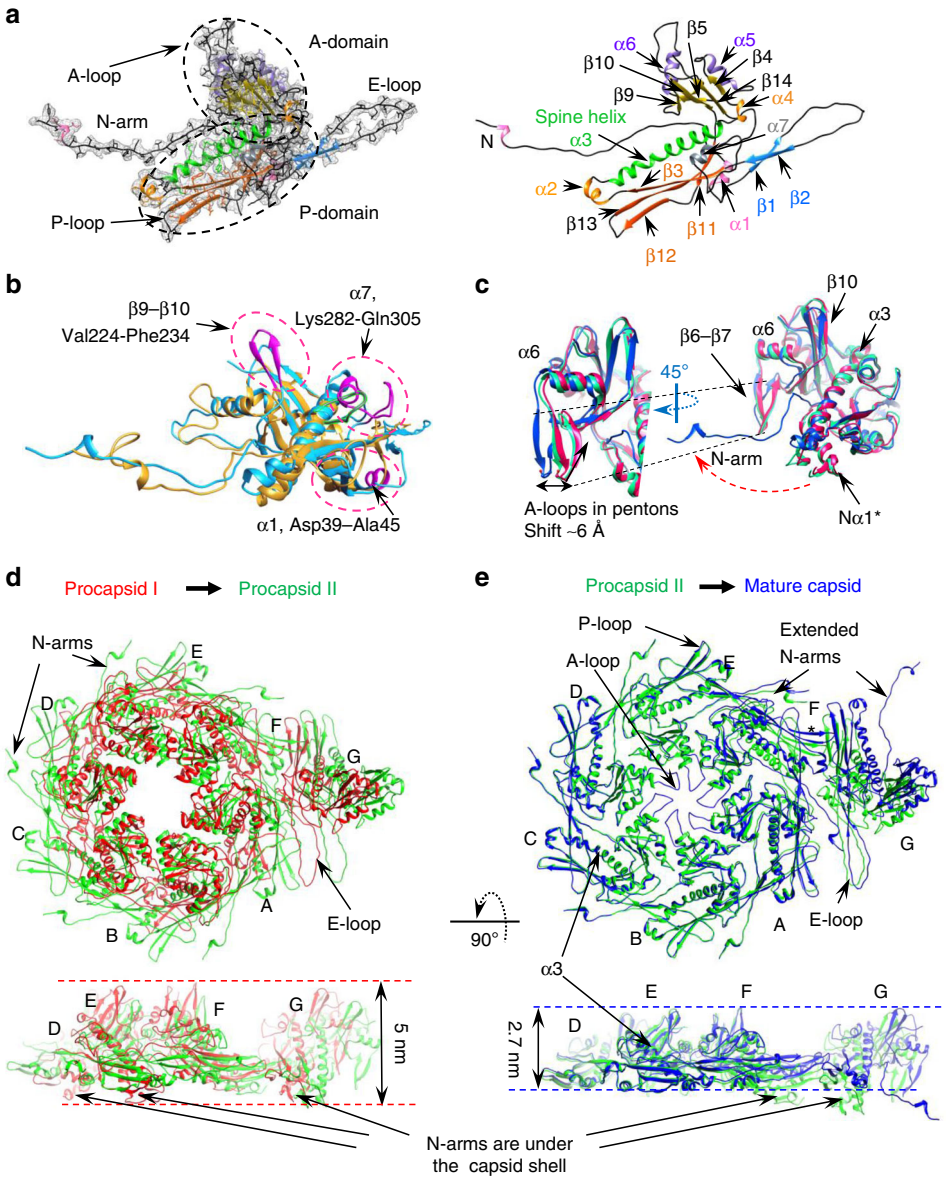


Fig. 2 Conformational plasticity of the capsid protein gp13 throughout assembly. **a** The gp13 atomic model is traced de novo within the cryo-EM density of a MCV subunit (left). Secondary structure elements are shown on the right panel. **b** Superposition of atomic models of SPP1 gp13 (in cyan) and phage HK97 gp5 (PDB 1OHG, in gold). The additional structural elements of gp13 are shown in magenta. **c** Superposition of the G subunits that form pentons in three capsid states: procapsid I in red, procapsid II in green, and MVC in blue. The β-hairpin β9-β10 adopts the most vertical position in procapsid I. **d** Overlay of the ASUs from procapsid I and procapsid II viewed from the outside (top) and from the side (bottom). **e** Overlay of the ASUs from procapsid II and MVC

Conformational changes in the MCP during phage maturation. The reconstructions of three capsid states reveal distinctive structural changes exposing the transient steps of the maturation process related specifically to SPs release and DNA packaging. The polypeptide chains of all seven subunits in the ASUs of the three capsid states were traced de novo (see Methods). Their conformational changes were assessed by superposition of equivalent subunits from the procapsids and the MVC. The subunits were aligned through the main spine helix α3 (Fig. 2c–e; Supplementary Fig. 4). Supplementary Movie 2 shows the motions undergone by structural elements of the hexon subunits A-E from the initial (procapsid I) to the final state (MVC) of the capsid assembly pathway (Note that this movie does not detail the sequential order of transitions through the intermediate state procapsid II; that sequence is presented in Supplementary Movies 3 and 4).

The A domain of the MCP has nearly the same conformation in all capsomeres except for hairpins β6–β7 and β-strands β9–β10. The A-loops form hairpins (β6–β7) pointing towards the centre of procapsid I and define the boundaries of the central opening of hexons (Supplementary Fig. 5). They move to a more vertical position in procapsid II. During transition from procapsid II to the post-DNA packaging capsid state (MVC) hairpin β6–β7 of the hexons unfolds to form the A-loop, which turns outwards closing the hexon central opening (Fig. 2e; Supplementary Figs. 3–5; Supplementary Movies 3 and 4). Interestingly, the β6–β7 hairpins of pentons (subunit G) do not undergo significant conformational changes during maturation, apart from a ~6 Å motion towards the penton centre between the procapsid II and MVC states (Fig. 2c; Supplementary Figs. 3–5). The gp13 A-loops in the MVC have well-defined shapes unlike those of HK97, and interact with the SPP1 auxiliary protein gp12

(see below). The β9–β10 strands move slightly towards the centres of capsomeres by 5 Å during the transition from procapsid I to procapsid II in subunits A, B, D, and E.

Subunits A-E of the ASU undergo significant conformational changes in the P domains and N-arms during transition from procapsid I to procapsid II (Supplementary Fig. 4a–e; Supplementary Movies 3 and 4). In procapsid I all MCP subunits are bound to the SP gp11 protein (Fig. 1d, upper panel and Supplementary Fig. 5) via the N-terminal helix α1* (Phe15-Leu26), positioned beneath the capsid shell and the N-terminus of the spine helix α3 from the same subunit (Fig. 2d; Supplementary Fig. 5). These interactions are disrupted in subunits A–E of procapsid II where helices α1* unfold to an extended strand similar to the one found in the MVC. The N-terminus arm moves upwards by 60° and stretches out from the P domain towards the hexon periphery, protruding to the outer surface (Fig. 2d; Supplementary Fig. 4; Supplementary Movies 3 and 4).

In procapsid I the P-loop (Ala261-Ser264) and parts of the β-sheets β12 (Thr257- Asn260) and β13 (Gln265-Leu268) have a curved shape with their ends pointing inwards to the capsid interior. They straighten out during transition to procapsid II (Fig. 2d; Supplementary Fig. 4a–e). The E-loops of subunits A–E move upwards by ~20° and become straighter. This movement is accompanied by their rotation in the facet plane of ~20° clockwise if we are looking at the capsid exterior (top views) (Supplementary Fig. 4a–e). The E-loop shift leads to formation of β-sheet β1 (Met56-Asn60) - β2 (Asn81-Asn85) oriented roughly parallel to the spine helix. As the E-loop unbends, helix α1 (Asp39-Ala45, Fig. 2a) positioned underneath the spine helix α3 shifts upwards by ~10 Å.

These cumulative changes in the MCP subunits A-E lead to a less skewed conformation of the ASU in procapsid II than in procapsid I making the hexon more symmetrical (Figs. 1c, d and 2d; Supplementary Movie 4). The A and P domains of subunits F and G keep a similar conformation in the two procapsid states, correlating with maintenance of their attachment to the SP (Fig. 1b–d), while the E-loop moves outwards making the overall conformation more flat (Fig. 2d; Supplementary Fig. 4f, g).

Subunits A to E remain nearly unchanged during transition from procapsid II to the MVC. In contrast, subunits F and G undergo a significant structural rearrangement similar to the transition of subunits A–E from procapsid I to procapsid II when the SP is released. The F subunit adopts a fold close to the one of the other hexon subunits while the penton G subunits acquire also an extended conformation but without noticeable changes in the A domain (Fig. 2e, Supplementary Fig. 4). These rearrangements increase further the overall length of the ASU by ~10 Å while the hexons preserve their sizes (Fig. 2e). The conformational changes of gp13 during the transition from procapsid I to the mature capsid are analogous to the movement of an umbrella as it opens upon release of the scaffolding proteins. Bending of the spine helices away from the A-domains gives room for expansion of the N-arms that project outwards making contacts with neighbour capsomeres (Supplementary Movie 4). These concerted rearrangements result in flattening of the overall shape of the capsid faces.

Interaction of SP gp11 with MCP gp13. The SPP1 gp11 SP is an elongated α-helical molecule found in solution as a very stable dimer and as a dimer of dimers[22]. It was proposed that the former complex is the one present in procapsids[22]. The low resolution of the bulky SP density bound to gp13 did not allow an ab initio tracing of the gp11 polypeptide chain (Fig. 1b–d; Supplementary Fig. 5). A structural model of the gp11 subunit generated using the protein fold recognition server PHYRE2[23] predicts a long α-helix fold with inserts of short unstructured elements while the N- and C-termini are predicted to be clusters of short α-helices (Supplementary Fig. 6a). The gp11 N-terminus model has two small helices α1 and α2 forming a hook which resembles the C-terminus of phage P22 gp8 SP[24] (PDB 2GP8) and the N-terminus of phage ϕ29 gp7 SP[25] (PDB 1NO4) that bind to their MCP counterparts[24–27] (Supplementary Fig. 6b). The two N-terminus helices of the gp11 model fit well into the bulk of density attached to gp13 in the cryo-EM map of procapsid I. Taking in account that gp11 is most likely a dimer[22] its two hooks link two independent gp13 subunits (see Discussion).

In both procapsids I and II the SP density is attached to the gp13 amino end of N-terminal helix α1* and to the amino end of spine helix α3 of the P domain (Supplementary Figs. 3 and 6c). An interaction of the MCP N-terminus with SPs is a common feature found in the procapsids of phages P22, T7, and α80 while the second contact of the MCP spine helix with the SP appears to be less conserved[10,11,14,27]. The EM maps suggest that residues Asp109-Gln112 of the gp13 P domain contact directly gp11 (Supplementary Fig. 6c). The amino acid substitution Gln112Glu in this region of gp13 is not detrimental while Asp109Asn or Pro110Ala impair procapsid assembly. The two latter mutations lead to polymerisation of gp13 into structures that lack gp11 (Supplementary Fig. 7a, b, f, g). A similar phenotype is found for substitution Tyr18Ala in the amino terminus of gp13 that is located in the α1* region of interaction between gp13 and the SP (Supplementary Figs. 6c, d and 7b, d).

The spine helix α3 of the MCP is straight when gp11 is bound. The release of gp11 from the P domain causes α3 to adopt a bent conformation where the segment of amino acids 110 to 119 of gp13 move upwards, turning by ~40°. The turn takes place at Ala119 and Ala120 (Supplementary Fig. 6d). Another dramatic change induced by the release of the SP is the unfolding of α1* and straightening of the gp13 N-terminus to form the N-arm. The N-arm moves outwards following the shape of the neighbour subunit of the capsomere (Figs. 1c, d, and 2d, e) and establishing inter-capsomere interactions (Fig. 3). These large motions conceivably result from disruption of the interaction with the SP initiating the overall programme of structural rearrangements that lead to capsid expansion.

Inter-capsomere interactions. Each face of the SPP1 capsid comprises three ASUs. In procapsid I the ASUs are held together by interactions of six gp13 subunits (from three adjacent ASUs (Fig. 3, Supplementary Fig. 8). The outer strands of the P domains from D subunits and the E-loops of E subunits from three ASUs outline a small opening ~15 Å in size at the threefold axis (Fig. 3a–c; Supplementary Fig. 8). The end of the E-loop of subunit Ei from ASUi, where i is the number of the ASU ($i = 1$, 2, or 3) has a strong connection with the end of the P-loop from subunit D($i + 1$) from ASU($i + 1$) (Fig. 3a–c). The structure suggests that Asp74 in the E-loops is involved in inter-capsomere interactions.

In procapsid II, nine MCP subunits are involved in inter-capsomere interactions around the exact and pseudo three-fold axes (Fig. 3d–f; Supplementary Fig. 8). Subunits C, D, and E from each ASU make connections with subunits of adjacent ASUs. During transition from procapsid I to procapsid II the N-arms from MCP Ci subunits threaded through a 20 Å opening between MCP D(i) of the same ASU and MCP E($i + 1$) of adjacent ASU($i + 1$). The distance between the tip of the E(i)-loop and the side of the E($i + 1$) loop increases from 3.5 Å in procapsid I to 12.5 Å in procapsid II (Fig. 3b, e; Supplementary Fig. 8). The N-arm of each subunit then passes through this crevice thus

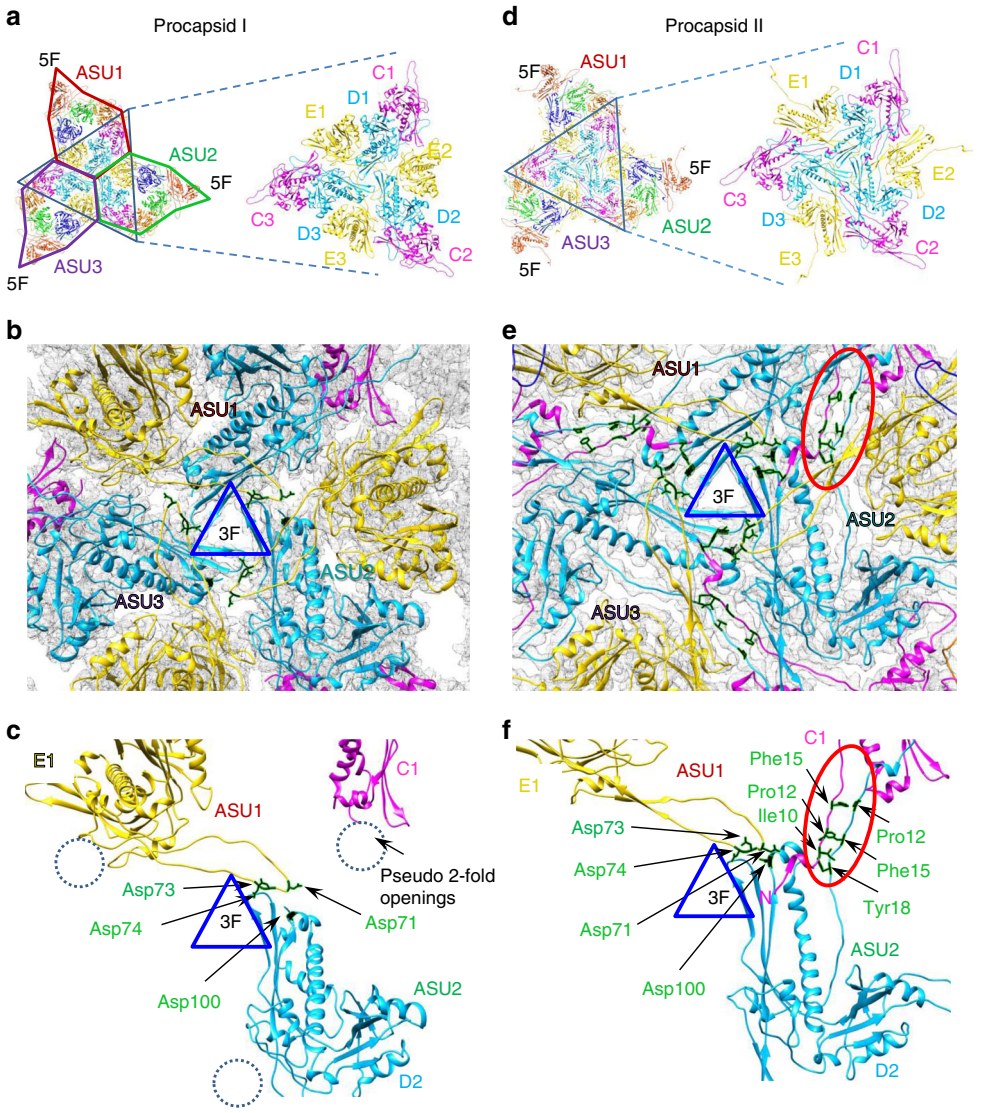


Fig. 3 Inter-capsomere interactions. **a** Schematic representation of subunit positions around the threefold axis in the procapsid I face. **b** Cryo-EM density of procapsid I with the gp13 MCP atomic model superimposed. **c** Points of interaction between subunits from different capsomeres are shown by arrows and the residues are labelled in green. **d** Schematic representation of subunits around the threefold axis in the procapsid II face. **e** gp13 MCP atomic models superimposed in the cryo-EM map of procapsid II. **f** Points of interaction between subunits of adjacent capsomeres. The red ovals in **e** and **f** show the region of antiparallel N-termini interaction. The residues involved in interactions are displayed in green

resulting in new contacts on the capsid outer surface between N-termini of subunits from adjacent ASUs (Fig. 3e, f).

Each N-arm of the MCP Ci subunit forms four contacts with the MCP D($i+1$) subunit from the neighbouring ASU in procapsid II. The first contact is between subunit Tyr3 of C*i* subunit and Gln265 in β13 of the P domain from subunit D($i+1$). The other interactions are established between N-arms of subunits from different capsomeres that run in an antiparallel direction: Ile10(C(i))-Tyr18(D($i+1$)); Pro12(C(i))-Phe15(D($i+1$)); Phe15(C(i))-Pro12(D($i+1$)); Tyr18(C(i))- Ile10(D($i+1$)) (Fig. 3f). These inter-capsomere links are observed both in pseudo and in exact 2-fold axes. Gp13 mutations Pro12Ala and Tyr18Phe do not impact on SPP1 capsid assembly while mutation Ile10Val makes gp13 less functional leading to low yields of infectious virions (Supplementary Fig. 7). The non-conservative substitution Tyr18Ala impairs more drastically procapsid formation but this phenotype results from a defect of gp13 interaction with the SP that occurs earlier in the assembly pathway (Supplementary Fig. 7a, b) (see above, Results sub-section Interaction of SP gp11 with MCP gp13). An additional connection is possibly formed during the transition from procapsid I to II between the E-loop end of subunit E*i* and helix-α2 in the P domain of subunit D($i+1$) (Fig. 3e, f). Mutation gp13 Asp100Ala likely affects such interaction. This substitution allows gp13 binding to gp11 but impairs assembly of functional procapsids with the portal protein (Supplementary Fig. 7a, b, e). The interactions between pentons and hexons in the MVC are identical to that of hexons in procapsid II (Supplementary Fig. 8; Supplementary Table 2).

Binding of the auxiliary protein gp12. Trimers of the 6.6-kDa SPP1 auxiliary protein gp12[18] bind to the centre of hexons[21]. Gp12 has a collagen-like fold on its central part and possibly α-helices at the ends[18]. The cryo-EM structure resolved only part of the gp12 attached to the capsid since its distal end is highly flexible. Gp12 binds to a ring of six Glu197 that converge from the tip of the A loop closing the hexamer central hole (Fig. 4a, b).

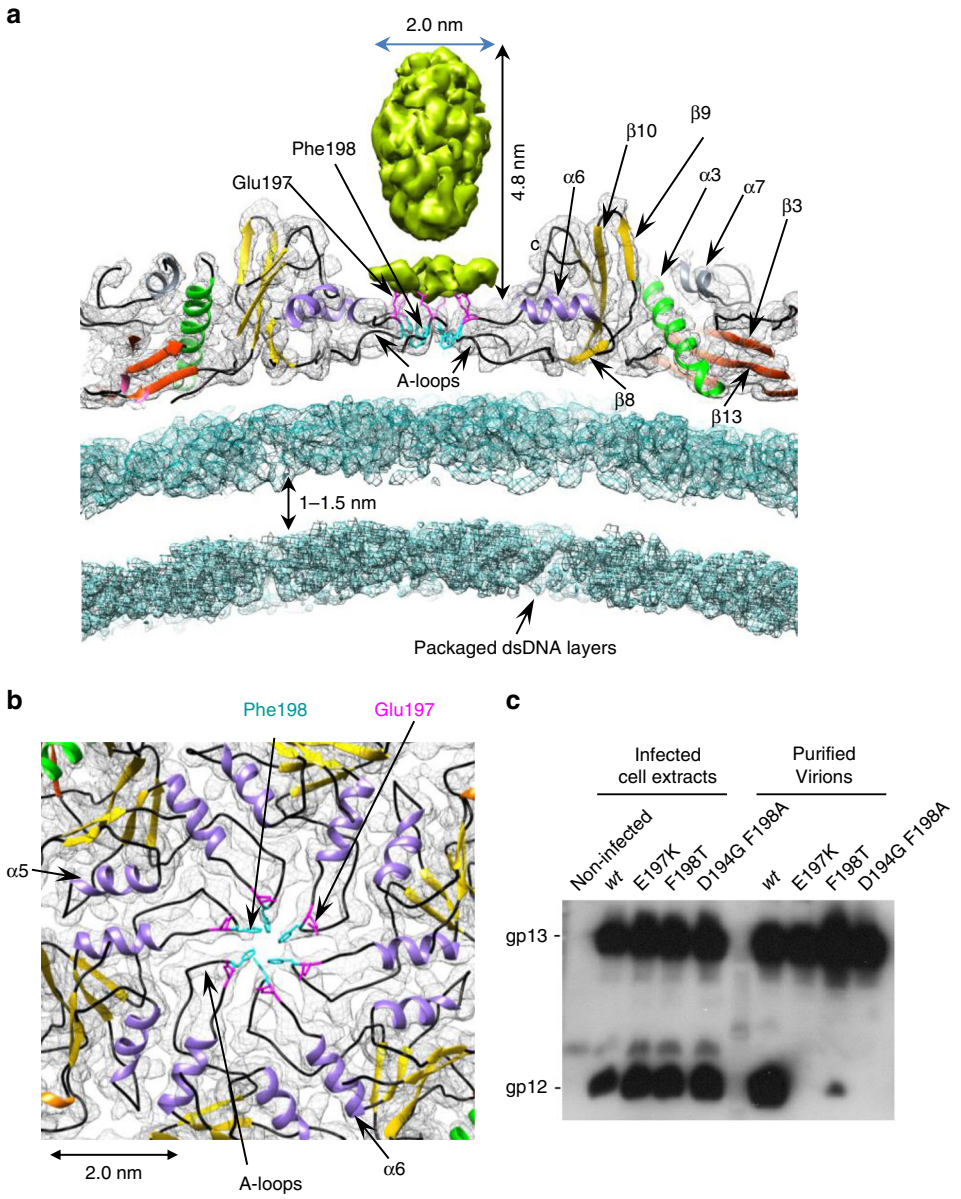


Fig. 4 Interaction of the accessory protein gp12 with the MCP A-loops in the hexons. **a** Central cross section through the hexon. **b** Arrangement of A-loops in the centre of the hexon. Residues involved in the interaction with gp12 are labelled. **c** Western blot of crude extracts of *B subtilis* bacteria infected with SPP1 phages carrying gp13 with the amino acid substitution(s) displayed above the gel lanes (left) and of purified SPP1 viral particles with the mutant MCP forms (right). Note that gp12 is produced in all infections but binds only stably to wild-type particles. The phage encoding gp13 Asp194Gly Phe198Ala is an escape mutant isolated from the poorly growing mutant phage SPP1gp13 Phe198Ala

Mutagenesis of this residue to lysine specifically abolished gp12 attachment to SPP1 DNA-filled capsids without impairing assembly of infectious SPP1 virions because gp12 is a non-essential component of the virus (Fig. 4c). A hydrophobic ring of six Phe198 residues positioned beneath the base of the gp12 trimer closes the centre of hexamers. Its substitution by lysine disrupts capsid formation. The mutation Phe198Ala reduces assembly efficiency (small phage plaque phenotype) but can be compensated by the second site substitution Asp194Gly that probably renders the A-loop more flexible. Both mutations impair stable binding of gp12 to the hexon centre (Fig. 4c).

Discussion

The conserved assembly pathway of icosahedral capsids from the tailed phages-herpesviruses lineage is characterised by dramatic rearrangements during transition from their initial procapsid state to the final mature genome-filled viral particles. The cryo-EM near-atomic structures reported here uncouples, for the first time, structural transitions caused by disruption of the scaffolding protein-procapsid association from the effect of force applied on the MCP lattice by DNA packing inside the capsid[6]. Conformational changes also occur most likely in the portal system surrounded by hexamers during procapsid maturation and DNA packaging[17,19]. These changes however were not revealed due to imposed icosahedral symmetry during the reconstruction procedure aimed to reveal details of the MCP at high resolution. This step of symmetrisation has obscured the features of the portal protein and its contacts with the capsid.

Procapsid I is the first structure formed during assembly. It exhibits full occupancy of SP proteins bound to each MCP subunit in a 1:1 ratio (Figs. 1d, top panel, and 5, left panel). The SP

interacts with the N-terminus of the MCP, similar to other phages[10,11,14,27], and with the end of spine helix α3 of the P domain. Disruption of the SP-MCP interaction during maturation leads to unfolding of the MCP Nter helix α1* to form the N-arm extended conformation. The N-arm projects through an opening between neighbour subunits, moving their E-loops apart, to establish antiparallel contacts with the N-arm from an adjacent capsomere (Fig. 3e, f). SP release and procapsid expansion is also associated with a change in spine helix α3 from the straight to the bent conformation (Supplementary Fig. 6d; Supplementary Movie 4). This differs from phage HK97 where the spine helix is bent in the non-expanded procapsid and straight in the mature capsid, a structural change that was proposed to promote procapsid expansion in the HK97 system[8].

The transition from procapsid I to procapsid II is induced by major structural rearrangements in the hexon subunits that become flattened resulting in expansion without major changes in the pentons (Fig. 2d). This is somewhat analogous to the capsomeres behaviour in phage HK97 when procapsid expansion was triggered in vitro[28,29]. In the SPP1 procapsid II intermediate state, SPs are associated only to MCP subunit F of hexons and to all subunits forming pentons (Figs. 1d and 5, central panel), demonstrating that this interaction is more stable than the one of SP bound exclusively to hexons. The SP density corresponds to the hook of a single SP subunit (Supplementary Fig. 6). Each SP hook is attached to one MCP subunit and tilted towards the capsomere centre (Fig. 1d and Supplementary Fig. 5). To interpret this pattern of interactions between the SP and the MCP, we need to take into account that gp11 dimers were the most effective form of gp11 for procapsid assembly in vitro[22]. That is likely the SP association state found in SPP1 procapsids. One dimer conceivably links a penton subunit to the F subunit of the neighbour hexon in procapsid II (Fig. 1d, central panel). This results in five dimers connecting each penton to the surrounding hexons in procapsid II, as represented by divergent arrowheads in the model shown in Fig. 5 (centre). Such organisation implies that gp11 dimers would also establish inter-capsomere contacts in procapsid I leading to an overall internal bridging of the icosahedral lattice (Fig. 5, left panel). Gp11 dimers act as double hooks between capsomeres chaperoning assembly to achieve their correct position for building procapsids[3].

The interaction with the SP likely maintains the MCP subunits in a strained conformation. Disruption of SP bonds with the MCP P-domain and helix α1* probably provides the energy necessary for α1* unfolding and spine helix bending that lead to flattening and widening of MCP hexons (Figs. 2d, 3 and 5, central panel), resulting in an overall energetically favourable expansion of the capsid lattice by 25% in volume (Fig. 1b–d, top and middle panels). Following the SP release, connections between capsomeres are established by extensive new inter-capsomere bonding that stabilise the expanded conformation (Fig. 3). A similar mechanism might explain the structural changes of Herpes Simplex Virus-I (HSV-1) round-shaped procapsids that spontaneously mature to a polyhedral capsid shape in absence of DNA packaging[30].

The transition from procapsid II to the MVC is linked to release of the SP from the capsid pentons regions and to DNA packaging. Departure of the SP leads to the same conformational changes in penton subunits as previously in hexons, when helix α1* became stretched out. The major structural change found in hexons of the mature capsids is the outward motion of the A loops, probably caused by DNA packaging inside the capsid, closing the central openings and creating the binding site for the gp12 trimeric auxiliary protein (Figs 2e and 4). Similarly, the openings in phage P22 procapsids[31,32] are closed by motion of the A domain during capsid maturation[10,33]. It was proposed that

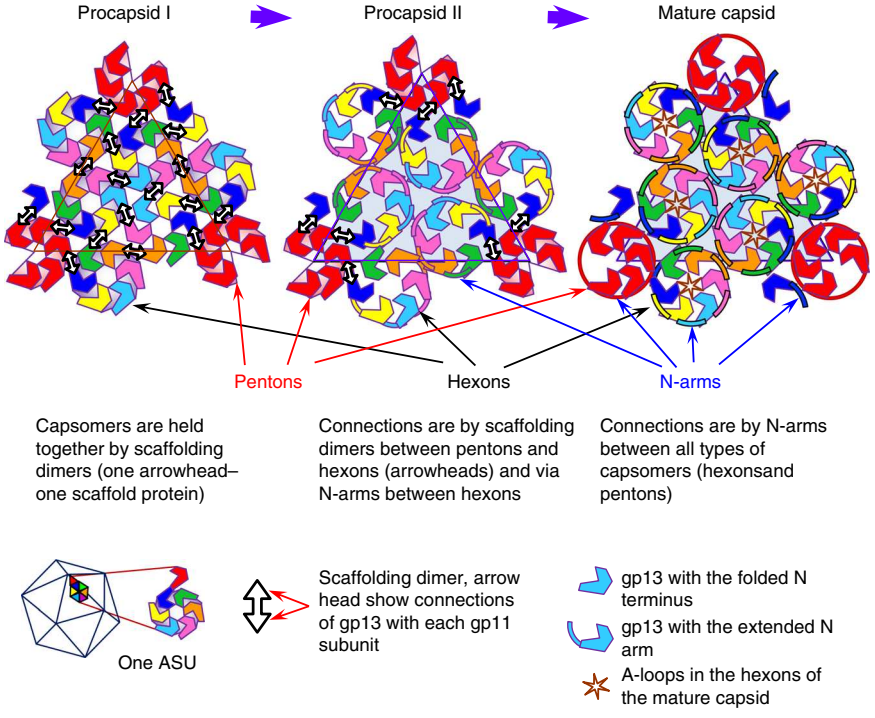


Fig. 5 Schematic representation of the maturation process of the SPP1 capsid. The left panel represents procapsid I where the capsomeres are fastened together by SP dimers (double head arrows). Pentons, located on the fivefold vertices of the capsid, are shown as red hexamers with one subunit omitted. The middle panel shows procapsid II where SP dimers remain only around pentons. Hexons are held together via extended N-arms. The MCV is shown on the right panel. Capsomeres are connected through extended N-arms. The central opening in hexons is closed by A loops indicated by stars. Colour coding of the MCP subunits is as in Fig. 1

these openings function as exits for scaffolding proteins in phages like P22[3,10,31–33], T7[11], and SPP1 (this work) that do not have an internal protease to degrade the scaffold[3]. The finding that SPP1 hexon openings are closed only at the mature capsid stage, after all of the gp11 has been released, is consistent with this functional assignment. The A-loops of SPP1 penton subunits remain pointing towards the capsid interior (Supplementary Fig. 5) explaining why gp12 binds exclusively to the hexon centre.

Our results demonstrate that the SP maintains the procapsid in a non-expanded state. Its release from procapsids directs the major structural rearrangements in the MCP forming an extensive inter-capsomere bonding network and leading to the stable expanded state. In contrast to previous models[8,10,14], such transition is independent of DNA packaging whose major impact is closure of the hexons central openings (Figs. 1c, d, 2e and 5, right panel). This sequence of structural transitions unravels the stepwise molecular mechanism engaged to engineer capsids withstanding high pressure. It is likely that similar mechanisms operate at capsid assembly of other tailed phages and the larger herpesviruses[15,16], but additional studies on other viral systems are necessary to validate this assertion.

Methods

Microbiological and genetic methods. The bacteriophage strains used were SPP1 wild type, SPP1*sus70* (defective in gene *1*), SPP1*sus31* (defective in gene *13*), and SPP1*sus31sus117* (defective in genes *11* and *16*)[34–38].

Gene *13* alleles coding for gp13 Glu197Lys and Phe198Ala were transferred from plasmids to the SPP1 genome by double cross-over during SPP1*sus31sus117* infection of the non-permissive host *B. subtilis* YB886 bearing constructs pPT291 or pPT275, respectively. These plasmids were obtained by site-directed mutagenesis of pCC40[38]. Individual phage clones multiplying in *B. subtilis* YB886 were screened by PCR and DNA sequencing to confirm presence of the desired mutations. SPP1 phages coding gp13 Phe198Ala had a small phage plaque phenotype. Revertant phages with normal phage plaque size arose during their amplification, normally after the confluent phage plaque lysate step[39]. Single revertant phage clones were isolated and sequenced to identify compensatory mutations.

Plasmid construction. Plasmid pPT290 was constructed by cloning a PCR fragment spanning genes *11* to *13* (coordinates 6863 to 8815 of the SPP1 genome sequence; GenBank accession number X97918.3[40]) downstream of the inducible promoter $P_{N25/0}$ present in the pHP13 derived plasmid pIV2[41]. The PCR product was cleaved at *Eco*RI and *Pst*I sites and ligated to the vector cut with the same restriction endonucleases. Restriction sites were engineered in the sequence of primers 11CSminus1 (GT*GAATTC***GCGTGAGGTGTGACACG**; the restriction site is shown in italics and the SPP1 sequence in bold) and 13NCS4 (CTA*CTG CAG***CTTTCAAAAAAAGAGAGGCG**) used for PCR amplification.

Plasmids pCC40[38] and pPT290 were used as templates for site-directed mutagenesis with the QuikChange Site-Directed Mutagenesis Kit (Stratagene) to engineer mutations in gene *13*. Primers for the mutagenesis reaction were designed according to the Kit instructions. All plasmid constructions were carried out in *Escherichia coli* DH5α or DH5α (pGB3). Selected clones were transformed into *B. subtilis* YB886 or YB886 (pEB104)[41] for genetic and functional infection experiments with SPP1.

Mutations phenotyping. The effect of gp13 mutations was characterised in strains expressing the gene *13* mutant alleles (bearing plasmid derivatives of pCC40) or co-expressing genes *11* to *13* in which gene 13 was mutagenized (bearing plasmid derivatives of pPT290). The capacity of gp13 to assemble biologically functional capsids was determined by complementation of SPP1*sus31*, a conditional lethal mutant in gene *13*, under non-permissive infection conditions[36,37]. Gp13 mutations defective in the complementation assay were further characterised. Gp11 and gp13 production in extracts of infected cells was assessed by western blot of 15% SDS-PAGE gels[17]. Gp11-gp13 complexes were partially purified in linear 10–30% glycerol gradients and the presence of procapsid-like structures was analysed by western blot and electron microscopy of negatively stained samples[17]. Production of gp12 in infected bacteria and its presence in purified SPP1 phage particles were analysed by western blot[18].

Sample purification. Wild-type SPP1 virions were produced by infection of *B subtilis* YB886 and purified by isopycnic centrifugation in a discontinuous CsCl gradient[39]. SPP1 procapsids were produced by infection of *B subtilis* YB886 with SPP1*sus70*, a mutant defective in the small terminase subunit (TerS) gp1, followed by purification on a 10–30% glycerol gradient and ion-exchange chromatography in a Resource Q column (GE Healthcare)[19]. The presence of procapsids I and II was monitored by electron microscopy throughout the purification procedure to confirm that procapsid II did not result from expansion of procapsid I in vitro (Supplementary Fig. 1a, b). Both structures were found in the pellet of particles sedimented from the SPP1*sus70* lysate, in the procapsids band of the 10–30% linear glycerol gradient, and in the procapsids peak on the ion-exchange chromatography.

Cryo sample preparation and data collection. Cryo-EM grids were prepared using a Vitrobot Mark II (FEI) set to 4 °C and 100% humidity. Samples of SPP1 procapsids and of SPP1 infectious viral particles (3.5 µL at a protein concentration of 1.9 mg/mL) were applied to freshly glow discharged 300 mesh Quantifoil R3/3 grids covered with an additional 2 nm carbon support film. After 45 s the grids were blotted for 2–4 s (blot force 1), plunged into liquid ethane and then stored in liquid nitrogen.

Data were collected using a Tecnai G2 Polara (FEI), operated at 300 kV, and equipped with a Gatan K2 Summit direct detector (Gatan, Inc.) at a defocus range from −0.8 to −3.5 µm. Images were acquired in super-resolution mode using Leginon[42] at ×31,000 nominal magnification corresponding to a 0.64-Å pixel size at object scale. The electron dose was set to 2.5 counts/pixel/s for using the K2 summit in counting super-resolution mode. For each exposure, 25 frames with 200 ms exposure time resulting in an electron dose of 1.1 $e-/Å^2$ per frame and a total dose of 27.5 $e^-/Å^2$ per 5 s exposure. Two independent data sets were collected, one of procapsids and another of mature viral particles.

Image processing and structure analysis. Movie frames from the K2 Summit were aligned using MOTIONCORR-2[43]. Data were 2 × 2 binned, yielding a pixel size of 1.28 Å. All 25 frames were included for motion correction and determination of CTF parameters. For image analysis frame correction was done only for frames 4–21, the first three and last four frames were excluded being affected by the beam induced movement and radiation damage, respectively. The total accumulated dose was 23.1 $e^-/Å^2$. Particles were selected manually using EMAN-2[44]. In all, 18,395 procapsid particles and 6000 mature capsid particles were extracted in frames of 800 × 800 pixels. Assessment of the contrast transfer function (CTF) of the microscope was done using CTFFIND4[45] and phase flipping was applied in SPIDER[46]. IMAGIC-5[20] was used in further image processing. Images of capsids were centred and subjected to multivariate statistical analysis (MSA[47]) that allowed us to detect heterogeneity in the procapsid population and to separate particle images into homogenous groups using hierarchical ascendant classification (HAC[47]). The two major groups of procapsids were identified as procapsid I (6078 particles) and procapsid II (11,745 particles). In total, 572 very small particles (nanocapsids) were discarded. The mature phage capsids population was homogenous (Supplementary Table 1). Orientations of single images were determined using the angular reconstitution technique applying icosahedral symmetry[47]. In all, 3D maps were computed using the filtered back projection method implemented in IMAGIC-5 applying icosahedral symmetry. Refinement was done iteratively using for the next round a 3D map obtained from the images with the lowest errors[47]. Approximately 75% images from each data set were used to reconstruct the final 3D maps, which were subsequently sharpened retaining spatial frequencies between 1/7 and ~1/2.75 Å[47] (Supplementary Table 1). The local resolution for each 3D map was evaluated using the RESMAP program[48].

Model building. Tracing of the SPP1 MCP polypeptide chain was done using one ASU of the mature capsid (~4.0 Å resolution). Extraction of the ASU was done in CHIMERA[49] using the fitting of the previous atomic model of gp13 (PDB 4AN5[21]). Each gp13 subunit of the ASU was then extracted individually. The initial model allowed defining the ASU boundaries sufficiently well although this was an incomplete model that missed the N-terminal arm (Met1—Thr37), the E-loop (Asp61—Lys79), and the C-terminus (Gly290—Ala 324). The EM map at 4 Å resolution (Supplementary Figs. 2 and 9) revealed side-chains of amino acids allowing to trace the complete polypeptide chain and to create the atomic model. The ASU subunit D (Fig. 1d, bottom panel, and 2; Supplementary Fig. 2) was used as the starting model to build the models of other subunits. Modelling was started by rigid body docking of the 4AN5 model followed by flexible fitting into the cryo-EM density using IMODFIT[50]. Once the major gp13 helices position was identified, COOT[51,52] was used for de novo tracing of the gp13 Cα chain. Large side chains were fitted followed by the smaller side chains. The de novo model of subunit D from the mature capsid was refined in PHENIX[53]. More than 90% of the gp13 amino acids were fit into the EM density (Supplementary Movie 1). The final cross-correlation of the fit reached 0.88. Chain D was then used as the starting model to refine individually the fit into the remaining six extracted subunits of the mature capsid ASU, which was done iteratively in COOT and PHENIX (Supplementary Table 1). The mature capsid gp13 chain D structure was also used as the starting model to refine individually the fit of all extracted subunits from procapsid I and procapsid II ASUs.

For each type of capsid the seven chains of the ASU, from chains A to G, were then refined as one single model iteratively in PHENIX and COOT. This procedure was also performed for the nine subunits which form the threefold axis i.e, chains C, D, E from ASU's 1, 2, 3 in the three capsid structures. Ramachandran plot and

results of Molprobity of the results of the atomic modelling presented in Supplementary Fig. 9 and Supplementary Table 1.

Figures were prepared using UCSF Chimera[49].

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data supporting the findings of this study are available within the paper and its supplementary information files and can be obtained from the corresponding authors upon reasonable request. Three-dimensional cryo-EM density maps of the SPP1 procapsid I, SPP1 procapsid II and SPP1 mature capsids have been deposited in the Electron Microscopy Data Bank under the accession numbers EMD-4717 [https://www.ebi.ac.uk/pdbe/entry/emdb/EMD-4717], EMD-10002 [https://www.ebi.ac.uk/pdbe/entry/emdb/EMD-10002], EMD-4716 [https://www.ebi.ac.uk/pdbe/entry/emdb/EMD-4716]. Atomic coordinates of the asymmetric units for each capsid state have been deposited in the RCSB Protein Data Bank under the accession codes 6R3B [https://www.ebi.ac.uk/pdbe/entry/pdb/6R3B], 6RTL [https://www.ebi.ac.uk/pdbe/entry/pdb/6RTL] and 6R3A [https://www.ebi.ac.uk/pdbe/entry/pdb/6R3A] respectively.

Received: 8 April 2019; Accepted: 25 September 2019;
Published online: 24 October 2019

References

1. Hendrix, R. W. Bacteriophage genomics. *Curr. Opin. Microbiol.* **6**, 506–511 (2003).
2. Prasad, B. V. & Schmid, M. F. Principles of virus structural organization. *Adv. Exp. Med. Biol.* **726**, 17–47 (2012).
3. Prevelige, P. E. & Fane, B. A. Building the machines: scaffolding protein functions during bacteriophage morphogenesis. *Adv. Exp. Med. Biol.* **726**, 325–350 (2012).
4. Oh, B., Moyer, C. L., Hendrix, R. W. & Duda, R. L. The delt of the HK97 major capsid protein is essential for assembly. *Virology* **456–457**, 171–178 (2014).
5. Huet, A., Conway, J. F., Letellier, L. & Boulanger, P. In vitro assembly of the T = 13 procapsid of bacteriophage T5 with its scaffolding domain. *J. Virol.* **84**, 9350–9358 (2010).
6. Rao, V. B. & Feiss, M. Mechanisms of DNA packaging by large double-stranded DNA viruses. *Annu. Rev. Virol.* **2**, 351–378 (2015).
7. Wikoff, W. R. et al. Topologically linked protein rings in the bacteriophage HK97 capsid. *Science* **289**, 2129–2133 (2000).
8. Gertsman, I. et al. An unexpected twist in viral capsid maturation. *Nature* **458**, 646–650 (2009).
9. Fokine, A. et al. Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry. *Proc. Natl Acad. Sci. USA* **102**, 7163–7168 (2005).
10. Chen, D. H. et al. Structural basis for scaffolding-mediated assembly and maturation of a dsDNA virus. *Proc. Natl Acad. Sci. USA* **108**, 1355–1360 (2011).
11. Guo, F. et al. Capsid expansion mechanism of bacteriophage T7 revealed by multistate atomic models derived from cryo-EM reconstructions. *Proc. Natl Acad. Sci. USA* **111**, E4606–E4614 (2014).
12. Chen, Z. et al. Cryo-EM structure of the bacteriophage T4 isometric head at 3.3-Å resolution and its relevance to the assembly of icosahedral viruses. *Proc. Natl Acad. Sci. USA* **114**, E8184–E8193 (2017).
13. Bayfield O.W. et al. Cryo-EM structure and in vitro DNA packaging of a thermophilic virus with supersized T = 7 capsids. *Proc. Natl Acad. Sci. USA*. **116**, 3556–3561 (2019).
14. Dearborn, A. D. et al. Competing scaffolding proteins determine capsid size during mobilization of Staphylococcus aureus pathogenicity islands. *Elife* **6**, pii: e30822 (2017).
15. Dai, X. & Zhou, Z. H. Structure of the herpes simplex virus 1 capsid with associated tegument protein complexes. *Science* **360**, pii: eaao7298 (2018).
16. Yuan, S. et al. Cryo-EM structure of a herpesvirus capsid at 3.1 Å. *Science* **360**, eaao7283 (2018).
17. Dröge, A. et al. Shape and DNA packaging activity of bacteriophage SPP1 procapsid: protein components and interactions during assembly. *J. Mol. Biol.* **296**, 117–132 (2000).
18. Zairi, M., Stiege, A. C., Nhiri, N., Jacquet, E. & Tavares, P. The collagen-like protein gp12 is a temperature-dependent reversible binder of SPP1 viral capsids. *J. Biol. Chem.* **289**, 27169–27181 (2014).
19. Oliveira, L., Alonso, J. C. & Tavares, P. A defined in vitro system for DNA packaging by the bacteriophage SPP1: insights into the headful packaging mechanism. *J. Mol. Biol.* **353**, 529–539 (2005).
20. van Heel, M., Harauz, G., Orlova, E. V., Schmidt, R. & Schatz, M. A new generation of the IMAGIC image processing system. *J. Struct. Biol.* **116**, 17–24 (1996).
21. White, H. E. et al. Capsid structure and its stability at the late stages of bacteriophage SPP1 assembly. *J. Virol.* **86**, 6768–6777 (2012).
22. Poh, S. L. et al. Oligomerization of the SPP1 scaffolding protein. *J. Mol. Biol.* **378**, 551–564 (2008).
23. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
24. Sun, Y. et al. Structure of the coat protein-binding domain of the scaffolding protein from a double-stranded DNA virus. *J. Mol. Biol.* **297**, 1195–1202 (2000).
25. Morais, M. C. et al. Bacteriophage phi29 scaffolding protein gp7 before and after prohead assembly. *Nat. Struct. Biol.* **10**, 572–576 (2003).
26. Cortines, J. R. et al. Decoding bacteriophage P22 assembly: identification of two charged residues in scaffolding protein responsible for coat protein interaction. *Virology* **421**, 1–11 (2011).
27. Cortines, J. R. et al. Highly specific salt bridges govern bacteriophage P22 icosahedral capsid assembly: identification of the site in coat protein responsible for interaction with scaffolding protein. *J. Virol.* **88**, 5287–5297 (2014).
28. Lata, R. et al. Maturation dynamics of a viral capsid: visualization of transitional intermediate states. *Cell* **100**, 253–263 (2000).
29. Conway, J. F. et al. A thermally induced phase transition in a viral capsid transforms the hexamers, leaving the pentamers unchanged. *J. Struct. Biol.* **158**, 224–232 (2007).
30. Heymann, J. B. et al. Dynamics of herpes simplex virus capsid maturation visualized by time-lapse cryo-electron microscopy. *Nat. Struct. Biol.* **10**, 334–341 (2003).
31. Prasad, B. V. et al. Three-dimensional transformation of capsids associated with genome packaging in a bacterial virus. *J. Mol. Biol.* **231**, 65–74 (1993).
32. Thuman-Commike, P. A. et al. Three-dimensional structure of scaffolding-containing phage p22 procapsids by electron cryo-microscopy. *J. Mol. Biol.* **260**, 85–98 (1996).
33. Parent, K. N. et al. P22 coat protein structures reveal a novel mechanism for capsid maturation: stability without auxiliary proteins or chemical crosslinks. *Structure* **18**, 390–401 (2010).
34. Riva, S., Polsinelli, M. & Falaschi, A. A new phage of Bacillus subtilis with infectious DNA having separable strands. *J. Mol. Biol.* **35**, 347–356 (1968).
35. Chai, S. et al. Molecular analysis of the Bacillus subtilis bacteriophage SPP1 region encompassing genes *1* to *6*. The products of gene 1 and gene 2 are required for *pac* cleavage. *J. Mol. Biol.* **224**, 87–102 (1992).
36. Becker, B. et al. Head morphogenesis genes of the Bacillus subtilis bacteriophage SPP1. *J. Mol. Biol.* **268**, 822–839 (1997).
37. Dröge, A. & Tavares, P. In vitro packaging of DNA of the Bacillus subtilis bacteriophage SPP1. *J. Mol. Biol.* **296**, 103–115 (2000).
38. Chaban, Y. et al. Structural rearrangements in the phage head-to-tail interface during assembly and infection. *Proc. Natl Acad. Sci. USA* **112**, 7009–7014 (2015).
39. Jakutytė, L. et al. Bacteriophage infection in rod-shaped gram-positive bacteria: evidence for a preferential polar route for phage SPP1 entry in Bacillus subtilis. *J. Bacteriol.* **193**, 4893–4903 (2011).
40. Godinho, L. et al. The revisited genome of Bacillus subtilis bacteriophage SPP1. *Viruses* **10**, pii: E705 (2018).
41. Isidro, A., Henriques, A. O. & Tavares, P. The portal protein plays essential roles at different steps of the SPP1 DNA packaging process. *Virology* **322**, 253–263 (2004).
42. Suloway, C. et al. Automated molecular microscopy: the new leginon system. *J. Struct. Biol.* **151**, 41–60 (2005).
43. Li, X. et al. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods* **10**, 584–590 (2013).
44. Tang, G. et al. EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
45. Rohou, A. & Grigorieff, N. CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
46. Shaikh, T. R. et al. SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. *Nat. Protoc.* **3**, 1941–1974 (2008).
47. van Heel, M. et al. Single-particle electron cryo-microscopy: towards atomic resolution. *Q. Rev. Biophys.* **33**, 307–369 (2000).
48. Kucukelbir, A. et al. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).
49. Pettersen, E. F. et al. UCSF Chimera -a visualization system for exploratory research and analysis. *J. Comput. Chem.* **13**, 1605–1612 (2004).

50. Lopéz-Blanco, J. R. & Chacón, P. iMODFIT: efficient and robust flexible fitting based on vibrational analysis in internal coordinates. *J. Struct. Biol.* **184**, 261–270 (2013).
51. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).
52. Brown, A. et al. Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallogr. D Biol. Crystallogr.* **71**, 136–153 (2015).
53. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol Crystallogr.* 66, (213–221 (2010).

Acknowledgements

We would like to thank I. Vinga for kindly providing plasmid pIV2, Dr. D. Houldershaw, Dr. R. Westlake and Y. Goudetsidis for computer support throughout the duration of the project. Dr. A. Pandurangan and L. Simons are acknowledged for assistance with building the atomic models. We would like to thank M. Ouldali for imaging electron microscopy of particles with mutant gp13 forms in Supplementary Fig. 7c–g, from the platform Cryo-EM of I2BC supported by the French Infrastructure for Integrated Structural Biology (FRISBI) ANR-10-INBS-05. We thank Dr. H. White for useful comments and for proof reading and Dr. S. Roche for the comments on the manuscript. This work was supported by the Biotechnology and Biological Sciences Research Council to E.V.O. and A.I., by CNRS institutional funding to P.T. and by MPI institutional funding to T.M.

Author contributions

S.B. purified complexes and prepared samples. M.E.S.F., S.B., and P.T. carried out mutagenesis and phenotyping; J.B. and T.M. prepared grids for cryo imaging and cryo-EM data acquisition; A.I. and E.V.O. analysed the cryo-EM data, obtained the structures, A.I. built the atomic models, and movies; A.I., M.T., E.V.O., and P.T. interpreted the results. E.V.O. and P.T. designed the experiments and supervised research. A.I., E.V.O., and P.T. wrote the manuscript. All authors discussed and commented on the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-019-12790-6.

Correspondence and requests for materials should be addressed to P.T. or E.V.O.

Peer review information *Nature Communications* thanks Robert Duda, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

viruses

MDPI

Article

The Revisited Genome of *Bacillus subtilis* Bacteriophage SPP1

Lia M. Godinho [1], Mehdi El Sadek Fadel [1], Céline Monniot [2], Lina Jakutyte [3], Isabelle Auzat [1], Audrey Labarde [1], Karima Djacem [1,†], Leonor Oliveira [1], Rut Carballido-Lopez [2], Silvia Ayora [4] and Paulo Tavares [1,*]

1 Institut de Biologie Intégrative de la Cellule (I2BC), French Alternative Energies and Atomic Energy Commission (CEA), Centre National de la Recherche Scientifique (CNRS), Univ Paris-Sud, Université Paris-Saclay, 91190 Gif-sur-Yvette, France; lia.marques-godinho@i2bc.paris-saclay.fr (L.M.G.); mehdi.elsadekfadel@i2bc.paris-saclay.fr (M.E.S.F.); Isabelle.AUZAT@i2bc.paris-saclay.fr (I.A.); audrey.labarde@i2bc.paris-saclay.fr (A.L.); karima.djacem@gmail.com (K.D.); Leonor.OLIVEIRA@i2bc.paris-saclay.fr (L.O.)

2 MICALIS, Institut National de la Recherche Agronomique (INRA), AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France; celine.monniot@jouy.inra.fr (C.M.); rut.carballido-lopez@inra.fr (R.C.-L.)

3 Unité de Virologie Moléculaire et Structurale (VMS), CNRS, 91198 Gif-sur-Yvette, France; lina.jakutyte@gmail.com

4 Centro Nacional de Biotecnología (CNB-CSIC), 28049 Madrid, Spain; sayora@cnb.csic.es

* Correspondence: paulo.tavares@i2bc.paris-saclay.fr

† Current address: Folium Science, Unit DX, St Philips Central, Albert road, Bristol BS2 0XJ, UK.

Received: 23 October 2018; Accepted: 6 December 2018; Published: 11 December 2018

Abstract: *Bacillus subtilis* bacteriophage SPP1 is a lytic siphovirus first described 50 years ago. Its complete DNA sequence was reported in 1997. Here we present an updated annotation of the 44,016 bp SPP1 genome and its correlation to different steps of the viral multiplication process. Five early polycistronic transcriptional units encode phage DNA replication proteins and lysis functions together with less characterized, mostly non-essential, functions. Late transcription drives synthesis of proteins necessary for SPP1 viral particles assembly and for cell lysis, together with a short set of proteins of unknown function. The extensive genetic, biochemical and structural biology studies on the molecular mechanisms of SPP1 DNA replication and phage particle assembly rendered it a model system for tailed phages research. We propose SPP1 as the reference species for a new SPP1-like viruses genus of the *Siphoviridae* family.

Keywords: Bacteriophage; *Siphoviridae*; SPP1; *Bacillus subtilis*; genome organization; DNA replication; virus assembly; virus DNA packaging; virus evolution

1. Introduction

The lytic bacteriophage SPP1 (Subtilis Phage Pavia 1) that infects the soil bacterium *Bacillus subtilis* was isolated in the Botanical Garden of Pavia in Italy [1]. The capacity of its purified DNA to transfect *B. subtilis* competent cells and the difference of density between its two DNA strands, rendering easy their separation following denaturation, were quickly recognized. They attracted initial interest to this phage system for studying DNA uptake into bacteria [2,3], mismatch repair [2,3] and recombination [4–6]. Establishment of the genetic [7] and physical [8,9] maps of the SPP1 genome then paved the way for studies of phage gene expression, DNA replication and assembly of the viral particle. The intensive research that followed on the molecular mechanisms supporting SPP1 infection rendered it one of the best characterized phages of Gram-positive bacteria. SPP1 generalized transduction was

also instrumental for fine genetic mapping of the *B. subtilis* chromosome, strain construction, and plasmid horizontal transfer [5,10–14].

SPP1 belongs to the family *Siphoviridae*. Its viral particle has an isometric icosahedral capsid with a diameter of 61 nm [15] and a 190 nm-long non-contractile tail [16]. The linear double-stranded DNA (dsDNA) molecule contained in the viral capsid has a length of ~45.9 kbp with a variation in size of ~2.5 kbp [17,18]. The phage genome size is 44,016 bp. The DNA molecules in a phage population are terminally redundant and partially circularly permuted, resulting from a headful packaging mechanism [19,20].

SPP1 infection is initiated by reversible adsorption of the viral particle to glycosylated teichoic acids [21] followed by irreversible binding of SPP1 to YueB. This integral membrane protein is a component of a type VII secretion system [22] that crosses the *B. subtilis* cell wall to be exposed at the bacterial surface [23–25]. The interaction of the SPP1 tail fiber with YueB triggers ejection of SPP1 DNA through its tail tube, committing the phage to infection. Phage DNA circularizes in the bacterial cytoplasm most probably by recombination between its redundant ends [26–30]. DNA replication then ensues in a discrete position of the bacterial cytoplasm [24]. DNA synthesis was proposed to initiate by theta replication of circular molecules followed by a switch to rolling circle (sigma) replication that generates concatemers of the SPP1 genome [30,31]. These are the substrate for phage DNA encapsidation into a preformed procapsid structure. Packaging is initiated by specific recognition and cleavage of a *pac* sequence within the SPP1 genome, followed by its translocation into the procapsid interior through a specialized portal vertex. Encapsidation is terminated by an imprecise [18], sequence-independent, endonucleolytic cleavage of the substrate concatemer when a threshold amount of DNA is reached inside the capsid (headful packaging mechanism) [19,20]. Subsequent encapsidation cycles follow processively along the concatemer. After packaging, DNA is retained inside the capsid by proteins that close the portal vertex and build the interface for attachment of the phage tail. Binding of the tail, which is assembled in an independent pathway, yields the infectious particle (virion). The SPP1 tail is formed by the adsorption apparatus, responsible for interaction with the host cell surface, which is connected by a 160-nm long helical tail tube to the tail tapered end that binds to the capsid portal [16,32]. Virions accumulate in the infected bacterium until lysis. Lysis is promoted by membrane proteins (holins) that concentrate in the cytoplasmic membrane leading to its disruption and release of an endolysin that digests the *B. subtilis* cell wall [33].

Here we revisit the sequence and organization of the bacteriophage SPP1 genome. A detailed updated annotation of its genes combining available experimental data and bioinformatics completes this comprehensive study of SPP1 genetics and biology.

2. Materials and Methods

The SPP1 genome was fully sequenced using Illumina sequencing of genome libraries at the I2BC NGS platform facility, as described [34,35]. Sanger sequencing of PCR fragments covering the SPP1 genome was carried out at GATC Biotech (Germany).

Open reading frames (ORFs) identification and visualization of the SPP1 chromosome organization was performed using the SnapGene Viewer software (GSL Biotech, Chicago, IL) and Fgenesb annotator (http://www.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb) [36]. A cut-off of ORFs initiated at AUG, GUG or UUG codons that code for putative proteins of at least 45 amino acids-long was used as criterion to identify the ORFs for annotation. Among those, there are 20 ORFs without ribosome binding site (RBS) that are embedded in longer ORFs. Such internal ORFs were eliminated from our downstream analysis because they are most likely not translated. The strong distribution bias of ORFs coded by the SPP1 DNA heavy strand together with transcriptional and functional data (compiled in [26]) support that this strand is used, possibly exclusively, as the coding DNA template. Therefore, the analysis described here is limited to 80 ORFs, defined according to the above criteria, which are encoded by the heavy strand. The assignments made in [26] and a survey of SPP1 functional studies were used as the primary information for ORF

annotation. Visual inspection of the ORF 5′ region was used to confirm the initiation codon and to assess if it is preceded by an RBS whose sequence is complementary to the *B. subtilis* 16 S rRNA 3′ sequence UCUUUCCUCCACUAG [37]. A spacing of 8–14 nucleotides was allowed between the center of the RBS (complementary to the U underlined) and the nucleotide preceding the initiation codon. Sequence and structural homology searches of the ORFs encoded proteins were carried out with BLASTp [38,39] and HHPred [40], respectively. Protein properties were analyzed with the Expasy ProtParam tool (https://web.expasy.org/protparam/). Prediction of transmembrane helices was carried out with the TMHMM server v.2.0 (http://www.cbs.dtu.dk/services/TMHMM/).

SPP1 putative early promoters were predicted using BPROM. BPROM recognizes promoter sequences bound by bacterial sigma factors of the σ^{70} family with about 80% accuracy and specificity (http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb) [36]. These include promoters recognized by the primary sigma factor σ^{A} of *B. subtilis*. Promoter sequence prediction was carried out within the regions where early promoters were previously mapped [26,41,42]. The −10 and −35 sequences of late promoter *P*L1 were used for word scanning of the SPP1 genome to search for putative late promoters without success. Rho-independent transcriptional terminators were determined using ARNold (http://rna.igmors.u-psud.fr/toolbox/arnold/index.php) [43–46], performing a whole sequence search analysis on the coding strand using two complementary programs, Erpin and RNAmotif. The free energy (ΔG°) of the predicted terminator stem-loop structure was computed with ARNold [44].

Codon usage bias of the 80 SPP1 ORFs was calculated using the Codon Usage Calculator from Biologics International Corp (https://www.biologicscorp.com/tools/CodonUsageCalculator/). The *B. subtilis* genome codon usage frequencies were obtained from [47,48].

The revised version of the complete SPP1 genome sequence and its annotation are in the process of submission to Genbank with accession code "X97918.3". Table 1 summarizes the changes identified relative to the previous sequence.

Table 1. Revision of the SPP1 sequence. Changes in the GenBank X97918.2 sequence relative to the GenBank X97918.3 revised version are listed. The position coordinates presented are the ones found in the GenBank X97918.3 sequence. The ORF concerned, the nucleotide sequence change(s), and the effect(s) on the gene product amino acid sequence are displayed.

Position (nt)	ORF	ORF Sequence Change	ORF Product Amino Acid Change
230	*1*	insertion of a G; frameshift	gp1 Cter sequence changed and shortened
766	*2*	T→C	none (silent mutation)
1997	*4*	insertion of an A; frameshift	gp4 Cter changed and lengthened
5001	intergenic region ORFs *7* and *8*	insertion of a T	none
23,794	*30*	deletion of a C; frameshift	gp30 Cter changed and lengthened
34,450	*37.1*	T→G	gp37.1 C_{18}→G_{18}
37,304 and 37,306	*40*	GGC→CGG	gp40 $R_{194}R_{195}$→$P_{194}G_{195}$
38,371 to 38,374	*41*	GTGT→TGTG	gp41 $K_{101}C_{102}$→$N_{101}V_{102}$
40,240	*44* and former *45*	insertion of a G; frameshift; ORFs fusion	gp44 Cter lengthened by former gp45 sequence
41,787	*48*	insertion of an A; frameshift	gp48 changed and shortened
42,750	*51*	insertion of an A; frameshift	gp51 changed and lengthened
42,778 to 42,781	*51*	CGCG→GCGC	gp51 $R_{56}E_{57}$→$A_{56}Q_{57}$
42,819	*51*	insertion of a G; frameshift	gp51 Cter lengthened

3. Results and Discussion

3.1. Properties of the SPP1 DNA Molecule

The genome of SPP1 was completely sequenced in the context of research projects on phage DNA replication and viral particle assembly, as well as by the necessity for an accurate genome annotation for downstream SPP1 "omics" studies. Very few changes were identified relative to the deposited sequence (Genbank accession code X97918) confirming the high quality of the original Sanger sequencing work [26] and its 2006 revision (X97918.2; [16]). Changes in the nucleotide sequence are listed in Table 1. The 44,016 bp genome has a GC content of 43.7% which is similar to the one found in the host *B. subtilis* genome (43.5%) [49]. The distribution of purines and pyrimidines in the two DNA strands is very different. This asymmetry provided the physical basis for their separation after denaturation by density using isopycnic centrifugation [1]. The heavy strand, whose purine (dA+dG) content is 58.4%, has a density (ρ) of 1.725 g/cm^3. It is the SPP1 genome coding strand. The light chain density is 1.713 g/cm^3 [1,3]. The strongest sequence bias are dA+dT rich-islands found in intergenic regions involved in transcriptional regulation (see Section 3.2) and in the two origins of SPP1 DNA replication *ori*R and *ori*L (see Section 3.5) (Figure 1).

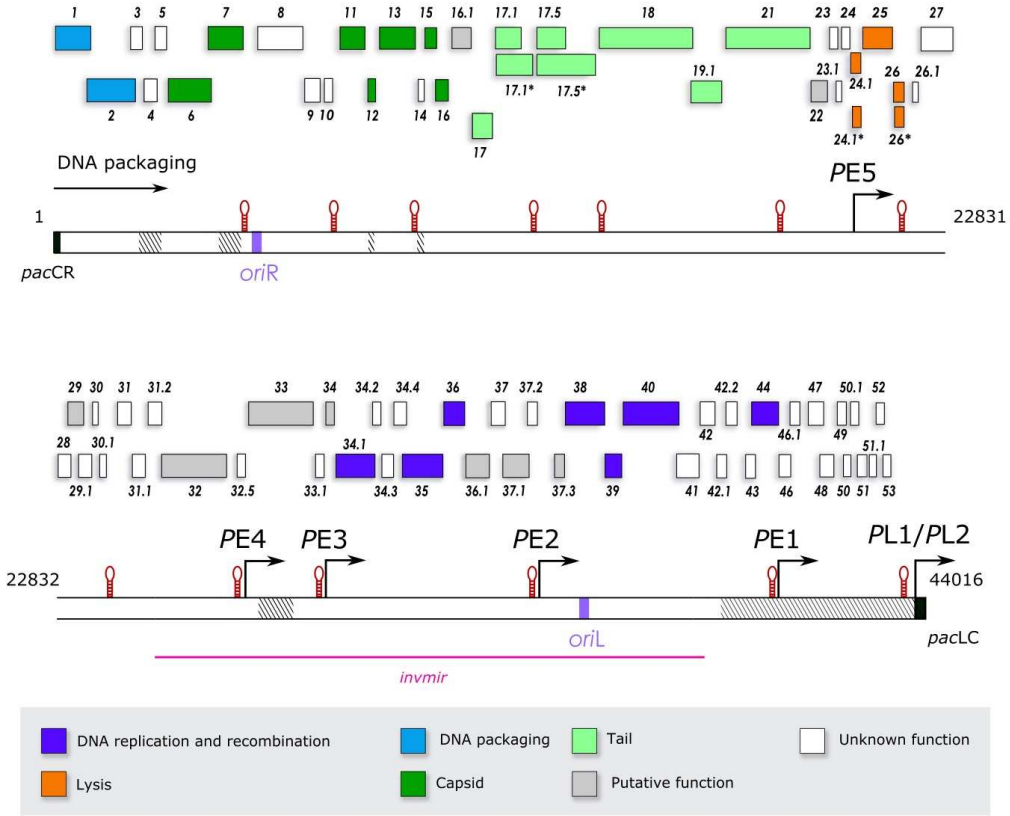


Figure 1. Organization of the SPP1 genome. The continuous bar represents the 44,016 bp-long genome where coordinate 1 is the main *pac* cleavage position [35]. The two origins of replication *ori*R and *ori*L (magenta), the DNA packaging signal *pac* (black) and non-essential regions of the SPP1 genome defined by deletions (dashed) are highlighted in the bar. The sequence inverted in SPP1*invmir* is displayed by a pink line underneath the genome bar. The position of promoters (Figure 2) and potential Rho-independent transcriptional terminators that form stem loops (red) in mRNA (Table 2) is displayed on top of the bar. Transcription is from left to right. DNA packaging initiated at *pac* occurs in the same direction (arrow on the top left). The set of SPP1 genes and ORFs, identified as described in Materials and Methods (see Section 2), are presented above the genome bar and colored according to their function assignment shown on the bottom legend.

3.2. Transcription and Translation of the SPP1 Genome

Transcription of the SPP1 genome was reported to occur from the DNA heavy strand [1,50–52]. This is not an absolute requirement, as phage SPP1*invmir* carries an inversion within its genome that leads to transcription from the light chain of a large genome segment which includes the DNA replication genes (Figure 1) [53]. The temporal gene expression program is defined by early and late transcription carried out by the *B. subtilis* RNA polymerase [54]. Five promoters with a canonical sequence recognized by *B. subtilis* σ^A direct transcription of early ORFs/genes (we name gene the ORFs for which a function was demonstrated experimentally or supported by substantial bioinformatics data) (Figures 1 and 2) [33,41,55,56]. These polycistronic transcriptional units are bracketed by the promoter and by putative Rho-independent transcriptional terminators coding for potential RNA stem-loop-forming sequences (Figure 1 and Table 2). The two strongest promoters (*P*E3 and *P*E2) drive transcription of gene sets that include essential DNA replication functions (Figure 1) [41,52,56]. *P*E1, also a strong promoter, and the weaker *P*E4 and *P*E5 promoters control expression of less characterized ORFs (see Section 3.4). Early genes/ORFs are encoded contiguously in the SPP1 genome. Transcription of the late genes segment requires translation of (an) early, yet unidentified, SPP1 factor(s) [55,57]. The only late promoters experimentally characterized are the adjacent *P*L1 and *P*L2 (Figure 1). *P*L1, which accounts for more than 95% of the transcriptional activity of the region upstream from gene *1*, has a canonical −10 sequence and an atypical −35 region ([57]; Figure 2). Several clusters of late genes are delimited by potential stem-loop transcription terminator sequences ([26]; Figure 1), mostly found in intergenic regions, but no consensual late promoter sequence related to *P*L1 was identified.

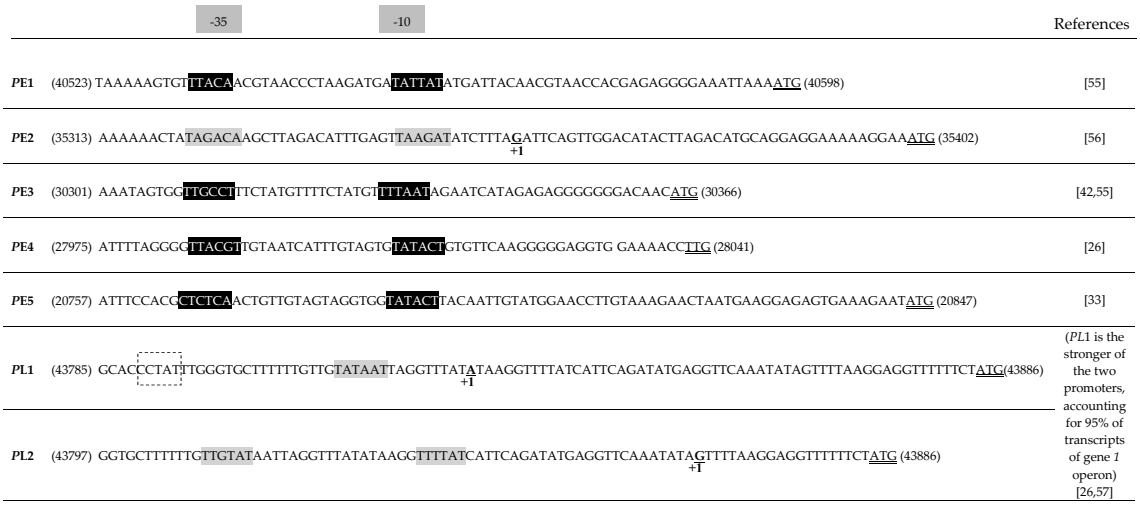


Figure 2. SPP1 promoters. The sequence of SPP1 promoters and the initiation codon of their downstream gene (double underline) are displayed. The −35 and −10 promoter regions are shaded in grey for promoters whose transcription start position (+1) was determined experimentally. A dashed box denotes the atypical −35 sequence of *P*L1. SPP1 putative early promoters identified by sequence similarity to *B. subtilis* vegetative promoters are highlighted in black with sequence characters in white. Note that the approximate position of transcription initiation and promoter strength was determined for all early promoters by electron microscopy of DNA-RNA polymerase complexes [41,42].

Table 2. Rho-independent transcriptional terminators of SPP1. The terminators were identified as described in Material and Methods (see Section 2). Putative stem (underlined) and loop (bold) regions are highlighted. The coordinates of the sequence presented are shown (nt—nucleotide). The ΔG° of the RNA stem loop is shown on the right column.

Terminator Sequence (Loop, Stem)	Start (nt)	End (nt)	ΔG° (kcal.mol^{-1})
AAAAGCGAGTTAACCGACG**TAAAAATG**CGTCGGTTTTTTTCGTGCTTC	4718	4765	−12.20
GTGTCAGTGCGCGGGTTC**AATTCCCG**GAGTCCGTTTTTACCGCCC	6780	6824	−5.50
TAAAAAGAGAAGAGGGGC**TAAC**GCCTCTCTTTTTTTGAAAG	8775	8815	−14.00
ATTGCCAGCAGAGAGCAC**GGGTTAATTCC**CGCGCTCTTTTTTTGTATTCA	11,255	11,304	−9.20
TGAATTAGACAGGGCCGC**GCAA**GTGGCTCTTTTTAATAGGT	12,009	12,049	−15.20
TGAAAAGACGCGGCGCCG**CTAA**CGGCGTCCGTTTTGAACATGA	16,711	16,753	−12.10
GATTGACGAAGTAAAGGGC**CATGT**GCCCTTTATTTTTTTGCAAA	21,897	21,940	−10.70
ACCTTCGCTTGCCGCCCG**GCTGAT**GGGCGGTTTTTTATTTTT	24,095	24,136	−13.20
TGTCCTAAAATCGGCCCGT**TCCCAG**TCGGGCCACTTTTTTTATTTTA	27,934	27,980	−13.60
GACTTTGAAAGGAACCGTT**CTCT**AACGGTTCTTTTTTTATTTC	30,205	30,247	−9.90
TATTTTGATTGAGTCGGGG**AAA**CCCGGCTTTTTTATTTTGGG	35,232	35,273	−14.90
ATCAAAGTTATGGTGGGA**GTAA**TCCCGCCTTTTTCTATTTT	40,466	40,506	−14.40
AACACAGAGAGGCACCC**TATTT**GGGTGCTTTTTTGTTGTA	43,774	43,813	−11.90

Genome-wide sequence analysis and subsequent individual inspection identified 80 ORFs encoded by the SPP1 wild type DNA heavy chain with a length longer than 45 codons (135 bp), as described in Material and Methods (see Section 2) (Table 3; Figure 1). A stringent criterion was used to annotate only ORFs that are most likely translated, an assignment further supported by the subsequent search for their 5′ RBS (see below). The ORFs defined cover 94% of the SPP1 sequence. Non-coding segments longer than 45 bp were characterized in most cases by the presence of transcriptional promoters and/or terminators sequences, but were also found in five cases between adjacent genes within a transcriptional unit (genes *24–24.1*, *26.1–27*, *31–31.1*, *46.1–47*, *48–49*). Only one ORF longer than 15 codons was identified in these intergenic regions. It has 33 codons starting by AUG and preceded by an RBS in the segment between ORFs *46.1* and *47* (not shown). It cannot be excluded at present that other regions of the genome code polypeptides smaller than 45 amino acids-long, as found for several phages [58,59]. We are pursuing proteomic and genetic studies for an accurate annotation of short SPP1 ORFs. An RBS whose sequence is complementary to the *B. subtilis* 16 S rRNA 3′ was found in the 5′ region of 77 from the 80 ORFs investigated here (Table 3). No RBS was identified for genes *2* and *16* that encode well-characterized proteins essential for phage particle assembly (Table 3). Note that in case of gene *16* the initiation codon was confirmed experimentally by amino-terminus sequencing of its encoded protein [60]. The initiation codon of gene *2* overlaps the stop codon of its 5′ gene *1* (**AUGA**; the initiation codon is shown in bold and the termination codon in double underlined) while the initiation codon of gene *16* is spaced by one nucleotide from the gene *15* termination codon (UAAG**AUG**). This organization possibly allows coupling initiation of translation of the genes lacking an RBS with translational termination of their upstream gene [61]. It might also provide a strategy to control the ratio between proteins encoded by adjacent genes. Those proteins interact directly during phage assembly: gene product 1 (gp1 (note that the designations gpX (gene product X) and GXP (gene *X* product) are synonymous in the SPP1 literature)) and gp2 form the DNA packaging terminase while gp15 and gp16 bind sequentially to the capsid portal after viral DNA packaging (see Section 3.8). ORF *42.2* also lacks a canonical RBS. Its initiation codon overlaps the termination codon of ORF *42.1* (**AUGA**) and is preceded by a GGGG sequence that could act as a weak RBS (Table 3), probably ensuring gp42.2 synthesis. It is possible that other SPP1 ORFs without RBS, a situation found for ~10% of the host *B. subtilis* ORFs [62], are expressed during SPP1 infection. These can be either smaller than 45 codons or embedded within SPP1 annotated ORFs (see Section 2).

Table 3. SPP1 ORFs. The 80 ORFs of the SPP1 transcribed heavy chain were assigned as described in Material and Methods (see Section 2) and numbered following the original nomenclature of Alonso et al. [26]. The RBS are sequences complementary to the *B. subtilis* 16 S rRNA 3′ sequence (mismatches are shown in small case; the position after which spacing is calculated between the RBS and the nucleotide preceding the initiation codon is underlined; n.d.—not determined). The ORFs coding regions coordinates are listed. Genes are defined essential when their inactivation in conditional lethal mutants prevents phage multiplication.The length, molecular mass (MM) and presence of putative transmembrane segments in proteins translated from the ORFs coding frame are listed on the right side of the Table. The coordinates of X-ray crystallography, NMR or cryo-electron microscopy structures available for individual proteins or their complexes are also provided. The function of individual proteins is based on experimental data while putative function is deduced from bioinformatics analysis. Structural components of the SPP1 viral particle (st.) are assigned based on biochemical, structural and/or robust bioinformatics data. Proteins are grouped according to function following the color code used in Figure 1.

			ORF					Protein					
ORF	**RBS (mRNA)**	**Spacing**	**Start**	**nt**	**Stop**	**nt**	**Essential Gene**	**Length (aa)**	**MM (kDa)**	**Predicted TMM Segments**	**3D Structure (PDB or EMD)**	**Protein Function**	**References**
1	AAGGAGGU	10	AUG	43,884	UGA	311	yes	147	16.3	no	3ZQQ [a] (Xtal [b])	small terminase subunit (TerS)	[57,63–69]
2	n.d. [c]		AUG [d]	308 [d]	UAG	1576	yes	422 [d]	48.8 [d]	1 [e]	2WBN [f]; 2WC9 [f] (Xtal [b])	large terminase subunit (TerL)	[57,67,70–74]
3	AAAGGAGG	11	AUG	1567	UAA	1782	no	71	8.5	no	n.d.	unknown	
4	GGgGGU	10	AUG	1782	UAA	2072	no	96	11.4	no	n.d.	unknown	
5	AAGGAGG	11	AUG	2065	UGA	2334	no	89	10.3	no	n.d.	unknown	
6	AGGAGGU	11	AUG	2336	UGA	3847	yes	503	57.3	no	2JES (Xtal [b]); 5A20, 5A21 (cryoEM [g])	st.; portal protein	[20,60,67,73–82]
7	AGGAGG	12	AUG	3804	UAA	4730	no	308	35.1	no	n.d.	st.; initiation of infection; binds to portal	[83,84]
8	AAAGGAG	12	AUG	5067	UGA	6215	n.d.	382	43.7	no	n.d.	unknown	
9	AAcGGAGG	9	AUG	6217	UAA	6555	n.d.	112	12.6	no	n.d.	unknown	
10	GGuGGUG [h]	12 [h]	AUG	6583	UAG	6750	n.d.	55	6.2	no	n.d.	unknown	
11	AGGAG	9	AUG	6917	UAA	7561	yes	214	23.4	no	n.d.	procapsid scaffolding protein	[85–87]
12	AAGGgGG	11	AUG	7576	UAA	7770	no	64	6.6	no	n.d.	st.; capsid accessory protein with collagen-like fold	[15,88]
13	AAAGGAG	9	AUG	7803	UAA	8777	yes	324	35.3	no	4AN5 (cryoEM [g])	st.; major capsid protein (MCP)	[15,85,86]
14	AAAGGAG	10	AUG	8828	UGA	9004	no	58	6.7	no	n.d.	unknown	
15	AAuGAGG	10	AUG	9015	UAA	9323	yes	102	11.6	no	2KBZ (NMR [i]); 5A20, 5A21 (cryoEM [g])	st.; connector adaptor protein	[60,78,81,89]
16	n.d. [c]		AUG [j]	9325 [j]	UAG	9654	yes	109 [j]	12.5 [j]	no	2KCA (NMR [i]) 5A20, 5A21 (cryoEM [g])	st.; connector stopper protein	[60,78,81,89]
16.1	AAaGAGG	11	AUG	9644	UGA	10,069	n.d.	141	15.9	no	n.d.	putative tail protein	
17	AGGAGGU	10	AUG	10,066	UGA	10,470	yes	134	15	no	2LFP (NMR [i])	st.; tail-to-head joining protein (THJP)	[32,90]
17.1	AGGAGG	10	AUG	10,484 [k]	UAA	11,017	yes	177	19.2	no	n.d.	st.; tail tube protein (TTP)	[16,91,92]

Table 3. *Cont.*

ORF	RBS (mRNA)	Spacing	ORF Start	ORF nt	ORF Stop	ORF nt	Essential Gene	Protein Length (aa)	Protein MM (kDa)	Protein Predicted TMM Segments	Protein 3D Structure (PDB or EMD)	Protein Function	References
17.1*	AGGAGG	10	AUG	10,484 [k]	UAA	11,279	no	264	28.2	no	n.d.	st.; tail tube protein; Cter FN3 motif	[91]
17.5	GAGG	12	AUG	11,363 [l]	UAA	11,884	n.d. [m]	173	20.2	no	n.d.	tail chaperone protein	[91]
17.5*	GAGG	12	AUG	11,363 [l]	UAG	12,255	n.d. [m]	297	34	no	n.d.	tail chaperone protein	[91]
18	AGGAGG	9	AUG	12,267	UGA	15,365	n.d. [m]	1032	110.9	4 [n]	n.d.	st.; tape measure protein (TMP)	[16]
19.1	GAGG	10	AUG	15,362	UAA	16,123	n.d. [m]	253	28.6	no	2X8K (Xtal [b])	st.; distal tail protein (Dit)	[93,94]
21	AAGaAGGUGA	10	UUG	16,137	UAA	19,463	n.d. [m]	1108	123.6	no	n.d.	st.; tail tip protein; Tal; anti-receptor protein	[93,95]
22	AAGGAGG	9	AUG	19,476	UAA	19,916	n.d.	146	16.7	no	2XC8 (Xtal [b])	putative tail protein	[96]
23	AGGAGGU	10	AUG	19,932	UGA	20,096	n.d.	54	6.1	no	n.d.	n.d.	
23.1	GGAG	9	AUG	20,089	UAA	20,244	n.d.	51	5.8	no	2XF7 (Xtal [b])	n.d.	[97]
24	GgGGUG [h]	10 [h]	AUG	20,237	UAG	20,467	n.d.	76	8.4	no	n.d.	n.d.	
24.1	AAAGGgGG	11	AUG	20,547	UAA	20,825	n.d.	92	10.6	1	n.d.	component of holin; cell lysis	[33,98]
24.1*	AGGAGGU	10	AUG	20,574	UAA	20,825	n.d.	83	9.5	1	n.d.	component of holin; cell lysis	[33]
25	AAGGAG	12	AUG	20,845	UAA	21,660	n.d.	271	29.9	no	n.d.	endolysin; cell lysis	[33,98]
26	AAAGGAG	8	AUG	21,662	UAA	21,910	n.d.	82	9.4	2	n.d.	component of holin; cell lysis	[33,98]
26*	AAAGGAG [o]	14 [o]	AUG	21,668	UAA	21,910	n.d.	80	9.1	2	n.d.	component of holin; cell lysis	[33]
26.1	AAGGgGG [o]	10 [o]	AUG	22,009	UAG	22,152	n.d.	47	5.8	no	n.d.	unknown	
27	AAGGAGG	12	UUG	22,277	UAA	22,831	n.d.	184	20.8	no	n.d.	unknown	
28	GGAGG	9	AUG	22,834	UGA	23,121	n.d.	95	10.8	no	n.d.	unknown	
29	AGGAGG	13	AUG	23,069	UGA	23,371	n.d.	100	12	no	n.d.	putative DNA binding protein	
29.1	AGGgGG	9	GUG	23,358	UGA	23,675	n.d.	105	12.3	no	n.d.	unknown	
30	AGGgGG	10	AUG	23,675	UAA	23,854	n.d.	59	7.2	1	n.d.	unknown	
30.1	AGGgGG	9	AUG	23,859	UGA	24,029	n.d.	56	6.4	no	n.d.	unknown	
31	AAcGGAGGU	12	AUG	24,209	UAA	24,493	n.d.	94	11	no	n.d.	unknown	
31.1	GAGG	12	AUG	24,589	UAA	24,951	n.d.	120	12.9	3	n.d.	unknown	
31.2	GgGGUG [h]	10 [h]	AUG	24,964	UGA	25,281	no [p]	105	11.5	2	n.d.	unknown	
32	GGAGGUG	8	AUG	25,278	UAA	27,788	no [p]	836	96.3	no	n.d.	putative ATP-binding protein	
32.5	GGAGGUG	11	UUG	28,039	UAA	28,209	n.d.	56	6.7	no	n.d.	unknown	
33	AAAaGgGGU	11	AUG	28,226	UAA	29,995	no	589	64.9	no	n.d.	putative bacteria surface binding protein	
33.1	AAcGGAGG	9	AUG	30,011	UAA	30,229	n.d.	72	8.4	no	n.d.	unknown	
34	AGGgGG	9	AUG	30,364	UAG	30,522	n.d.	52	6.3	no	n.d.	putative transcriptional repressor	
34.1	AGGAGG	9	AUG	30,534	UGA	31,469	no	311	35.9	no	n.d.	5′-3′ exonuclease	[13,99]
34.2	GGAGG	12	AUG	31,466	UAA	31,639	n.d.	57	6.7	no	n.d.	unknown	
34.3	AAGGAGG	11	AUG	31,641	UAA	31,895	n.d.	84	9.8	no	n.d.	unknown	
34.4	GGAGG	11	AUG	31,897	UAA	32,187	n.d.	96	11.1	no	n.d.	unknown	
35	GGAGGU	11	AUG	32,177	UAG	33,040	yes	287	32	no	n.d.	recT-like recombinase	[13,100]
36	AGGAGGgGA	10	AUG	33,033	UAA	33,512	no	159	17.1	no	n.d.	SSB	[101]
36.1	AAaGgGGUGA [h]	10 [h]	AUG	33,537	UGA	34,028	n.d.	163	18.9	no	n.d.	putative HNH endonuclease	
37	GGAGG	10	AUG	34,032	UGA	34,406	n.d.	124	14.3	no	n.d.	unknown	

Table 3. *Cont.*

ORF	RBS (mRNA)	Spacing	ORF Start	ORF nt	ORF Stop	ORF nt	Essential Gene	Protein Length (aa)	Protein MM (kDa)	Protein Predicted TMM Segments	Protein 3D Structure (PDB or EMD)	Protein Protein Function	References
37.1	GGAGG	11	AUG	34,399	UGA	34,992	n.d.	197	22.3	no	n.d.	putative poly-gamma-glutamate hydrolase	
37.2	GaAGG	14	AUG	34,989	UGA	35,243	n.d.	84	9.7	no	n.d.	unknown	
37.3	AGGAGG	11	AUG	35,400	UAG	35,573	n.d.	57	6.7	no	n.d.	putative DNA binding protein	
38	AAGGAGG	13	AUG	35,580	UGA	36,350	yes	256	30	no	n.d.	SPP1 origin binding protein and replication re-start (PriA-like)	[56,102]
39	AGGAGG	9	AUG	36,347	UGA	36,727	yes	126	14.6	no	1NO1 (Xtal [b])	gp40 helicase loader	[56,103,104]
40	GAGG	11	AUG	36,724	UAA	38,052	yes	442	49.7	no	3BGW (Xtal [b])	replicative DNA helicase; binds host DnaG and DnaX	[56,103,105–109]
41	AAAGGgGG	10	AUG	38,069	UGA	38,569	n.d.	166	19.1	no	n.d.	unknown	
42	AAAGGAG	11	AUG	38,566	UAA	38,961	no	131	16	no	n.d.	unknown	
42.1	GGAGG	9	AUG	38,964	UGA	39,134	no	56	6.5	no	n.d.	unknown	
42.2	GgGG [c, q]	12 [c, q]	AUG	39,131 [q]	UAA	39,427	no	98 [q]	10.7 [q]	no	n.d.	unknown	
43	GGAGG	10	GUG	39,431	UGA	39,784	no	117	14.2	no	n.d.	unknown	
44	AAGGAG	11	AUG	39,777	UAA	40,487	no	236	27.5	no	n.d.	Holliday junction resolvase	[13,31]
46	GAGG	12	AUG	40,596	UAA	40,898	no	100	11.5	no	n.d.	unknown	
46.1	AGGAGG	9	AUG	40,898	UAA	41,209	no	103	11.7	3	n.d.	unknown	
47	AGGgGG	9	AUG	41,304	UGA	41,663	no	119	13.7	no	n.d.	unknown	
48	GGAG	13	AUG	41,645	UGA	41,995	no	116	13.2	no	n.d.	unknown	
49	GGAGG	9	GUG	42,075	UGA	42,248	no	57	6.5	no	n.d.	unknown	
50	AAGGAGG	9	GUG	42,245	UAA	42,418	no	57	6.8	no	n.d.	unknown	
50.1	AAAGGAGG	9	GUG	42,434	UGA	42,616	no	60	6.7	2	n.d.	unknown	
51	AAGGAGG	9	AUG	42,613	UAA	43,014	no	133	14.7	1	n.d.	unknown	
51.1	AAGGAG	9	AUG	43,027	UGA	43,182	no	51	6.1	1	n.d.	unknown	
52	AAAGGAG	10	AUG	43,179	UGA	43,421	no	80	9.5	no	n.d.	unknown	
53	AAAGGAG	10	AUG	43,405	UGA	43,611	no	68	7.4	1	n.d.	unknown	

[a] structure of gp1 from the SPP1-related phage SF6 (83% amino acid sequence identity with SPP1 gp1); [b] X-ray crystallography structure PDB access code; [c] No RBS identified according to the criteria defined in Material and Methods (see Section 2). In case of ORF *42.2* a G-rich sequence is identified as a potential site for ribosome binding; [d] the gene *2* beginning and the resulting length of gp2 is based exclusively on the position of the initiation codon assigned during annotation as no RBS was identified (see Section 3.2 for details) [e] gp2 is not a membrane protein according to presently available biochemical data; [f] Structures of the gp2 nuclease domain (residues 232 to 422 of the gp2 amino acid sequence); [g] cryo-electron microscopy structure EMD access code; [h] The RBS sequence is compatible with different spacings relative to the ORF initiation codon. [i] NMR structure PDB access code; [j] the gene *16* beginning and the resulting length of gp16 is based on the position of the initiation codon assigned during annotation, as no RBS was identified, and on the amino terminus sequencing of gp16 (see Section 3.2 for details); [k] genes *17.1* and *17.1** have the same 5′ sequence because the product of *17.1** results from a +1 frameshift at the end of their common reading frame [91] (see Section 3.2 for details); [l] genes *17.5* and *17.5** have the same 5′ sequence as the product of *17.5** results from a putative −1 frameshift within their common reading frame [91] (see Section 3.2 for details); [m] the essential nature of SPP1 genes coding for the tail chaperones, TMP, Dit and Tal proteins was not demonstrated experimentally but their functional homologs are essential in the phage systems presently characterized; [n] the SPP1 TMP gp18 features four predicted transmembrane segments but it is an anticipated component of the phage particle occupying the tail tube internal space [16] (see Section 3.8 for details); [o] the same RBS is used for translation at the initiation codons of genes *26* and *26** [33] (see Section 3.6 for details); [p] the sequence inversion in SPP1*invmir* (Figure 1) disrupts ORF *31.2* and renders gene *32* promoterless indicating that they are non-essential; [q] poor RBS sequence which might not ensure putative ORF *42.2* translation.

AUG is the translation initiation codon of 72 ORFs (90%) of the SPP1 genome. It is found in all genes coding proteins of known function with the exception of the tail spike gene *21* that starts with UUG (Table 3). ORFs *27* and *32.5* also initiate with UUG while five other ORFs start with GUG (6%). These percentages differ somehow from the ones found for *B. subtilis* ORFs but AUG remains the initiation codon most frequently used (78%) by the host bacterium [110]. A significantly higher frequency of UGA termination codons (39%) is found in the SPP1 genome when compared to *B. subtilis* (24%). This is accounted by a reduction of UAA codon usage from 62% in *B. subtilis* to 50% in SPP1 (Table 4). UAG remains the rarest codon in the two genomes. The overrepresentation of UGA in SPP1 is surprising because there is an abnormally high 6% read-through of this codon in *B. subtilis* that extends the length of polypeptide chains [111]. However, the frequency of UGA is reduced to 29% when considering the SPP1 genes of known function that play key functions in DNA replication, viral particle assembly, and cell lysis (the "core" genome). In case of gene *6*, which encodes the essential capsid portal protein, correct polypeptide chain termination is ensured by an arrangement of stop codons in tandem (UGAUAA). Furthermore, all genes coding proteins required in high amounts for phage particle assembly feature a UAA stop codon (hundreds of copies of the (pro)capsid proteins gp11, gp12, gp13 as well as of the major tail tube protein (TTP (note that TTPs are also designated major tail proteins (MTPs)) gp17.1 are used for assembly of one viral particle (see Section 3.8)). The global codon usage frequency shows also some differences between SPP1 and *B. subtilis* (frequency differences above 10% are highlighted in bold and rare codons (<10% frequency) are underlined in Table 4).

There are two documented cases of programmed translational frameshifts in SPP1. These recoding events result from slippage of ribosomes into a different coding frame during translation of mRNA. The frequency of the frameshift dictates a constant ratio of the two proteins synthesized which have an identical amino terminus but a different carboxyl region sequence. The two TTPs gp17.1 and gp17.1* share an identical sequence, but the gp17.1* carboxyl terminus is extended by 87 additional amino acids [91]. The +1 translational frameshift results of ribosomes pausing at CCC rare proline codons and their shift to the overlapping frequent CCU proline codon in the CCCUAA sequence which codes also for the gene *17.1* termination codon [91]. The amount of gp17.1/gp17.1* synthesized is compatible with their 3:1 ratio found in SPP1 tails. The other programmed frameshift was identified by bioinformatics in genes *17.5* and *17.5** which encode functional analogs of phage lambda gpG/gpGT [91,112]. These are chaperones of tail tube assembly [113,114]. In SPP1, the putative −1 frameshift occurs at a UUUUUUC heptanucleotide slippage sequence within gene *17.5* that leads some ribosomes to change coding frame yielding gp17.5* [91]. Gp17.5* has the first 112 amino acid sequence identical to gp17.5.

Genes with two initiation codons preceded by canonical RBSs for the same coding frame and spaced by a few codons were identified in the two holin genes (genes *24.1*/*24.1** and *26*/*26**) [33]; see Section 3.6).

Table 4. Codon usage bias in SPP1 and *B. subtilis*. The total number of codons (No) used in the complete set of ORFs of SPP1 (this work) or *B. subtilis* [47,48] and the fraction of each codon used to code a specific amino acid are listed. When the fraction value differs by more than 0.1 this variation of codon usage frequency is highlighted in bold. Rare codons (fraction of usage below 0.1) are underlined.

Amino Acid	Codon		Fraction	No	Amino Acid	Codon		Fraction	No	Amino Acid	Codon		Fraction	No
Ala	GCG	SPP1	0.26	268	Gly	GGA	SPP1	0.31	318	Pro	CCU	SPP1	0.22	107
		B.s	0.26	24,574			*B.s*	0.31	26,381			*B.s*	0.29	12,824
Ala	GCA	SPP1	0.27	272	Gly	GGU	SPP1	0.24	245	Pro	CCC	SPP1	0.08	42
		B.s	0.28	26,416			*B.s*	0.18	15,457			*B.s*	0.09	4001
Ala	GCU	SPP1	0.27	279	Gly	GGC	SPP1	0.25	262	Ser	AGU	SPP1	0.16	123
		B.s	0.25	23,062			*B.s*	0.34	28,493			*B.s*	0.11	8096
Ala	GCC	SPP1	0.20	201	**His**	**CAU**	**SPP1**	**0.53**	131	Ser	AGC	SPP1	0.22	170
		B.s	0.21	19,342			***B.s***	**0.67**	18,610			*B.s*	0.23	17,226
Arg	AGG	SPP1	0.18	130	**His**	**CAC**	**SPP1**	**0.47**	119	Ser	UCG	SPP1	0.10	79
		B.s	0.10	4788			***B.s***	**0.33**	9019			*B.s*	0.10	7717
Arg	AGA	SPP1	0.28	198	**Ile**	**AUA**	**SPP1**	**0.27**	261	Ser	UCA	SPP1	0.25	193
		B.s	0.26	13,077			***B.s***	**0.13**	11,517			*B.s*	0.24	18,053
Arg	CGG	SPP1	0.12	83	**Ile**	**AUU**	**SPP1**	**0.35**	335	Ser	UCU	SPP1	0.17	132
		B. s	0.15	7329			***B.s***	**0.50**	45,181			*B.s*	0.20	15,615
Arg	CGA	SPP1	0.10	74	Ile	AUC	SPP1	0.38	365	Ser	UCC	SPP1	0.11	83
		B. s	0.11	5115			*B.s*	0.37	32,872			*B.s*	0.13	9757
Arg	CGU	SPP1	0.18	125	Leu (s)	UUG	SPP1	0.24	246	Thr	ACG	SPP1	0.26	211
		B. s	0.18	8755			*B.s*	0.16	18,745			*B.s*	0.27	17,693
Arg	CGC	SPP1	0.14	99	Leu	UUA	SPP1	0.24	251	Thr	ACA	SPP1	0.47	379
		B. s	0.20	9444			*B.s*	0.20	23,338			*B.s*	0.41	27,117
Asn	**AAU**	**SPP1**	**0.45**	344	**Leu**	**CUG**	**SPP1**	**0.10**	107	Thr	ACU	SPP1	0.15	120
		B.s	**0.57**	27,137			***B.s***	**0.24**	28,295			*B.s*	0.16	10,620
Asn	**AAC**	**SPP1**	**0.55**	425	Leu	CUA	SPP1	0.13	139	Thr	ACC	SPP1	0.12	94
		B.s	**0.43**	20,861			*B.s*	0.05	6030			*B.s*	0.16	10,497
Asp	**GAU**	**SPP1**	**0.52**	442	Leu	CUU	SPP1	0.20	203	Trp	UGG	SPP1	1.00	190
		B.s	**0.64**	40,291			*B.s*	0.24	28,226			*B.s*	1.00	12,571
Asp	**GAC**	**SPP1**	**0.48**	415	Leu	CUC	SPP1	0.08	87	**Tyr**	**UAU**	**SPP1**	**0.50**	276
		B.s	**0.36**	22,699			*B.s*	0.11	13,232			***B.s***	**0.65**	27,650
Cys	UGU	SPP1	0.54	50	Lys	AAG	SPP1	0.38	460	**Tyr**	**UAC**	**SPP1**	**0.50**	278
		B.s	0.45	4429			*B.s*	0.30	25,647			***B.s***	**0.35**	14,673
Cys	UGC	SPP1	0.46	42	Lys	AAA	SPP1	0.62	760	Val (s)	GUG	SPP1	0.23	211
		B.s	0.55	5322			*B.s*	0.70	60,072			*B.s*	0.26	21,585

Table 4. *Cont.*

Amino Acid	Codon		Fraction	No	Amino Acid	Codon		Fraction	No	Amino Acid	Codon		Fraction	No
Gln	CAG	SPP1	0.38	192	Met (s)	AUG	SPP1	1.00	412	Val	GUA	SPP1	0.26	240
		B.s	0.46	22,750			*B.s*	1.00	32,918			*B.s*	0.20	16,296
Gln	CAA	SPP1	0.62	319	**Phe**	**UUU**	**SPP1**	**0.48**	270	Val	GUU	SPP1	0.32	292
		B.s	0.54	23,889			***B.s***	**0.68**	37,445			*B.s*	0.28	23,440
Glu	GAG	SPP1	0.39	452	**Phe**	**UUC**	**SPP1**	**0.52**	294	Val	GUC	SPP1	0.18	168
		B.s	0.32	28,211			***B.s***	**0.32**	17,253			*B.s*	0.26	21,143
Glu	GAA	SPP1	0.61	704	Pro	CCG	SPP1	0.44	219	**End**	**UGA**	**SPP1**	**0.39**	31
		B.s	0.68	59,808			*B.s*	0.43	19,421			***B.s***	**0.24**	965
Gly	GGG	SPP1	0.21	215	Pro	CCA	SPP1	0.26	127	End	UAG	SPP1	0.11	9
		B.s	0.16	13,670			*B.s*	0.19	8541			*B.s*	0.14	591
										End	**UAA**	**SPP1**	**0.50**	40
												B.s	**0.62**	2542

3.3. Organization and Function of the SPP1 Genes

The 80 ORFs annotated here have a compact arrangement leaving only 6% non-coding sequences in the SPP1 genome. The longest intergenic regions carry transcriptional regulation sequences and, one of them, *ori*R (Figure 1).

The core SPP1 genome is presently composed of 28 genes: (**i**) 13 essential genes for which conditional lethal mutations (suppressor sensitive mutations (*sus*) or temperature sensitive mutations (*ts*)) were obtained by chemical mutagenesis of the overall SPP1 genome [7] (genes *1*, *2*, *6*, *11*, *13*, *15*, *16*, *17*, *17.1*, *35*, *38*, *39*, *40*; see Sections 3.5 and 3.8); (**ii**) 11 genes coding proteins that are anticipated to act in phage tail assembly (*17.5*, *17.5**, *18*, *19.1*, *21*, *22*; see Section 3.8) or in host cell lysis (*24.1*, *24.1**, *25*, *26* and *26**; see Section 3.6); and (**iii**) 4 non-essential genes whose inactivation is detrimental for SPP1 multiplication (*7*, *34.1*, *36*, and *44*; see Sections 3.5 and 3.8). Genes *12* and *17.1** are excluded from this group because their inactivation has no detectable effect in SPP1 fitness in laboratory conditions (see Section 3.8). The core genes identified include most of the minimal genetic set necessary for lytic phage multiplication ensuring viral DNA replication, viral particle assembly, and host lysis. A present omission in SPP1 research is dissection of the genetic circuitry that controls genome transcription and, in particular, late genes expression. Several putative DNA-binding proteins encoded by early genes are good candidates to be transcriptional regulators (Table 3; see Section 3.4). Bioinformatics allowed us to assign putative function and/or biochemical activity to genes *16.1*, *29*, *32*, *33*, *34*, *36.1*, *37.1*, and *37.3* (Figure 1; Table 3; see Sections 3.4 and 3.8). In total, 38 genes were functionally annotated (48% of SPP1 ORFs). They cluster in segments within four early transcriptional units, coding DNA replication and early lysis proteins together with uncharacterized polypeptides, and in one large genome region transcribed late. In this late region the synteny of genes encoding structural proteins of the SPP1 virion and late lysis genes is conserved when compared to the genomes of other *Siphoviridae* (Figure 1).

The function of other SPP1 ORF products is poorly understood. Genome deletions or inversions showed that ORFs *3–5*, *7*, *12*, *14*, *31.2*, *32*, *33* and *42–53* are non-essential (Figure 1; [26,35,53,55,83,88]; this work; P.T., unpublished) which corresponds to 19% of the SPP1 genome. Among these, a function was assigned experimentally only to the products of genes *7*, *12* (see Section 3.8) and *44* (see Section 3.5) while bioinformatics revealed a putative role for gp32 and gp33 (see Section 3.4). Within the SPP1 genome regions not tested by deletion analysis, the (putative) function of the products from ORFs *8*, *9*, *10*, *23*, *23.1*, *24*, *26.1*, *27*, *28*, *29.1*, *30*, *30.1*, *31*, *31.1*, *32.5*, *33.1*, *34.2*, *34.3*, *34.4*, *37*, *37.2*, and *41* remains unknown. Most of these 22 ORFs are probably non-essential for SPP1 multiplication. In total, 42 ORFs (52% of the total SPP1 ORFs) have no assigned function. Protein sequence (BLASTp) and structural (HHPred) homology searches revealed 18 orphan ORFs among those (*8*, *10*, *14*, *23*, *23.1*, *24*, *28*, *30.1*, *32.5*, *33.1*, *34.3*, *34.4*, *37*, *37.2*, *42.1*, *46.1*, *51.1*, and *52*) whose products are unrelated to protein sequences in databases. The other SPP1 ORF protein products are homologous in their vast majority to proteins with no known function from *Bacillus* spp. or from their phages, suggesting that they belong to a mobile gene pool of bacilli and their viruses. Most of the less characterized SPP1 ORFs are likely additions to the lytic phage core genome resulting from insertions of DNA that add "more on to it" ("morons" [115–117]). Morons designated originally complete transcriptional units added to phage genomes [115] but include presently also gene insertions that are unique or found in a limited set of genomes. They frequently encode beneficial features for phage adaptation to its host and environment. We consider SPP1 morons the ORFs *3–5*, *8–10*, *12*, *14*, *23–24*, *26.1–28*, *29.1–31.2*, *32.5*, *33.1*, *34.2–34.4*, *37*, *37.2*, *41–43*, and *46–53*.

The SPP1 ORFs length is highly variable, encoding proteins with an average length of 179 $\pm$ 197 amino acids and an average molecular mass of 20.3 $\pm$ 21.9 kDa. The tail tape measure protein and in some cases the tail tip protein, both with more than 1000 residues in SPP1, is (are) the longest protein(s) encoded by phages with long tails. They are landmarks to map the phage tail genes which, when combined with the synteny of genes coding virion assembly proteins, provide a first approximation to the genomic organization of late genes. The SPP1 core genes (most coding proteins >100 amino acids-long) tend to be longer than moron genes (most coding proteins <100 amino acids-long).

A BLASTn of the complete SPP1 genome showed extensive nucleotide sequence homology only to phages of the SPP1 group rho15, SF6, 41c [9,118], and the recently identified Lurz phage series [35]; P.T. unpublished) defining the genetic basis for a SPP1-like *genus*. Strong DNA sequence homology hits to other phages and to *Bacillus* spp. genomes (probably to prophages, defective phages or other elements of the bacilli mobile genome) were limited to a few individual genes (*5*, *9*, *10*, *17.1**, *25*, *33*, *48*, *49*, *50.1*, and *51*). An interesting case is *B. subtilis* (natto) phage PM1 [119] that has four blocks of DNA homology to four SPP1 ORFs/genes (*9*, *25*, *33* and *51*) which are separated by unrelated sequences. Multiple horizontal gene transfer events thus occur within the genetic mobile landscape of *Bacillus* spp. and its phages.

DNA nucleotide homology provides a sensitive criterion to detect recent genetic exchanges, while protein amino acid homology, found for a much larger set of SPP1 gene products, assesses far-reaching functional relationships with proteins from other phages. This is particularly well illustrated by bacteriophage GBK2 that infects the termophilic bacterium *Geobacillus kaustophilus*. Its genome shows no nucleotide sequence homology to SPP1 which infects a mesophilic bacillus. However, a large region encodes proteins homologous to SPP1 gp26 (holin), gp29, gp31, gp31.1, gp32, known DNA replication proteins (gp34.1, gp35, gp36, gp39, gp40), gp42, gp42.2 and gp43 [120] (see Sections 3.4 and 3.5). Their genes order is conserved in the two phage genomes, being spaced by genes of unrelated proteins. In contrast, phage particle assembly proteins of the two phages lack significant homology, apart from SPP1 proteins gp16.1 and gp17 which have GBK2 homologs, while the ensemble of their coding genes conserves the synteny found in siphoviruses [120] (this study). Thus, the genome of both phages is assembled from a common ancient genome module of early genes, traceable by protein homology, and of an evolutionarily distinct module coding the viral particle assembly proteins. The first module diversified by acquisition and eventual loss of different morons (single genes and also transcriptional units like the one controlled by *P*E4 in SPP1; Figure 1) while the viral particle assembly module acquired the genes *16.1–17* cluster, most probably from horizontal exchange into a conserved position of the module. The genomes of SPP1 and GBK2 are a compelling case of tailed phages evolution by combination of modules, horizontal gene transfer, and diversification by morons acquisition/loss [115–117,121,122].

In the following Sections we describe current knowledge on SPP1 genes and ORFs with particular attention to those that are less characterized. In-depth reviews on SPP1 biology [123], DNA replication [30], and viral particle assembly [124–126] are available.

3.4. The SPP1 Genes Set. I. Uncharacterized Early Genes

Putative moron genes are spread throughout the five early SPP1 transcriptional units (Figure 1).

The dispensable set of short ORFs *46* to *53*, under the control of *P*E1, has no known or putative function deduced from bioinformatics. ORFs *46.1*, *53*, and the sequential ORFs *50.1*, *51*, *51.1* have predicted transmembrane segments (Table 3) indicating that these polypeptides insert most likely in the cytoplasmic membrane to achieve early roles in the host cell. Segments of DNA homologous to different combinations of ORFs *48* through *51* are found in different SPP1 phages or *Bacillus* spp. strains, suggesting common functions.

Protein sequence analysis and predicted structural homology led to functional assignments to several genes in transcriptional units controlled by promoters *P*E4 and *P*E5. *P*E4 drives expression of ORFs *32.5* and *33.1*, whose function is unknown, together with the dispensable gene *33*. Gp33 is a 589-long protein that shares 88% amino acid sequence identity with a protein of *B. subtilis* (natto) phage PM1. It is also highly homologous to a large number of proteins from *Bacillus* spp. annotated to have cell wall hydrolysis activity or as proteins of *Bacillus* phages with a predicted right-handed parallel β-helix repeat fold. HHPred extends this structural homology to adhesins binding to the cell wall surface as well as to phage tail spikes of *B. subtilis* phage phi29 and of enterobacteria phages like sf6, HK620, and LKA1. Gp33 could be the trace of a tail fiber used by an ancestor of SPP1, like the Ur-lambda tail fibers lost during laboratory evolution [127]. However, this does not explain why SPP1

maintained gene *33* expressed early during infection in an apparently functional form. We privilege the hypothesis that gp33 accumulates in the cytoplasm and its release upon lysis acts on the wall of the infected bacterium to facilitate lysis and/or on the envelope of other *B. subtilis* cells to facilitate their subsequent infection.

*P*E5 controls expression of 14 ORFs including genes *25*, *26* and *26** involved in cell lysis (see Section 3.6). Most of the other ORFs encode short polypeptides of unknown function. Gp29 and its homologous hypothetical protein of *Geobacillus* phage GBK2 give strong hits in HHPred to DNA-binding proteins with a winged helix fold like transcriptional regulators and excionases. ORFs *30*, *31.1*, and *31.2* products have predicted transmembrane segments suggesting an early role in the cytoplasmic membrane during infection. Gp31.1 and gp31.2, which are encoded by contiguous ORFs, are homologous to numerous proteins of *Bacillus* spp. and their phages, likely sharing a widespread function. The non-essential gp32 (Table 3) is a basic 836-long protein with 75% amino acid sequence identity to a protein of unknown function from *Geobacillus* phage GBK2. Gp32 has also extensive similarity to ATP-binding proteins from *Bacillus* spp. and its phages. HHPred reveals structural homology of its carboxyl terminus, with high confidence scores, to the conjugation protein TrwB, to VirB4 of type IV secretion systems and to proteins of the FtsK family. These machines translocate DNA or proteins across the bacterial membrane. Since gp32 has no predicted transmembrane segments, it could act together with gp31.1 and/or gp31.2 to build a trans-membrane translocon of macromolecules. The synteny of genes coding for proteins homologous to gp31, gp31.1 and gp32 is found also in phage GBK2 [120] suggesting a conserved activity. The exact protein composition, functionality and role of such potential machine remain to be established.

The strongest early promoters, *P*E3 and *P*E2 [41], drive expression of DNA replication genes (see 3.5) that alternate with genes coding uncharacterized proteins (Figure 1). Bioinformatics allowed us to assign putative activities to four proteins of the latter group. Gp34 is a putative 6.3 kDa basic polypeptide with predicted structural homology to DNA binding proteins, whose strong hits are transcriptional repressors. Its small size advises, nevertheless, some caution on this functional assignment. Gp36.1 is highly homologous to HNH endonucleases of Gram-positive bacteria, and HHPred uncovered a relationship to the structure of *B. subtilis* phage SPO1 HNH homing endonuclease I-HmuI (PDB accession number 1U3E) [128]. Proteins containing the HNH motif carry out intron homing in phages T4, SPO1 and SP82 [129,130], and are essential for DNA packaging in a group of *cos*-phages [131,132]. In other phages they play dispensable functions like in *Staphylococcus aureus* phage 80α [133]. Gp37.3 is a small basic polypeptide whose amino acid sequence and structure prediction show relatedness to numerous bacterial and phage DNA binding proteins. SPP1 gp34, gp36.1, gp37.3 together with gp29, and possibly other early polypeptides that did not deliver robust bioinformatics hits yet, are strong candidates to participate in SPP1 DNA metabolism and in gene expression regulation processes calling for further research. Gene *37.1* that is found within this set of genes codes for an enzyme with a distinct function. Gp37.1 is highly homologous to poly-γ-glutamate (γ-PGA) hydrolases from phages of Gram-positive bacteria, mainly infecting *Bacillus* spp. like vB BsuM-Goe3, BSP10, Grass, BSNPO1, PM1, phiNIT1, PBS1, AR9 and others. HHPred provided a strong hit to the structure of the γ-PGA hydrolase PghP from phage phiNIT1 (PDB accession number 3A9L). The supposed role of the enzyme in phages is to degrade the γ-PGA polymer layer. This layer forms a shield at the surface of several microorganisms, mainly *Bacillus* spp., to protect them from environmental attacks as diverse as phagocytosis or phage infection [134,135]. Release of γ-PGA hydrolase from phage-infected bacteria would thus open the way for infection of new host bacteria protected by a γ-PGA layer. Experimental demonstration of γ-PGA hydrolase activity from *B. subtilis* phages phiNIT1 (protein PghP; [136]), BSP10 [137], as well as from prophages SPβ and prophage-like element 5 (YokZ and YndL, respectively) [135] support this hypothesis.

The large number of early ORFs not conserved among SPP1 and other tailed phages, whose majority has an unknown function, is a rich and diverse patrimony. We anticipate that this genetic pool is mostly dedicated to strategies engaged at the beginning of infection to take over the host cell [138]

(and references therein), possibly to mediate superinfection exclusion of the infected bacterium by other phages [139,140] (and references therein), and/or to accumulate molecules that will be released upon lysis to support subsequent infection of new host bacteria [134–137]. Most of the phage effectors involved are dispensable (e.g., the proteins encoded by transcriptional units under the control of *P*E1 and *P*E4 in SPP1) but their combination might provide a determinant fitness advantage, globally or in specific environmental settings, to SPP1-like phages.

3.5. The SPP1 Genes Set. II. DNA Replication Early Genes

The bacteriophage SPP1 DNA replication is a well-characterized process. SPP1 was reported to have two origins of replication, *ori*R and *ori*L (Figure 1), localized ~13 kbp apart in the circularized genome. The two sequences have a similar organization composed of direct repeats and an AT-rich region that acts as a DNA unwinding element (DUE) [102]. The SPP1 origin binding protein gp38 binds to the direct repeats of *ori*L, positioned within gene *38* (Figure 1), and of *ori*R with nanomolar affinity [102]. Replication initiation requires gp38, the helicase loader gp39 [103,104] (PDB accession number 1NO1), and the replicative helicase gp40 that belongs to the DnaB family [103,105,106] (PDB accession number 3BGW). After melting of the origin sequence and unwinding, the gp40 hub interacts with the host DnaG [107,108] and with the DnaX subunit of the clamp loader [109]. These interactions recruit the host replisome. The theta replication reaction was reconstituted in vitro with a supercoiled plasmid bearing *ori*L and purified SPP1 gp38, gp39 and gp40 together with *B. subtilis* DNA polymerases PolC and DnaE, the processivity clamp DnaN, the hetero-pentameric clamp loader complex (3xDnaX+YqeN+HolB), the primase DnaG, DNA gyrase, and a single-stranded DNA binding protein (SSB) (bacterial SsbA, or SPP1 gp36) [141]. Cellular SsbA can replace the non-essential SPP1-encoded SSB gp36 in SPP1 DNA replication. In contrast, gp36 does not support *B. subtilis* DNA replication and acts in vitro as an inhibitor of the cellular process [101].

After a few rounds of circular molecules DNA synthesis, SPP1 DNA replication switches to generate linear head-to-tail concatemers [28,29] which are the substrate for DNA encapsidation into viral particles. The switch was proposed to be triggered by stalling of the replication fork, probably when it collides with gp38 firmly bound to *ori*R [30,142]. Four SPP1 proteins are likely involved in the process of resuming DNA replication after stalling: the Holliday junction resolvase gp44 [31], the 5′→3′ exonuclease gp34.1 [99], the ATP-independent single-strand annealing recombinase gp35 [100], and the SSB gp36 [101]. The current model is that they would act coordinately to process the stalled replication fork and make a double-strand break, followed by homology-directed recombination generating the template for DNA replication by a rolling circle mode (sigma-type) [30,31,142]. A DNA substrate assembled in vitro which mimics the sigma template was used to show that subsequent rolling circle replication is achieved by the same set of proteins that carry out theta-type DNA replication, with the exception of DNA gyrase that is not necessary [101]. The assays also highlighted that gp38 may act like the bacterial PriA protein, because no specific DNA region (*ori*) is needed for this reaction and DNA replication can start at any site after fork pausing [101]. Gp35 recombinase is the only SPP1 essential protein among the switch putative effectors. The lack of gp34.1 and gp36 is detrimental while deletion of the gp44-encoding gene has only a marginal effect in SPP1 viability [13]. Functionally related cellular proteins most likely (partially) compensate for the roles of those viral effectors. Homologs of the SPP1 DNA replication proteins are found in many phages, suggesting that such recombination-driven strategy is a common mechanism to produce concatemers.

The SPP1 replication proteins are encoded by two transcriptional units under the control of promoters *P*E2 and *P*E3 (Figure 1), as in phage GBK2 that shares a similar operon organization and gene synteny [120]. Genes *38*, *39*, and *40* whose products initiate DNA replication form a cluster, but the other DNA replication genes are spaced by ORFs coding proteins with putative DNA binding, host takeover, or unknown functions (see Section 3.4). The selective pressure to cluster genes encoding proteins that closely interact [143] appears stronger for the viral particle assembly genes module (see Section 3.8) than for the DNA replication genes module.

3.6. The SPP1 Genes Set. III. Lysis Early and Late Genes

Tailed phage lysis cassettes code typically for a holin that inserts in the cytoplasmic membrane to create holes and for an endolysin that hydrolyses the cell wall leading to efficient bacterial disruption [33,98,144]. Interestingly, The SPP1 genome codes two holin proteins that localize in the cytoplasmic membrane (gp24.1 and gp26) and an endolysin (gp25 or LysSPP1). They act together to lyse *B. subtilis* at the end of the infectious cycle allowing viral particles to escape from the host [33]. Strong nucleotide homology was only detected between gene *25* and the endolysin gene of phage PM1 [119] but proteins homologous to gp24.1, gp25 and gp26 are encoded by numerous phages of Gram-positive bacteria.

A lysis system with two holins (XhlA and XhlB) combined with an endolysin (XlyA) was also described for the defective *B. subtilis* phage PBSX [145]. XhlA and XhlB are homologous to SPP1 gp24.1 and gp26, respectively. In SPP1 their coding genes flank the endolysin gene *25*, while the two holin genes precede the endolysin gene in the PBSX genome. Gp24.1 and XhlA feature a predicted transmembrane helix in their carboxyl terminus while gp26 and XhlB have the canonical organization of phage holins with two transmembrane segments. Each individual holin of these phages does not appear detrimental to *B. subtilis*, while their co-production leads to immediate cell death followed by lysis, showing that the two holins cooperate for disruption of the bacterial membrane [33,145].

Both SPP1 genes *24.1* and *26* have one in-frame internal initiation codon AUG preceded by a correctly spaced RBS (Table 3). Translation started at these AUG codes gp24.1* and gp26* which lack the first 9 and 2 amino terminus amino acids when compared to gp24.1 and gp26, respectively. Such dual start motifs leading to synthesis of two highly related proteins might be involved in lysis regulation as found for the phage lambda S protein whose longer polypeptide has an antiholin function that counteracts the shorter polypeptide holin activity [146]. In contrast to most phages, SPP1 genes *25* and *26* are transcribed early during infection from promoter *P*E5 localized within the gene *24.1* coding sequence [33]. It is likely that the production of gp24.1, which results from late transcription, defines the tempo for the three SPP1 lysis proteins to concur for disruption of the infected bacterium.

3.7. The SPP1 Genes Set. IV. Uncharacterized Late Genes

The SPP1 late genome region has an order of genes coding proteins involved in assembly of the viral particle that is conserved among numerous tailed phages. This arrangement is interrupted by a genome segment, bracketed between transcriptional signals, which codes for the origin of replication *ori*R and three moron genes (Figure 1). Gp8 and gp10 are unrelated to known proteins, while gp9 is highly homologous to a protein from *B. subtilis* (natto) phage PM1 [119] and to numerous hypothetical proteins encoded by *Bacillus* spp. The function of these SPP1 proteins and if they are necessary for phage amplification remain to be established.

The genome segment between the tail fiber gene *21* and the lysis proteins encodes also four proteins whose precise role is unknown. Gp22 has homology to a group of *Bacillus* spp. putative proteins and its structure (PDB access code 2XC8) reveals a fold similar to a domain of the tail receptor binding protein from lactococcal phage p2 [96] (see Section 3.8). The structure of the gp23.1 hexamer was also determined (PDB access code 2XF7) but provided no conclusive insight on protein function [97]. Gp23, gp23.1 and gp24 gave no strong homology hits to proteins in the databank.

The role of the dispensable short ORFs *3*, *4*, *5* and *14* that separate essential genes for SPP1 capsid assembly is not known. The products of ORFs *3* and *4* show homology to a hypothetical protein of *Bacillus* phages vB_BsuM-Goe3 and PM1, respectively, while gp5 has similarity to a large number of bacterial and phage uncharacterized proteins. ORF *14* is an orphan.

3.8. The SPP1 Genes Set. V. Viral Particle Assembly Late Genes

SPP1 devotes ~40% of its genome information to assembly of the SPP1 viral particle, a process that was extensively studied. DNA-filled capsids (nucleocapsids) and tails are built in two independent

assembly pathways. The two structures then join in a final reaction to yield the infectious particle (virion), like in all studied phages with a long tail.

Construction of the SPP1 nucleocapsid follows the similar assembly pathway used by viruses of the tailed phages-herpesviruses lineage [147,148] (and references therein). A spherically shaped icosahedral DNA-free procapsid with a diameter of ~55 nm is formed first. It is built by polymerization of the major capsid protein (MCP) gp13 that is chaperoned by the internal scaffolding protein gp11 [85–87]. One of the 12 vertexes of the procapsid icosahedron is a specialized structure defined by presence of a cyclical dodecamer of the portal protein gp6 ([20,60,75–81]; PDB access code 2JES). Assembly of the procapsid initiates most likely at the portal vertex by co-interaction between gp6, gp11 and gp13. In absence of gp6, gp11 and gp13 assemble procapsids of normal and of a smaller size, revealing that the portal ensures correct size determination of the procapsid [86]. Co-production of gp6, gp11 and gp13 is sufficient and necessary for assembly of procapsids functional for DNA packaging, showing that they are the minimal set of essential components to build these structures [86]. Gp6 makes a strong interaction with gp7 that targets this protein to the procapsid interior in a few number of copies [84]. This strategy was likely exploited in numerous lysogenic phages for targeting to their capsid interior a toxin fused to the carboxyl terminus of gp7-like proteins (also designated Mufs) [149]. The toxin was proposed to be subsequently delivered to the host at the beginning of infection [149,150]. SPP1 gp7 is a dispensable component of the virion. Phage assembly is not affected in its absence, but only ~25% of virions lacking gp7 are infectious, showing that it supports initiation of SPP1 infection [83].

Head-to-tail concatemers of the SPP1 genome synthesized during phage DNA replication (see Section 3.5) are the substrate for encapsidation into procapsids. The *pac* sequence in the SPP1 genome [151] (Figure 1) is specifically recognized and cleaved by the SPP1 terminase complex to initiate DNA packaging [70,152]. The terminase is composed of the small subunit (TerS) gp1 that binds to *pac* [57,63–66] (PDB access code 3ZQQ of gp1 from the SPP1-related phage SF6) and of the large subunit (TerL) gp2, a two-domain protein with ATPase and nuclease activities [57,67,70–72] (PDB access codes 2WBN and 2WC9 of the gp2 nuclease domain). Re-analysis of gene *1* sequence showed that gp1 is shorter (Table 1) and more similar to the TerS from other SPP1-related phages [118] than initially reported. The *pac* sequence overlaps the *P*L1/*P*L2 promoters (see Section 3.2), the RBS, and the sequence of gene *1* encoding the gp1 DNA-binding domain (Figure 1) [63]. Gp1 binds to the *pac*L and *pac*R regions of *pac* [63,68] and recruits gp2 for making a double-strand cleavage within *pac*C, a sequence that is flanked by *pac*L and *pac*R [70]. Gp2 has a non-specific nuclease activity [57,67,70–72] that is controlled by its appropriate positioning in the gp1-gp2-DNA complex [35]. Cleavage at *pac* is auto-regulated [69] leaving most *pac* sequences uncut in the concatemer used for packaging. This allows encapsidation of DNA molecules longer than the unit-genome length during processive headful packaging (see Section 1) [17,19]. The terminal redundancy of SPP1 encapsidated DNA molecules is essential for their re-circularization at the beginning of infection. The auto-regulated cleavage of *pac* was reproduced in a plasmid minimal system bearing genes *1*, *2* and *pac*, showing that gp1 and gp2 are necessary and sufficient to carry out this reaction in *B. subtilis* [72].

The gp1-gp2-SPP1 DNA complex docks at the procapsid portal vertex to assemble the DNA packaging motor. DNA translocation powered by the ATPase activity of gp2 involves an intricate cross-talk between gp1, gp2, and gp6 [67,73,74,82]. The scaffolding protein gp11 leaves the capsid interior and the capsid lattice undergoes a major conformational change leading to its expansion, maximizing the space for phage DNA packing [85,86]. This process uncovers binding sites at the capsid surface for attachment of gp12 trimers that have a central collagen-like fold [88]. The DNA packaging reaction was characterized in vitro, either using extracts of infected cells [153] or purified linear DNA, gp1, gp2 and procapsids [154]. When a threshold amount of DNA is packaged inside the capsid, the portal protein senses the level of DNA headfilling and triggers the sequence-independent headful cleavage of the DNA concatemer that is most likely achieved by the gp2 nuclease domain [20,77,125,155]. Disassembly of the DNA packaging motor is coordinated with sequential binding of the head completion protein gp15 (PDB access code 2KBZ) to gp6, extending the portal channel, and of gp16 (PDB access code

2KCA) that closes the channel, preventing release of the packaged DNA [60,78,89]. The gp6-gp15-gp16 complex is named connector.

The SPP1 long non-contractile tail has a building plan similar to the long tails of *Siphoviridae* and *Myoviridae*. These structures are structurally related to a variety of cellular, tube-like, delivery devices used in bacterial warfare like phage tail-like bacteriocins (PTLBs), phage tail-like complexes that confer toxicity against eukaryotic cells, or type VI secretion systems [156,157].

The SPP1 tail features a host adsorption apparatus that binds selectively to the *B. subtilis* surface and promotes DNA transfer across the bacterial envelope. This ~31 nm-long structure binds selectively to glycosylated teichoic acids [21], which facilitates its subsequent strong interaction with the membrane protein YueB [23,158], committing the phage particle to infection. The known components of the adsorption apparatus are the tail fiber (or Tal) gp21 [93], which binds YueB [95], and the distal tail protein (Dit) gp19.1 (PDB access code 2X8K) [94]. The gp19.1 hexamer forms a complex with the trimeric gp21 amino terminus that closes the tail tube in the virion and opens at the beginning of infection [93]. Gp19.1 defines one end of the ~160 nm-long tail tube built by a helical array of the TTPs gp17.1/gp17.1* that are found at a ratio of ~3:1 in the tube [91]. Similar tubes can be formed exclusively by gp17.1 in vivo [91] and in vitro [92]. The TTPs form hexameric rings organized around the tape measure protein (TMP) gp18 [16]. The tail tube of sipho and myoviruses is tapered by (a) tail completion protein(s), which provide(s) the interface for tail attachment to the connector. In SPP1 it is the tail-to-head joining protein (THJP) gp17 (PDB access code 2LFP) that interacts with the capsid connector in the viral particle [32,90]. The connector-tail completion proteins structure is named head-to-tail interface [81].

SPP1 tail assembly is anticipated to follow the pathway of phages with long tails typified by T4 [159] and lambda [160]. In such case, the tail adsorption apparatus is formed first providing a platform for initiation of gp17.1/gp17.1* helical polymerization around the TMP gp18. During this reaction gp18 is probably pre-shielded by the chaperones gp17.5 and gp17.5*, as proposed for phage lambda [114]. When the tail tube reaches a defined length, determined by gp18, polymerization stops and tail completion proteins bind to the tube end. Gp17 that is exposed at this end joins the tail to the capsid, a reaction that was characterized in vitro [32].

The function of proteins engaged in assembly of the SPP1 nucleocapsid was defined. They have a large number of homologous proteins from phages or prophages of Gram-positive bacteria that play similar roles in viral particle assembly. The exception is the non-essential capsid accessory protein gp12, whose collagen-like motif is found only in some capsid associated proteins [88] and in tail fibers [161]. The order of genes coding proteins necessary for nucleocapsid assembly follows the conserved genome organization of siphoviruses [162–165] with the particularity that ORFs *3–5*, *8–10* and *14* interrupt the usual gene order. The DNA packaging/portal module (coding gp1, gp2, (-, -, -,), gp6, (gp7) spaced by ORFs *8–10*, is followed by the icosahedral capsid building module (coding gp11, (gp12), gp13) which is separated by ORF *14* from the head completion proteins module (coding gp15, gp16). Note that in this protein list "-" identifies a protein encoded by a moron gene while proteins that have a known function but are non-essential for SPP1 multiplication are shown within brackets.

Gene *16* is followed by *16.1* and a set of genes whose synteny and protein products are conserved in the tail module of siphoviruses [16,91,162–167]. The function of SPP1 gp16.1 and its homologous proteins that are widespread among phages with long tails is yet unknown. Their coding gene is consistently found between the capsid connector and the tail proteins coding genes [166]. The tail module codes for gp17, gp17.1, (gp17.1*), gp17.5, gp17.5*, gp18, gp19.1, and gp21 (Figure 1) [16]. They all have extensive homology to proteins from Gram-positive bacteria and their phages. Gp17.1* and gp17.5*, which result from programmed translational frameshifts (see Section 3.2), feature two domains defined by similarity to different proteins. The amino terminus domains boundaries are roughly delimited by their region of identity with gp17.1 and gp17.5, respectively. The gp17.1* amino terminus shows robust similarity to TTP annotated proteins, while its non-essential carboxyl terminus is homologous to FN3 motifs of proteins from *Bacillus* spp. that are involved in binding to cell

surfaces [91]. Their exposure in phage structures suggests a role in adhesion to bacteria [91,168]. The amino and carboxyl domains of gp17.5* have similarity to two distinct sets of proteins with unknown function from Gram-positive bacteria or from their phages. The full-length gp21 Tal has numerous homologs. In addition, a subset of proteins has similarity only to its amino terminus (~400 residues), which is the conserved region of Tal proteins that closes the tail tube [93]. The gp21 long carboxyl terminus interacts with the SPP1 bacterial receptor [95]. This organization is consistent with the modular organization of Tal proteins whose carboxyl termini differs both in length and sequence. Such variation results, namely, from whether they carry or not a domain for interaction with the bacterium and of the selective pressure on this domain to generate diverse strategies to target a specific host [169,170] (and references therein). It remains to be established if proteins gp22, gp23, gp23.1 and/or gp24 (see Section 3.7) which are encoded by genes downstream of gene *21* play a role in tail assembly that is specific for the SPP1 system. Note that the three-dimensional structure of the gp22 monomer shows structural similarity to the shoulder of the tail receptor binding protein of lactococcal phage p2 [96] supporting that gp22 is a SPP1 tail component.

4. Conclusions

This reannotation of the SPP1 genome provides an actualized view of our understanding of this phage genetic patrimony and how it supports its multiplication. Phage DNA replication and assembly of the viral particle are particularly well studied, rendering SPP1 one of the forefront systems to understand the molecular basis of these processes in tailed bacteriophages. The available knowledge and tools are excellent assets to pursue research on both themes. Recent work provided also insight on the SPP1 lysis mechanism. The genes involved in these essential steps of the virus cycle occupy ~60% of the SPP1 genome. Much less is known about the function of other genes in spite of 50 years of SPP1 research. One or several of them code yet unidentified proteins that regulate SPP1 gene expression circuitry, a less studied aspect of SPP1 molecular biology. Apart from such regulators, a majority of uncharacterized genes likely code effectors that participate in host cell hijacking to optimize viral multiplication, that provide immunity to super-infection by other phages, or that support phage dissemination in the natural environment. These functions are frequently non-essential for phage survival, having subtle effects that render their study difficult. Transcriptomics, proteomics, and metabolomics combined with systematic knock-outs and sensitive phenotyping assays in different infection settings appears as a promising approach to deliver a complete functional map of SPP1. Those studies will provide insights on the wild side of SPP1 that remains to be explored.

Current knowledge on this phage system and its genetic landscape, that is distinct from other model phages, clearly recommend SPP1 as the reference virus for a new SPP1-like virus genus of the *Siphoviridae* family. The genus includes SPP1-related phages rho15, SF6, 41c [9,118], and the recently identified Lurz phage series [35] (P.T. unpublished) whose DNA sequence and genome organization are similar to SPP1. Phages GBK2 and PM1 have modules of the genome that are evolutionarily linked to SPP1 but other modules of essential genes code proteins without detectable similarity to SPP1 gene products. This mosaic genome organization brings to debate if such type of relatedness [122] has enough taxonomic value to include GBK2 and PM1 in the same taxon as SPP1. Therefore, we limit at present the SPP1-like genus proposal to its close genetic neighborhood until more robust phylogenetic and biological evidence is obtained to expand the genus to other phages.

Author Contributions: Conceptualization, L.M.G., M.E.S.F., S.A. and P.T.; Investigation, L.M.G., M.E.S.F., C.M., L.J., A.L., K.D. and S.A.; Supervision, S.A., R.C.-L. and P.T.; Validation, L.M.G., M.E.S.F., I.A., K.D., L.O., R.C.-L., S.A. and P.T.; Vizualization, L.M.G., M.E.S.F. and P.T.; Writing—original draft P.T.; Writing—review and editing L.M.G., M.E.S.F., I.A., L.O., K.D., A.L. and S.A.

Funding: Research in our teams was funded by the CNRS, INRA, MINECO grant BFU2015-67065-P (to S.A.), ANR grants 09-BLAN-0149-0 (to P.T.), ANR-12-Blanc-BSV3-0021 (to P.T. and R.C.L.), ANR-15-CE11-0 010-01 (to P.T. and R.C.L.), and an "Equipe de la Fondation Médicale (FRM)" grant (to P.T.).

Acknowledgments: We thank past and present researchers working on bacteriophage SPP1 for their contributions to our current understanding of this exciting biological system.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Riva, S.; Polsinelli, M.; Falaschi, A. A new phage of *Bacillus subtilis* with infectious DNA having separable strands. *J. Mol. Biol.* **1968**, *35*, 347–356. [CrossRef]
2. Riva, S.; Polsinelli, M. Relationship between competence for transfection and for transformation. *J. Virol.* **1968**, *2*, 587–593. [PubMed]
3. Spatz, H.C.; Trautner, T.A. One way to do experiments on gene conversion? *Mol. Gen. Genet.* **1970**, *109*, 84–106. [CrossRef] [PubMed]
4. Spatz, H.C.; Trautner, T.A. The role of recombination in transfection of *B. subtilis*. *Mol. Gen. Genet.* **1971**, *113*, 174–190. [CrossRef] [PubMed]
5. Deichelbohrer, I.; Alonso, J.C.; Lüder, G.; Trautner, T.A. Plasmid transduction by *Bacillus subtilis* bacteriophage SPP1: Effects of DNA homology between plasmid and bacteriophage. *J. Bacteriol.* **1985**, *162*, 1238–1243.
6. Alonso, J.C.; Lüder, G.; Trautner, T.A. Intramolecular homologous recombination in *Bacillus subtilis* 168. *Mol. Gen. Genet.* **1992**, *236*, 60–64. [CrossRef]
7. Behrens, B.; Lüder, G.; Behncke, M.; Trautner, T.A.; Ganesan, A.T. The genome of *B. subtilis* phage SPP1: Physical arrangement of phage genes. *Mol. Gen. Genet.* **1979**, *175*, 351–357. [CrossRef]
8. Ratcliff, S.W.; Luh, J.; Ganesan, A.T.; Behrens, B.; Thompson, R.; Montenegro, M.A.; Morelli, G.; Trautner, T.A. The genome of *Bacillus subtilis* phage SPP1: The arrangement of restriction endonuclease generated fragments. *Mol. Gen. Genet.* **1979**, *168*, 165–172. [CrossRef]
9. Santos, M.A.; Almeida, J.; de Lencastre, H.; Morelli, G.; Kamke, M.; Trautner, T.A. Genomic organization of the related *Bacillus subtilis* bacteriophages SPP1, 41c, rho 15, and SF6. *J. Virol.* **1986**, *60*, 702–707.
10. Yasbin, R.E.; Frank, E. Young Transduction in *Bacillus subtilis* by bacteriophage SPP1. *J. Virol.* **1974**, *14*, 1343–1348.
11. Ferrari, E.; Canosi, U.; Galizzi, A.; Mazza, G. Studies on Transduction Process by SPP1 Phage. *J. Gen. Virol.* **1978**, *41*, 563–572. [CrossRef] [PubMed]
12. Canosi, U.; Lüder, G.; Trautner, T.A. SPP1-mediated plasmid transduction. *J. Virol.* **1982**, *44*, 431–436. [PubMed]
13. Valero-Rello, A.; López-Sanz, M.; Quevedo-Olmos, A.; Sorokin, A.; Ayora, S. Molecular mechanisms that contribute to horizontal transfer of plasmids by the bacteriophage SPP1. *Front. Microbiol.* **2017**, *8*, 1–13. [CrossRef] [PubMed]
14. Alonso, J.C.; Lüder, G.; Trautner, T.A. Requirements for the formation of plasmid-transducing particles of *Bacillus subtilis* bacteriophage SPP1. *EMBO J.* **1986**, *5*, 3723–3728. [CrossRef] [PubMed]
15. White, H.E.; Sherman, M.B.; Brasilès, S.; Jacquet, E.; Seavers, P.; Tavares, P.; Orlova, E.V. Capsid structure and its stability at the late stages of bacteriophage SPP1 assembly. *J. Virol.* **2012**, *86*, 6768–6777. [CrossRef] [PubMed]
16. Plisson, C.; White, H.E.; Auzat, I.; Zafarani, A.; São-José, C.; Lhuillier, S.; Tavares, P.; Orlova, E.V. Structure of bacteriophage SPP1 tail reveals trigger for DNA ejection. *EMBO J.* **2007**, *26*, 3720–3728. [CrossRef] [PubMed]
17. Tavares, P.; Lurz, R.; Stiege, A.; Rückert, B.; Trautner, T.A. Sequential headful packaging and fate of the cleaved DNA ends in bacteriophage SPP1. *J. Mol. Biol.* **1996**, *264*, 954–967. [CrossRef] [PubMed]
18. Humphreys, G.O.; Trautner, T.A. Maturation of bacteriophage SPP1 DNA: Limited precision in the sizing of mature bacteriophage genomes. *J. Virol.* **1981**, *37*, 832–835.
19. Morelli, G.; Fisseau, C.; Behrens, B.; Trautner, T.A.; Luh, J.; Ratcliff, S.W.; Allison, D.P.; Ganesan, A.T. The genome of *B. subtilis* phage SPP1: The topology of DNA molecules. *Mol. Gen. Genet.* **1979**, *168*, 153–164. [CrossRef]
20. Tavares, P.; Santos, M.A.; Lurz, R.; Morelli, G.; De Lencastre, H.; Trautner, T.A. Identification of a gene in *Bacillus subtilis* bacteriophage SPP1 determining the amount of packaged DNA. *J. Mol. Biol.* **1992**, *225*, 81–92. [CrossRef]

21. Baptista, C.; Santos, M.A.; São-José, C. Phage SPP1 reversible adsorption to *Bacillus subtilis* cell wall teichoic acids accelerates virus recognition of membrane receptor YueB. *J. Bacteriol.* **2008**, *190*, 4989–4996. [CrossRef] [PubMed]
22. Baptista, C.; Barreto, H.C.; São-José, C. High levels of DegU-P activate an Esat-6-Like secretion system in *Bacillus subtilis*. *PLoS ONE* **2013**, *8*, e67840. [CrossRef] [PubMed]
23. São-José, C.; Lhuillier, S.; Lurz, R.; Melki, R.; Lepault, J.; Santos, M.A.; Tavares, P. The ectodomain of the viral receptor YueB forms a fiber that triggers ejection of bacteriophage SPP1 DNA. *J. Biol. Chem.* **2006**, *281*, 11464–11470. [CrossRef] [PubMed]
24. Jakutyte, L.; Baptista, C.; São-José, C.; Daugelavičius, R.; Carballido-López, R.; Tavares, P. Bacteriophage infection in rod-shaped gram-positive bacteria: Evidence for a preferential polar route for phage SPP1 entry in *Bacillus subtilis*. *J. Bacteriol.* **2011**, *193*, 4893–4903. [CrossRef] [PubMed]
25. Jakutyte, L.; Lurz, R.; Baptista, C.; Carballido-López, R.; São-José, C.; Tavares, P.; Daugelavičius, R. First steps of bacteriophage SPP1 entry into *Bacillus subtilis*. *Virology* **2012**, *422*, 425–434. [CrossRef] [PubMed]
26. Alonso, J.C.; Lüder, G.; Stiege, A.C.; Chai, S.; Weise, F.; Trautner, T.A. The complete nucleotide sequence and functional organization of *Bacillus subtilis* bacteriophage SPP1. *Gene* **1997**, *204*, 201–212. [CrossRef]
27. Burger, K. Biochemische und genetische Untersuchungen zur DNA-Replikation des *B. subtilis*-Bakteriophagen SPP1. Ph.D. Thesis, Freien Universität, Berlin, Germany, 1978.
28. Burger, K.J.; Trautner, T.A. Specific labelling of replicating SPP1 DNA: Analysis of viral DNA synthesis and identification of phage DNA-genes. *Mol. Gen. Genet.* **1978**, *166*, 277–285. [CrossRef] [PubMed]
29. Mastromei, G.; Riva, S.; Fietta, A.; Pagani, L. SPP1 DNA replicative forms: Growth of phage SPP1 in *Bacillus subtilis* mutants temperature-sensitive in DNA synthesis. *Mol. Gen. Genet.* **1978**, *167*, 157–164. [CrossRef]
30. Lo Piano, A.; Martínez-Jiménez, M.I.; Zecchi, L.; Ayora, S. Recombination-dependent concatemeric viral DNA replication. *Virus Res.* **2011**, *160*, 1–14. [CrossRef]
31. Zecchi, L.; Lo Piano, A.; Suzuki, Y.; Cañas, C.; Takeyasu, K.; Ayora, S. Characterization of the Holliday Junction resolving enzyme encoded by the *Bacillus subtilis* bacteriophage SPP1. *PLoS ONE* **2012**, *7*, e48440. [CrossRef]
32. Auzat, I.; Petitpas, I.; Lurz, R.; Weise, F.; Tavares, P. A touch of glue to complete bacteriophage assembly: The tail-to-head joining protein (THJP) family. *Mol. Microbiol.* **2014**, *91*, 1164–1178. [CrossRef] [PubMed]
33. Fernandes, S.; São-José, C. Probing the function of the two holin-like proteins of bacteriophage SPP1. *Virology* **2017**, *500*, 184–189. [CrossRef] [PubMed]
34. Van Dijk, E.L.; Auger, H.; Jaszczyszyn, Y.; Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **2014**, *30*, 418–426. [CrossRef] [PubMed]
35. Djacem, K.; Tavares, P.; Oliveira, L. Bacteriophage SPP1 pac cleavage: A precise cut without sequence specificity requirement. *J. Mol. Biol.* **2017**, *429*, 1381–1395. [CrossRef] [PubMed]
36. Solovyev, V.; Salamov, A. Automatic annotation of microbial genomes and metagenomic sequences. In *Metagenomics and Its Applications in Agriculture, Biomedicine, and Environmental Studies*; Li, R.W., Ed.; Nova Science Publisher's: Hauppauge, NY, USA, 2011; pp. 61–78, ISBN 978-1-61668-682-6.
37. Moran, C.P.; Lang, N.; LeGrice, S.F.J.; Lee, G.; Stephens, M.; Sonenshein, A.L.; Pero, J.; Losick, R. Nucleotide sequences that signal the initiation of transcription and translation in *Bacillus subtilis*. *Mol. Gen. Genet.* **1982**, *186*, 339–346. [CrossRef] [PubMed]
38. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
39. Madden, T. The BLAST Sequence Analysis Tool. Available online: https://www.ncbi.nlm.nih.gov/books/NBK153387/ (accessed on 22 August 2018).
40. Zimmermann, L.; Stephens, A.; Nam, S.-Z.; Rau, D.; Kübler, J.; Lozajic, M.; Gabler, F.; Söding, J.; Lupas, A.N.; Alva, V. A completely reimplemented MPI Bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.* **2018**, *430*, 2237–2243. [CrossRef] [PubMed]
41. Stüber, D.; Morelli, G.; Bujard, H.; Montenegro, M.A.; Trautner, T.A. Promoter sites in the genome of *B. subtilis* phage SPP1. *Mol. Gen. Genet.* **1981**, *181*, 518–521. [CrossRef] [PubMed]
42. Tailor, R.; Bensi, G.; Morelli, G.; Canosi, U.; Trautner, T.A. The genome of *Bacillus subtilis* phage SPP1: structure of an early promoter. *J. Gen. Microbiol.* **1985**, *131*, 1259–1262. [CrossRef]
43. Gautheret, D.; Lambert, A. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J. Mol. Biol.* **2001**, *313*, 1003–1011. [CrossRef] [PubMed]

44. Hofacker, I.L.; Fontana, W.; Stadler, P.F.; Bonhoeffer, L.S.; Tacker, M.; Schuster, P. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie* **1994**, *125*, 167–188. [CrossRef]
45. Lesnik, E.A.; Sampath, R.; Levene, H.B.; Henderson, T.J.; McNeil, J.A.; Ecker, D.J. Prediction of rho-independent transcriptional terminators in *Escherichia coli*. *Nucleic Acids Res.* **2001**, *29*, 3583–3594. [CrossRef] [PubMed]
46. Macke, T.J.; Ecker, D.J.; Gutell, R.R.; Gautheret, D.; Case, D.A.; Sampath, R. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.* **2001**, *29*, 4724–4735. [CrossRef] [PubMed]
47. Moszer, I.; Rocha, E.P.C.; Danchin, A. Codon usage and lateral gene transfer in *Bacillus subtilis*. *Curr. Opin. Microbiol.* **1999**, *2*, 524–528. [CrossRef]
48. Gupta, S.K.; Ghosh, T.C. CUCG: A non-redundant codon usage database from complete genomes. *Curr. Sci.* **2000**, *78*, 28–29.
49. Kunst, F.; Ogasawara, N.; Moszer, I.; Albertini, A.M.; Alloni, G.; Azevedo, V.; Bertero, M.G.; Bessières, P.; Bolotin, A.; Borchert, S.; et al. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **1997**, *390*, 249–256. [CrossRef] [PubMed]
50. Riva, S.C. Asymetric transcription of *B. subtilis* phage SPP1 DNA in vitro. *Biochem. Biophys. Res. Commun.* **1969**, *34*, 824–830. [CrossRef]
51. Chenciner, N.; Milanesi, G. Restriction fragment analysis of bacteriophage SPP1 in vitro transcription by host RNA polymerase. *J. Virol.* **1978**, *28*, 95–105. [PubMed]
52. Montenegro, M.A.; Trautner, T.A. In vivo transcription of *Bacillus subtilis* bacteriophage SPP1. *Mol. Gen. Genet.* **1981**, *181*, 512–517. [CrossRef]
53. Desmyter, A.; Reeve, J.N.; Morelli, G.; Trautner, T.A. Inversion and deletion mutants in *Bacillus subtilis* bacteriophage SPP1 as a consequence of cloning. *Mol. Gen. Genet.* **1985**, *198*, 537–539. [CrossRef] [PubMed]
54. Milanesi, G.; Cassani, G. Transcription after bacteriophage SPP1 infection in *Bacillus subtilis*. *J. Virol.* **1972**, *10*, 187–192. [PubMed]
55. Chai, S.; Szepan, U.; Liider, G.; Trautner, T.A.; Alonso, J.C. Sequence analysis of the left end of the *Bacillus subtilis* bacteriophage SPP1 genome. *Gene* **1993**, *129*, 41–49. [CrossRef]
56. Pedré, X.; Weise, F.; Chai, S.; Lüder, G.; Alonso, J.C. Analysis of cis and trans acting elements required for the initiation of DNA replication in the *Bacillus subtilis* bacteriophage SPP1. *J. Mol. Biol.* **1994**, *236*, 1324–1340. [CrossRef]
57. Chai, S.; Bravo, A.; Lüder, G.; Nedlin, A.; Trautner, T.A.; Alonso, J.C. Molecular analysis of the *Bacillus subtilis* bacteriophage SPP1 region encompassing genes 1 to 6. The products of gene 1 and gene 2 are required for pac cleavage. *J. Mol. Biol.* **1992**, *224*, 87–102. [CrossRef]
58. Pope, W.H.; Jacobs-Sera, D.; Russell, D.A.; Rubin, D.H.F.; Kajee, A.; Msibi, Z.N.P.; Larsen, M.H.; Jacobs, W.R.; Lawrence, J.G.; Hendrix, R.W.; et al. Genomics and proteomics of mycobacteriophage patience, an accidental tourist in the *Mycobacterium* neighborhood. *MBio* **2014**, *5*, e02145. [CrossRef] [PubMed]
59. Erez, Z.; Steinberger-Levy, I.; Shamir, M.; Doron, S.; Stokar-Avihail, A.; Peleg, Y.; Melamed, S.; Leavitt, A.; Savidor, A.; Albeck, S.; et al. Communication between viruses guides lysis-lysogeny decisions. *Nature* **2017**, *541*, 488–493. [CrossRef] [PubMed]
60. Lurz, R.; Orlova, E.V.; Günther, D.; Dube, P.; Dröge, A.; Weise, F.; van Heel, M.; Tavares, P. Structural organisation of the head-to-tail interface of a bacterial virus. *J. Mol. Biol.* **2001**, *310*, 1027–1037. [CrossRef] [PubMed]
61. Sprengel, R.; Reiss, B.; Schaller, H. Translationally coupled initiation of protein synthesis in *Bacillus subtilis*. *Nucleic Acids Res.* **1985**, *13*, 893–909. [CrossRef] [PubMed]
62. Ma, J.; Campbell, A.; Karlin, S. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J. Bacteriol.* **2002**, *184*, 5733–5745. [CrossRef]
63. Chai, S.; Lurz, R.; Alonso, J.C. The small subunit of the terminase enzyme of *Bacillus subtilis* bacteriophage SPP1 forms a specialized nucleoprotein complex with the packaging initiation region. *J. Mol. Biol.* **1995**, *252*, 386–398. [CrossRef]
64. Gual, A.; Alonso, J.C. Characterization of the small subunit of the terminase enzyme of the *Bacillus subtilis* bacteriophage SPP1. *Virology* **1998**, *242*, 279–287. [CrossRef] [PubMed]
65. Büttner, C.R.; Chechik, M.; Ortiz-Lombardía, M.; Smits, C.; Ebong, I.-O.; Chechik, V.; Jeschke, G.; Dykeman, E.; Benini, S.; Robinson, C.V.; et al. Structural basis for DNA recognition and loading into a viral packaging motor. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 811–816. [CrossRef] [PubMed]

66. Greive, S.J.; Fung, H.K.H.; Chechik, M.; Jenkins, H.T.; Weitzel, S.E.; Aguiar, P.M.; Brentnall, A.S.; Glousieau, M.; Gladyshev, G.V.; Potts, J.R.; et al. DNA recognition for virus assembly through multiple sequence-independent interactions with a helix-turn-helix motif. *Nucleic Acids Res.* **2016**, *44*, 776–789. [CrossRef] [PubMed]
67. Camacho, A.G.; Gual, A.; Lurz, R.; Tavares, P.; Alonso, J.C. *Bacillus subtilis* bacteriophage SPP1 DNA packaging motor requires terminase and portal proteins. *J. Biol. Chem.* **2003**, *278*, 23251–23259. [CrossRef]
68. Chai, S.; Alonso, J.C. Distamycin-induced inhibition of formation of a nucleoprotein complex between the terminase small subunit G1P and the non-encapsidated end (pacL site) of *Bacillus subtilis* bacteriophage SPP1. *Nucleic Acids Res.* **1996**, *24*, 282–288. [CrossRef] [PubMed]
69. Chai, S.; Szepan, U.; Alonso, J.C. *Bacillus subtilis* bacteriophage SPP1 terminase has a dual activity: It is required for the packaging initiation and represses its own synthesis. *Gene* **1997**, *184*, 251–256. [CrossRef]
70. Gual, A.; Camacho, A.G.; Alonso, J.C. Functional analysis of the terminase large subunit, G2P, of *Bacillus subtilis* Bacteriophage SPP1. *J. Biol. Chem.* **2000**, *275*, 35311–35319. [CrossRef]
71. Smits, C.; Chechik, M.; Kovalevskiy, O.V.; Shevtsov, M.B.; Foster, A.W.; Alonso, J.C.; Antson, A.A. Structural basis for the nuclease activity of a bacteriophage large terminase. *EMBO Rep.* **2009**, *10*, 592–598. [CrossRef]
72. Cornilleau, C.; Atmane, N.; Jacquet, E.; Smits, C.; Alonso, J.C.; Tavares, P.; Oliveira, L. The nuclease domain of the SPP1 packaging motor coordinates DNA cleavage and encapsidation. *Nucleic Acids Res.* **2013**, *41*, 340–354. [CrossRef]
73. Oliveira, L.; Henriques, A.O.; Tavares, P. Modulation of the viral ATPase activity by the portal protein correlates with DNA packaging efficiency. *J. Biol. Chem.* **2006**, *281*, 21914–21923. [CrossRef]
74. Oliveira, L.; Cuervo, A.; Tavares, P. Direct interaction of the bacteriophage SPP1 packaging ATPase with the portal protein. *J. Biol. Chem.* **2010**, *285*, 7366–7373. [CrossRef] [PubMed]
75. Dube, P.; Tavares, P.; Lurz, R.; Van Heel, M. The portal protein of bacteriophage SPP1: a DNA pump with 13-fold symmetry. *EMBO J.* **1993**, *1*, 1303–1309. [CrossRef]
76. Jekow, P.; Behlke, J.; Tichelaar, W.; Lurz, R.; Regalla, M.; Hinrichs, W.; Tavares, P. Effect of the ionic environment on the molecular structure of bacteriophage SPP1 portal protein. *Eur. J. Biochem.* **1999**, *264*, 724–735. [CrossRef] [PubMed]
77. Orlova, E.V.; Dube, P.; Beckmann, E.; Zemlin, F.; Lurz, R.; Trautner, T.A.; Tavares, P.; Van Heel, M. Structure of the 13-fold symmetric portal protein of bacteriophage SPP1. *Nat. Struct. Biol.* **1999**, *6*, 842–846. [CrossRef] [PubMed]
78. Orlova, E.V.; Gowen, B.; Dröge, A.; Stiege, A.; Weise, F.; Lurz, R.; van Heel, M.; Tavares, P. Structure of a viral DNA gatekeeper at 10 Å resolution by cryo-electron microscopy. *EMBO J.* **2003**, *22*, 1255–1262. [CrossRef] [PubMed]
79. Isidro, A.; Henriques, A.O.; Tavares, P. The portal protein plays essential roles at different steps of the SPP1 DNA packaging process. *Virology* **2004**, *322*, 253–263. [CrossRef]
80. Isidro, A.; Santos, M.A.; Henriques, A.O.; Tavares, P. The high-resolution functional map of bacteriophage SPP1 portal protein. *Mol. Microbiol.* **2004**, *51*, 949–962. [CrossRef]
81. Chaban, Y.; Lurz, R.; Brasilès, S.; Cornilleau, C.; Karreman, M.; Zinn-Justin, S.; Tavares, P.; Orlova, E.V. Structural rearrangements in the phage head-to-tail interface during assembly and infection. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 7009–7014. [CrossRef]
82. Cuervo, A.; Vaney, M.-C.; Antson, A.A.; Tavares, P.; Oliveira, L. Structural rearrangements between portal protein subunits are essential for viral DNA translocation. *J. Biol. Chem.* **2007**, *282*, 18907–18913. [CrossRef]
83. Vinga, I.; Dröge, A.; Stiege, A.C.; Lurz, R.; Santos, M.A.; Daugelavičius, R.; Tavares, P. The minor capsid protein gp7 of bacteriophage SPP1 is required for efficient infection of *Bacillus subtilis*. *Mol. Microbiol.* **2006**, *61*, 1609–1621. [CrossRef]
84. Stiege, A.C.; Isidro, A.; Dröge, A.; Tavares, P. Specific targeting of a DNA-binding protein to the SPP1 procapsid by interaction with the portal oligomer. *Mol. Microbiol.* **2003**, *49*, 1201–1212. [CrossRef] [PubMed]
85. Becker, B.; de la Fuente, N.; Gassel, M.; Günther, D.; Tavares, P.; Lurz, R.; Trautner, T.A.; Alonso, J.C. Head morphogenesis genes of the *Bacillus subtilis* bacteriophage SPP1. *J. Mol. Biol.* **1997**, *268*, 822–839. [CrossRef] [PubMed]
86. Dröge, A.; Santos, M.A.; Stiege, A.C.; Alonso, J.C.; Lurz, R.; Trautner, T.A.; Tavares, P. Shape and DNA packaging activity of bacteriophage SPP1 procapsid: protein components and interactions during assembly. *J. Mol. Biol.* **2000**, *296*, 117–132. [CrossRef] [PubMed]

87. Poh, S.L.; el Khadali, F.; Berrier, C.; Lurz, R.; Melki, R.; Tavares, P. Oligomerization of the SPP1 scaffolding Protein. *J. Mol. Biol.* **2008**, *378*, 551–564. [CrossRef] [PubMed]
88. Zairi, M.; Stiege, A.C.; Nhiri, N.; Jacquet, E.; Tavares, P. The collagen-like protein gp12 is a temperature-dependent reversible binder of SPP1 viral capsids. *J. Biol. Chem.* **2014**, *289*, 27169–27181. [CrossRef] [PubMed]
89. Lhuillier, S.; Gallopin, M.; Gilquin, B.; Brasilès, S.; Lancelot, N.; Letellier, G.; Gilles, M.; Dethan, G.; Orlova, E.V.; Couprie, J.; et al. Structure of bacteriophage SPP1 head-to-tail connection reveals mechanism for viral DNA gating. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 8507–8512. [CrossRef] [PubMed]
90. Chagot, B.; Auzat, I.; Gallopin, M.; Petitpas, I.; Gilquin, B.; Tavares, P.; Zinn-Justin, S. Solution structure of gp17 from the Siphoviridae bacteriophage SPP1: Insights into its role in virion assembly. *Proteins Struct. Funct. Bioinform.* **2012**, *80*, 319–326. [CrossRef]
91. Auzat, I.; Dröge, A.; Weise, F.; Lurz, R.; Tavares, P. Origin and function of the two major tail proteins of bacteriophage SPP1. *Mol. Microbiol.* **2008**, *70*, 557–569. [CrossRef]
92. Langlois, C.; Ramboarina, S.; Cukkemane, A.; Auzat, I.; Chagot, B.; Gilquin, B.; Ignatiou, A.; Petitpas, I.; Kasotakis, E.; Paternostre, M.; et al. Bacteriophage SPP1 tail tube protein self-assembles into β-structure-rich tubes. *J. Biol. Chem.* **2015**, *290*, 3836–3849. [CrossRef]
93. Goulet, A.; Lai-Kee-Him, J.; Veesler, D.; Auzat, I.; Robin, G.; Shepherd, D.A.; Ashcroft, A.E.; Richard, E.; Lichière, J.; Tavares, P.; et al. The opening of the SPP1 bacteriophage tail, a prevalent mechanism in Gram-positive-infecting siphophages. *J. Biol. Chem.* **2011**, *286*, 25397–25405. [CrossRef]
94. Veesler, D.; Robin, G.; Lichière, J.; Auzat, I.; Tavares, P.; Bron, P.; Campanacci, V.; Cambillau, C. Crystal structure of bacteriophage SPP1 distal tail protein (gp19.1): A baseplate hub paradigm in gram-positive infecting phages. *J. Biol. Chem.* **2010**, *285*, 36666–36673. [CrossRef] [PubMed]
95. Vinga, I.; Baptista, C.; Auzat, I.; Petipas, I.; Lurz, R.; Tavares, P.; Santos, M.A.; São-José, C. Role of bacteriophage SPP1 tail spike protein gp21 on host cell receptor binding and trigger of phage DNA ejection. *Mol. Microbiol.* **2012**, *83*, 289–303. [CrossRef] [PubMed]
96. Veesler, D.; Blangy, S.; Spinelli, S.; Tavares, P.; Campanacci, V.; Cambillau, C. Crystal structure of *Bacillus subtilis* SPP1 phage gp22 shares fold similarity with a domain of lactococcal phage p2 RBP. *Protein Sci.* **2010**, *19*, 1439–1443. [CrossRef] [PubMed]
97. Veesler, D.; Blangy, S.; Lichière, J.; Ortiz-Lombardía, M.; Tavares, P.; Campanacci, V.; Cambillau, C. Crystal structure of *Bacillus subtilis* SPP1 phage gp23.1, a putative chaperone. *Protein Sci.* **2010**, *19*, 1812–1816. [CrossRef] [PubMed]
98. Fernandes, S.; São-José, C. More than a hole: The holin lethal function may be required to fully sensitize bacteria to the lytic action of canonical endolysins. *Mol. Microbiol.* **2016**, *102*, 92–106. [CrossRef] [PubMed]
99. Martínez-Jiménez, M.I.; Alonso, J.C.; Ayora, S. *Bacillus subtilis* bacteriophage SPP1-encoded gene *34.1* product is a recombination-dependent DNA replication protein. *J. Mol. Biol.* **2005**, *351*, 1007–1019. [CrossRef]
100. Ayora, S.; Missich, R.; Mesa, P.; Lurz, R.; Yang, S.; Egelman, E.H.; Alonso, J.C. Homologous-pairing activity of the *Bacillus subtilis* bacteriophage SPP1 replication protein G35P. *J. Biol. Chem.* **2002**, *277*, 35969–35979. [CrossRef]
101. Seco, E.M.; Zinder, J.C.; Manhart, C.M.; Lo Piano, A.; McHenry, C.S.; Ayora, S. Bacteriophage SPP1 DNA replication strategies promote viral and disable host replication in vitro. *Nucleic Acids Res.* **2013**, *41*, 1711–1721. [CrossRef]
102. Missich, R.; Weise, F.; Chai, S.; Lurz, R.; Pedré, X.; Alonso, J.C. The replisome organizer (G38P) of *Bacillus subtilis* bacteriophage SPP1 forms specialized nucleoprotein complexes with two discrete distant regions of the SPP1 genome. *J. Mol. Biol.* **1997**, *270*, 50–64. [CrossRef]
103. Ayora, S.; Stasiak, A.; Alonso, J.C. The *Bacillus subtilis* bacteriophage SPP1 G39P delivers and activates the G40P DNA helicase upon interacting with the G38P-bound replication origin. *J. Mol. Biol.* **1999**, *288*, 71–85. [CrossRef]
104. Bailey, S.; Sedelnikova, S.E.; Mesa, P.; Ayora, S.; Waltho, J.P.; Ashcroft, A.E.; Baron, A.J.; Alonso, J.C.; Rafferty, J.B. Structural analysis of *Bacillus subtilis* SPP1 phage helicase loader protein G39P. *J. Biol. Chem.* **2003**, *278*, 15304–15312. [CrossRef] [PubMed]
105. Ayora, S.; Weise, F.; Mesa, P.; Stasiak, A.; Alonso, J.C. *Bacillus subtilis* bacteriophage SPP1 hexameric DNA helicase, G40P, interacts with forked DNA. *Nucleic Acids Res.* **2002**, *30*, 2280–2289. [CrossRef] [PubMed]

106. Mesa, P.; Alonso, J.C.; Ayora, S. *Bacillus subtilis* Bacteriophage SPP1 G*40*P helicase lacking the N-terminal domain unwinds DNA bidirectionally. *J. Mol. Biol.* **2006**, *357*, 1077–1088. [CrossRef] [PubMed]
107. Ayora, S.; Langer, U.; Alonso, J.C. *Bacillus subtilis* DnaG primase stabilises the bacteriophage SPP1 G40P helicase-ssDNA complex. *FEBS Lett.* **1998**, *439*, 59–62. [CrossRef]
108. Wang, G.; Klein, M.G.; Tokonzaba, E.; Zhang, Y.; Holden, L.G.; Chen, X.S. The structure of a DnaB-family replicative helicase and its interactions with primase. *Nat. Struct. Mol. Biol.* **2008**, *15*, 94–100. [CrossRef] [PubMed]
109. Martínez-Jiménez, M.I.; Mesa, P.; Alonso, J.C. *Bacillus subtilis* tau subunit of DNA polymerase III interacts with bacteriophage SPP1 replicative DNA helicase G40P. *Nucleic Acids Res.* **2002**, *30*, 5056–5064. [CrossRef] [PubMed]
110. Rocha, E.P.C.; Danchin, A.; Viari, A. Translation in *Bacillus subtilis*: Roles and trends of initiation and termination, insights from a genome analysis. *Nucleic Acids Res.* **1999**, *27*, 3567–3576. [CrossRef] [PubMed]
111. Lovett, P.S.; Ambulos, N.P.; Mulbry, W.; Noguchi, N.; Rogers, E.J. UGA can be decoded as tryptophan at low efficiency in *Bacillus subtilis*. *J. Bacteriol.* **1991**, *173*, 1810–1812. [CrossRef]
112. Xu, J.; Hendrix, R.W.; Duda, R.L. Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. *Mol. Cell* **2004**, *16*, 11–21. [CrossRef]
113. Xu, J.; Hendrix, R.W.; Duda, R.L. A balanced ratio of proteins from gene G and frameshift-extended gene GT is required for phage Lambda tail assembly. *J. Mol. Biol.* **2013**, *425*, 3476–3487. [CrossRef]
114. Xu, J.; Hendrix, R.W.; Duda, R.L. Chaperone–protein interactions that mediate assembly of the bacteriophage lambda tail to the correct length. *J. Mol. Biol.* **2014**, *426*, 1004–1018. [CrossRef] [PubMed]
115. Juhala, R.J.; Ford, M.E.; Duda, R.L.; Youlton, A.; Hatfull, G.F.; Hendrix, R.W. Genomic sequences of bacteriophages HK97 and HK022: Pervasive genetic mosaicism in the lambdoid bacteriophages. *J. Mol. Biol.* **2000**, *299*, 27–51. [CrossRef] [PubMed]
116. Hendrix, R.W.; Lawrence, J.G.; Hatfull, G.F.; Casjens, S. The origins and ongoing evolution of viruses. *Trends Microbiol.* **2000**, *8*, 504–508. [CrossRef]
117. Cumby, N.; Davidson, A.R.; Maxwell, K.L. The moron comes of age. *Bacteriophage* **2012**, *2*, e23146. [CrossRef]
118. Chai, S.; Kruft, V.; Alonso, J.C. Analysis of the *Bacillus subtilis* bacteriophages SPP1 and SF6 gene *1* product: a protein involved in the initiation of headful packaging. *Virology* **1994**, *202*, 930–939. [CrossRef]
119. Umene, K.; Shiraishi, A. Complete nucleotide sequence of *Bacillus subtilis* (natto) bacteriophage PM1, a phage associated with disruption of food production. *Virus Genes* **2013**, *46*, 524–534. [CrossRef] [PubMed]
120. Marks, T.J.; Hamilton, P.T. Characterization of a thermophilic bacteriophage of *Geobacillus kaustophilus*. *Arch. Virol.* **2014**, *159*, 2771–2775. [CrossRef] [PubMed]
121. Botstein, D. A theory of modular evolution for bacteriophages. *Ann. N. Y. Acad. Sci.* **1980**, *354*, 484–491. [CrossRef]
122. Hendrix, R.W.; Smith, M.C.; Burns, R.N.; Ford, M.E.; Hatfull, G.F. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 2192–2197. [CrossRef]
123. Alonso, J.C.; Tavares, P.; Lurz, R.; Trautner, T.A. Bacteriophage SPP1. In *The Bacteriophages*; Calendar, R., Ed.; Oxford University Press: New York, NY, USA, 2006; Volume 54, p. 746, ISBN 9780195148503.
124. Tavares, P.; Zinn-Justin, S.; Orlova, E.V. Genome gating in tailed bacteriophage capsids. In *Viral Molecular Machines*; Rossmann, M.G., Rao, V.B., Eds.; Advances in Experimental Medicine and Biology; Springer: Boston, MA, USA, 2012; Volume 726, pp. 585–600, ISBN 978-1-4614-0979-3.
125. Oliveira, L.; Tavares, P.; Alonso, J.C. Headful DNA packaging: bacteriophage SPP1 as a model system. *Virus Res.* **2013**, *173*, 247–259. [CrossRef]
126. Tavares, P. The bacteriophage head-to-tail interface. In *Subcellular Biochemistry*; Harris, J.R., Bhella, D., Eds.; Springer: Singapore, 2018; Volume 88, pp. 305–328, ISBN 978-981-10-8456-0.
127. Hendrix, R.W.; Duda, R.L. Bacteriophage lambda PaPa: Not the mother of all lambda phages. *Science* **1992**, *258*, 1145–1148. [CrossRef] [PubMed]
128. Shen, B.W.; Landthaler, M.; Shub, D.A.; Stoddard, B.L. DNA binding and bleavage by the HNH Homing Endonuclease I-HmuI. *J. Mol. Biol.* **2004**, *342*, 43–56. [CrossRef] [PubMed]
129. Edgell, D.R.; Gibb, E.A.; Belfort, M. Mobile DNA elements in T4 and related phages. *Virol. J.* **2010**, *7*, 290. [CrossRef]

130. Landthaler, M.; Lau, N.C.; Shub, D.A. Group I intron homing in *Bacillus* phages SPO1 and SP82: A gene conversion event initiated by a nicking homing endonuclease. *J. Bacteriol.* **2004**, *186*, 4307–4314. [CrossRef] [PubMed]
131. Quiles-Puchalt, N.; Carpena, N.; Alonso, J.C.; Novick, R.P.; Marina, A.; Penadés, J.R. Staphylococcal pathogenicity island DNA packaging system involving cos-site packaging and phage-encoded HNH endonucleases. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 6016–6021. [CrossRef] [PubMed]
132. Kala, S.; Cumby, N.; Sadowski, P.D.; Hyder, B.Z.; Kanelis, V.; Davidson, A.R.; Maxwell, K.L. HNH proteins are a widespread component of phage DNA packaging machines. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 6022–6027. [CrossRef]
133. Neamah, M.M.; Mir-Sanchis, I.; López-Sanz, M.; Acosta, S.; Baquedano, I.; Haag, A.F.; Marina, A.; Ayora, S.; Penadés, J.R. Sak and Sak4 recombinases are required for bacteriophage replication in *Staphylococcus aureus*. *Nucleic Acids Res.* **2017**, *45*, 6507–6519. [CrossRef]
134. Ogunleye, A.; Irorere, V.U.; Williams, C.; Hill, D.; Bhat, A.; Radecka, I. Poly-γ-glutamic acid: production, properties and applications. *Microbiology* **2015**, *161*, 1–17. [CrossRef]
135. Mamberti, S.; Prati, P.; Cremaschi, P.; Seppi, C.; Morelli, C.F.; Galizzi, A.; Fabbi, M.; Calvio, C. γ-PGA hydrolases of phage origin in *Bacillus subtilis* and other microbial genomes. *PLoS ONE* **2015**, *10*, 1–17. [CrossRef]
136. Kimura, K.; Itoh, Y. Characterization of poly-γ-glutamate hydrolase encoded by a bacteriophage genome: possible role in phage infection of *Bacillus subtilis* encapsulated with poly-γ-glutamate. *Appl. Environ. Microbiol.* **2003**, *69*, 2491–2497. [CrossRef]
137. Ghosh, K.; Senevirathne, A.; Kang, H.; Hyun, W.; Kim, J.; Kim, K.-P.; Ghosh, K.; Senevirathne, A.; Kang, H.S.; Hyun, W.; et al. Complete nucleotide sequence analysis of a novel *Bacillus subtilis*-infecting bacteriophage BSP10 and its effect on poly-gamma-glutamic acid degradation. *Viruses* **2018**, *10*, 240. [CrossRef] [PubMed]
138. Molshanski-Mor, S.; Yosef, I.; Kiro, R.; Edgar, R.; Manor, M.; Gershovits, M.; Laserson, M.; Pupko, T.; Qimron, U. Revealing bacterial targets of growth inhibitors encoded by bacteriophage T7. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 18715–18720. [CrossRef] [PubMed]
139. Bondy-Denomy, J.; Qian, J.; Westra, E.R.; Buckling, A.; Guttman, D.S.; Davidson, A.R.; Maxwell, K.L. Prophages mediate defense against phage infection through diverse mechanisms. *ISME J.* **2016**, *10*, 2854–2866. [CrossRef] [PubMed]
140. Dedrick, R.M.; Jacobs-Sera, D.; Bustamante, C.A.G.; Garlena, R.A.; Mavrich, T.N.; Pope, W.H.; Reyes, J.C.C.; Russell, D.A.; Adair, T.; Alvey, R.; et al. Prophage-mediated defence against viral attack and viral counter-defence. *Nat. Microbiol.* **2017**, *2*, 16251. [CrossRef] [PubMed]
141. Seco, E.M.; Ayora, S. *Bacillus subtilis* DNA polymerases, PolC and DnaE, are required for both leading and lagging strand synthesis in SPP1 origin-dependent DNA replication. *Nucleic Acids Res.* **2017**, *45*, 8302–8313. [CrossRef]
142. Alonso, J.C.; Ayora, S.; Canosa, I.; Weise, F.; Rojo, F. Site-specific recombination in Gram-positive theta-replicating plasmids. *FEMS Microbiol. Lett.* **1996**, *142*, 1–10. [CrossRef]
143. Stahl, F.W.; Murray, N.E. The evolution of gene clusters and genetic circularity in microorganisms. *Genetics* **1966**, *53*, 569–576.
144. Young, R. Phage lysis: Do we have the hole story yet? *Curr. Opin. Microbiol.* **2013**, *16*, 790–797. [CrossRef]
145. Krogh, S.; Jørgensen, S.T.; Devine, K.M. Lysis genes of the *Bacillus subtilis* defective prophage PBSX. *J. Bacteriol.* **1998**, *180*, 2110–2117.
146. Bläsi, U.; Young, R. Two beginnings for a single purpose: The dual-start holins in the regulation of phage lysis. *Mol. Microbiol.* **1996**, *21*, 675–682. [CrossRef]
147. Abrescia, N.G.A.A.; Bamford, D.H.; Grimes, J.M.; Stuart, D.I. Structure unifies the viral universe. *Annu. Rev. Biochem.* **2012**, *81*, 795–822. [CrossRef] [PubMed]
148. Veesler, D.; Cambillau, C. A common evolutionary origin for tailed-bacteriophage functional modules and bacterial machineries. *Microbiol. Mol. Biol. Rev.* **2011**, *75*, 423–433. [CrossRef] [PubMed]
149. Jamet, A.; Touchon, M.; Ribeiro-Gonçalves, B.; Carriço, J.A.; Charbit, A.; Nassif, X.; Ramirez, M.; Rocha, E.P.C. A widespread family of polymorphic toxins encoded by temperate phages. *BMC Biol.* **2017**, *15*, 75. [CrossRef] [PubMed]
150. Jamet, A.; Charbit, A.; Nassif, X. Antibacterial toxins: Gram-positive bacteria strike back! *Trends Microbiol.* **2018**, *26*, 89–91. [CrossRef] [PubMed]

151. Deichelbohrer, I.; Messer, W.; Trautner, T.A. Genome of *Bacillus subtilis* bacteriophage SPP1: structure and nucleotide sequence of *pac*, the origin of DNA packaging. *J. Virol.* **1982**, *42*, 83–90. [PubMed]
152. Bravo, A.; Alonso, J.C.; Trautner, T.A. Functional analysis of the *Bacillus subtilis* bacteriophage SPP1 pac site. *Nucleic Acids Res.* **1990**, *18*, 2881. [CrossRef] [PubMed]
153. Dröge, A.; Tavares, P. In vitro packaging of DNA of the *Bacillus subtilis* bacteriophage SPP1. *J. Mol. Biol.* **2000**, *296*, 103–115. [CrossRef] [PubMed]
154. Oliveira, L.; Alonso, J.C.; Tavares, P. A defined in vitro system for DNA packaging by the bacteriophage SPP1: Insights into the headful packaging mechanism. *J. Mol. Biol.* **2005**, *353*, 529–539. [CrossRef]
155. Tavares, P.; Dröge, A.; Lurz, R.; Graeber, I.; Orlova, E.; Dube, P.; Van Heel, M. The SPP1 connection. *FEMS Microbiol. Rev.* **1995**, *17*, 47–56. [CrossRef]
156. Bönemann, G.; Pietrosiuk, A.; Mogk, A. Tubules and donuts: a type VI secretion story: MicroReview. *Mol. Microbiol.* **2010**, *76*, 815–821. [CrossRef]
157. Scholl, D. Phage Tail–Like Bacteriocins. *Annu. Rev. Virol.* **2017**, *4*, 453–467. [CrossRef] [PubMed]
158. São-José, C.; Baptista, C.; Santos, M.A. *Bacillus subtilis* operon encoding a membrane receptor for bacteriophage SPP1. *J. Bacteriol.* **2004**, *186*, 8337–8346. [CrossRef] [PubMed]
159. King, J. Assembly of the tail of bacteriophage T4. *J. Mol. Biol.* **1968**, *32*, 231–262. [CrossRef]
160. Katsura, I.; Kühl, P.W. Morphogenesis of the tail of bacteriophage lambda: III. Morphogenetic pathway. *J. Mol. Biol.* **1975**, *91*, 257–273. [CrossRef]
161. Ghosh, N.; McKillop, T.J.; Jowitt, T.A.; Howard, M.; Davies, H.; Holmes, D.F.; Roberts, I.S.; Bella, J. Collagen-like proteins in pathogenic *E. coli* strains. *PLoS ONE* **2012**, *7*, e37872. [CrossRef] [PubMed]
162. Brüssow, H.; Desiere, F. Comparative phage genomics and the evolution of Siphoviridae: Insights from dairy phages. *Mol. Microbiol.* **2001**, *39*, 213–222. [CrossRef] [PubMed]
163. Casjens, S.R. Comparative genomics and evolution of the tailed-bacteriophages. *Curr. Opin. Microbiol.* **2005**, *8*, 451–458. [CrossRef]
164. Stockdale, S.R.; Collins, B.; Spinelli, S.; Douillard, F.P.; Mahony, J.; Cambillau, C.; van Sinderen, D. Structure and assembly of TP901-1 virion unveiled by mutagenesis. *PLoS ONE* **2015**, *10*, e0131676. [CrossRef]
165. Murphy, J.; Bottacini, F.; Mahony, J.; Kelleher, P.; Neve, H.; Zomer, A.; Nauta, A.; van Sinderen, D. Comparative genomics and functional analysis of the 936 group of lactococcal Siphoviridae phages. *Sci. Rep.* **2016**, *6*, 21345. [CrossRef]
166. Lopes, A.; Tavares, P.; Petit, M.-A.; Guérois, R.; Zinn-Justin, S. Automated classification of tailed bacteriophages according to their neck organization. *BMC Genom.* **2014**, *15*, 1027. [CrossRef]
167. Zivanovic, Y.; Confalonieri, F.; Ponchon, L.; Lurz, R.; Chami, M.; Flayhan, A.; Renouard, M.; Huet, A.; Decottignies, P.; Davidson, A.R.; et al. Insights into bacteriophage T5 structure from analysis of its morphogenesis genes and protein components. *J. Virol.* **2014**, *88*, 1162–1174. [CrossRef] [PubMed]
168. Fraser, J.S.; Yu, Z.; Maxwell, K.L.; Davidson, A.R. Ig-Like Domains on bacteriophages: a tale of promiscuity and deceit. *J. Mol. Biol.* **2006**, *359*, 496–507. [CrossRef] [PubMed]
169. Spinelli, S.; Veesler, D.; Bebeacua, C.; Cambillau, C. Structures and host-adhesion mechanisms of lactococcal siphophages. *Front. Microbiol.* **2014**, *5*, 3. [CrossRef] [PubMed]
170. Dowah, A.S.A.; Clokie, M.R.J. Review of the nature, diversity and structure of bacteriophage receptor binding proteins that target Gram-positive bacteria. *Biophys. Rev.* **2018**, *10*, 535–542. [CrossRef] [PubMed]

université PARIS-SACLAY

ÉCOLE DOCTORALE
Innovation thérapeutique
du fondamental à l'appliqué
(ITFA)

Title : Molecular mechanisms of selective viral DNA recognition and *pac*kaging by a bacterial virus

Keywords : viral infection of bacteria, viral assembly, bacteriophage, protein and nucleic acids biochemistry, protein-DNA interaction, structure of nucleoprotein complexes

Abstract : In tailed bacteriophages and herpesviruses, recognition of double stranded genomic DNA and its encapsidation within a preformed procapsid is a crucial step of the viral replication cycle. Tailed bacteriophages use a protein complex known as the terminase to recognise and package their DNA. The terminase complex specifically binds the phage genome, cleaves and then packages the DNA inside the procapsid via a specialised portal vertex. Sometimes however, mistakes lead to packaging of bacterial instead of viral DNA. Such mistakes can lead to generalized transduction in which the virus becomes a vector of horizontal gene transfer. Bacteriophage SPP1 is a model system for studying the encapsidation process among tailed bacteriophages. The SPP1 terminase complex (composed of gp1 and gp2) recognises a specific '*pac*' sequence on the viral genome.

We characterized the interaction between the terminase complex and *pac* sequence to better understand how the phage discriminates its own genome from its host's DNA during encapsidation. We induced some mutations on two sites within *pac*; *pac*L and *pac*R, which are upstream and inside the coding sequence of gp1, respectively. We demonstrate that a large portion (~100 bp) of *pac*L was not necessary for specific recognition of the *pac* sequence and that in fact a poly-A sequence of *pac*R was crucial for specific SPP1 DNA recognition. However, deletions greater than 46 bp in *pac*L led to defects in phage infectivity and revertants arose during subsequent phage amplification via modification of transcription of the terminase operon . In *pac*R mutants, suppressors emerged via amino-acid substitutions in gp1. These mutations led to a decrease of specificity of interaction between gp1 and viral DNA and a dramatic increase in the frequency of bacterial DNA transduction.

Maison du doctorat de l'Université Paris-Saclay
2ème étage aile ouest, Ecole normale supérieure Paris-Saclay
4 avenue des Sciences,
91190 Gif sur Yvette, France

université PARIS-SACLAY

ÉCOLE DOCTORALE
Innovation thérapeutique
du fondamental à l'appliqué
(ITFA)

Titre : Mécanismes moléculaires régissant la reconnaissance sélective et l'encapsidation du génome d'un virus bactérien

Mots clés : bactériophage, assemblage viral, complexes nucléoprotéiques, infection virale de bactéries, interaction protéine-ADN, biochimie des protéines et des acides nucléiques.

Résumé : Chez les bactériophages caudés et les herpèsvirus, la reconnaissance et l'encapsidation d'un génome à ADN double brin dans une procapside préformée est une étape clé du cycle viral. Chez les bactériophages caudés, la reconnaissance et l'encapsidation de l'ADN viral requièrent l'assemblage d'un complexe moléculaire, la terminase. Il reconnaît spécifiquement le génome phagique qu'il clive et encapside à l'intérieur de la procapside au travers d'un pore formé par la protéine portal. De rares erreurs du complexe terminase aboutissent à la reconnaissance et à l'encapsidation d'ADN bactérien. On parlera alors de transduction généralisée, le virus devenant ainsi un vecteur de transfert horizontal de matériel génétique. SPP1 est un système modèle utilisé dans l'étude de l'encapsidation de l'ADN viral chez les bactériophages caudés. Chez ce virus, le complexe terminase, composé de gp1 et gp2, reconnait une séquence appelée *pac*. Afin de comprendre les mécanismes par lesquels le phage discrimine son ADN de celui de l'hôte, la spécificité d'interaction entre la terminase et *pac* a été étudiée.

Nous avons généré des mutations dans deux régions de *pac*, *pac*L et *pac*R qui se localisent respectivement juste en amont et dans le gène codant pour gp1. Nous avons ainsi prouvé que ~100 pb dans *pac*L n'étaient pas nécessaires à la reconnaissance de l'ADN phagique et qu'une séquence poly-A de *pac*R était importante pour la reconnaissance de *pac* par gp1. Cependant, des délétions supérieures à 46pb dans *pac*L ont entrainé des défauts chez les phages conduisant à l'apparition de suppresseurs. Des modifications de la transcription de l'opéron codant pour les terminases sont à l'origine de la réversion. Dans le cas de *pac*R, des suppresseurs apparaissent avec des mutations induisant des changements dans la séquence en acides aminés de gp1. Ces changements de résidus dans la protéine induisent une baisse de spécificité entre gp1 et l'ADN viral, ce qui conduit à une augmentation drastique de la fréquence de transduction d'ADN bactérien.

Maison du doctorat de l'Université Paris-Saclay
2ème étage aile ouest, Ecole normale supérieure Paris-Saclay
4 avenue des Sciences,
91190 Gif sur Yvette, France