



HAL
open science

GP12: a collagen-like protein that binds to the SPP1 capsid

Mohamed Zairi

► **To cite this version:**

Mohamed Zairi. GP12: a collagen-like protein that binds to the SPP1 capsid. Structural Biology [q-bio.BM]. Université Paris Saclay (COMUE), 2019. English. NNT : 2019SACLS140 . tel-03092296

HAL Id: tel-03092296

<https://theses.hal.science/tel-03092296>

Submitted on 2 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Gp12 : a collagen-like protein that binds to the bacteriophage SPP1 capsid

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

Ecole doctorale N°569
Innovation thérapeutique : du fondamental à l'appliqué
Spécialité de doctorat: Biochimie et biologie structurale

Thèse présentée et soutenue à Gif-sur-Yvette le 11/06/2019, par

Mohamed Zairi

Composition du Jury :

Mr. Herman van Tilbeurgh Professeur, Université Paris-Sud, Orsay (I2BC)	Président
Mme. Sylvie Ricard-Blum Professeur, Université Claude Bernard Lyon 1 (ICBMS)	Rapporteuse
Mme. Cécile Breyton Directrice de Recherche, CNRS Grenoble (IBS)	Rapporteuse
Mr. Christian Roumestand Professeur, Université Montpellier 1 (CBS)	Examineur
Mme. Sophie Zinn-Justin Directrice de Recherche, CEA Saclay (I2BC)	Examinatrice
Mr. Paulo Tavares Directeur de Recherche, CNRS Gif-sur-Yvette (I2BC)	Directeur de thèse

**Gp12 : a collagen-like protein that binds to the
bacteriophage SPP1 capsid**

Mohamed Zairi

Thesis director : Paulo Tavares

Remerciements

Je remercie, tout d'abord, l'Ecole Doctorale Innovation thérapeutique du fondamental à l'appliqué pour son financement, sa confiance et pour m'avoir autorisé à soutenir ma thèse. J'exprime ma reconnaissance en particulier à Herman van Tilbeurgh.

Je remercie aussi les membres du jury de thèse qui ont accepté d'évaluer ce travail.

Je remercie mon directeur de thèse, Paulo Tavares pour son accompagnement et ses encouragements tout au long de ce travail. Je le salue particulièrement pour ses qualités pédagogiques et humaines.

Je remercie les membres l'équipe Bactériophages de bactéries Gram-positives, surtout Isabelle Auzat, Sandrine Brasilès et Charlène Cornilleau qui m'ont formé aux techniques de laboratoire. Je les remercie aussi pour tous les bons moments que j'ai passé entre eux.

Je remercie Stéphane Roche et Stéphane Bressanelli pour leurs conseils et toutes les discussions qu'on a eu lors de la réalisation de ce travail.

Je remercie les membres des plateformes de biophysique des protéines et de qPCR, en particulier Eric Jacquet, qui m'ont accompagné lors de la réalisation de certaines expériences de ce travail.

Je remercie mes parents Fethi Zairi et Fetiha Zairi, mes sœurs Sonia Zairi et Sana Zairi, ainsi que mon oncle Lotfi Benfekihali et sa femme Imen Benfekihali pour tous leurs encouragements et et leur support moral.

Je finis par remercier toutes les personnes qui m'ont écouté, conseillé et encouragé pour conclure ce travail.

Abbreviations

CD : Circular dichroism

CDD : Conserved domain detection

CDSs : Coding DNA sequences

CMCPs : Collagen motif containing proteins

CP : Coat protein

cryoEM : Cryo-electron microscopy

dsDNA : Double-stranded DNA

dsRNA : double-stranded RNA

EM : Electron microscopy

FBTSA : Fluorescence-based thermal shift assay

Gp12 : Gene product *12*

Hoc : Highly immunogenic outer capsid protein

HSV-1 : Herpes Simplex Virus-1

LF : Lethal factor

MCP : Major capsid protein

MTP : Major tail protein

ORF : Open reading frame

RT : Reverse transcribing virus

Scl1 and Scl2 : Streptococcus collagen-like proteins 1 and 2

SEC : Size exclusion chromatography

SFP : Scaffolding protein

Soc : Small outer capsid protein

ssDNA : single-stranded DNA

ssRNA : Single-stranded RNA

T : Triangulation number

TEV : Tobacco etch virus

TMV : Tobacco mosaic virus

Table of Contents

French Extended Abstract

Résumé.....	i
1 Introduction.....	iii
1.1 Les bactériophages caudés.....	vii
1.1.1 Classification des virus bactériens.....	vii
1.1.2 Structure de la capsid virale icosahédrique.....	viii
1.1.3 Assemblage des phages caudés.....	x
1.2 Le bactériophage SPP1.....	xiii
1.3 Les protéines collagène et "collagène-like".....	xiv
2 Problématique de la thèse et objectifs.....	xvii
3 Résultats et Discussion.....	xviii
3.1 Propriétés de gp12.....	xviii
3.2 Repliement et association de gp12.....	xix
3.3 Interaction de gp12 avec la capsid virale.....	xxi
3.4 Identification et organisation modulaire des protéines proches de gp12.....	xxiii
3.5 Distribution et diversité des protéines avec motifs collagène dans les procaryotes et dans les virus.....	xxvi
3.6 Conclusions et perspectives.....	xxviii
4 Références.....	xxix

PhD Thesis

Abstract	1
1 Introduction part I : What are viruses ?.....	3
1.1 The discovery of viruses, an historical account.....	3
1.2 Virus diversity.....	7

1.3	Viruses classification and taxonomy.....	10
1.4	Tailed bacteriophages classification and their relationship with herpesviruses...	13
1.5	Assembly of tailed bacteriophages virions.....	16
1.6	Capsid auxiliary proteins.....	19
1.7	The <i>Bacillus subtilis</i> phage SPP1.....	22
1.8	Gp12, the SPP1 capsid auxiliary protein: a collagen motif containing protein.....	26
2	Introduction part II : What is collagen?.....	27
2.1	The collagen family.....	27
2.2	Prokaryotic Collagen Motif Containing Proteins (CMCPs).....	30
2.3	Structural organization of collagen.....	33
3	Thesis Aims.....	36
4	Materials and Methods.....	37
4.1	Cloning procedures and creation of the gene 12 knock-out SPP1 strain	37
4.2	Production and purification of SPP1 virions and DNA-filled capsids	38
4.3	Production and purification of tag-gp12.....	39
4.4	Mass spectrometry	40
4.5	Digestion of tag-gp12 with collagenase	41
4.6	Analytical size exclusion chromatography (SEC)	41
4.7	Analytical ultracentrifugation	42
4.8	Circular dichroism (CD) measurements	42
4.9	Fluorescence-based thermal shift assay (FBTSA)	43
4.10	Gp12 chimerisation experiments	43
4.11	Binding of tag-gp12 to SPP1 12 virions	43
4.12	Binding of tag-gp12 to HΔ12 capsids analysed by a trypsin protection assay or FBTSA	44
4.13	Search for gp12 close relatives and sequence based analysis.....	44
4.14	Genomic context analysis of genes coding gp12 relatives	45
4.15	Troglodyte, the CMCPs detection tools	45
5	Results.....	48

5.1 Gp12 has a collagen-like sequence motif.....	48
5.2 Gp12 is an elongated trimer in solution.....	48
5.3 Gp12 has a collagen-like fold.....	49
5.4 Gp12 dissociates and unfolds reversibly at physiological temperature	51
5.5 Binding of gp12 to the capsid lattice is reversible and increases the trimer thermal stability of 20°C.....	54
5.6 Native and unfolded gp12 bind to SPP1 capsids in a distinct way	58
5.7 Identification and modular organization of gp12-related proteins	62
5.8 Distribution of CMCP proteins in prokaryotes and their viruses	66
6 General Discussion.....	68
7 Perspectives.....	75
8 References.....	77

French Extended Abstract

Résumé

Les virus sont d'une extrême diversité. Par contre, l'étude de leurs cycles infectieux, des structures de leurs protéines, et des voies d'assemblage de leurs particules virales ont mis en évidence des relations qui ont permis de tracer des liens évolutifs entre eux. L'assemblage d'un virion mature et infectieux est le but ultime de l'infection lytique. Les nouveaux virions qui sont libérés dans l'environnement vont, à leur tour, initier un nouveau cycle infectieux. L'assemblage des particules virales est une étape essentielle impliquant des acteurs d'origine virale et de la cellule hôte. Les éléments composant la particule virale sont assemblés suivant un programme bien défini d'interactions macromoléculaires séquentielles.

Gp12 est la protéine auxiliaire du bactériophage SPP1 qui infecte la bactérie Gram-positive *Bacillus subtilis*. Elle est exposée à la surface de la capsid virale se fixant spécifiquement au centre de chaque hexamère de gp13, la protéine majoritaire de la capsid. Seules les capsides qui ont encapsidé de l'ADN peuvent fixer gp12. Des phages mutants n'ayant pas gp12 sont viables et infectieux dans des conditions de laboratoire.

La séquence de gp12 est caractérisée par la succession de 8 motifs GXY. La répétition de ce motif est la signature des protéines de type collagène qui forment une triple hélice intramoléculaire. Le collagène eucaryote animal a été très étudié à cause de son abondance et son rôle dans l'organisme. La découverte de protéines de type collagène d'origine procaryote est relativement récente. Il a été démontré expérimentalement que des segments protéiques de type collagène d'origine procaryote ou synthétique sont capables de former une triple hélice stable. L'absence de modifications post-traductionnelles, retrouvées chez les eucaryotes, est compensée par d'autres mécanismes résultant dans une stabilité thermique semblable à celle observée chez le collagène eucaryote.

Dans ce travail de thèse, nous montrons que gp12 de SPP1 est une protéine virale stable de type collagène. La protéine isolée est un trimère allongé en solution.

Malgré la courte longueur de la répétition GXY, le profil de dichroïsme circulaire de gp12 montre la présence de la signature du motif de type collagène. Ce motif est coupé par la collagénase VII à un site spécifique à l'intérieur de la séquence (GXY)₈. La protéine peut être dénaturée-dissociée et puis renaturée-associée sous l'effet de la température. La fixation de gp12 à la capsidie augmente significativement sa stabilité thermique. Cependant, à des températures supérieures à 50°C, gp12 se dissocie de la capsidie, effet réversible lorsque la température est baissée. En fonction de la température, les protéines gp12 native (trimère) ou gp12 dénaturée (monomère) peuvent se fixer/refixer à la capsidie, mais avec des profils d'interaction différents.

Les propriétés de gp12 lui confèrent un potentiel pour la nano-ingénierie. Nous avons pu attacher à des capsides de SPP1 de la gp12 avec un long peptide fusionné à son terminal aminique. Des capsides de SPP1 sans gp12 peuvent donc être utilisées comme des plateformes pour fixer gp12 fusionnée à des protéines cibles. Cette propriété, associée à la multivalence de gp12 dans la capsidie et à son fort pouvoir immunogène, sont prometteurs pour conduire, par exemple, des essais de vaccination. Les propriétés thermiques de gp12 apportent une flexibilité additionnelle pour la fixation réversible de protéines à la surface de la capsidie de SPP1.

Une étude bio-informatique a permis d'identifier des protéines dont la séquence présente une similarité avec gp12. Elles ont une organisation modulaire commune avec un segment centrale portant la répétition (GXY)_n qui lie les régions amino- et carboxyl-terminales de la protéine. Ces modules ont une taille variable. Les protéines sont codées par des prophages de bactéries du genus *Bacillus*. Leur gène est localisé à proximité du gène codant la protéine majoritaire de la capsidie suggérant qu'elles sont aussi des protéines auxiliaires de capsidie.

Une recherche de protéines procaryotes et virales avec des segments collagène a montré qu'elles sont abondantes parmi les bactéries et virus. Le motif est rare parmi les archées et leurs virus. Ces résultats montrent l'importance des protéines avec des séquences collagène dans le monde non-eucaryote et de développer leur étude biochimique et fonctionnelle.

1 Introduction

Les virus n'ont été décrits que vers la fin du dix-neuvième siècle malgré leur implication en tant qu'agents infectieux responsables de plusieurs maladies et leur contribution au façonnement de l'histoire de l'évolution de la vie. A cette période, les micro-organismes responsables de maladies étaient décrits comme des germes retenus par filtration, cultivables dans des milieux nutritifs et visibles à l'aide du microscope optique¹.

La naissance de la virologie moderne est liée à la découverte du virus de la mosaïque du tabac²⁻⁴. En 1886, Adolf Mayer a publié une communication à propos de ses travaux sur la maladie de la mosaïque du tabac. Il a remarqué que la maladie peut être transmise des plantes malades aux plantes saines par inoculation d'un extrait des plantes malades. Étant incapable de cultiver l'agent infectieux, il a essayé de reproduire la maladie en utilisant les micro-organismes connues à cette époque-là. Aucun d'eux n'était capable d'infecter les plantes. Mayer a posé l'hypothèse que l'agent infectieux pourrait être soit un organisme inconnu soit un agent de type enzyme. Il a inclus, alors, une étape de filtration par un filtre en papier avant l'inoculation de plantes saines. Le jus filtré des plantes malades était suffisant pour reproduire la maladie par inoculation des plantes saines. Après plusieurs étapes de filtration, le filtrat clair était devenu stérile. Mayer a conclu que l'agent infectieux responsable de la maladie de mosaïque du tabac ne pouvait être qu'un nouveau type de bactérie.

Quelques années plus tard, Dimitri Ivanofsky a répété les expériences de Mayer mais en remplaçant le filtre en papier par un filtre Chamberland qui a des pores plus petits que la taille des bactéries. Ivanofsky a obtenu les mêmes résultats que Mayer. Par contre, il a conclu que l'agent responsable de la maladie est une toxine. Un avancement majeur a été apporté par Martinus Beijerinck. En 1898, il a observé que le filtrat dilué de jus des plantes malades, en utilisant un filtre de Chamberland, est capable de restaurer son infectivité après répllication dans des plantes vivantes. Cette expérience montre, pour la première fois, le caractère parasite obligatoire du pathogène

et explique pourquoi Mayer et Ivanofsky ont été incapables de le cultiver. Il l'a nommé virus. Le terme *virus* est dérivé du latin pour poison^{2,5}. Après plusieurs années de débat, la forme, la taille et la nature du pathogène ont été résolues par un cliché de microscopie électronique en 1939⁶. Le virus de la mosaïque du tabac était le premier virus découvert, et la virologie était née.

La découverte des virus qui infectent les bactéries remonte aux publications de Twort en 1915⁷ et d'Hérelle en 1917⁸. Des études antérieures avaient déjà fourni des pistes sur l'existence d'entités semblables aux virus avec des activités antibactériennes⁹. D'Hérelle les a nommés bactériophage, pour mangeur de bactérie. Il les a décrits également comme un agent thérapeutique capable d'éradiquer les bactéries. Ceci était l'un des thèmes de recherche les plus excitants pendant les années 1920. Le premier cliché de microscopie électronique montrant l'ultrastructure d'un bactériophage, également abrégé phage, a été publié en 1940¹⁰.

Au cours du dernier siècle, le nombre d'espèces virales isolées a augmenté rapidement. La co-évolution de la virologie et de la technologie moderne constituent un élément qui a poussé les connaissances sur ces agents infectieux^{2,11}. L'évolution des techniques biochimiques a permis la préparation de particules virales (virions) purifiées ouvrant la voie à leur étude en utilisant des approches biochimiques et biophysiques. Ces travaux ont abouti, notamment, à la découverte de la structure moléculaire des virions en utilisant la diffraction des rayons X ou la cryo-microscopie électronique. La RMN a permis de caractériser des composants individuels des virions. La morphologie de la particule virale est devenue un critère majeur dans l'identification des virus tandis que leurs structures atomiques ont permis d'établir des relations phylogénétiques, quelques fois inattendues, entre les virus infectant les trois Domaines du Vivant^{12,13}.

La découverte de virus géants a brisé la barrière de la taille entre le monde bactérien et virale. Par exemple, la capsidie icosaédrique des Mimivirus a un diamètre de 400nm protégeant un génome de 1,2Mb codant pour près de 911 gènes (Figure 1f). Ils sont plus volumineux que certaines petites bactéries comme *Mycoplasma genitalium* qui fait 300nm de long avec un génome de 580kb et codant pour moins de 500 gènes^{14,15}. Les virus ne sont pas, non plus, les seuls parasites obligatoires dans la nature.

Certaines bactéries, comme *Rickettsiae* et *Chlamydiae*, sont devenues tellement dépendantes de leur hôte que leur cycle extracellulaire ne peut durer qu'une courte période, sinon elles perdent leur viabilité^{16,17}. De nos jours, la seule propriété distinctive entre les virus et la cellule est l'absence complète de métabolisme et de ribosomes^{1,11}. Les virus ne grandissent pas et leur multiplication est strictement liée à l'infection de leurs hôtes^{1,11}.

Les virus qui infectent les bactéries (bactériophages ou phages) constituent une partie majeure de la Virosphère. Ils sont présents dans les écosystèmes où se retrouvent leurs hôtes comme le sol, les océans ou les organismes multicellulaires colonisés par les bactéries. Les infections phagiques jouent un rôle majeur dans la dynamique de toutes ces populations bactériennes.

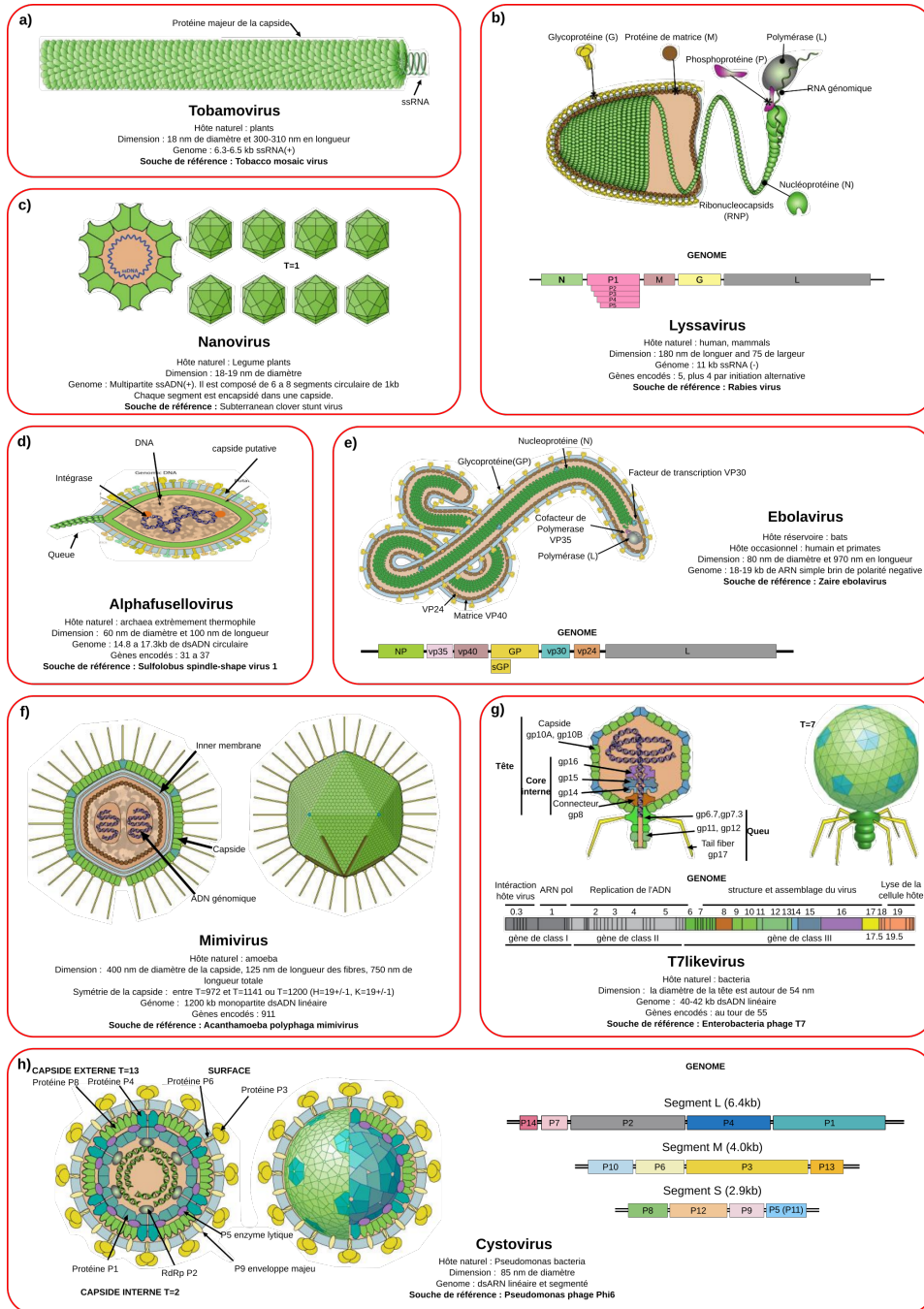


Figure 1: Morphologie, structure et organisation génomique de quelques virus. Les figures ont été adaptées d'ExpASY ViralZone (<http://viralzone.expasy.org/>). Chaque panneau décrit les propriétés d'un genre viral représenté par sa souche de référence. **a)** virus de la mosaïque du tabac ("TMV virus"). **b)** virus de la rage ("Rabies virus"). **c)** "Subterranean clover stunt virus". Le génome viral en ADN simple-brin ("ssDNA") est segmenté en 8 éléments. Chaque segment est encapsulé individuellement dans une capside virale. **d)** "Sulfolobus spindle-shape virus 1". **e)** "Zaire ebolavirus". **f)** "Acanthamoeba polyphaga Mimivirus". **g)** "Enterobacteria phage T7". **h)** "Pseudomonas phage Phi6". Le génome viral en ARN double-brin ("dsRNA") est segmenté en 3 éléments. Ceux-ci sont encapsulés dans la même capside virale.

1.1 Les bactériophages caudés

1.1.1 Classification des virus bactériens

De nos jours, plusieurs milliers de virus ont été décrits. Face à ce nombre important, leur classification est une nécessité majeure¹⁷. En effet, les virus sont très divers ce qui rend leur classification une des tâches les plus compliquées en virologie. Plusieurs systèmes de classification ont été décrits. Les plus anciens distinguent les virus en fonction de leur pathogénicité, de leur hôtes ou encore de leur voie de transmission^{17,18}. La classification de Baltimore est l'un des systèmes les plus connus où les virus sont classifiés en fonction de la nature de leurs génomes ainsi que de leur mode de réplication et transcription. De nouvelles approches basées sur la morphologie, la génétique, la biochimie, la structure et plus récemment la séquence complète des génomes viraux sont largement utilisées¹⁹⁻²¹. En pratique, chaque système a ses avantages et ses limitations. Actuellement, plusieurs systèmes sont combinés pour décrire les virus.

Les bactériophages sont les entités biologiques les plus abondantes sur Terre²². Parmi plus de 5500 particules virales de phages différents qui ont été examinées par microscopie électronique, près de 96% sont composés d'une capsidie icosaédrique et d'une queue²³. Ce sont des virus à ADN double brin (Baltimore classe 1). Un lien évolutif entre ces phages caudés a été démontré en se basant sur l'alignement de séquences de certaines protéines fonctionnelles et de leur structure^{12,24}. Ils constituent l'ordre Caudovirales^{23,25,26}. Cet ordre est divisée en trois familles: Podoviridae, Siphoviridae, et Myoviridae. L'élément distinctif entre ces familles est la structure de la queue. Les Podoviridae ont une queue courte non contractile, les Siphoviridae ont une queue longue non contractile et les Myoviridae ont une queue longue contractile.

1.1.2 Structure de la capsid virale icosaédrique

La géométrie icosaédrique de la capsid virale permet d'assembler des structures de différents diamètres pour protéger le génome viral en utilisant un seul type de sous-unité (Figure 2). Dans le cas le plus simple le matériel génétique est protégé par un conteneur formé d'une couche unique de copies multiples d'une seule protéine. D'autres virus sont plus complexes, pouvant contenir des lipides et de nombreuses protéines différentes. Les capsides dont toutes les sous-unités établissent des contacts équivalents sont composées par 60 copies de la protéine majeure de la capsid ("major capsid protein"; MCP) qui forment un icosaèdre simple avec 12 sommets pentamériques (Figure 1c et 2). Caspar et Klug ont prédit que des capsides avec plus de 60 sous-unités peuvent exister si elles forment des interactions quasi-équivalentes entre elles²⁷. L'organisation de ces structures icosaédriques est définie par un nombre de triangulation T qui peut être calculé pour chaque cas en utilisant l'équation: $T=h^2+k^2+hk$, dans laquelle T est le nombre de triangulation tandis que h et k ce sont des entiers positifs²⁷⁻²⁹ (Figures 2 et 3). T prends toujours des valeurs discrètes (1, 3, 4, 7, 9, 13, 16, 25, etc) et augmente avec le nombre de sous-unités de la MCP. Par exemple, pour $T=7$, la capsid est formée de 420 sous-unités de la MCP. Il y a, cependant, de nombreux virus qui ont un système portal pour l'entrée et la sortie du génome viral de la capsid. Celui-ci remplace un pentamère dans un des sommets de l'icosaèdre réduisant ainsi le nombre de MCPs de 5 sous-unités^{28,29}.




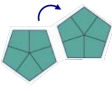



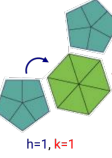
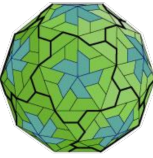


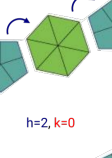
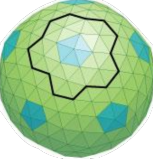


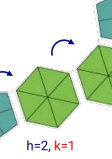
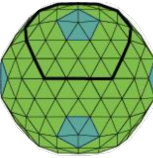


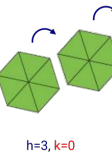
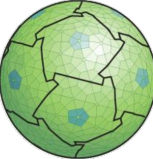


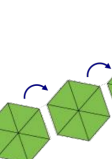
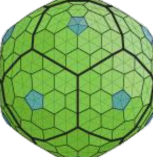


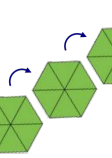
T	Nombre de MCPs	Structure de la capside icosaoédrique	Construction de la capside icosaoédrique par		Calcul de T selon les règles de Caspar et Klug $T=h^2 + k^2 + hk$	Exemple de virus
			12 sous unité	60 sous unité		
T = 1	60 12 pentamères				 h=1, k=0	ssDNA Circoviridae Parvoviridae Anelloviridae Geminiviridae Nanoviridae Microviridae dsRNA Partitiviridae Chrysoviridae ssRNA+ Hepeviridae Barnaviridae Ourmivirus
T = 3	180 12 pentamères 20 hexamères				 h=1, k=1	dsRNA Picobirnaviridae ssRNA+ Astroviridae Nodaviridae Caliciviridae Leviviridae Luteoviridae Bromoviridae (not all genera) Tombusviridae Tymoviridae Sobemovirus Polemavirus Umbravirus
T = 4	240 12 pentamères 30 hexamères				 h=2, k=0	ssDNA+ Tetraviridae Togaviridae dsDNA(RT) Hepadnaviridae
T = 7	420 12 pentamères 60 hexamères				 h=2, k=1	dsDNA Polyomaviridae Papillomaviridae Siphoviridae Podoviridae dsDNA(RT) Caulimoviridae
T = 9	540 12 pentamères 80 hexamères				 h=3, k=0	N4likevirus
T = 13	780 12 pentamères 120 hexamères				 h=3, k=1	dsRNA Bimaviridae Cystoviridae Reoviridae
T = 16	960 12 pentamères 150 hexamères				 h=4, k=0	dsDNA Herpesviridae SP01-like viruses

Figure 2: Géométrie des capsides virales en fonction de l'équivalence (T=1) et quasi-équivalence (T>1) des interactions entre les sous-unités de la MCP. Cette figure ne montre que les cas les plus fréquents. Des capsides avec un numéro de triangulation plus élevé existent. Les figures ont été adaptées de ViralZone.

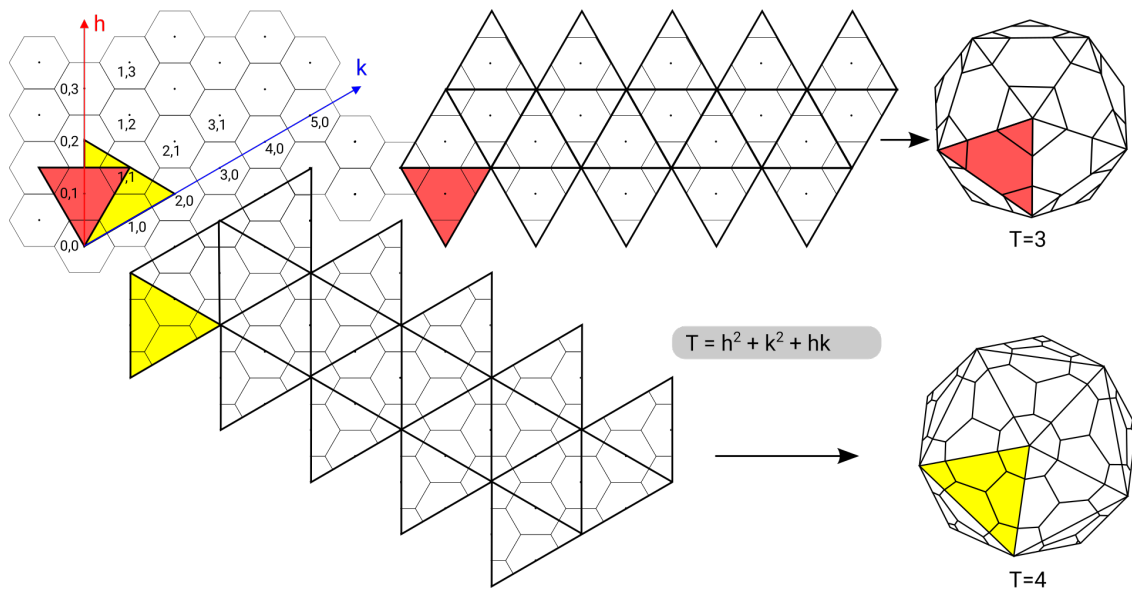


Figure 3: Construction d'un icosàdre de géométrie T=3 et T=4 en utilisant le même réseau plat d'hexagones selon les règles de Caspar et Klug. Pour calculer T on choisi un hexon arbitrairement. Ensuite on dessine les axes h et k qui se croisent au centre de l'hexon avec un angle de 60°. Le point d'intersection représente le point de référence (h=0, k=0). Puis, on dessine un triangle équilatéral en commençant par le point de référence jusqu'au point d'insertion du sommet de l'icosàdre le plus proche (axe de symétrie de 5). Les sommets du triangle doivent se localiser au centre des hexons. Enfin, les triangles sont assemblés pour former la capsidie icosahédrique. T est calculé utilisant l'équation $T=h^2+k^2+hk$. Les figures ont été adaptées de Baker et al. 1999²⁹.

1.1.3 Assemblage des phages caudés

L'assemblage de la capsidie chez les bactériophages caudés et les virus herpès suit un schéma global commun^{12,13,28,30-32} (Figure 4). Ses composants essentiels sont la MCP, la protéine d'échafaudage et la protéine portale. Des sous-unités de ces protéines co-assemblent pour former un icosàdre avec une morphologie sphérique, la procapsidie. Lors du démarrage de l'encapsidation du génome viral, les protéines d'échafaudage sont relâchées pour laisser la procapsidie vide. Puis, le complexe terminase lié à l'ADN phagique se fixe sur le système portal pour démarrer l'encapsidation du génome viral (Figure 4).

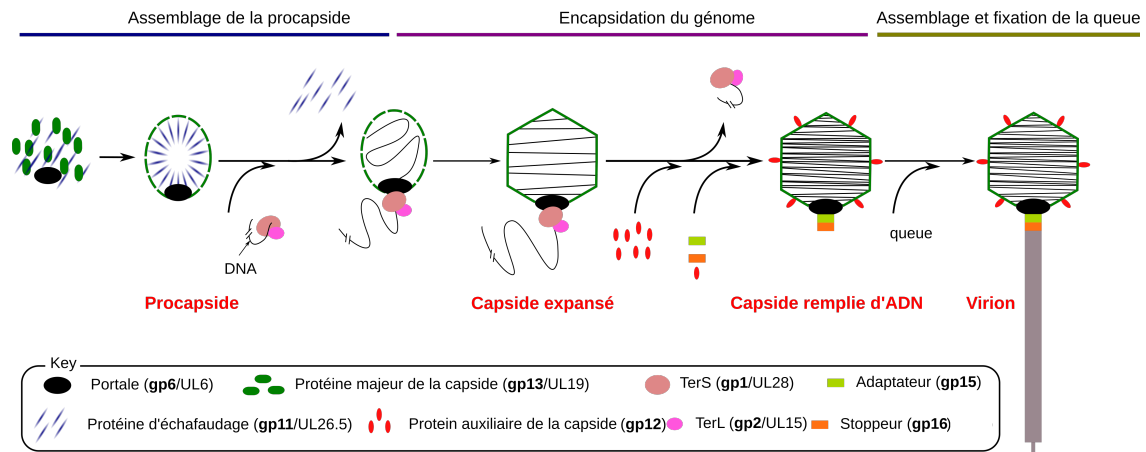


Figure 4: Représentation schématique de la voie d'assemblage des bactériophages à queue. Le bactériophage SPP1, objet de cette étude, en est un exemple. Les étapes d'assemblage de la capsid sont communes entre les bactériophages à queue et les virus herpès. Les protéines impliquées dans l'assemblage de la nucléocapsid du bactériophage SPP1 et de Herpes Simplex Virus 1 (HSV-1) sont présentées dans la légende. Les protéines de SPP1 commencent par gp (en gras) et les protéines du HSV-1 commencent par UL. TerS est la petite sous-unité de la terminase. TerL est la grande sous-unité de la terminase. La protéine gp1 de SPP1 s'assemble en nonamères alors que gp2 reste monomérique.

Au cours de l'assemblage, la capsid subit une expansion permettant l'augmentation de son volume interne, de sa stabilité et, dans de nombreux virus, aussi l'exposition de sites de fixation pour des protéines auxiliaires de la capsid³³⁻⁴². Après la fin de l'encapsidation de l'ADN, le système portal est fermé par la fixation d'autres protéines à la capsid virale. Cette stratégie d'assemblage est similaire pour la capsid des phages caudés qui infectent des bactéries et des virus herpès qui infectent des animaux (Figure 4). L'analyse de la structure des protéines effectrices du processus confirme que ces virus ont une origine évolutive commune^{12,13}. Dans le cas des virus à queue, le sommet portal de la tête mature sert comme point pour l'assemblage d'une queue courte (Podoviridae) ou pour fixation d'une queue longue formée dans une voie d'assemblage indépendante (Siphoviridae et Myoviridae). La queue est la structure qui permet l'interaction avec la surface de la bactérie et le transfert du génome phagique de la capsid vers le cytoplasme bactérien. En contraste, les capsides matures des virus herpès sont entourées par le tégument et par une enveloppe lipidique avec des glycoprotéines. Celles-ci permettent la fusion avec la membrane plasmique de la cellule eucaryote hôte conduisant à l'entrée de la capsid dans le cytoplasme. Par la suite, la capsid cible le pore nucléaire pour livrer l'ADN viral dans le noyau.

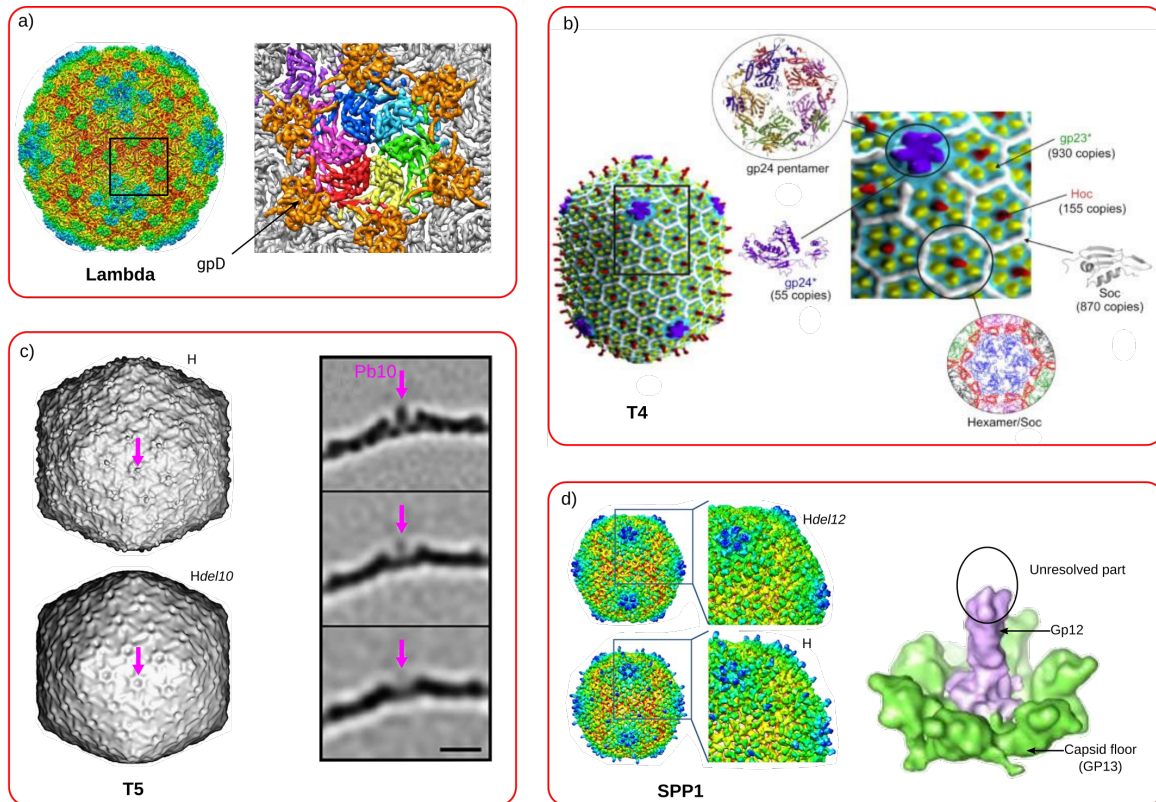


Figure 5: Localisation des protéines auxiliaires de la capside dans différentes capsides virales. Les différentes capsides virales ne sont pas présentées à l'échelle. **a)** La carte à résolution sous-nanométrique de la capside du bactériophage lambda (gauche) a été colorée radialement depuis le centre du phage (du rouge au bleu). La reconstruction montre la symétrie T=7 laevo de la capside et la protéine de décoration gpD qui émerge de la capside³⁶. **b)** Reconstruction de la capside du phage T4. Le carré correspond à une vue agrandie montrant gp23 (sous-unité jaune), gp24 (sous-unité mauve), Hoc (sous-unité rouge) et Soc (sous-unité blanche)³⁸. **c)** Le panneau à gauche montre une reconstruction tridimensionnelle de la capside du phage T5 (H) et la capside du phage T5 sans pb10 (*Hdel10*). Le panneau à droite montre une section d'un hexamère de la capside à différents temps de traitement avec du hydrochlorure de guanidinium. Ce traitement permet de décrocher pb10 de la capside (haut: 0h, centre: 1h, bas: 3h)⁴¹. **d)** Le panneau à gauche montre une reconstruction tridimensionnelle de la capside du phage SPP1 (H) et la capside de SPP1 sans la protéine gp12 (*Hdel12*). Le panneau à droite montre la localisation de gp12 au centre d'un hexamère de gp13. La différence entre deux capsomères est colorée en magenta. La surface verte montre l'organisation commune des hexamères de la capside³⁴.

Durant le processus de maturation, la capside subit des changements structuraux qui conduisent souvent à la fixation de protéines auxiliaires (Figures 4 et 5). Elles sont également appelées protéines de cimentation ou protéines de décoration, ceci dépendant de leurs fonctions. Chez certains virus comme le phage lambda, la protéine gpD est nécessaire pour la stabilité de la capside (fonction de cimentation)^{36,37}. Chez

d'autres virus elles ne sont nécessaires ni pour stabilité de la capsidie ni pour la viabilité des virions, comme les protéines Hoc et Soc du phage T4 ou pb10 de T5 (protéines de décoration)³⁸⁻⁴². Par contre, il a été démontré que les protéines de T4 jouent un rôle protecteur dans des conditions extrêmes comme la haute température ou un pH de 11³⁸⁻⁴⁰.

A cause de leur abondance et leur distribution régulière au tour de la capsidie, les protéines auxiliaires de la capsidie sont d'excellentes candidates pour la nano-ingénierie de particules virales.

1.2 Le bactériophage SPP1

Le bactériophage SPP1, sujet de cette étude (Figures 4 et 5d), est un siphovirus lytique qui infecte la bactérie Gram-positif *Bacillus subtilis*^{43,44}. SPP1 est composé d'une capsidie icosaédrique avec un nombre de triangulation $T=7$ et une queue longue non-contractile. La capsidie de 60nm de diamètre contient une molécule d'ADN double-brin de ~45.9kbp. Les facettes de la capsidie sont formées de hexamères de la protéine majoritaire de la capsidie, la protéine gp13. La reconstruction par cryo-microscopie électronique de la capsidie de SPP1 a montré que la protéine gp12 est présente au centre de chaque hexamère de gp13 sous forme d'un tube allongé^{34,45}. La capsidie et la queue sont liées par le connecteur formé par le système portal (dodécamère de la protéine gp6) qui occupe l'un des sommets pentamériques de la capsidie, un adaptateur (la protéine gp15) et un stoppeur (la protéine gp16) (Figure 4)⁴⁶. Le connecteur joue le rôle d'une porte à travers laquelle l'ADN est pompé vers l'intérieur de la capsidie et qui permet sa sortie lors de l'infection⁴⁶. La queue de SPP1 est construite par l'arrangement de gp17.1 et gp17.1* dans une structure hélicoïdale formant un tube flexible. Les deux protéines ont la même séquence d'acides aminés dans leur partie N-terminale mais gp17.1* a un terminal carboxylique plus long⁴⁷. L'extrémité de la queue distale de la capsidie se termine par une fibre qui permet la reconnaissance de l'hôte⁴⁷⁻⁵⁰.

La fonction exacte de gp12 n'est toujours pas connue. L'élément le plus caractéristique de gp12 est la présence dans sa partie centrale d'un motif (GXY)₈. Le

motif $(GXY)_n$ est la signature des protéines de type collagène, où G est une glycine, X et Y peuvent être n'importe quel acide aminé, et n est le nombre de répétitions successives du motif.

1.3 Les protéines collagène et "collagène-like"

Le collagène est la famille de protéines la plus abondante chez les vertébrés. Il est l'un des composants les plus importants de la matrice extracellulaire jouant un rôle majeur dans la stabilité des tissus. Il existe 28 types de collagène qui sont décrits par des numéros romains. En se basant sur leur organisation structurale et supramoléculaire, ces types sont classés en sous-familles comme présenté dans la Figure 6⁵¹⁻⁵⁵. Par exemple, les collagènes de type I et II sont caractérisés par leur assemblage en fibres. Le collagène de type I est le plus étudié et le plus abondant dans plusieurs tissus comme la peau et les ligaments, constituant également 90% de la masse organique des os⁵²⁻⁵⁵.

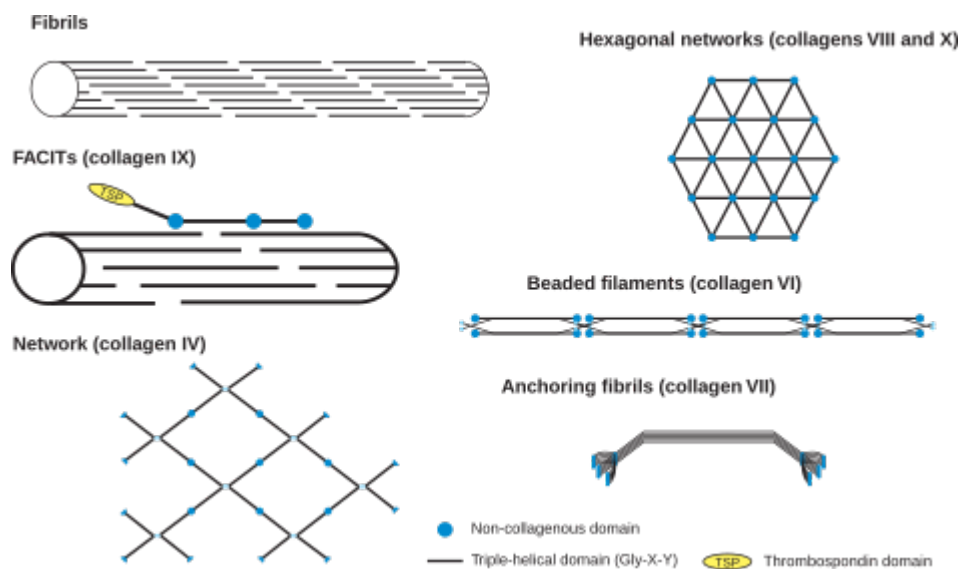


Figure 6 : Structures supramoléculaires formées par les différents types de collagène. Reproduit de Ricard-Blum. 2011⁵³.

Il a été démontré que des mutations dans la séquence du collagène type I ou des anomalies dans sa voie de synthèse peuvent conduire à un assemblage défectueux des fibres, ce qui altère les propriétés de la matrice des os et peut causer des maladies

comme l'ostéoporose^{56,57}. A côté de son rôle structural, le collagène a également des propriétés fonctionnelles. Par exemple le collagène IV joue un rôle dans la filtration^{52,53}.

Les molécules collagène des animaux alternent des segments caractérisés par la répétition (GXY)_n avec d'autres motifs ou domaines. Ces combinaisons confèrent des propriétés biochimiques et biophysiques nécessaires à leur fonction^{53,58}. Les segments (GXY)_n forment une triple hélice intramoléculaire résultant dans une structure allongée compacte. La position Y du motif GXY est occupée fréquemment par une proline qui est modifiée en hydroxyproline. Cette modification post-traductionnelle est essentielle pour la stabilité du collagène eucaryote⁵⁹.

Les critères pour définir le collagène ne sont pas clairement définis. Le terme collagène est utilisé pour définir des protéines qui ont une triple hélice et sont déposées dans la matrice extracellulaire. Une exception, ce sont des protéines transmembranaires qui ont une triple hélice extracellulaire. Il existe également des protéines animales ayant la triple hélice telles que MARCO ("macrophage receptor"), l'acétylcholinestérase, l'emiline, ou l'ectodysplasine mais qui ne sont pas considérées comme des collagènes⁵³. Elles sont décrites comme des « collagen-like ». Chez les vertébrés, ces protéines peuvent être sécrétées ou rester ancrées à la membrane. Elles ont des fonctions diverses dans l'environnement extracellulaire⁵³.

Chez les procaryotes plusieurs gènes codant pour des protéines de type collagène caractérisées par des répétitions (GXY)_n ont été décrits. Leur nombre a augmenté significativement grâce au séquençage à haut débit qui a fourni une grande quantité de données permettant la recherche de gènes codant pour ce type de protéines⁶⁰⁻⁶¹. L'absence dans ces protéines des modifications post-traductionnelles retrouvées dans les systèmes eucaryotes a soulevé la question des mécanismes de stabilisation de la triple-hélice intramoléculaire du collagène d'origine procaryote⁶². Les fonctions de ces protéines sont aussi généralement plus méconnues. Durant les 15 dernières années, plusieurs protéines de type collagène d'origine procaryote ont été étudiées. Les protéines Scl1 et Scl2 de *Streptococcus* sont parmi les mieux connues. Ces protéines de surface de la bactérie jouent un rôle dans l'interaction de *Streptococcus* pathogènes avec des cellules de mammifère et dans la formation de

biofilms^{63,64}. Elles sont stables malgré qu'elles soient dépourvues de modifications post-traductionnelles qui stabilisent le collagène eucaryote montrant que des mécanismes alternatifs permettent d'assembler des triple-hélices intramoléculaires avec des propriétés de collagène⁶². L'utilisation d'acides aminés chargés impliquant des interactions entre chaînes latérales est l'un des mécanismes les plus connus. Chez les virus, plusieurs gènes codant pour des protéines de type collagène ont été décrites^{60,61}. Par contre, leurs structures et leurs fonctions restent mal connues.

2 Problématique de la thèse et objectifs

L'objectif de cette thèse a été l'étude des propriétés biochimiques et structurales de la protéine gp12 du bactériophage SPP1. Nous avons d'abord étudié la structure quaternaire de gp12 et analysé expérimentalement si la protéine a une triple hélice collagène résultant de son motif (GXY)₈. Ensuite nous nous sommes intéressés à l'association de gp12 et, enfin, à l'interaction entre gp12 et la capsid virale de SPP1. Nos travaux ce sont conclus par une analyse bioinformatique de la distribution de protéines "collagen-like" parmi les procaryotes et les virus.

3 Résultats et Discussion

3.1 Propriétés de gp12

La chaîne polypeptidique de gp12 a une masse moléculaire de 6,6kDa. Sa région carboxyl-terminale est prédite en hélice alpha et sa région centrale a une signature de séquence d'un motif de type collagène ((GXY)₈) (Figure 7a). Nous avons produit la protéine gp12 couplée à une étiquette hexahistidine du côté amino-terminal (tag-gp12 ou his6-gp12) afin d'augmenter la production et faciliter la purification de la protéine. La forme et l'état d'oligomérisation de tag-gp12 ont été déterminés par ultracentrifugation analytique (Figure 7b,c). Tag-gp12 se comporte comme une espèce unique avec un coefficient de sédimentation de 1,7S, une masse moléculaire de 31,3kDa et un coefficient de friction de 1,79. Ainsi, tag-gp12 est un trimère allongé en solution.

Le trimère de tag-gp12 s'est avéré sensible à la collagénase VII qui coupe gp12 spécifiquement entre la Gln-19 et la Gly-20 (Figure 7a,d) mettant en évidence un repliement du type collagène. Le spectre de dichroïsme circulaire (CD) de tag-gp12 enregistré à 10°C montre deux minima (Figure 8a). Le premier minimum profond à 200nm est spécifique du repliement poly-proline II. Ce minimum est caractéristique des protéines du type collagène. Le second minimum correspond à l'hélice alpha prédite dans la région carboxyl-terminale de gp12 (Figure 7a). Ce profil montre que gp12 est formée d'un motif de type collagène occupant la région centrale de la protéine et des hélices alpha dans les extrémité C-terminales des sous-unités du trimère.

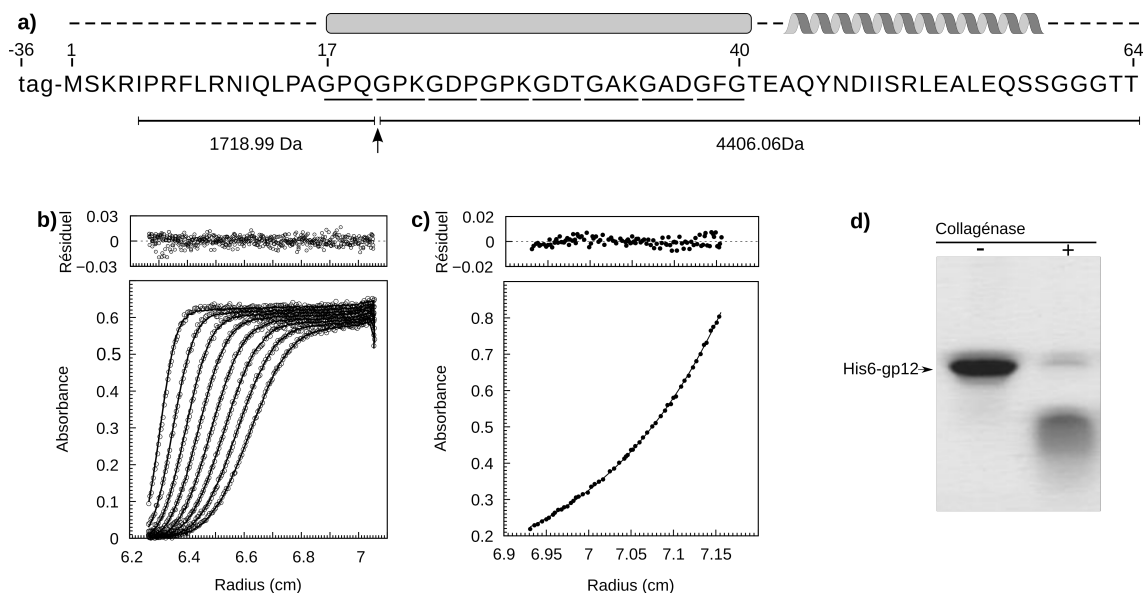


Figure 7: Propriétés de la protéine tag-gp12. **a)** Séquence de la protéine gp12 montrant la position prédite de l'hélice alpha, les triplets GXY (souligné) et le tag hexahistidine présent sur tag-gp12. La position de la coupure de la collagénase établie par spectrométrie de masse est montrée au-dessous de la séquence. **b)** Ultracentrifugation analytique par sédimentation à 16°C, à une vitesse de 220000 *g* et une concentration de 1 mg/mL tag-gp12. **c)** Ultracentrifugation analytique par équilibre à 16°C, à une vitesse de 16300 *g* et une concentration de 0,8 mg/mL tag-gp12. **d)** Digestion de tag-gp12 à la collagénase VII analysée par SDS-PAGE (droite). La protéine non-digérée est à gauche dans le gel.

3.2 Repliement et association de gp12

Nous avons enregistré plusieurs spectres de CD de tag-gp12 à différentes températures. On observe une perte de la structure de la protéine entre 30 et 40°C. Les spectres enregistrés entre 45 et 80°C sont caractéristiques d'une protéine dénaturée (Figure 8a). Afin de mieux étudier cette transition, nous avons enregistré un spectre CD à 200nm en fonction de la température pour suivre la dissociation et réassociation de la triple hélice de collagène (Figure 8b). On observe une seule transition avec un T_m de 41°C. Après dénaturation à 80°C puis refroidissement à 10°C, la protéine récupère la totalité de sa structure secondaire et quaternaire (Figure 8). Le profil de dénaturation et le profil de renaturation de la triple hélice intramoléculaire sont identiques. Ces résultats montrent que la dénaturation et dissociation de gp12 sont réversibles. Cependant on doit noter que la dissociation implique une molécule (trimère de gp12

associé) alors que l'association implique trois molécules (monomères de gp12). Ceci engendre deux cinétiques différentes. Le profil CD, seul, ne permet pas de les distinguer.

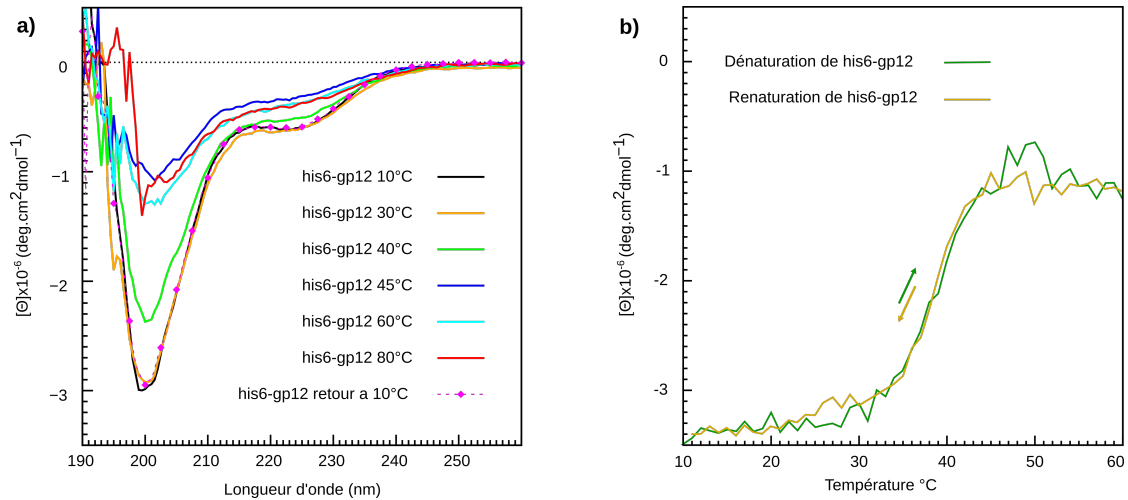


Figure 8: Réversibilité de la dénaturation et dissociation de tag-gp12. **a)** Spectre CD de tag-gp12 à différentes températures en utilisant le même échantillon. La protéine a été ensuite maintenue à 80°C pendant 30min puis refroidie à 10°C et un nouveau spectre a été enregistré (ligne rose pointillée). **b)** Dénaturation et renaissance de la protéine tag-gp12 enregistrées à 200nm en fonction de la température à une vitesse de 1°C par minute.

Afin de déterminer si les trois chaînes polypeptidiques du trimère se séparent physiquement lors de la dénaturation, nous avons conduit une expérience de chimérisation entre gp12 et tag-gp12. La protéine tag-gp12 se fixe fortement à la colonne d'affinité de nickel grâce à son étiquette hexahistidine. Tag-gp12 élué de la colonne en présence de 500mM d'imidazole. La protéine gp12 se fixe également à la colonne d'affinité mais son élution a lieu en présence de 100mM d'imidazole. Les deux protéines ont été mélangées puis soit incubées à 16°C, soit chauffées à 60°C, et finalement refroidies à 16°C. Dans le premier cas, les deux protéines éluent indépendamment l'une de l'autre. En contraste, lorsqu'elles ont été dénaturées à 60°C et réassociées ensemble, une fraction de gp12 co-élué avec tag-gp12⁴⁵. Ce résultat confirme la séparation physique des trois polypeptides du trimère lors de la dénaturation, condition nécessaire pour la formation de hétéro-trimères pendant l'étape de réassociation.

3.3 Interaction de gp12 avec la capsid virale

La protéine gp12 se fixe à la capsid de SPP1. Afin d'étudier cette interaction nous avons produit par génie génétique des capsides qui ont subi l'expansion et encapsidé de l'ADN viral mais qui n'ont pas la protéine gp12. Ces capsides fixent gp12 et tag-gp12 alors que les capsides sauvages qui contiennent gp12 ne lient aucune des deux protéines. Ces résultats montrent que la protéine gp12 se fixe uniquement et spécifiquement aux sites dédiés sur la capsid au centre de ses hexamères^{34,45}.

Nous avons par la suite étudié la stabilité de gp12 en utilisant la technique de "Fluorescence-based thermal shift assay" (FBTSA). Dans cette technique on enregistre l'augmentation de la fluorescence du fluorophore Sypro Orange qui se fixe aux régions hydrophobes exposées lors de la dénaturation de la protéine. Dans le cas de gp12, on constate un comportement inverse au principe de la technique. Le Sypro Orange se fixe à la protéine native et la fluorescence diminue en fonction de la température (Figure 9a).

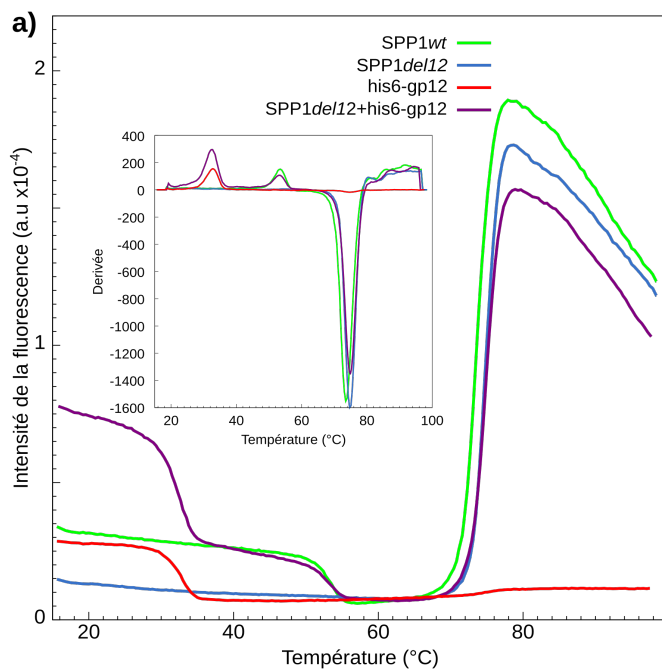


Figure 9: Stabilité de la protéine tag-gp12 isolée et liée à la capsid. **a)** Analyse par FBTSA de tag-gp12 (ligne rouge), SPP1del12 (ligne bleue), SPP1wt (ligne verte) et SPP1del12 en présence d'un excès de tag-gp12 (ligne mauve). L'insert montre l'opposé de la première dérivée des courbes de fluorescence pour déterminer les T_m . **b)** Etat de tag-gp12 et de la protéine majoritaire de la capsid en fonction de la température.

b)

	His6-gp12 isolée	his6-gp12 fixée à la capsid	Protéine majeure de la capsid
16°C	Repliée	Repliée	Repliée
45°C	Dénaturée	Repliée	Repliée
60°C	Dénaturée	Dénaturée	Repliée

Une seule transition est observable à 33,4°C pour la gp12 isolée. L'analyse par la même technique de particules virales sauvages, montre deux transitions. La première a lieu à 54°C qui présente la signature de gp12 (Figure 9a). La deuxième à 75°C correspond à la dénaturation de la protéine majoritaire de la capsid (Figure 9a)³⁴. Ce résultat montre que la capsid stabilise la protéine gp12 de plus de 20°C. Grâce à cette propriété, on a pu caractériser le comportement de la protéine en utilisant trois températures : 16°C, 45°C et 60°C (Figure 9b). A 16°C, la protéine gp12 isolée, la protéine gp12 fixée à la capsid et la protéine majoritaire de la capsid sont toutes repliées. À 45°C, la protéine gp12 couplée à la capsid et la protéine majoritaire de la capsid sont repliées alors que gp12 isolée est dénaturée. À 60°C, seule la protéine majoritaire de la capsid est stable. Ainsi, nous avons pu étudier l'interaction entre la capsid et gp12 native ou gp12 dénaturée.

Pour caractériser l'interaction gp12-capsid nous avons pris avantage du fait que la décoration de la capsid par gp12 modifie la charge de la surface de la capsid résultant dans une différence de migration sur gel d'agarose entre les capsides nues et les capsides décorées (Figure 10). A 16°C, les capsides sont soit entièrement décorées, soit non décorées. Nous n'avons observé aucun intermédiaire. L'interaction entre le trimère et la capsid est coopérative dans une expérience de dépendance de concentration de gp12 native se fixant sur des capsides sans gp12 (Figure 10a,b,d).

A 45°C, nous avons observé l'interaction entre gp12 dénaturée et la capsid sans gp12. Le profil sur gel d'agarose témoigne d'une interaction non-coopérative avec présence de capsides partiellement occupées par gp12 (Figure 10c,e). Ces résultats montrent également que la capsid joue un rôle de plateforme de nucléation pour l'assemblage du trimère de gp12 (Figure 10c,e)⁴⁵.

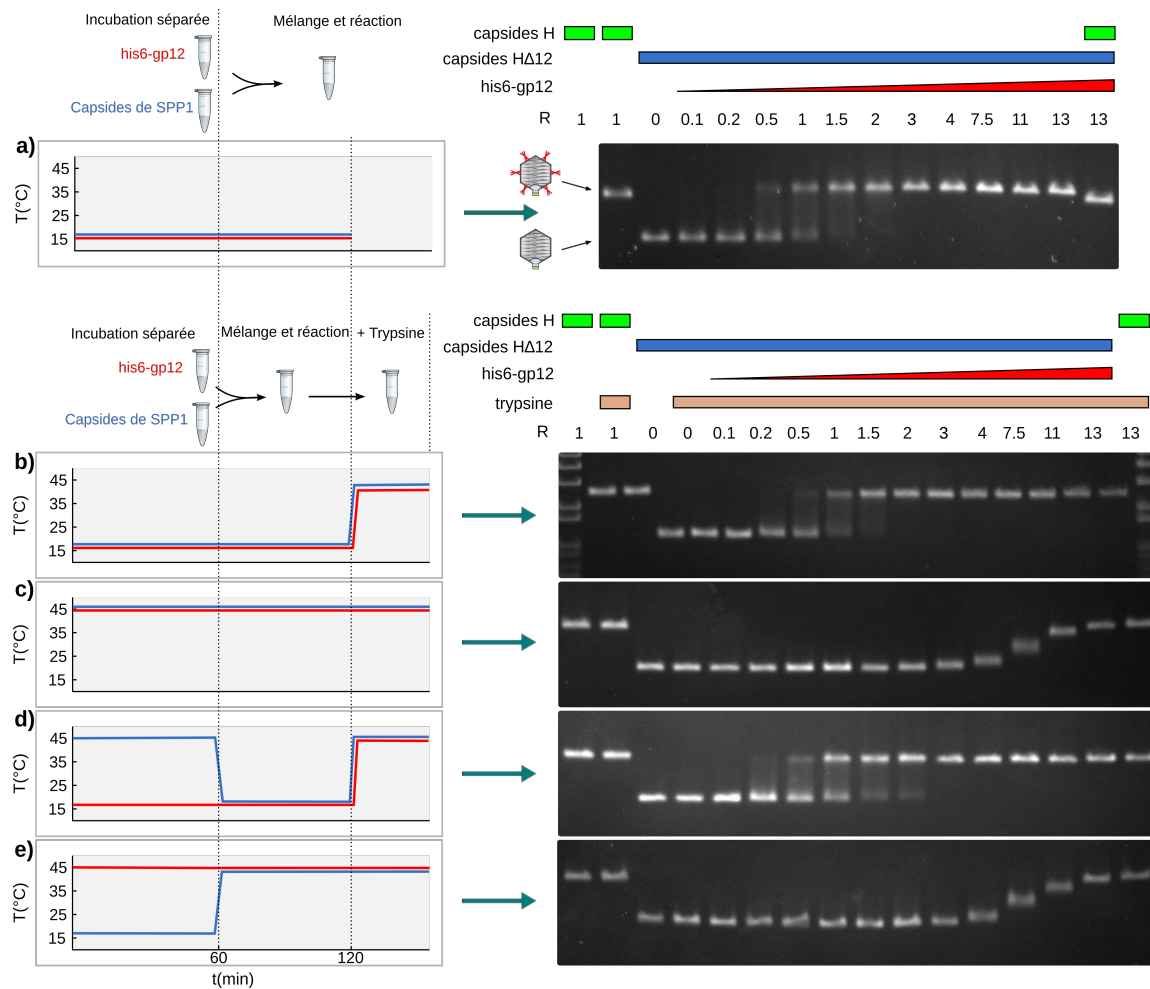


Figure 10: Fixation de tag-gp12 (rouge) native et dénaturée à la capside H Δ 12 (bleu), dépourvue de gp12. Les capsides et la protéine ont été incubées séparément à différentes températures, puis mixées (avec ou sans changement de température, panneaux de gauche) et puis traitées à la trypsine à 45°C. La trypsine dégrade la gp12 libre mais pas celle liée à la capside. Ce traitement permet donc d'empêcher la liaison de gp12 libre aux capsides lors de leur dépôt sur gel d'agarose. Les capsides sont ensuite analysées par électrophorèse sur gel d'agarose (panneaux de droite).

3.4 Identification et organisation modulaire des protéines proches de gp12

Nos travaux montrent que gp12 est une protéine virale de type collagène avec un motif (GXY)₈ qui occupe la partie centrale de la protéine. Une analyse avec l'outil pBLAST a permis d'identifier 5 protéines nommées B, C, D, E et F (A étant gp12) qui

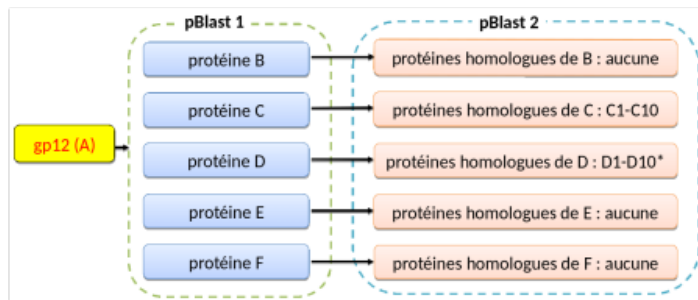
présentent une similarité de séquence avec gp12 dans les banques de données en 2013 (Figure 11a,b). Par la suite, nous avons réalisé un deuxième pBLAST pour chacune de ces protéines proches de gp12 permettant de faire apparaître 10 nouvelles protéines qui sont homologues des protéines nommées C et D identifiées dans le pBlast 1 (Figure 11a,c). Elles ont une relation plus lointaine avec gp12 (Figure 11).

Toutes ces protéines présentent la même organisation modulaire avec un domaine N-terminal, un domaine central avec un nombre variable de répétitions du motif GXY et un domaine C-terminal. L'alignement de leurs séquences fait apparaître deux sous-groupes. Le premier groupe est lié à gp12 avec une extrémité C-terminale conservée et une extrémité N-terminale variable (Figure 11b). Le deuxième groupe, homologue des protéines C et D, a une extrémité N-terminale conservée tandis que le nombre de répétitions GXY ainsi que la séquence distale de l'extrémité C-terminale sont variables (Figure 11c).

Gp12 est une protéine virale de type collagène qui se fixe à la capsid virale. Chez SPP1, le gène codant pour la protéine gp12 est localisé juste avant le gène 13 qui code pour la protéine majoritaire de la capsid dans un groupe de gènes codant pour les protéines de la capsid virale. Cette organisation génomique est conservée chez les bactériophages caudés. Les gènes codant pour des protéines qui interagissent entre elles lors du cycle virale sont normalement regroupées dans des unités transcriptionnelles où l'ordre des gènes est conservée⁶⁵.

Nous avons étudié par la suite le contexte génomique des gènes codant les protéines proches de gp12 précédemment identifiées. Dans tous les cas nous avons observé qu'ils se trouvent dans des régions codant pour des protéines structurales de bactériophages caudés. La protéine de type collagène est codée par un gène localisé à côté de celui de la protéine majoritaire de la capsid. Ces gènes sont retrouvés sur des prophages dont le mieux connu est celui du phage tempéré phi105 de *B. subtilis* (protéine C8 dans la Figure 11c). L'ensemble des données suggère que les protéines que nous avons identifiées ce sont des protéines auxiliaires de la capsid de ces prophages.

a)



b)

```

      ..... 10 ..... 20 ..... 30 ..... 40 ..... 50 ..... 60 ..... 70
A) NP_690673.1 1 .....-MSKR
B) YP_003919540.1 1 -----MAITDDQKR RLNESMPVVA DLKLGDI IQE
C) BAI85394.1 1 MEKEFLNKG NVWTASEMDT DGKPTTRVYL GGNSEENPLF -IKGMQGEQG PAGPQGPKG PGEQGPKGD
D) ZP_10509484.1 1 MAEEFLNKG NVWTASEMDT DGKPTTRVYL GGNSEENPLF -IKGMQGEQG PAG---FKGD PGEQGPKGD
E) NP_244394.1 1 MAEVLQTFDG K--SVSDKRP LPVKIIGDVG GGGSSMRPIT GTSDPSTNDG QPDVPLNTS TGDIFSNVNG
F) ZP_10863162.1 1 ---MYTTKNY KEPGRRWVI EGELALVGDG RITKDGEEVF FVGLPSDRIT AGQSAYEIAV KHGFEGTEEE

      ..... 80 ..... 90 ..... 100 ..... 110 ..... 120 ..... 130
A) NP_690673.1 5 IPRFLRNILQ PAGPQGPKG DGKPTTRVYL GGNSEENPLF YNDIISRLEA LE-QSSGGGT T--
B) YP_003919540.1 30 LQEG--GGTAG PKGDKDGTGP QGPKGDTGPK GADGFGTKAQ YDDIARLKA LE--GAGS--
C) BAI85394.1 70 GPQGPQKGTG AQPQGEAGP QGPKGEGKDP GHDGFGTEEQ YNELVSRLEA LE-QAAQAK--
D) ZP_10509484.1 67 GPQGPQKGTG AQPQGEAGP QGPKGDKGDS KDGFGTEEQ YNELVSRLEA LE-KAAQAK--
E) NP_244394.1 69 TWQKQGNLRG IQGPEGPQP EGPRGPEGPE GPPGFGTEEQ YNEIISRLEA LEGEVFGGD S--
F) ZP_10863162.1 68 WLTSLIGSTG EAGVQGEKGE TGAKGSAGAK GDFGFTKEE WDKLVARVEE LEGANGSKSA EQT
  
```

c)

```

      ..... 10 ..... 20 ..... 30 ..... 40 ..... 50 ..... 60 ..... 70
C) BAI85394.1 1 ME--KEFLNK SGNVWTASEM DTDGKPTTRV YLGNSEENP LFIKGMQGEQ GPAG-----
C1) ZP_10510839.1 1 MA--EDYLYE SGGVKTSEK GADGKAITPV YLKNSEENP VYVKGKLGDP GPQGPQGG--E
C2) EID48456.1 1 MA--DQFLNQ SNGVYTS AED DGTGKPVTPV YLKNSEENP LYIKGMQGEQ GPQGP-----
C3) YP_080705.1 1 MA--DQFLNQ SNGVYTS AED DGTGKPVTPV YLKNSEENP LYIKGMQGEQ GPQGP-----
C4) ZP_08001397.1 1 MA--DQFLNQ SNGVYTS AED DGTGKPVTPV YLKNSEENP LYIKGMQGEQ GPQGP-----
C5) ZP_06873031.1 1 MA--EDYLYE SGGVKTSEK GADGKAITPV YLKNSEENP VYVKGKLGEP GPQG-----E
C6) YP_003866273.1 1 MA--EDYLYE SGGVKTSEK GADGKAITPV YLKNSEENP VYVKGKLGEP GPQG-----E
C7) ZP_08000033.1 1 MA--EQFLNE SNGVYTS AED DGTGKPVTPV YLKNSEENP LYIKGMQGEQ G-----
C8) ADF59141.1 1 MA--EDYLYE SNGVKTSEK GADGKAITPV YLKNSEENP LEVKGKLGKE GEKDKGDTG KQGPQGEPE
C9) ZP_03055863.1 1 MA--KDYLF E SNGVLTSAE GADGKPTTPV YLKNSEENP LFIKGMQGEK GPKGDTG--
C10) YP_005421487.1 1 MAKLNILNE SNGVLTSAE NGKGTPTTDI SVADNSEENP LYVKGKLGDP GEQGPKGDG

      ..... 80 ..... 90 ..... 100 ..... 110 ..... 120 ..... 130 ..... 140
C) BAI85394.1 52 ---PQGPK GDFGEGPKG DTGP----- Q GPQKGTGAQG PQ-----G EAGPQGPKE
C1) ZP_10510839.1 57 TGPQGPQGEK GETGPQGPKG DKGDTEGQP QGEAGPQGPK GEKGDPAVIA DGSITHEMLL EKSVRSNKIG
C2) EID48456.1 53 ---QGPKGDK GDTGPQGPQG EPGPQGP--- K GDKGDPADIG EKSITHEMLL DKSVRSNKIG
C3) YP_080705.1 53 ---QGPKGDK GDTGPQGPQG EPGPQGP--- K GDKGDPADIG EKSITHEMLL EKVVRSNKIG
C4) ZP_08001397.1 53 ---QGPKGDK GDTGPQGPQG EPGPQGP--- K GDKGDPADIG DGSITHEMLL EKVVRSNKIG
C5) ZP_06873031.1 54 PGPQGEPEPQ GEPGPGQ--- EPGPQGEPEP QGEPGQGEPE GPKGDPAVIE EGSITHEMLL DKSVRSNKIG
C6) YP_003866273.1 54 PGPQGEPEPQ GEPGPGQ--- EPGPQGEPEP QGEPGQ--- -PKGDPAVIE EGSITHEMLL DKSVRSNKIG
C7) ZP_08000033.1 49 ---PKGDK GDKGEPGPKG D----- K GDKGDPADIG EKSITHEMLL DNIVRSNKIG
C8) ADF59141.1 69 PGPQGPQGEK GEPGEGPQG EPGPAGPKGD TGEQGPQGEK GDKGDPADIG ESSITYEMLA EKSVRSNKIG
C9) ZP_03055863.1 55 ---PQGPQGEK GDFGEGPQG EPGPAGPKGD TGEQGP--- K GDKGDAVIE SSVKNEHLS DKSVRSNKIG
C10) YP_005421487.1 60 ---AQGPQGEK GDFGEGPQG EPGPAGPKGD TGEQGP--- K GEKGDPAVIG EGSITHEMLL DKSVRSNKIG

      ..... 150 ..... 160 ..... 170 ..... 180
C) BAI85394.1 96 KGDGPKDGFQ ---TEEQYNEL VKRIEALQEA AQAK-----
C1) ZP_10510839.1 127 TGSVMDHLN AEVKAVFDKL QNQIDELKNE VETLKGTDDEA PQE---
C2) EID48456.1 109 TGSVMDHLN SEVKAVFEGE LKQIDELKGG ASS-----
C3) YP_080705.1 109 TGSVMDHLN SEVKAVFEDL LKQIDELKGG TSS-----
C4) ZP_08001397.1 106 TGSVMDHLN SEVKAVFDSL QNQIDELREK VAGSDSANN EPQE---
C5) ZP_06873031.1 121 TGSVMDHLN SDVKT VFNQL QNQIDELKNE VQTLKGTDDEA PQE---
C6) YP_003866273.1 115 TGSVMDHLN SDVKT VFNQL QNQIDELKNE VQTLKGTDDEA PQE---
C7) ZP_08000033.1 97 TGSVMDHLN SEVKAVFDDL QNQIDELKGS QASS-----
C8) ADF59141.1 139 TGSVMDHLN SEITKVLDEL KQKMNLESD LAALKGTEEE PTE---
C9) ZP_03055863.1 103 TGSVMDHLN SAVKDLIEL QNKVEALENP KSE-----
C10) YP_005421487.1 126 TGSVMDHLN SEVKAVLDGM QNQIDELKST TPAE-----
  
```

■ Identique
■ Similaire

Figure 11 : Recherche par pBLAST de protéines proches de gp12. **a)** Protocole opératoire de recherche bioinformatique. **b)** Alignement de séquences de protéines identifiés par similarité avec gp12 (pBlast 1). **c)** Alignement de séquences des protéines identifiées par similarité avec la protéine C (BAI85394.1) (pBlast 2). *-les protéines C1-C10 (panneau c) sont identiques à D1-D10 due à la similarité des protéines C et D.

3.5 Distribution et diversité des protéines avec motifs collagène dans les procaryotes et dans les virus

La présence de motifs de type collagène a été retrouvée dans les protéines virales gp12 et de ces protéines proches (Figure 11). Nous nous sommes interrogés sur la présence de ce type de protéines ("Collagen motif containing proteins"; CMCPs) dans l'ensemble des organismes procaryotes et des virus dont la séquence est connue. Ainsi nous avons analysé l'intégralité de protéines procaryotes et virales disponibles dans la base de données RefSeq (NCBI Reference Sequence). RefSeq est caractérisée par sa non redondance, une annotation riche et une mise à jour régulière par l'équipe de NCBI et leurs collaborateurs. Par conséquent, elle constitue une source de choix pour l'extraction de données et l'analyse protéomique comparative.

Vu le volume considérable des données, le traitement manuel n'a pas été possible. Ainsi, nous avons développé un ensemble d'outils informatiques, appelé Troglodyte, afin d'automatiser l'extraction et traitement des données. D'abord, les données génomiques de bactérie, archaea et virus ont été téléchargées depuis le serveur FTP de NCBI (<ftp://ftp.ncbi.nlm.nih.gov/refseq/release/>) à la date du 15 juillet 2014. Les séquences des protéines codées par les génomes ont été extraites et insérées dans une base de données afin faciliter l'accès et l'analyse des données. En se basant sur les données taxonomiques, nous avons généré 5 groupes de données : les bactéries, les virus bactériens, les archaea, les virus d'archaea et les virus (Figure 12a). Le groupe de virus contient tous les virus y compris les virus eucaryotes. Pour chaque groupe, nous avons généré deux ensembles de données : des données natives et des données normalisées. La normalisation des données permet d'éliminer des données redondantes en se basant sur l'alignement des séquences protéiques via le logiciel CD-HIT. Nous avons fixé le seuil d'identité $\leq 80\%$ pour retenir une séquence protéique. Afin d'éviter l'élimination de petites protéines, nous avons imposé une couverture des séquences d'une longueur de 80 % comme critère supplémentaire lors de leur alignement sur CD-HIT.

Pour identifier les CMCPs, le logiciel scanne la séquence de chaque protéine à

la recherche d'un nombre minimum de répétitions GXY. Nous avons réalisé 100 cycles où on incrémente, à chaque cycle, le nombre minimum de GXY de 1 (Figure 12b). Par exemple, une protéine ayant 20 répétitions successives du motif GXY sera positive pour un $(GXY)_{\min}$ allant de 1 à 20 et sera négative au-delà. A chaque cycle, le nombre de protéines positives sont comptées puis attribuées à leur groupe taxonomique.

Nos données montrent la présence de CMCPs dans tous les groupes de taxonomiques étudiées. Leur fréquence diminue rapidement étant inférieure à 0.6 % quand n_{\min} atteint 5. Après cette chute, la fréquence se stabilise avec une diminution plus ou moins prononcée en fonction du groupe taxonomique (Figure 12b). Les virus ont le taux le plus élevé de CMCPs (0,2 % du nombre total de protéines pour $n_{\min}=10$). Presque la moitié de ces protéines appartient à des virus bactériens. La majorité de ces protéines phagiques ont des répétitions GXY courtes et des motifs avec plus de 65 GSY consécutifs sont très rares, au contraire de ce qui est observé pour les bactéries. Chez les archaea et leurs virus, les CMCPs sont beaucoup plus rares et les répétitions GXY particulièrement courtes (Figure 12b).

a)

	Protéines totales				
	Natives		Normalisées		Gly %
	Protéines	Génomes	Protéines	Génomes	
Archaea	855 021	352	429 687	352	7.67
Virus d'archaea	3 923	64	3 086	64	6.85
Bactéries	49 809 020	14 927	10 015 887	14 698	7.54
Virus bactériens	125 889	1 301	74 367	1 295	7.07
Virus	187 375	3 959	121 614	3 732	6.26

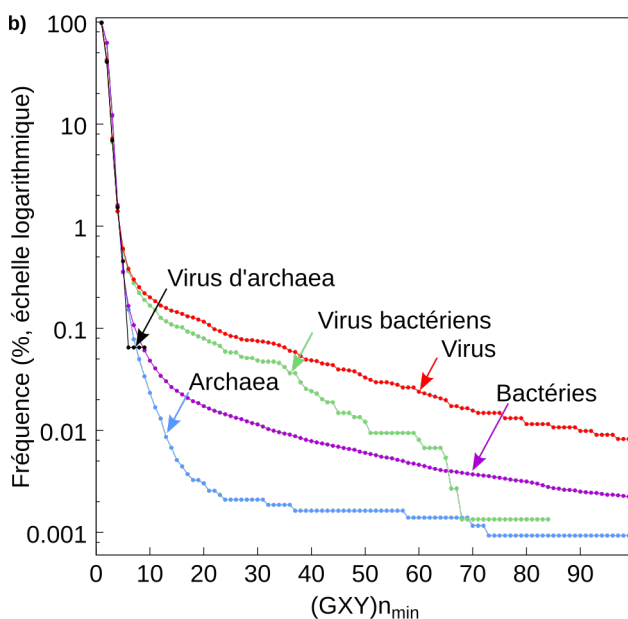


Figure 12 : Des protéines contenant des motifs de type collagène sont présentes dans tous les procaryotes et dans les virus. **a)** Nombre total de protéines et de génomes dans chaque groupe taxonomique présentes dans RefSeq (natives) et après élimination de séquences redondantes avec CD-HIT (normalisées). Le groupe virus inclut les virus des procaryotes et des eucaryotes. La fréquence de glycine est présentée pour les données normalisées. **b)** Fréquences des protéines de type collagène dans les différents groupes taxonomiques en fonction du nombre de $(GXY)_{\min}$. Chaque valeur de n_{\min} est le résultat d'un cycle d'analyse indépendant de chaque banque de données normalisée (voir texte pour détails). Les fréquences sont présentées sous forme de pourcentage.

3.6 Conclusions et perspectives

Nos travaux permettent de conclure que gp12 est une protéine virale qui présente un motif de type collagène court avec 8 triplets GXY. La protéine est stable sans aucune modification post-traductionnelle détectée. La stabilité de gp12 est fortement augmentée lors de son interaction avec la capsid. La fixation de gp12 modifie également les propriétés de surface de la capsid. Celle-ci joue un rôle de plateforme de nucléation du trimère lorsque gp12 est dénaturée. En modifiant la température on peut séparer gp12 de la capsid ou la refixer. Cette propriété peut être exploitée pour l'ingénierie de la nanoparticule virale sous forme d'un verrou thermique. On peut également suggérer l'utilisation de la surface de la capsid de SPP1 comme plateforme d'exposition de ligand ou antigène par couplage à la protéine gp12.

L'analyse des séquences et du contexte génomique des protéines proches de gp12 montrent qu'il s'agit très probablement de protéines auxiliaires de la capsid virale de prophages. Des cas similaires à gp12 de SPP1 existent donc dans la nature démontrant l'intérêt de nos études pour la compréhension du comportement moléculaire et de la fonction de ce type de protéines.

L'analyse à large échelle de banques de données de séquences montrent qu'il existe un nombre important de protéines avec des motifs collagène dans les procaryotes et dans les virus. Leur présence et conservation pendant l'évolution prouvent l'importance du repliement de type collagène pour le Vivant. Cependant, les propriétés et fonctions des CMCPs restent beaucoup moins étudiées que celles du collagène eucaryote.

4 Références

1. Dimmock, N., Easton, A. & Leppard, K. *Introduction to Modern Virology*. (Wiley, 2007).
2. Levine, A.J. The origins of Virology. *Fields Virology*, 4th ed., vol. 1. (Lippincott Williams & Wilkins, 2001).
3. Kung, S. & Yang, S.-F. *Discoveries in Plant Biology*. (World Scientific, 1998).
4. Creager, A. N. H., Scholthof, K.-B. G., Citovsky, V. & Scholthof, H. B. Tobacco Mosaic Virus: Pioneering Research for a Century. *Plant Cell Online* **11**, 301–308 (1999).
5. Douglas, Harper. Virus. *The Online Etymology Dictionary* Available at: http://www.etymonline.com/index.php?term=virus&allowed_in_frame=0.
6. Kausche, G. A., Pfankuch, E. & Ruska, H. Die Sichtbarmachung von pflanzlichem Virus im Übermikroskop. *Naturwissenschaften* **27**, 292–299 (1939).
7. Twort, F. W. An investigation on the nature of ultra-microscopic viruses. *The Lancet* **186**, 1241–1243 (1915).
8. D’Herelle F. Sur un microbe invisible antagoniste des bacillus dysentériques. *Acad Sci. Paris* **165**, 373–375 (1917).
9. Abedon, S. T., Thomas-Abedon, C., Thomas, A. & Mazure, H. Bacteriophage prehistory. *Bacteriophage* **1**, 174–178 (2011).
10. Ackermann, H.-W. Ruska H. Visualization of bacteriophage lysis in the hypermicroscope. *Naturwissenschaften* **1940**; **28**:45–6. *Bacteriophage* **1**, 183–185 (2011).
11. Cann, A. *Principles of Molecular Virology*. (Academic Press, 2012).
12. Bamford, D. H., Grimes, J. M. & Stuart, D. I. What does structure tell us about virus evolution? *Curr. Opin. Struct. Biol.* **15**, 655–663 (2005).

13. 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).
14. Svenstrup, H. F., Fedder, J., Abraham-Peskir, J., Birkelund, S. & Christiansen, G. Mycoplasma genitalium attaches to human spermatozoa. *Hum. Reprod.* **18**, 2103–2109 (2003).
15. Taylor-Robinson, D. & Jensen, J. S. Mycoplasma genitalium: from Chrysalis to Multicolored Butterfly. *Clin. Microbiol. Rev.* **24**, 498–514 (2011).
16. Becker, Y. Chlamydia. in *Medical Microbiology* (ed. Baron, S.) (University of Texas Medical Branch at Galveston, 1996).
17. Murphy, F. A. Virus Taxonomy and Nomenclature. in *Laboratory Diagnosis of Infectious Diseases Principles and Practice* 153–176 (Springer New York, 1988).
18. Gibbs, A. J. Viral taxonomy needs a spring clean; its exploration era is over. *Virology* **10**, 254 (2013).
19. Yu, C. *et al.* Real time classification of viruses in 12 dimensions. *PloS One* **8**, e64328 (2013).
20. Muhire, B. M., Varsani, A. & Martin, D. P. SDT: A Virus Classification Tool Based on Pairwise Sequence Alignment and Identity Calculation. *PLoS One* **9**, e108277 (2014).
21. Simmonds, P. Methods for virus classification and the challenge of incorporating metagenomic sequence data. *J. Gen. Virol.* **96**, 1193–1206 (2015).
22. Van Twest, R. & Kropinski, A. M. Bacteriophage enrichment from water and soil. *Methods Mol. Biol. Clifton NJ* **501**, 15–21 (2009).
23. Ackermann, H.-W. 5500 Phages examined in the electron microscope. *Arch. Virol.* **152**, 227–243 (2007).
24. Ackermann, H. W. Tailed bacteriophages: the order caudovirales. *Adv. Virus Res.* **51**, 135–201 (1998).

25. Krupovic, M., Prangishvili, D., Hendrix, R. W. & Bamford, D. H. Genomics of Bacterial and Archaeal Viruses: Dynamics within the Prokaryotic Virosphere. *Microbiol. Mol. Biol. Rev.* **75**, 610–635 (2011).
26. Ackermann, H.-W. Phage classification and characterization. *Methods Mol. Biol.* **501**, 127–140 (2009).
27. Caspar, D. L. & Klug, A. Physical principles in the construction of regular viruses. *Cold Spring Harb. Symp. Quant. Biol.* **27**, 1–24 (1962).
28. Prasad, B. V. V. & Schmid, M. F. Principles of Virus Structural Organization. *Adv. Exp. Med. Biol.* **726**, 17–47 (2012).
29. Baker, T. S., Olson, N. H. & Fuller, S. D. Adding the third dimension to virus life cycles: three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs. *Microbiol. Mol. Biol. Rev.* **63**, 862–922 (1999).
30. Rixon, F. J. & Schmid, M. F. Structural similarities in DNA packaging and delivery apparatuses in Herpesvirus and dsDNA bacteriophages. *Curr. Opin. Virol.* **5**, 105–110 (2014).
31. Baker, M. L., Jiang, W., Rixon, F. J. & Chiu, W. Common Ancestry of Herpesviruses and Tailed DNA Bacteriophages. *J. Virol.* **79**, 14967–14970 (2005).
32. Dai, X. & Zhou, Z.H. Structure of the herpes simplex virus 1 capsid with associated tegument protein complexes. *Science* **360**, pii: eaa07298 (2018).
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. White, H. E. *et al.* Capsid structure and its stability at the late stages of bacteriophage SPP1 assembly. *J. Virol.* **86**, 6768–6777 (2012).
35. Newcomer RL *et al.* The phage L capsid decoration protein has a novel OB-fold and an unusual capsid binding strategy. *Elife* pii: e45345 (2019).

36. Lander, G. C. *et al.* Bacteriophage lambda stabilization by auxiliary protein gpD: timing, location, and mechanism of attachment determined by cryo-EM. *Structure* **16**, 1399–1406 (2008).
37. Hernando-Pérez M, Lambert S, Nakatani-Webster E, Catalano CE, de Pablo PJ. Cementing proteins provide extra mechanical stabilization to viral cages. *Nat Commun.* **29**;5:4520 (2017).
38. Rao, V. B. & Black, L. W. Structure and assembly of bacteriophage T4 head. *Virology* **7**, 356 (2010).
39. Qin, L., Fokine, A., O'Donnell, E., Rao, V. B. & Rossmann, M. G. Structure of the small outer capsid protein, Soc: a clamp for stabilizing capsids of T4-like phages. *J. Mol. Biol.* **395**, 728–741 (2010).
40. Robertson, K., Furukawa, Y., Underwood, A., Black, L. & Liu, J. L. Deletion of the Hoc and Soc capsid proteins affects the surface and cellular uptake properties of bacteriophage T4 derived nanoparticles. *Biochem. Biophys. Res. Commun.* **418**, 537–540 (2012).
41. Effantin, G., Boulanger, P., Neumann, E., Letellier, L. & Conway, J. F. Bacteriophage T5 structure reveals similarities with HK97 and T4 suggesting evolutionary relationships. *J. Mol. Biol.* **361**, 993–1002 (2006).
42. Vernhes E. *et al.* High affinity anchoring of the decoration protein pb10 onto the bacteriophage T5 capsid. *Sci Rep.* **7**, 41662 (2017).
43. Jakutyte, L. *et al.* First steps of bacteriophage SPP1 entry into *Bacillus subtilis*. *Virology* **422**, 425–434 (2012).
44. Cvirkaite-Krupovic, V., Carballido-López, R. & Tavares, P. Virus evolution toward limited dependence on nonessential functions of the host: the case of bacteriophage SPP1. *J. Virol.* **89**, 2875–2883 (2015).
45. 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).
46. Tavares P. The bacteriophage head-to-tail interface. *Subcell. Biochem.* **88**, 305–328 (2018).
 47. Auzat, I., Dröge, A., Weise, F., Lurz, R. & Tavares, P. Origin and function of the two major tail proteins of bacteriophage SPP1. *Mol. Microbiol.* **70**, 557–569 (2008).
 48. Plisson, C. *et al.* Structure of bacteriophage SPP1 tail reveals trigger for DNA ejection. *EMBO J.* **26**, 3720–3728 (2007).
 49. Langlois, C. *et al.* Bacteriophage SPP1 Tail Tube Protein self-assembles into β -structure rich tubes. *J. Biol. Chem.* **290**, 3836–3849 (2015).
 50. Goulet, A. *et al.* The opening of the SPP1 bacteriophage tail, a prevalent mechanism in Gram-positive-infecting siphophages. *J. Biol. Chem.* **286**, 25397–25405 (2011).
 51. Gordon MK, Hahn RA. Collagens. *Cell Tissue Res.* **339**(1):247–57 (2010)
 52. Kadler, K. E., Baldock, C., Bella, J. & Boot-Handford, R. P. Collagens at a glance. *J. Cell Sci.* **120**, 1955–1958 (2007).
 53. Ricard-Blum, S. The collagen family. *Cold Spring Harb Perspect Biol.* **3**, a004978 (2011).
 54. Hulmes, D. J. S. Collagen Diversity, Synthesis and Assembly. in *Collagen* (ed. Fratzl, P.) 15–47 (Springer US, 2008).
 55. Gelse, K., Pöschl, E. & Aigner, T. Collagens-structure, function, and biosynthesis. *Adv. Drug Deliv. Rev.* **55**, 1531–1546 (2003).
 56. Garnero, P. *et al.* Extracellular post-translational modifications of collagen are major determinants of biomechanical properties of fetal bovine cortical bone. *Bone* **38**, 300–309 (2006).

57. Myllyharju, J. & Kivirikko, K. I. Collagens, modifying enzymes and their mutations in humans, flies and worms. *Trends Genet.* **20**, 33–43 (2004).
58. Gordon M.K. & Hahn R.A. Collagens. *Cell Tissue Res.* **339**, 247-257 (2010).
59. Bella J. Collagen structure: new tricks from a very old dog. *Biochem J.* **473**, 1001-1025 (2016).
60. Ghosh, N. *et al.* Collagen-Like Proteins in Pathogenic E. coli Strains. *PLoS One* **7**, e37872 (2012).
61. Rasmussen, M., Jacobsson, M. & Björck, L. Genome-based identification and analysis of collagen-related structural motifs in bacterial and viral proteins. *J. Biol. Chem.* **278**, 32313–32316 (2003).
62. Yu, Z., An, B., Ramshaw, J. A. M. & Brodsky, B. Bacterial collagen-like proteins that form triple-helical structures. *J. Struct. Biol.* **186**, 451–461 (2014).
63. Chen, S.-M. *et al.* Streptococcal collagen-like surface protein 1 promotes adhesion to the respiratory epithelial cell. *BMC Microbiol.* **10**, 320 (2010).
64. Oliver-Kozup, H. A. *et al.* The streptococcal collagen-like protein-1 (Scl1) is a significant determinant for biofilm formation by group A Streptococcus. *BMC Microbiol.* **11**, 262 (2011).
65. Lopes, A., Tavares, P., Petit, M.A., Guérois, R. & Zinn-Justin, S. Automated classification of tailed bacteriophages according to their neck organization. *BMC Genomics* **15**, 1027 (2014).

PhD Thesis

Abstract

Viruses are of an extreme diversity. However, the study of their structures, their infectious cycles and their assembly pathways revealed relatedness that allowed their classification and establishment of evolutionary relationships between them. Assembly of new infectious viral particles (virions) is the final goal of the lytic viral cycle. These virions can then initiate a new infection cycle. Assembly of the viral particle is an essential step that involves actors from both viral and host origins. Their building process follows a precise program in which its components interact in a precisely defined, orderly, manner.

Gp12 is the SPP1 capsid auxiliary protein. It binds specifically to the center of all hexamers of the major capsid protein gp13. Only expanded capsids that packaged DNA can bind gp12. Phages lacking gp12 are viable and infectious in laboratory conditions.

The sequence of gp12 is marked by 8 GXY repeats, which is the sequence signature of collagen. The presence of proteins containing collagen motifs in eukaryotes was extensively described because of their abundancy and important role in vertebrates. The presence of collagen in prokaryotes is a recent finding. Experimental and theoretical studies showed that collagenous segments from prokaryotic proteins and synthetic peptides are able to form stable trimers that build the collagen super helix. The absence of post-translational modifications of collagen, essential in eukaryotes, is compensated by other mechanisms that result in a thermal stability near to the one observed for eukaryotic collagen.

In this work we show experimentally that gp12 is a stable collagen motif containing viral protein (CMCvP). The isolated protein is an elongated trimer in solution. Despite the shortness of the GXY repetition, the gp12 circular dichroism profile reveals the presence of the collagen motif signature. This motif is cut by collagenase VII in a precise site inside the (GXY)₈ motif. The protein can reversibly

unfold-dissociate and refold-reassociate under the effect of temperature. Binding of gp12 to the SPP1 capsid increases significantly the thermal stability of the capsid bound protein. Upon heating, gp12 dissociates from the capsid and unfolds-dissociates. Depending on temperature, both native trimers and unfolded monomers bind back to the capsid, but following different interaction profiles.

The gp12 properties, taken together, render it a suitable candidate for nano-engineering. We were able to bind tag-gp12, the gp12 protein with an amino-terminal fused peptide, to capsids lacking gp12. The highly immunogenic properties of gp12 make it also an interesting platform to fuse antigens at the SPP1 capsid surface for vaccination purposes. Its thermal properties add further plasticity to applications exploring its temperature-dependent association and release from SPP1 capsids.

Proteins with segments of sequence homology to gp12, detected by similarity searches, have a modular organization featuring GXY repetition motifs. Their encoding genes are adjacent to a major capsid protein gene in prophages of *Bacilli* genomes, as in case of SPP1 gene 12. This common genomic organization suggests a function similar to gp12.

We have then made a global genome-wide survey in prokaryotes and viruses revealing the diverse size and structural organization of their CMCPs. Archaea and archaeal viruses have the lowest frequency of CMCPs in their proteome. Bacteriophages have a higher frequency of CMCPs $\langle(GXY)_{65}$ repeats than their host but bacteria feature longer GXY stretches that are rare in phages. These observations reveal major roles of CMCPs in prokaryotic biology that call for further research on their function and potential exploitation in biotechnological applications.

1 Introduction part I : What are viruses ?

1.1 The discovery of viruses, an historical account

Current knowledge suggests that every living organism has a virus that it could be infected with^{1,2}. Although viruses are the causative agents of many diseases and they largely shape cellular evolution, modern Science had to wait until the late nineteenth century for the discovery of these pathogens. At that age of Science, micro-organisms responsible for diseases were defined as germs that could be retained by filters, that could be cultured in a nutrient broth and that could be seen using an optical microscope². Bacteria, fungi, and protozoa were the only known micro-organisms¹.

The birth of modern virology is linked to the history of the tobacco mosaic virus (TMV) discovery^{1,3,4}. In 1886, Adolf Mayer published a communication about his work on the tobacco mosaic disease. He noticed that the disease could be passed from infected plants to healthy ones by inoculation using the juice extracted from sick plants. Since he was unable to cultivate the causal agent, he tried to reproduce the disease using known micro-organisms. None of them was able to give rise to sick plants. First, he hypothesized that the infectious agent could be either a soluble enzyme-like contagion or an unknown micro-organism. Thus, he included a filtration step before inoculation of the plants using a filter paper. He noticed that the germ passed through the filter and that the filtrated inoculum reproduced the plant illness. Upon several filtration steps, the clear filtrate became sterile. Even though he was unable to satisfy the germ theory postulate, Mayer concluded that the tobacco mosaic disease agent is most likely a new type of bacteria. At that Science age, the concept of a self-reproducing enzyme-like unit was not believed to exist.

A few years later, Dimitri Ivanofsky repeated the same experiment but used a Chamberland filter that had pores smaller than bacteria, instead of the paper filter. The Chamberland filter was one of the best bacteria-proof filters at that time. Ivanofsky

reached the same result as Mayer being unable to culture the germ. In contrast, he concluded that the cause of the tobacco mosaic disease is some sort of toxin. A major advance was made by Martinus Beijerinck. He observed in 1898 that the diluted and filtrated sap of infected plants (using the Chamberland filter) is able to regain its strength after replication in living plant tissues. This experiment showed, for the first time, the obligate parasite property of the pathogen and explained why Mayer and Ivanofsky were unable to culture it. He named it a virus. The term virus was derived from the Latin for poison^{1,5}. Beijerinck made many other studies to investigate the nature of the tobacco mosaic disease agent. After many years of debate, the shape, the size and the nature of the pathogen was revealed by an electron micrograph in 1939⁶, just one year after the first micrographs of animal viruses were published⁷. The tobacco mosaic virus (TMV) was the first virus discovered, and virology was born.

The discovery of viruses that infect bacteria is traditionally traced to the publications of Twort in 1915⁸ and d'Hérelle in 1917⁹ although earlier studies already hinted for the existence of a virus-like antibacterial activity¹⁰. D'Hérelle gave the name of bacteriophage ("bacteria eater") to bacterial viruses. He rapidly recognized them as therapeutic agents to eradicate bacteria and his extensive work brought research on phages to the spotlight as one of the most exciting biological themes in the 1920s. The first electron micrographs revealing the ultrastructure of bacteriophages, also later abbreviated as phages, were published in 1940¹¹.

During the last century, the number of viral species identified increased rapidly. The co-evolution of virology and modern technology was determinant to drive fast knowledge progress about these infectious agents^{1,12}. Advances in purification techniques allowed the preparation of pure viral particles (or virions) for biochemical and biophysical studies that led, namely, to uncovering of their molecular structures by X-ray crystallography and electron microscopy. NMR was also instrumental to characterize individual components of the virion. The morphology of the viral particle became a central criterion for virus identification while the atomic structures of their components allowed establishing phylogenetic relationships, sometimes unexpected, between viruses infecting the three Domains of Life^{13,14}.

The earlier definition of viruses as tiny biological entities smaller than bacteria that pass through the Chamberland filter and that are invisible in the optical microscope was recently challenged¹⁵. The discovery of giant viruses broke the size barrier between the microbial and the viral worlds¹⁵⁻¹⁸. The icosahedral capsid of Mimivirus has a diameter of about 400nm protecting a 1.2Mb genome coding for around 911 genes¹⁸ (Figure 1f). This virus is bigger than some small bacteria in both particle and genome sizes. The bacterium *Mycoplasma genitalium*, for example, is 300nm long with a 580kb genome size coding for fewer than 500 genes^{19,20}. During the evolution process some bacterial species, such as *Rickettsiae* and *Chlamydiae*, became so dependent on their host that their extracellular cycle cannot exceed a short period without compromising their viability. *Rickettsiae* can only be cultivated on living eukaryotic host cells²¹, while *Chlamydiae* are defective in several metabolic and biosynthetic pathways relying on host cells for survival²². Their existence depends on ecological relationships with other organisms^{21,22}. Thus, viruses are not the only obligate parasites in Nature.

The most distinctive properties of viruses is, at the present state of knowledge, the complete absence of metabolism and ribosomes^{2,12}. Viruses do not grow and cannot replicate by their own. They always need a host cell that they hijack²³. The host cell machinery is then manipulated to replicate the viral genome and to produce viral building units coded by the viral genome. Those components assemble together in a precise order to produce new infectious viral particules (also named virions). This is a second distinctive property of viruses: cells replicate by division, viruses do not^{2,12}.

Bacteriophages are the largest population of the Virosphère. They are most likely present in all ecosystems where bacteria are found, as soil, oceans, freshwater, and multicellular organisms colonized by bacteria. Phage infection plays a major role in the dynamics of all those bacterial populations.

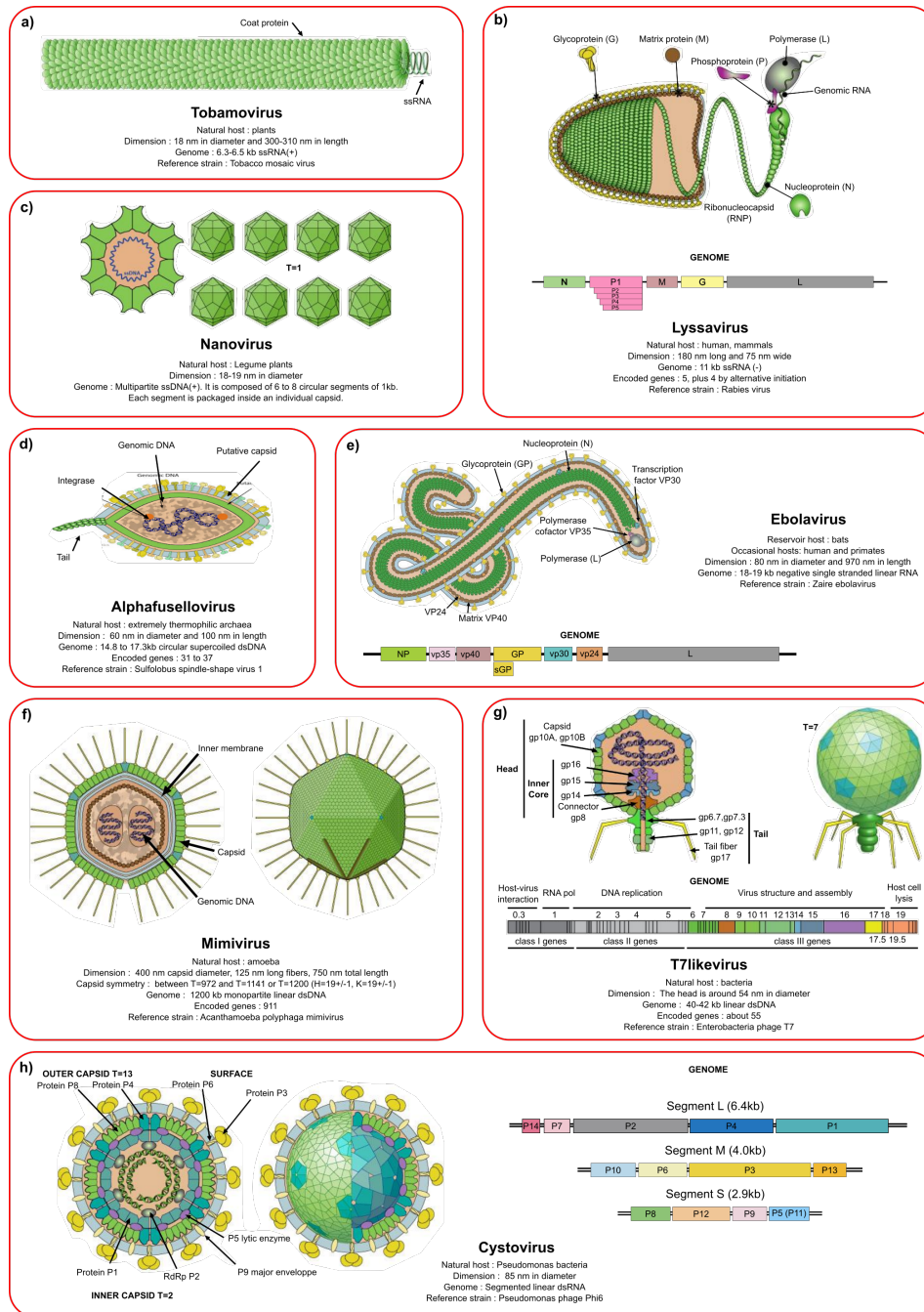


Figure 1 : Shape, genomic and structural organization of some viruses. Each panel (a-h) describes the characteristic features of a viral genus represented by its reference strains. **a)** TMV virus. **b)** Rabies virus. **c)** Subterranean clover stunt virus. The viral ssDNA genome is segmented in 8 parts. Each ssDNA is encapsulated individually in a viral capsid. **d)** Sulfolobus spindle-shape virus 1. **e)** Zaire ebolavirus. **f)** *Acanthamoeba polyphaga* Mimivirus. **g)** Enterobacteria phage T7. **h)** Pseudomonas phage Phi6. Its dsRNA genome is segmented in three elements that are encapsidated in the same viral capsid. Figures were assembled using different references from ViralZone, a part of ExPASy, the SIB Bioinformatics Resource Portal. Web site: <http://viralzone.expasy.org/>

1.2 Virus diversity

Studies of viral particles allowed the detailed description of their shapes and composition. One of the most impressive features of viruses is their extreme structural diversity, as shown in Figure 1. In a general manner, virions are made of genetic material protected by a container. The viral genetic material can be DNA or RNA, single (ss) or double-stranded (ds), segmented or not, circular or linear. In the most basic case, a single layer of multiple copies of one protein protects the nucleic acid. It is the case of TMV in which the single-stranded RNA (ssRNA) is coated with 2130 identical copies of a 158 amino acid-long protein, called the coat protein (CP)²⁴. The RNA-CPs complex forms a right-handed helix that defines the shape of the virus (Figure 1a). The virion carries its own polymerase associated to its genetic material. Despite the large difference in the shape and the size of particles, this global organization of TMV is close to the one of the more complex Ebolavirus genus²⁵ that infects mammals. Ebola virus differs from TMV also by being a flexible rod with a membrane envelope and by the polarity of its ssRNA (Figure 1e).

Cystoviruses are bacterial enveloped viruses that have a segmented genome. The three dsRNA segments are packaged inside the same particle, as shown in Figure 1h²⁶. The viral genome is not always packed inside a single particle. For example, plant nanoviruses have a multipartite ssDNA²⁷. Each circular ssDNA molecule is packed inside an individual icosahedral capsid and codes for one protein (with one exception). Each protein coded by the ssDNA in individual particles has a precise role. Since all the multiprotein machinery is needed to produce new nanoviruses virions, cells must be co-infected by different viral particles to provide the complete set of genes required for nanovirus multiplication.

T number	Number of CPs	Structure of the Icosahedral capsid	Building of the Icosahedral capsid from		T calculation following Caspar and Klug system $T=h^2 + k^2 + hk$	Viruses examples
			12 subunits	60 subunits		
T = 1	60 12 pentamers				$h=1, k=0$ 	ssDNA Circoviridae Parvoviridae Anelloviridae Geminiiviridae Nanoviridae Microviridae dsRNA Partitiviridae Chrysoviridae ssRNA+ Hepeviridae Barnaviridae Oourmiavirus
T = 3	180 12 pentamers 20 hexamers				$h=1, k=1$ 	dsRNA Picobirnaviridae ssRNA+ Astroviridae Nodaviridae Caliciviridae Leviviridae Luteoviridae Bromoviridae (not all genera) Tombusviridae Tymoviridae Sobemovirus Polemovirus Umbravirus
T = 4	240 12 pentamers 30 hexamers				$h=2, k=0$ 	ssDNA+ Tetraviridae Togaviridae dsDNA(RT) Hepadnaviridae
T = 7	420 12 pentamers 60 hexamers				$h=2, k=1$ 	dsDNA Polyomaviridae Papillomaviridae Siphoviridae Podoviridae dsDNA(RT) Caulimoviridae
T = 9	540 12 pentamers 80 hexamers				$h=3, k=0$ 	N4likevirus
T = 13	780 12 pentamers 120 hexamers				$h=3, k=1$ 	dsRNA Birnaviridae Cystoviridae Reoviridae
T = 16	960 12 pentamers 150 hexamers				$h=4, k=0$ 	dsDNA Herpesviridae SP01-like viruses

Figure 2: Geometry of the viral icosahedral capsid according to the equivalency ($T=1$) and quasi-equivalency ($T>1$) interaction of the CP subunits following to the Caspar and Klug rules^{28,29}. This figure shows only the most frequent cases found for viral capsids. Capsids with a higher triangulation number are known to exist. Figures were assembled from ViralZone. Web site: <http://viralzone.expasy.org/> RT stands for reverse transcribing viruses.

A large number of viruses have an icosahedral capsid¹²⁸. This geometry allows assembly of structures of different diameters to protect the viral genome using one

single type of subunit (Figure 1c,f-h and 2). In the simplest case, 60 CP subunits are precisely arranged by equivalent interactions to build a basic icosahedron with vertices made of CP pentamers (Figures 1c and 2). In such case all capsomers are pentamers. Caspar and Klug predicted that larger systems made of multiples of 60 subunits can also be built by allowing a minimum distortion from strict equivalent subunit interactions. In such systems, CP subunits are placed in a quasi-equivalent organization and have the ability to adapt their conformations depending on their local environment²⁹. The resulting icosahedron vertices are always made of CP pentamers. However, depending on the total number of CP subunits per capsid, the icosahedron faces contain one or several CP hexamers. A triangulation number T is computed for each case following the equation $T=h^2+k^2+hk$, where T is the triangulation number, and h and k are positive integers²⁸⁻³⁰ (Figures 2 and 3). To determine h and k , a hexagonal lattice of the same unit size as the capsid hexagonal capsomers is drawn with a 60° angle crossing h and k axes at the center of an arbitrary chosen hexon (Figure 3). An icosahedron facet is aligned on the hexagonal lattice placed with one 5-fold axis at the origin point (0,0). The position (h,k) of the closest 5-fold vertex gives h and k values to be used for T calculation. The T value takes a discrete value (1, 3, 4, 7, 9, 13, 16, 25, etc). As shown in Figure 2, the number of CP subunits per capsid grows along with the T number. For example, a $T=7$ icosahedral capsid is made of 420 CP subunits. The molecular basis of the quasi-equivalent interactions was revealed by a large number of sub-nanometer and atomic structures of icosahedral viral capsids^{13,14,28,30-32} and references therein.

Many viruses have a portal system for entry and exit of the genome in the capsid. This structure replaces one capsid pentamer, reducing the number of CP subunits by 5 units. In such systems, the portal vertex plays an essential role during both viral assembly and host infection^{30,33}. It constitutes a specialized gate for dsDNA entry and exit during those processes, respectively. Genome encapsidation requires a molecular engine, made of viral proteins, that pumps the genome to the interior of the capsid through the portal pore. At the beginning of infection, the same gate is used for genome exit³³. Besides the capsid, some viruses like tailed bacteriophages developed tails used as tubes to deliver the viral genome inside the host cell leading to a complex

virion morphology (Figure 1g). Virions of Archaea can also exhibit a complex structural organization (Figure 1d).

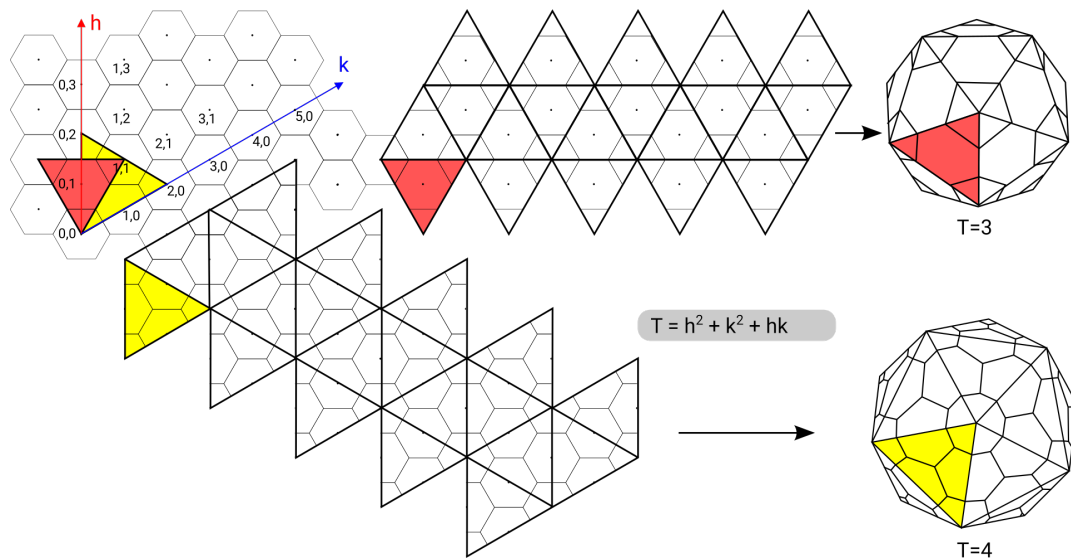


Figure 3: Building of an icosahedron of geometry T=3 and T=4 using the same flat hexagonal network according to the Caspar and Klug rules²⁹. To calculate the T number, an arbitrary hexon is chosen first. Then h and k axes are drawn with a 60° angle crossing the center of the hexons. The axes intersection point at a 5-fold vertex of the capsid indicates the zero reference coordinates (h=0; k=0). Next, an equilateral triangle is drawn starting from the reference point to the point of insertion of the closest 5-fold vertex. Triangle vertices localize at the center of a hexon. Finally, triangles are assembled to form the icosahedral capsid lattice. T is calculated using the equation $T = h^2 + k^2 + hk$. For the examples depicted, h=1; k=1 for T=3 particles and h=0; k=2 for T=4 particles. Figure adapted from Baker et al. 1999²⁸.

1.3 Viruses classification and taxonomy

Viruses are of an extreme diversity rendering their classification one of the hardest, but necessary, tasks in Virology. Over the years, many systems were established. Early classification features were based on pathogenic properties, hosts that viruses infect, and their mode of transmission^{34,35}. Thus, viruses that cause hepatitis (hepatitis A virus, hepatitis B virus, yellow fever virus, Rift Valley fever virus) could be grouped together as the hepatitis viruses. However, these viruses are very different and genetically unrelated. This approach suffered from other serious limitations. First, not all viruses cause diseases. Second, some viruses cause more than one disease. For example, the primo-infection of varicella zoster causes chickenpox and causes shingles

when reactivated later². Third, classification of viruses based on their host is also problematic. Some viruses infect more than one host and can cause disease in each one of them. The Ebolavirus infects both humans and non-human primates²⁵.

The Baltimore scheme is one of the most widespread classification systems. It is based on the nature of the viral genome, its mode of replication and its mode of gene expression (transcription and translation)^{2,36}. The actualized scheme includes seven classes, noted class 1 to 7 or I to VII as shown in Figure 4.

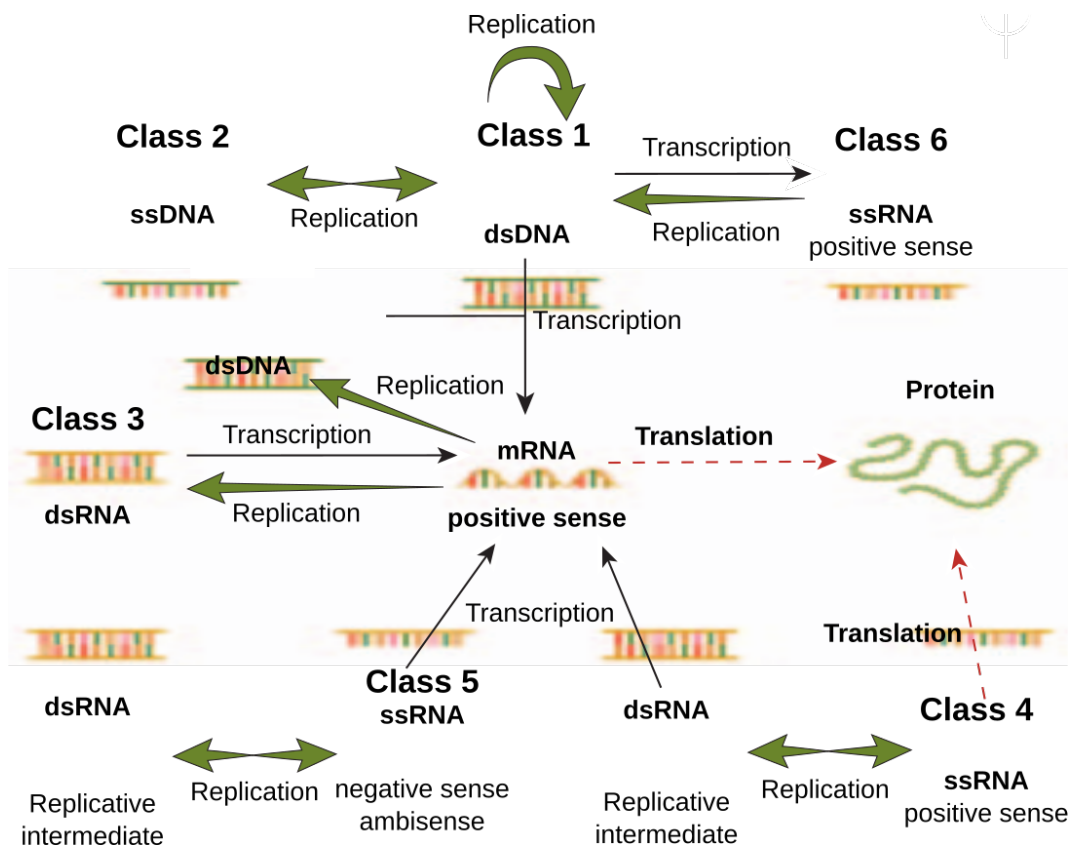


Figure 4: The Baltimore classification scheme. Messenger RNA (mRNA) is designated as positive sense RNA. Translation of protein from mRNA and positive sense RNA virus genomes is indicated with red arrows. The path of production of mRNA from double-stranded templates is shown with black arrows. Green arrows show the steps in the replication of the various types of genome with double-headed arrows, indicating production of a double-stranded intermediate from which single-stranded genomes are produced. The Baltimore class 1 contains viruses that have dsDNA. Class 2 contains viruses with ssDNA. Class 3 contains viruses that have dsRNA. Class 4 contains viruses with positive ssRNA which can be translated. Class 5 contains viruses with negative ssRNA. Class 6 contains viruses that have ssRNA but generate a replication dsDNA intermediate. Class 7 is a new class that groups viruses generating their genomic dsDNA from a positive ssRNA intermediate. Figure adapted from 6th edition of "Introduction to modern virology" by Dimmock et al. 2007².

Advances in biochemical and biophysical techniques played an important role in the classification of viruses. Electron microscopy (EM) was instrumental since its development to investigate the viral particle shape and organization^{6,7,10,12,24,28,30-33} and references therein. Viruses could be separated into two major groups according to the viral particle features: viruses wrapped with a host derived lipid envelope (called enveloped viruses) and naked viruses (called non-enveloped viruses)³⁷. Based on the viral capsid morphology, viruses could be further classified as filamentous (Figure 1a), isometric (Figure 1c,f,h) or complex². Despite the high diversity of viral major capsid protein sequences, comparison of their structural topologies and of their organization within capsid lattices revealed similarities allowing their classification into structurally related viral lineages for a variety of viruses^{13,14} (Figure 5a). The HK97 lineage encloses tailed bacteriophages and herpesviruses whose major capsid protein shares a common structural topology typified by bacteriophage HK97 gp5 (Figure 5b).

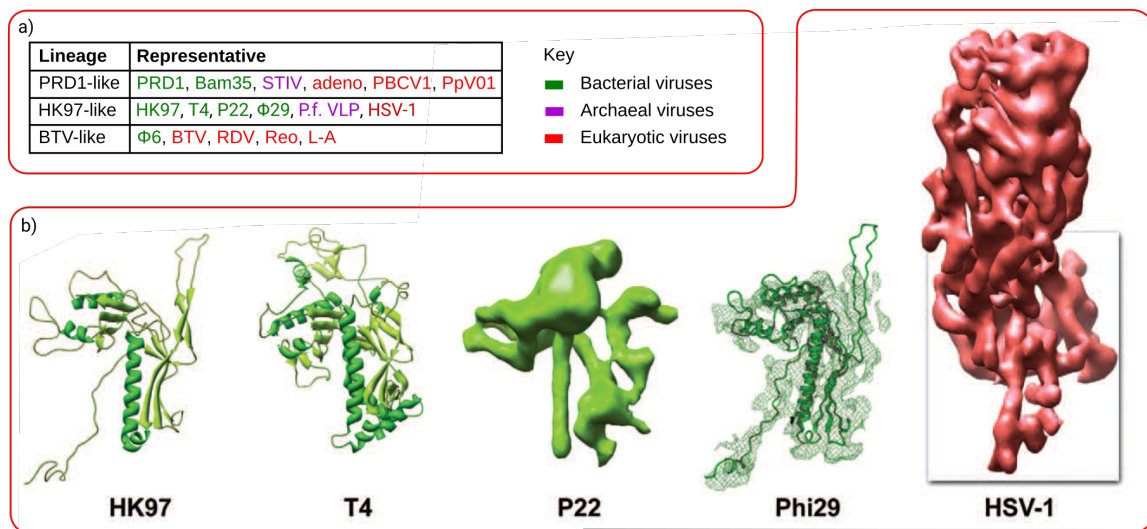


Figure 5: Examples of structurally related lineages of viruses with an icosahedral capsid. a) Viral lineages and their representative viral species. Viruses names are colored, in function of their host, as shown in the key on the right. b) "The HK97-like lineage". A gallery of structures of HK97-related coat proteins, determined using X-ray crystallography (HK97 gp5 and phage T4 gp24) or cryo-EM (P22 gp5, ϕ 29 gp5 and HSV-1 VP5). The view is perpendicular to the capsid surface. The box for HSV-1 VP5 indicates the portion of the molecule comprising the floor domain engaged in viral shell formation that has a HK97 fold. Adapted from Bamford et al. 2005¹³.

All the classification methods described above are based on biological and structural knowledge. Unfortunately, such type of information is available for a very

small number among all isolated viruses. The development of molecular biology, the explosion of the number of sequenced viral genomes, and the advent of the genomics and proteomics era was the origin for the development of new classification systems that use the viral genetic information to establish relatedness among viruses. Several systems exist. Some of them are based exclusively on the analyses of genomes sequence³⁸. Others are based on protein analyses^{39,40} or combine the proteomic with the genomic data for more reliable results⁴¹. In all cases, the most significant advantage of those systems is the ability to trace evolutionary relationship between viruses³⁸.

There are presently different possible criteria to classify viruses. Every system has its own advantages and disadvantages. The result can be conflicting. Thus, the International Committee on Taxonomy of Viruses (ICTV) provided rules aiming to homogenize the classification and nomenclature of viruses^{36,37,42,43}. The taxonomical hierarchy is made of order (virales), family (viridae), sub-family (virinae), genus (virus) and species⁴². Over the last 40 years, the ICTV succeeded to classify 2827 viral species into 455 genera, 22 sub-family, 103 families, and 7 orders.

Practically, a combination of one or more classification system is used to define a virus. For example, the Enterobacteria phage T7 virus is defined as a tailed virus (order Caudovirales) because it features a tail (Figure 1g), it is called phage (or bacteriophage) because it is a bacterial virus and the Enterobacteria suffix indicates its host^{1,44}.

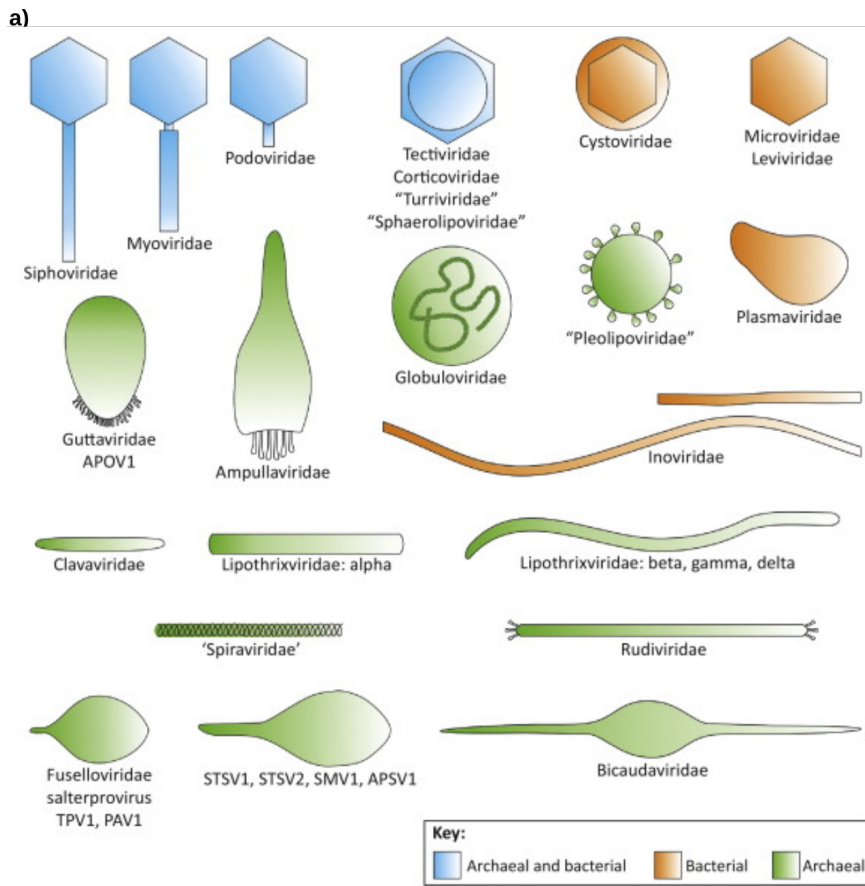
1.4 Tailed bacteriophages classification and their relationship with herpesviruses

The vast majority of known prokaryotic viruses infect bacteria and a few percent infects archaea⁴⁵. As shown in Figure 6, they have various genome types and capsid morphologies⁴⁵⁻⁴⁷. Among those, bacteriophages are the most abundant biological entities on Earth with an estimated number of 10^7 to 10^9 particles per gram of soil and around 10^7 particles per milliliter of water from oceans and freshwaters⁴⁸. More than 5500 phages were examined by electron microscope and 96% of them are tailed,

belonging to the order Caudovirales^{44,46,49}. Based on the alignment of amino acid sequences of some functional proteins (DNA polymerases, integrases, and peptidoglycan hydrolases), an evolutionary link was proposed to exist between all tailed phages⁴⁹. This link is further corroborated by structural data showing similar folds of protein components of the viral particle like the major capsid protein (Figure 5).

Members of the Caudovirales order are dsDNA viruses (Baltimore class 1), with an icosahedral capsid and a tail^{44,46,47}. This order has three families: Myoviridae, Siphoviridae and Podoviridae. The distinctive feature between those families is the structure of the tail: Myoviridae have a long contractile tail, Siphoviridae have a long non-contractile tail, and Podoviridae have a short non-contractile tail.

Building of the capsid and tail of viruses from the Siphoviridae and Myoviridae families follows distinct pathways that join at the last step of assembly to attach the two structures together forming an infectious virion. In contrast, the Podoviridae short tail is assembled at the capsid portal vertex. Tailed phages and herpesviruses share a similar capsid assembly pathway^{13,14,32,50,51} (Figure 7, see also section 1.5). This common strategy correlates with structural homology of the capsid building proteins, showing that tailed phages and herpesviruses form a viral lineage^{14,50} (Figures 5 and 7).



b)

Family	Capsid morphology	Additional feature(s)	Genome type	No. of complete genomes	Example
Myoviridae	Icosahedral	Tail (contractile)	dsDNA, L	134	T4
Siphoviridae	Icosahedral	Tail (long noncontractile)	dsDNA, L	268	λ
Podoviridae	Icosahedral	Tail (short noncontractile)	dsDNA, L	98	T7
Tectiviridae	Icosahedral	Internal membrane	dsDNA, L	4	PRD1
Corticoviridae	Icosahedral	Internal membrane	dsDNA, C	1	PM2
Plasmaviridae	Pleomorphic	Enveloped	dsDNA, C	1	L2
Microviridae	Icosahedral	Nonenveloped	ssDNA, C	15	Φ X174
Inoviridae	Filamentous	Long flexible or short rigid	ssDNA, C	29	M13
Cystoviridae	Icosahedral	Enveloped, multilayered	dsRNA, L, S	5	Φ 6
Leviviridae	Icosahedral	Nonenveloped	ssRNA, L	7	MS2

bacterial (blue and orange) viruses. Abbreviations: APOV1, Aeropyrum pernix ovoid virus 1; APSV1, Aeropyrum pernix spindle-shaped virus 1; PAV1, Pyrococcus abyssi virus 1; SMV1, Sulfolobus monocaudavirus 1; STSV1, Sulfolobus tengchongensis spindle-shaped virus 1; STSV2, Sulfolobus tengchongensis spindle-shaped virus 2; TPV1, Thermococcus prierii virus 1. Figure reproduced from Pietilä et al. 2014⁴⁵. **b)** Prokaryotic viruses families features. The number of complete genome sequences was obtained from GenBank and does not include the genome sequences available for different isolates of the same virus strain. Families Myoviridae, Siphoviridae, and Podoviridae are grouped into the order Caudovirales. Genome types: L, linear; C, circular; S, segmented. From Krupovic et al. 2011⁴⁷.

Figure 6: Prokaryotic viruses families. Phages and archaeal viruses shapes are represented schematically in panel (a) and their properties are compiled in (b). **a)** Virion morphotypes of prokaryotic viruses. Names of viral genera or families based on the International Committee on Taxonomy of Viruses (ICTV) classification are indicated below the scheme of the virus particles. If an archaeal virus has not been assigned to any genus or family, individual virus names are given. Virions are not drawn to scale. The key color code shows the morphotypes or archaeal (blue and green) and

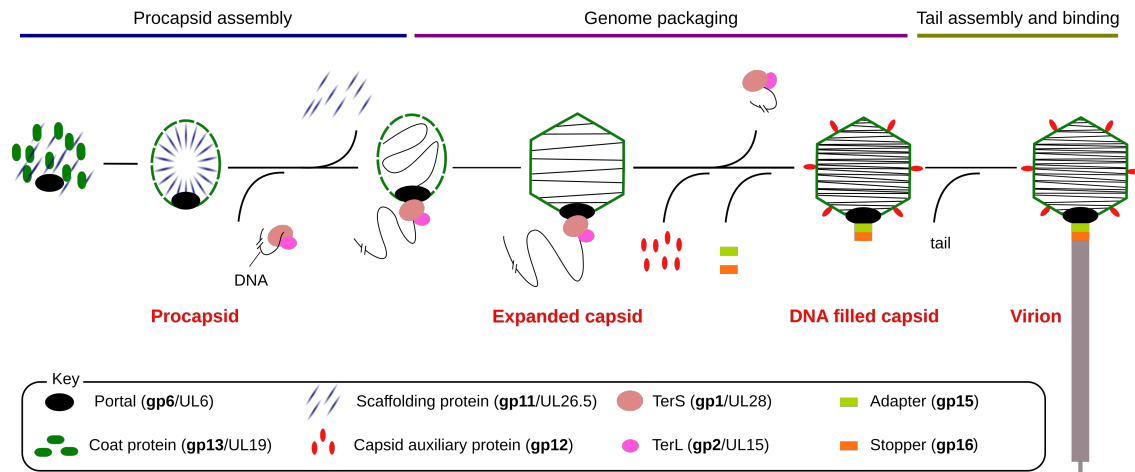


Figure 7: Schematic representation of the tailed phages assembly pathway and its similarity to herpesviruses nucleocapsid assembly. The capsid assembly steps are common between the two types of viruses. Bacteriophage SPP1 and Herpes Simplex Virus-1 (HSV-1) component proteins are shown as example. Proteins are listed in the key panel. SPP1 protein names start with gp and HSV-1 proteins start with UL. TerS stands for terminase small subunit, and TerL stands for terminase large subunit. The SPP1 gp1 protein assembles into nonamers while gp2 remains monomeric in solution.

1.5 Assembly of tailed bacteriophage virions

Tailed bacteriophages are large nucleoprotein complexes of homogeneous size and shape with more of 20 MDa mass. Their construction follows a strict order of sequential interactions between its components that are present simultaneously in the infected bacterium. Therefore, the program of conformational changes of phage proteins and the specificity of their macromolecular interactions ensures assembly following a defined pathway⁵², as illustrated in Figure 7. The pathway was initially defined by studying assembly intermediates that accumulate in non-permissive infections the impair synthesis of virion proteins⁵². The determination of structures of virion proteins and of their complexes from different bacteriophages showed that a conserved structure of the viral components correlates with their common viral particles assembly plan^{14,53 and references therein}. They also unraveled the common evolutionary origin of tailed phages and herpesviruses nucleocapsids^{13,14,32,50,51,53} (Figures 5 and 7) while phage long tails were shown to be related to cellular type VI secretion systems and to other delivery devices used in bacterial warfare^{53,54,55 and references therein}.

The bacteriophage nucleocapsid is a protein container that protects the linear dsDNA viral genome. A DNA-free procapsid, also named prohead, is constructed first. The major capsid protein (MCP) subunits establish quasi-equivalent interactions (section 1.2; Figures 2 and 3) in a reaction controlled by an internal chaperone, the scaffolding protein (SFP), that ensures correct positioning of the subunits to build an icosahedral lattice of homogeneous size. The SFP can be an elongated independent protein (e.g. phages P22, SPP1, phi29, herpesviruses)⁵⁶ or fused to the amino-terminus of the MCP (e.g. phages HK97 and T5)^{57,58}. Procapsids of the tailed phages-herpesviruses lineage are characterized by the presence of a specialized vertex formed by a dodecamer of the portal protein (section 1.2; Figure 7)³³. Procapsid assembly likely initiates at this vertex, ensuring the asymmetric incorporation of the portal in the procapsid^{33,59-62}. The portal dodecamer has a central channel through which phage DNA enters and exits the capsid. The SFP, either proteolyzed or intact, exits through channels in the procapsid lattice after procapsid assembly⁵⁶. Release of the SFP and/or initiation of DNA packaging through the portal trigger a major conformation change of the MCP lattice that increases in size, and becomes thinner. The resulting structure is highly robust and resists to internal pressures as high as ~60 atm applied by the dsDNA tight packing^{63,64}.

The portal vertex provides a platform for binding of the terminase-phage DNA complex leading to assembly of the DNA packaging motor. The terminase complex is normally composed of a small (TerS) and a large subunit (TerL)⁶³. TerS binds specifically phage DNA while TerL is a two-domain protein with endonuclease and ATPase activities. The bacteriophage DNA substrate for encapsidation that is most frequently a concatemer of viral genomes generated by phage DNA replication. After selective binding of TerS to phage DNA, TerL interacts with the TerS-DNA complex to cleave the phage DNA substrate to generate a free DNA end^{63,65,66}. The terminase-DNA complex then docks at the procapsid portal vertex to assemble the DNA packaging motor (Figure 7). The motor initiates DNA packaging at the DNA free end bound to the TerS-TerL complex. Mechanical translocation of phage DNA to the capsid interior involves an intimate cross-talk between TerL and the portal protein in a reaction energized by the TerL ATPase activity^{63,67}. At late stages of encapsidation, the DNA

concentration inside the capsid can reach ~500 mg/mL⁶⁸. The tight packing of DNA helices exerts a significant pressure on the capsid lattice and requires that the DNA packaging motor exerts forces as strong as 100 pN to terminate packaging⁶⁸. Consequently, DNA packaging slows-down as the capsid fills⁶⁹. This energy-driven strategy to encapsidate the viral genome optimizes the amount of genetic information carried in the virion. DNA packaging correlates with the binding of auxiliary proteins at the capsid external surface that are distributed symmetrically in the structure (Figure 8; section 1.6). Auxiliary proteins can provide additional stability to the capsid and/or change its surface properties to adapt to the extracellular environment (section 1.6).

Genome packaging is normally terminated by an endonucleolytic cleavage of the concatemeric DNA by TerL, followed by release of the terminase-DNA complex to initiate a new genome encapsidation cycle into a procapsid. Disassembly of the DNA packaging motor is coordinated with binding of one or several head completion proteins to avoid leakage of the packaged DNA^{33,70,71} and references therein (Figure 7). These proteins extend (adaptors) or close reversibly (stoppers) the portal channel. Their complex together with the portal protein is named connector. Its capsid distal region provides the interface for assembly of a short tail (Podoviridae) or for binding of a tail assembled in an independent pathway (Siphoviridae and Myoviridae)^{33,71}. Such step differs from herpesviruses whose DNA-filled capsid is surrounded with a tegument and a lipid envelope with glycoproteins. These glycoproteins engage fusion with the eukaryotic cell membrane leading to release of the nucleocapsid and tegument in the cytoplasm allowing tegument proteins to initiate hijacking of the host cell⁷² and references therein.

Building of long tails of Siphoviridae and Myoviridae starts by assembly of an adsorption apparatus that recognizes and binds host cell surface receptors. This apparatus is the platform to initiate helical polymerization of the major tail protein (MTP) around a tape measure protein. The tape measure is a molecular ruler that defines the length of the tail⁷³. The tail tube is surrounded by a contractile sheath in Myoviridae⁷⁴. Tail completion proteins bind to the tail tube end distal from the adsorption apparatus to terminate its assembly^{75,76} and references therein. This region of the tail

binds to the connector forming the head-to-tail interface, a final reaction that yields the infectious phage particle.

1.6 Capsid auxiliary proteins

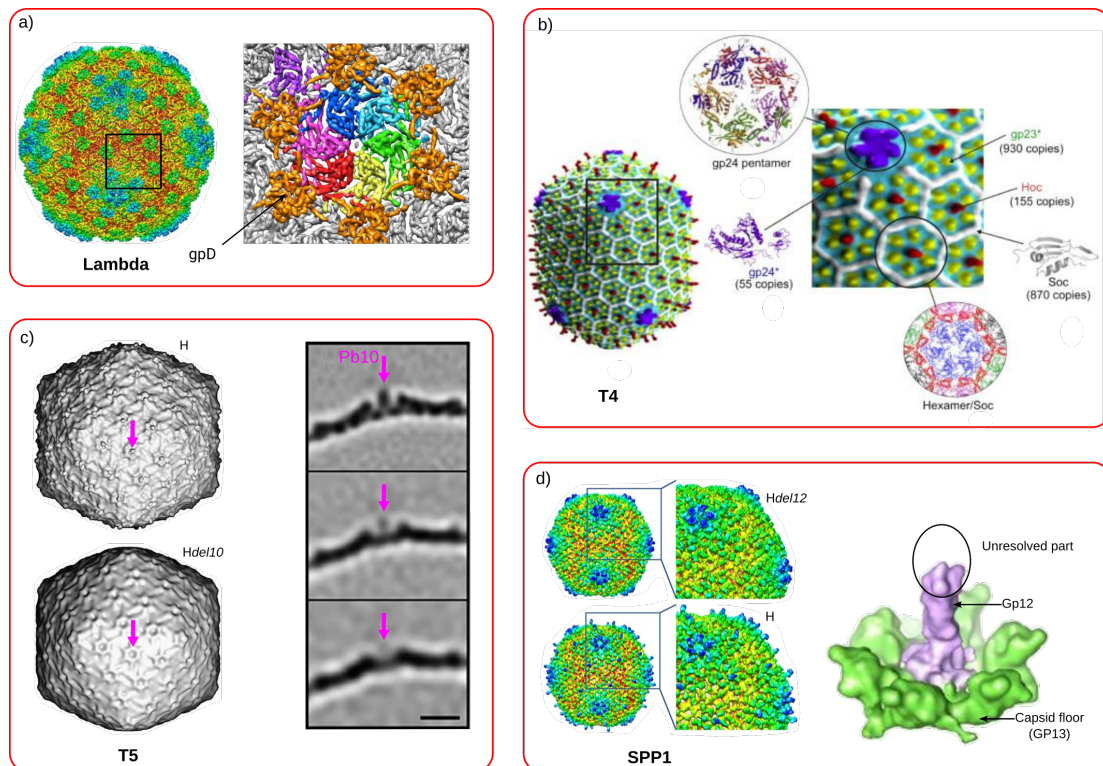


Figure 8: Localization of capsid auxiliary proteins in different viral capsids. Viral capsids are not drawn to scale. **a)** Subnanometer reconstruction of bacteriophage lambda DNA-filled capsid colored radially from the phage center (red to blue) (left). The capsid has a $T = 7$ laevo symmetry. The auxiliary protein gpD appears as protruding densities at the quasi 3-fold vertices of the capsid. A close-up segmented view (right) of the seven subunits that make up the icosahedral asymmetric unit was colored by subunit. The hexamer subunits are shown in the center while the subunit belonging to the pentamer is seen in the upper left-hand corner in purple. Six gpD trimers colored orange are present at the 3-fold symmetry axes. From Lander et al. 2008⁷⁷ **b)** Cryo-EM reconstruction of the phage T4 capsid. The rectangle on the right side is an enlarged view of the structure showing the major capsid protein gp23, that forms the structure hexamers (yellow subunits), gp24, which forms the pentameric vertexes (purple subunits), and the auxiliary proteins Hoc at the center of hexons (red subunits) and Soc which lines the periphery of the hexons (white subunits). From Rao et Black 2010⁷⁸. **c)** Three-dimensional reconstructions of the T5 capsid (H) and of the T5 capsid lacking the auxiliary protein pb10 (Hdel10) which is localized at the center of the capsid hexamer (left). The right panel shows sections through a hexameric capsomer of a time course experiment in which gp10 was removed by capsid treatment with guanidine-hydrochloride (top: 0h; center: 1h; bottom: 3h). From Effantin et al. 2006⁷⁹. **d)** Three-dimensional reconstructions of the SPP1 capsids (H)

and of the SPP1 capsid lacking the auxiliary protein gp12 (Hdel12). The right panel shows the localization of gp12 at the center of a gp13 hexon. The difference map between two hexameric capsomers is shown in light magenta. The green surface shows the capsid hexamer. Adapted from White et al. 2012⁸⁰.

Early steps in viral (pro)capsid assembly process require often weak interactions providing free energy minimization and error-free assembly, while later steps reinforce the capsid structure to withstand the internal pressure generated by tight packing of the DNA genome^{68,69,77,81-86}.

During maturation, the capsid undergoes structural changes that creates or exposes binding sites at its external surface for capsid auxiliary proteins (Figures 7 and 8). In the literature, they are also called capsid cementing proteins or capsid decoration proteins according to their function in the virus particle.

In phage lambda, capsid cementing protein trimers of protein gpD bind between capsomers at the 3-fold vertices of the expanded capsid (Figure 8a) reinforcing it to stand against 60 atmospheres internal pressure^{77,82,85,86,87}. Capsids lacking gpD are fragile and unable to package the entire phage genome, implying that gpD binds to capsids during the packaging process⁸⁶. GpD is a symmetric trimer with an N-terminal β -tulip domain and an α/β subdomain that binds to the capsid⁸⁷. This is a common fold found in the auxiliary proteins of phages 21's SHP⁸⁸, P74-26's gp87⁸⁹, and TW1 gp56⁹⁰. Triplexes that sit between herpesviruses capsomers were reported to have also a domain of interaction with the capsid lattice similar to gpD³². Triplexes are heterotrimers composed of one VP19c (or Tri1) and two VP23 (or Tri2) subunits in HSV-1. However, in contrast to the phage systems, triplexes are essential components of herpesviruses procapsids probably due to more demanding stability requirements to build large herpes procapsids. The protein Dec of phage L, a P22-like phage, is a trimer with a fold different from gpD⁹¹. The Dec trimer adopts an asymmetric organization to bind selectively to the quasi three-fold sites between capsomers while the capsid three-fold sites remain unoccupied, a feature unique to Dec at present⁹¹. Binding of Dec to phage L capsids increase their stability but, in contrast to gpD of lambda, it is not required to withstand the pressure exerted by DNA encapsidation⁹². A different mode of interaction of an auxiliary protein between capsomers is typified by 9-kDa Soc (small outer

protein) of phage T4 and its relatives that lines the periphery of the hexons, clamping neighbor capsomers (Figure 8b)^{78,93,94}. Soc is non-essential but adds stability to the T4 capsid in extreme conditions (e.g. pH 11 and high temperature) and changes its surface properties⁹⁵. In general, auxiliary proteins that bind between capsomers appear to have a cementing function that enhances the capsid stability. Depending on the phage system, this role can vary from essential (e.g. gpD of lambda) to accessory, helping the virion survive extreme environments (e.g. Soc of T4). Elaborations of the MCP HK97-fold⁹⁶ that strengthen inter-capsomer bonding can compensate for the absence of cementing proteins^{81,91}.

Other auxiliary proteins bind to the center of capsomers. Phage T4 Hoc binds to the capsid lattice in addition to the intercapsomer binder Soc (Figure 8b). Deletion of one or both genes coding for those proteins does not affect phage viability, infectivity or production, under laboratory conditions⁷⁸, but changes the capsid surface properties⁹⁵. The highly immunogenic Hoc binds to the center of capsid hexamers but does not stabilize the capsid shell. This multi-domain protein is characterized by an immunoglobulin-like domain that could provide the capsid with additional binding properties to the surface of host bacteria⁹⁷. Phage T5 (Figure 8c) carries 120 copies of a capsid decoration protein, gp10, distributed symmetrically around the capsid. One gp10 binds at the center of each MCP hexamer⁷⁹. The polypeptide is formed by an amino-terminus domain that binds to the T5 MCP hexamer and of a carboxyl terminus Ig-like domain that could bind to bacterial surfaces⁹⁸. Gp10 is not essential for phage viability or host infection^{79,98}. The dispensable gp12 of bacteriophage SPP1 binds also to the center of the MCP hexamers (Figure 8d)⁸⁰. A more rare case is the presence of two different auxiliary proteins, one that binds to capsid hexamers and the other to pentamers in the T=12 bacteriophage SIO-2⁹⁹. Apart from the less-studied SIO-2 phage, the other auxiliary capsid proteins that bind to the center of hexamers do not contribute to capsid stability but might rather act to provide adhesion properties to phage capsids.

Due to their abundance and their regular exposure around the surface of the capsid shell, capsid auxiliary proteins are a target of choice for nanoengineering. Phage lambda, T4 and L capsids were used as display systems^{87,100-104}. Active proteins (beta-

lactamase, IgG-binding domains of the *Staphylococcus aureus* protein A, and β -galactosidase) were exposed at the surface of the phage lambda capsid by fusion to the protein gpD using a peptide linker (N and C terminus)¹⁰⁰. Antigens were displayed at the surface of phage T4 using an in vitro assembly system^{101,102}. Both Soc and Hoc proteins were targeted. Anthrax toxin was exposed at the T4 capsid surface. Fusion proteins were constructed first by coupling the anthrax lethal factor (LF) to Hoc or Soc. Then, purified engineered proteins mixed with purified T4 capsid lacking decoration proteins lead to assembly in vitro of particles carrying LF fused to the auxiliary protein¹⁰¹. These T4 nanoparticles were recently used in vaccination trials¹⁰³. Phage L Dec trimers were also used to display functional proteins¹⁰⁴

1.7 The *Bacillus subtilis* phage SPP1

The SPP1 bacteriophage, subject of this study (Figures 7,8d,9), is a lytic siphovirus isolated in the Botanical garden of Pavia (Italy) that infects the Gram-positive bacterium *Bacillus subtilis*^{105,106}. The SPP1 virus is composed of a capsid and a long non-contractile tail (Figure 9). The SPP1 icosahedral capsid is about 60nm in diameter⁸⁰ and the tail is ~190 nm-long¹⁰⁷. The capsid encloses and protects a highly compacted ~45.9kbp dsDNA linear viral DNA molecule. Packaged DNA molecules have a terminal redundancy and are partially circularly permuted, resulting from a headful packaging mechanism¹⁰⁸. The SPP1 genome is 40,016 bp-long¹⁰⁹.

SPP1 infection is initiated by reversible adsorption to glycosylated teichoic acids at the bacterial surface¹¹⁰ followed by irreversible binding to YueB, a component of the type VII secretion system of *B. subtilis*¹¹¹⁻¹¹³. Interaction of the SPP1 tail adsorption apparatus with YueB triggers ejection of the phage genome¹¹¹ from the phage capsid to the bacterial cytoplasm¹¹⁴. Phage DNA then circularizes. SPP1 DNA replication occurs in a defined focus of the cytoplasm¹¹² generating concatamers of the SPP1 genome that are the substrate for DNA packaging into preformed procapsids¹⁰⁶.

Building of SPP1 viral particles follows the generic assembly pathway of tailed bacteriophages (Figure 7). The spherically shaped icosahedral procapsid with a

diameter of ~55 nm has a triangulation number $T=7$ and is built from 415 copies of the MCP gp13^{60,80}. Polymerization of gp13 is chaperoned by elongated dimers of the SFP gp11¹¹⁵. The dodecameric portal protein gp6 forms a specialized vertex where assembly initiates⁶⁰. Gp6 interacts also with gp7 targeting this minor protein to the procapsid interior in a few copies¹¹⁶. Gp7 is non-essential but its absence in viral particles reduces their infectivity ~4-fold¹¹⁷. Before or when DNA packaging initiates, the SFP gp11 exits the procapsid through channels in the gp13 lattice and the procapsid expands increasing in size to a diameter of ~60 nm, which maximizes the internal space for viral DNA packing.

The *pac* sequence in SPP1 genome concatemers is specifically recognized by the SPP1 TerS gp1 and subsequently cleaved by the endonuclease domain of TerL gp2 (Figure 7)^{65,66,118,120}. The terminase-SPP1 DNA complex then binds the portal vertex to assemble the packaging motor that translocates the DNA to the capsid interior^{67,120}. During packaging the capsid auxiliary protein gp12 binds to the center of each capsid hexamer (section 1.8)^{80,121}. When a threshold amount of DNA is reached inside the capsid TerL carries a sequence-independent cleavage of SPP1 DNA to terminate the packaging process (headful packaging mechanism)⁶⁶. After DNA packaging termination, the sequential binding of gp15 (adaptor) and gp16 (stopper) to the portal protein gp6 prevents exit of the packaged DNA (Figure 7)¹²². The gp6-gp15-gp16 complex forms the SPP1 connector. This structure acts as a gate used for DNA entry during the encapsidation process and for DNA exit during the infection process¹²³. It also provides the nucleocapsid interface for tail binding.

Assembly of the SPP1 tail was proposed to start by formation of the adsorption apparatus that features a cap (gp19.1) in which the gp21 tail fiber is anchored^{107,126}. This complex provides the platform to initiate polymerization of the MTPs gp17.1 and gp17.1* organized in a helical array forming the tail tube around the tape measure protein gp18^{107,124,125}. Gp17.1 and gp17.1* share the same N-terminus but gp17.1* has a longer carboxyl terminus resulting from a programmed translational frameshift¹²⁴. At the end of tail assembly the tail-to-head joining protein gp17 binds to the extremity of the tail tube distal from the adsorption apparatus to create the interface for interaction

with the capsid connector⁷⁶. Tail-to-head binding yields the infectious SPP1 particle.

The SPP1 virion recognizes its host cell through an interaction between the tail fiber and *B. subtilis* surface receptors¹¹⁰⁻¹¹³. Interaction of gp21 with the YueB receptor commits SPP1 to DNA ejection^{107,111,126,127} generating a signal^{75,107,128,129} that is communicated to the connector for DNA release from the phage capsid¹²³. Phage DNA is then delivered from the SPP1 capsid to the *B. subtilis* cytoplasm through the tail tube.

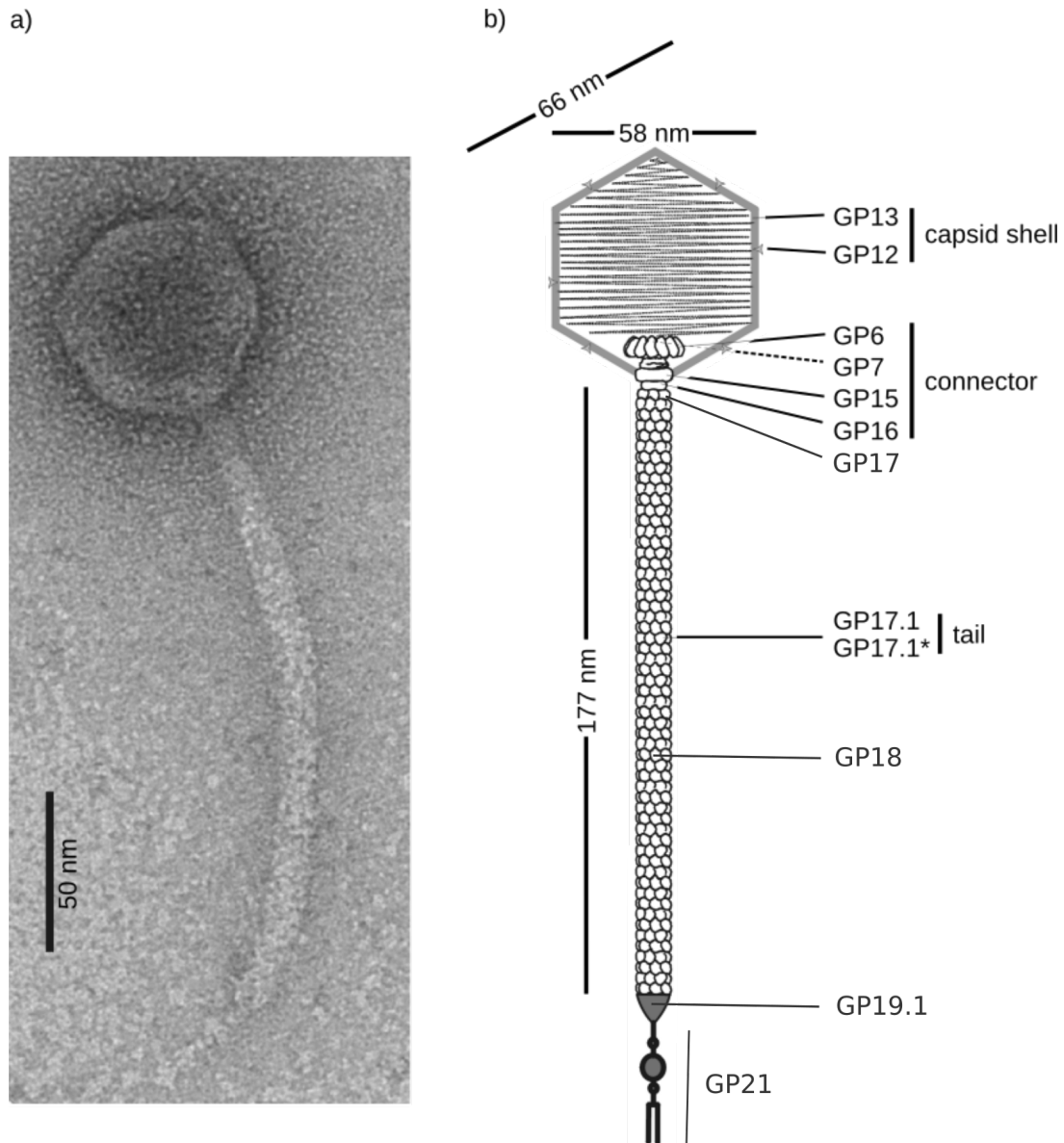


Figure 9: Bacteriophage SPP1. **a)** Visualization of a SPP1 virion by electron microscopy after negative staining with uranyl acetate. The bar represents 50 nm. **b)** The scheme of the mature phage compiles current knowledge of the particle structural organization. The icosahedral capsid is formed by the MCP gp13 and the auxiliary protein gp12. The connector complex between the capsid and the phage tail is composed of the portal protein gp6 and the head completion proteins gp15 and gp16. The location of the minor capsid protein gp7 in the mature phage capsid is not known but the protein binds gp6 at early stages of SPP1 capsid assembly. Gp17 is a tail completion protein found at the capsid-tail interface. The two major proteins of the tail are gp17.1 and gp17.1* that have an identical amino-terminus but gp17.1* has a longer carboxyl terminus¹²⁴. These two proteins form a helicoidal tube around the tape measure protein gp18. The tail tube is capped by gp19.1 that is extended by the tail fiber gp21. Adapted from Alonso et al 2006¹⁰⁶.

1.8 Gp12, the SPP1 capsid auxiliary protein: a collagen motif containing protein

The gp12 protein, subject of this study, is a 6.6kDa capsid auxiliary protein that assembles into trimers and binds at the center of the 60 hexamers of the major capsid protein gp13 (Figures 7 and 8d)^{80,130}. The function of gp12 is still not known.

The most characteristic feature of gp12 is the presence, at its central region, of a (GXY)₈ segment identified by amino acid sequence inspection¹³¹. The (GXY)_n motif (where n is the number of successive repeats, G is a glycine, and X and Y can be any amino acid) is a sequence signature of proteins from the collagen super family. The relevance of this motif for gp12 structure and behavior is addressed in this PhD thesis.

2 Introduction part II : What is collagen?

2.1 The collagen family

In the literature, both collagen and collagen-like proteins are described¹³². They are characterized by the presence of the (GXY)_n motif which leads to assembly of the intramolecular collagen triple helix (see 2.3)¹³³⁻¹³⁶. The designation of collagen is normally used for proteins that have the triple helix and are deposited in the extracellular matrix¹³². An exception are transmembrane collagen proteins that expose the triple helix to the extracellular environment^{136,137}. This extracellular portion can be released from the cell surface by shedding¹³⁶.

The collagen superfamily is a complex group of proteins with a large spectrum of functional and structural properties. In vertebrates they are widespread across the body, being the most abundant proteins in the extracellular matrix of connective tissues where they play an essential role in tissue stability. Twenty-eight different types of vertebrate collagens are known to exist. They are described by Roman numerals according to their discovery order (Table 1)¹³⁶⁻¹⁴⁰. Collagens are grouped in subfamilies based on their structural and supramolecular organization. Table 1 and Figure 10 detail those major animal collagen families¹³⁶⁻¹⁴⁰. For example, collagen types I and II belong to the group of fibril forming collagens. Members of this subfamily are characterized by their ability to assemble into fibers which are a highly oriented supramolecular aggregates, responsible for the tensile strength of tissues (Figure 10). Collagen type I is the most studied and the most abundant collagen. It is the major collagen in many tissues (skin, tendons, ligaments ...) and constitutes 90% of the organic mass of bone¹³⁵⁻¹⁴⁰. It has been shown that mutations in the sequence of collagen type I, or abnormalities on its synthesis pathway, could lead to a defective fiber assembly which alters the bone matrix properties and causes diseases such as Osteoporosis^{139,141,142}.

Besides their structural function, collagens have many functional properties.

For example, network forming collagens (Figure 10) are found in basement membranes and ensure an important filtration function (collagen IV)^{135,136,138}.

Table 1 : Vertebrate collagens. The 28 types of collagens are listed. Classes are defined according to the collagen supramolecular organization (Figure 10). Their subunit composition and isoforms are presented in the third column. The distribution in the body and pathologies resulting from defects in the collagen types presented are listed in the fourth and fifth columns, respectively. The Table compiles information from Shoulders & Raines 2009¹³⁹ and others¹⁴³⁻¹⁵⁷.

Type	Class	Composition	Distribution	Pathology
I	Fibrillar	$\alpha 1[\text{I}]_2 \alpha 2[\text{I}]$	Abundant and widespread: dermis, bone, tendon, ligament	Pathology OI, Ehlers–Danlos syndrome, osteoporosis
II	Fibrillar	$\alpha 1[\text{II}]_3$	Cartilage, vitreous	Osteoarthritis, chondrodysplasias
III	Fibrillar	$\alpha 1[\text{III}]_3$	Skin, blood vessels, intestine	Ehlers–Danlos syndrome, arterial aneurysms
IV	Network	$\alpha 1[\text{IV}]_2 \alpha 2[\text{IV}] \alpha 3[\text{IV}] \alpha 4[\text{IV}] \alpha 5[\text{IV}] \alpha 5[\text{IV}]_2 \alpha 6[\text{IV}]$	Basement membranes	Alport syndrome
V	Fibrillar	$\alpha 1[\text{V}]_3 \alpha 1[\text{V}]_2 \alpha 2[\text{V}] \alpha 1[\text{V}] \alpha 2[\text{V}] \alpha 3[\text{V}]$	Widespread: bone, dermis, cornea, placenta	Ehlers–Danlos syndrome
VI	Network	$\alpha 1[\text{V}] \alpha 2[\text{V}] \alpha 3[\text{V}] \alpha 1[\text{VI}] \alpha 2[\text{VI}] \alpha 3[\text{VI}] \alpha 1[\text{VI}] \alpha 2[\text{VI}] \alpha 4[\text{VI}]$	Widespread: bone, cartilage, cornea, dermis	Bethlem myopathy
VII	Anchoring fibrils	$\alpha 1[\text{VII}]_2 \alpha 2[\text{VII}]$	Dermis, bladder	Epidermolysis bullosa acquisita
VIII	Network	$\alpha 1[\text{VIII}]_3 \alpha 2[\text{VIII}]_3 \alpha 1[\text{VIII}]_2 \alpha 2[\text{VIII}]$	Widespread: dermis, brain, heart, kidney	Fuchs endothelia corneal dystrophy
IX	FACIT	$\alpha 1[\text{IX}] \alpha 2[\text{IX}] \alpha 3[\text{IX}]$	Cartilage, cornea, vitreous	Osteoarthritis, multiple epiphyseal dysplasia
X	Network	$\alpha 1[\text{X}]_3$	Cartilage	Chondrodysplasia
XI	Fibrillar	$\alpha 1[\text{XI}] \alpha 2[\text{XI}] \alpha 3[\text{XI}]$	Cartilage, intervertebral disc	Chondrodysplasia, osteoarthritis
XII	FACIT	$\alpha 1[\text{XII}]_3$	Dermis, tendon	Extracellular matrix-related myopathy ¹⁴⁴ , congenital myopathies ¹⁴⁵
XIII	MACIT	$\alpha 1[\text{XIII}]_3$	Endothelial cells, dermis, eye, heart,	Congenital myasthenic

			neuro-muscular junction	syndrome ^{146,147}
XIV	FACIT	$\alpha 1[XIV]_3$	Widespread: bone, dermis, cartilage	
XV	MULTIPLEXIN	$\alpha 1[XV]_3$	Capillaries, testis, kidney, heart	
XVI	FACIT	$\alpha 1[XVI]_3$	Dermis, kidney	
XVII	MACIT	$\alpha 1[XVII]_3$	Hemidesmosomes in epithelia	Generalized atrophic epidermolysis bullosa
XVIII	MULTIPLEXIN	$\alpha 1[XVIII]_3$	Basement membrane, liver	Knobloch syndrome
XIX	FACIT	$\alpha 1[XIX]_3$	Basement membrane, hippocampal synapses ¹⁴⁸	
XX	FACIT	$\alpha 1[XX]_3$	Cornea (chick)	Striate palmoplantar keratoderma ¹⁴⁹
XXI	FACIT	$\alpha 1[XXI]_3$	Stomach, kidney	
XXII	FACIT	$\alpha 1[XXII]_3$	Tissue junctions	Intracranial aneurysms ¹⁵⁰
XXIII	MACIT	$\alpha 1[XXIII]_3$	Heart, retina	Cancer ¹⁵¹
XXIV	Fibrillar	$\alpha 1[XXIV]_3$	Bone, cornea	
XXV	MACIT	$\alpha 1[XXV]_3$	Brain, heart, testis, amyloid formation	Alzheimer, congenital cranial dysinnervation disorder ¹⁵² , congenital ptosis ¹⁵³
XXVI	FACIT	$\alpha 1[XXVI]_3$	Testis, ovary	
XXVII	Fibrillar	$\alpha 1[XXVII]_3$	Cartilage	Steel syndrome ^{154,155}
XXVIII	-	$\alpha 1[XXVIII]_3$	Dermis, sciatic nerve, Ranvier nodes	Neurodegenerative disease

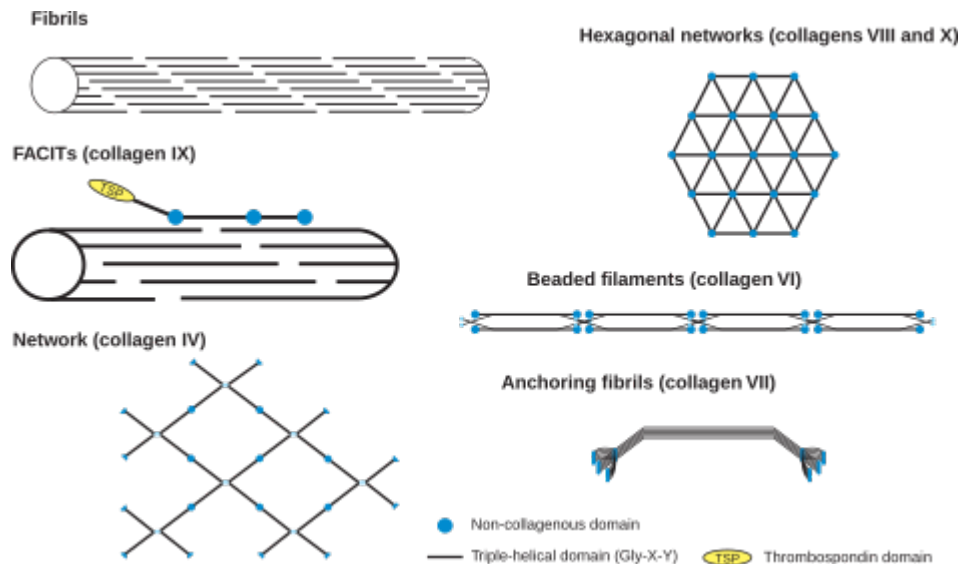


Figure 10: Supramolecular assemblies formed by collagens. The structures built by collagen molecules from different classes (Table 1) are presented schematically. The Figure is from Ricard-Blum, 2011¹³⁶.

There are also a large number of animal proteins having a triple helix segment in addition to the collagens. They are designated “collagen-like proteins” which include secreted and membrane proteins with an extracellular triple helix. Their functions are diverse¹³⁶. For example, transmembrane proteins such as the macrophage scavenger receptors class A and B are essential for efficient target binding¹⁵⁸. Secreted collagen-like proteins include lectins, emilins and complement proteins¹³⁶.

2.2 Prokaryotic Collagen Motif Containing Proteins (CMCPs)

The revolution raised by high throughput genome sequencing and computing power provided scientists with a huge amount of genomic data. Genes coding for proteins containing a $(GXY)_n$ repeated pattern were found in numerous prokaryotes^{159,160}. The absence of the postranslational modifications that are essential for animal collagen stability (see section 2.3) and poor understanding of their function in prokaryotes pushed scientists to investigate the structural organization of those proteins, their stability and their function.

Over the past 15 years, many bacterial collagen-like proteins were discovered

(Figure 11). Studies of virulent bacteria showed a putative role of those proteins in host recognition and cell adhesion¹⁶¹. The Streptococcal collagen-like proteins (Scl) are one of the best characterized collagen-like proteins in bacteria. The Scl family is a group of transmembrane collagen motif containing proteins whose most studied members are Scl1 and Scl2¹⁶²⁻¹⁶⁴. The collagen motif located in the central region of the proteins is exposed at the surface of the bacteria and is attached to the membrane via a transmembrane domain. Biochemical and biophysical studies on those proteins confirmed the triple helical nature of the (GXY)_n segment. The collagenous segment is stable with a melting temperature of 36.4 and 37.6°C for Scl1 and Scl2, respectively¹⁶³. Those values are remarkably close to the ones observed for animal collagen^{165,166}. In order to understand the function of Scl1 and Scl2, Oliver-Kozup et al. studied mutants lacking those proteins showing their implication in cell surface adhesion and biofilm formation¹⁶⁷.

Collagen-like proteins were also reported to be exposed at the surface of *Bacillus anthracis* spores, the causative agent of anthrax. BclA, the collagen-like protein of *B. anthracis* (BclA) forms the external hair-like nap filaments of the spore¹⁶⁸. Its collagenous segments occupy the central region of the protein and, depending on the bacterial strain, the number of successive GXY repeats varies from 17 to 91¹⁶⁸. The thermal stability of BclA exhibits unusual values around 90°C, which is extremely high for a collagen-like protein.

Based on sequence analysis, proteins with (GXY)_n repeats were also described in phages, prophages, and other viruses. Their role is not as clear as in bacteria. However, it is believed to be related to host recognition and surface attachment^{159,160}. Ghosh et al. identified several genes coding for collagen-like proteins¹⁵⁹. Those genes are present in the enterohaemorrhagic *E. coli* O157:H7 strain, but absent in the non-pathogenic K-12 strain. They are encoded by prophages. The proteins studied (Figure 11b) have the triple helical structure of collagen and are stable with a melting temperature slightly higher than vertebrate collagens¹⁵⁹. One or two collagenous segments are found per protein. They are coupled to a carboxyl-terminal trimerisation domain and to phage tail related domains. These observations led the authors to suggest

that those proteins are a component of tail fibers that help host recognition and cell adhesion.

a)

Bacterium	Gene	N-terminal domain*	Collagen domain*	C-terminal domain*	Calculated pI	Tm (°C) (CL domain)	Triple-helix validation
Streptococcus pyogenes	SclA/ Scl1	68	150	93	5.1	36.4	CD and trypsin
Streptococcus pyogenes	SclB/ Scl2	74	237	100	5.4	37.6	CD and trypsin
Bacillus anthracis	wt BclA	19	228	134	3.1	37.0	CD and trypsin
Legionella pneumophila	Lcl	n.d.	105	n.d.	5.3	n.d.	Trypsin
Clostridium perfringens		53	189	162	4.7	38.8	CD and trypsin
Solibacter usitatus		42	246	147	5.6	38.5	CD and trypsin
Rhodopseudomonas palustris		9	117	86	9.3	37.0	CD and trypsin
Methylobacterium sp4-46		102	147	74	8.6	35.0	CD and trypsin

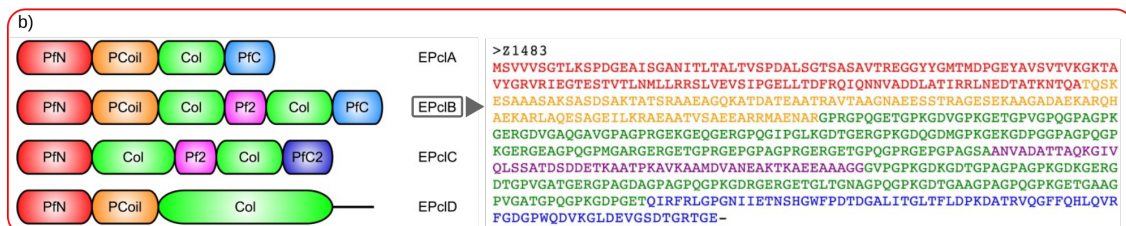


Figure 11: Bacterial collagen-like proteins. a) Properties of some bacterial collagen-like proteins¹⁶⁰. * number of amino acid of each domain. CD : circular dichroism profile signature of collagen. Trypsin : validation by resistance of the triple helix to trypsin. b) Collagen-like proteins from prophages embedded in the genomes of *E. coli* O157:H7 and other EHEC strains, EPclA to EPclD (EHEC Prophage collagen-like A to D)¹⁵⁹. The collagen triple helical domains are labelled “Col”, and domains predicted to adopt an α -helical coiled-coil conformation are labelled “PCoil” (for phage coiled-coils). Key to other domain labels : PfN, phage fiber N-terminal domain; PFC, phage fiber C-terminal domain; PFC2, phage fiber C-terminal domain, variant 2; Pf2, phage fiber repeat 2. The sequence on the right is a representative collagen-like protein with EPclB architecture (Z1483), from the genome of *E. coli* O157:H7 EDL933. Amino acid sequences corresponding to the different predicted domains are colour-coded as in the left panel. From Ghosh et al.¹⁵⁹.

Putative collagen-like proteins were also described as a part of the phage capsid. Bamford and colleagues identified short collagenous segments in the central region of the P5 protein from tectivirus PRD1^{169,170}. P5 is a multi-domain protein with a carboxyl-terminal trimerisation domain, a short collagen-like spacer, and an N-terminal domain that binds to the capsid pentameric vertex. However, no evidence was provided that its short GXY repeats fold into a collagen helix. Putative long collagen-like spikes arise also from the Mimivirus capsid. They are exposed at the surface of this giant virus particule that infects eukaryotic amoeba¹⁷¹. In contrast to all studied collagen-like from prokaryotes, two enzymes modifying collagen were identified in Mimiviruses: lysyl hydroxylase and glycosyltransferase¹⁷².

Prokaryotic collagen-like proteins show several common properties when compared to animal collagens such as the triple helix formation conferred by the GXY repeats and its associated properties like thermal stability, trypsin resistance and presence of a trimerization domain (Table 2). However, prokaryotic proteins have shorter successive GXY repeats and lack hydroxyproline, using alternative mechanisms for triple helix stabilization (Table 2).

Table 2: Comparison of bacterial and phage collagen-like proteins that have been characterized with mammalian collagens. Similarities and differences of their sequences as well as biophysical and biochemical properties are listed. Table is adapted from Yu et al. 2014¹⁶¹.

		Mammalian collagens	Bacterial collagens
Similarities	Gly-Xaa-Yaa repeats	Yes	Yes
	Triple-helix	Yes	Yes
	Trypsin resistance	Yes	Yes
	Thermal stability of collagen-like region	$T_m = \sim 37^\circ\text{C}$	$T_m = \sim 35\text{-}39^\circ\text{C}$
	Calorimetric enthalpy	Very high	Relatively high, but lower than mammalian collagen
	Trimerization domain	At either end	At either end
	Interruptions	In some of them	In putative phage fiber proteins
Differences	Length	~ 350 triplets for fibrillar collagens	$\sim 35\text{-}95$ triplets (characterized so far)
	Hydroxyproline	Yes	No
	Heterotrimer	Some of them	Not found yet
	Sequence and stabilization	Pro/Hyp rich, important for stabilization	Strategies include electrostatic interactions; glycosylation of Thr residues; very high Pro
	Types	28 different types in human	Species dependent – can be many species variants
	Fibril formation	<i>In vivo</i>	Probably limited

2.3 Structural organization of collagen

The nature of collagen was first studied using X-ray fiber diffraction. Its structural organization was established later at atomic resolution using X-ray crystallography of collagen peptides¹⁷³⁻¹⁷⁶. The collagen super helix (Figure 12) is an

elongated right-handed helix built from parallel polypeptide chains shifted by one amino-acid¹⁷⁵⁻¹⁷⁸. Each monomer is a succession of GXY motifs folded in a left-handed polyproline II helix (3 residues/turn). The three polypeptide chains twist together, cooperatively, to form the collagen right-handed helix^{179,180}. The protein sequence of subunits can be identical (homotrimer) or different (heterotrimer). Both cases were found in animal collagen¹⁷⁵. Because of the tight structure packing, the residues backbone is exposed to the solvent. The collagen motif is stabilized by various internal and external mechanisms^{176,181-183}. First, an interchain hydrogen bond is formed between the carboxyl group of the glycine (the first position of the GXY motif) and the amino group of the corresponding glycine from the opposite polypeptide chain. Second, side chains of amino acids in positions X and Y could interact together forming extra electrostatic bonds around the triple helix. Finally, a hydration shell is organized around the protein that stabilizes the all structure (Figure 12).

The nature of amino acids at positions X and Y within the GXY motif and the number of successive repeats are important features that affect the thermal stability of collagen¹⁸²⁻¹⁸⁵. In eukaryotes, the amino acid at position Y is most often a proline post-translationally modified to a hydroxyproline by the Prolyl-4-hydroxylase^{135-138,141,177}. The proline coupled hydroxyl group provides additional hydrogen bonding that stabilizes the collagen structure. Inhibition of hydroxylation leads to a decrease of the collagen thermal stability of around 15°C¹⁸⁶. Prolyl-3-hydroxylase is also known to exist which modifies the proline at the X position to 3-hydroxyproline¹⁷⁷.

Hydroxyproline is absent in prokaryotes due to the lack of these post-translational modifications^{175,187}. However, prokaryotic collagen-like proteins are stable and have melting temperatures usually close to animal collagens. Compared to animal collagen, collagen-like proteins have a higher amount of charged residues occupying the X and Y positions of the GXY motif¹⁸⁷. Thus, electrostatic interactions between side chains probably compensate the lack of hydroxylation.

In nature, the collagen motif can be made of identical or different subunits (homotrimers and heterotrimers, respectively). Usually, one or multiple collagenous segments are associated to other domains. Those non-collagenous domains play

distinct roles. Some of them are trimerization domains that promote and ensure the correct folding of the collagen motif by bringing together components of the trimer with the appropriate sequence and stoichiometry¹⁸⁸. In vertebrates, the collagen triple helix nucleates and then assembles in a zipper-like way from the carboxyl to the amino terminal domain. The exception is transmembrane collagens that nucleate and assemble from the membrane associated amino terminal domain to the surface exposed carboxyl-terminal domain in a zipper-like way¹³⁶. The most frequent collagen coupled structural elements are α -helical coiled-coils^{188,189} that act as a trimerization domain for both collagen and collagen-like proteins^{175,188}.

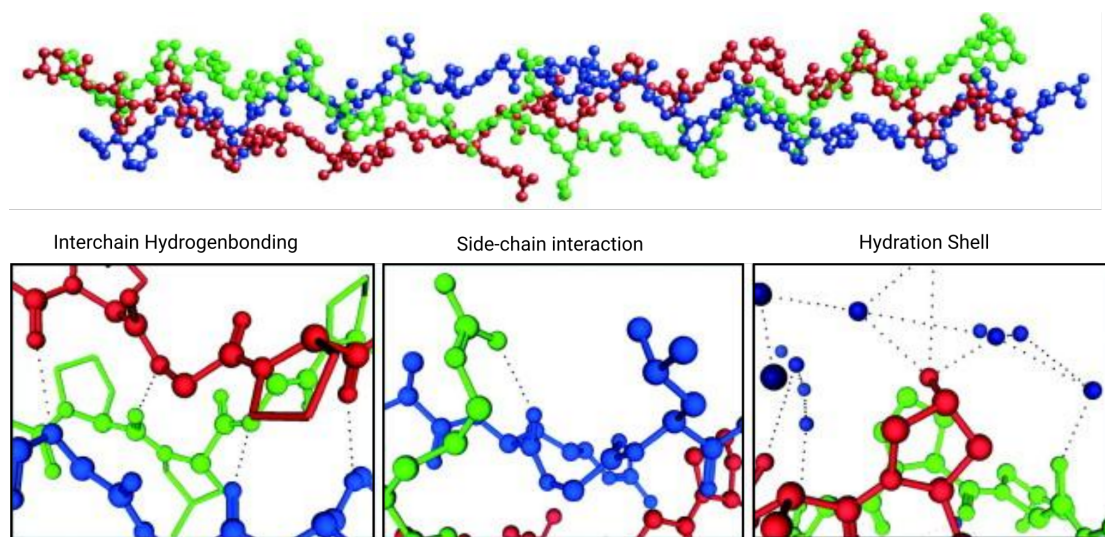


Figure 12: Structure of a collagen triple helix formed by a polyproline model peptide. From Brodsky and Persikov¹⁷⁵.

3 Thesis Aims

The aim of this PhD thesis is the study of the biochemical and structural properties of protein gp12 from bacteriophage SPP1. Inspection of the gp12 amino acid sequence had previously revealed the presence of 8 consecutive repeats of the motif GXY in the center of this small SPP1 capsid protein¹³¹. We have first studied the quaternary structure of gp12 and investigated experimentally if the protein has a collagen-like fold resulting from its (GXY)₈ motif. We have then analyzed the gp12 folding and association in solution to compare its behavior with other collagen-like proteins. This analysis provided the experimental conditions to characterize the reversible interaction of unfolded and of trimeric gp12 with the SPP1 capsid. We then used bioinformatics to identify gp12-related proteins in other phages. A search of the collagen motif in prokaryotes and viruses showing its widespread distribution in non-eucaryotes concludes the PhD thesis.

4 Material and Methods

4.1 Cloning procedures and creation of the gene 12 knock-out SPP1 strain

Gene 12 was cloned into plasmid pRSET A (Invitrogen) for protein overproduction in *Escherichia coli* using the strategy described by Lurz et al¹⁹⁰. The resulting plasmid pBT453 codes for tag-gp12 in which the gp12 amino terminus is fused in-frame to a 36 amino acid-long tag that includes a hexahistidine sequence. In order to engineer a cleavage site for Tobacco Etch Virus (TEV) protease (ENLYFQG) between the tag and the gp12 amino acid sequence, gene 12 was amplified from pBT453 DNA with oligonucleotides TGS (TAAGGTACCGGATCCGAGAATCTGTACTTCCAGGGCATGTCTAAGCGTATAC **CGCGTTTCTTGC**; the BamHI site is underlined; the sequence coding for a TEV protease cleavage site is in italics and the beginning of gene 12 is in bold) and TGA (ATACTCGAGTACCAGCTGCAGTTATTAAGTCGTTCC; the PstI site is underlined; the gene 12 complementary coding sequence is in bold and stop codons double underlined). The PCR fragment was then cleaved with *Bam*HI-PstI and cloned into pRSET A generating pMZ1.

Plasmid pBT450 was constructed in two steps. First a *S*fiI fragment bearing genes 15 and 16 of SPP1 (coordinates 8830 to 9787 of the SPP1 genome sequence; GenBank accession number X97918¹⁹¹) was treated with Klenow fragment to produce blunt ends and cloned in the *S*maI site of pBluescript SK- (Stratagene). Secondly, the resulting plasmid was used to clone a *N*ruI-EarI blunt-ended fragment (coordinates 6699 to 8778 of the SPP1 sequence), bearing genes 11 to 13, in the *H*incII site of the pBluescript SK- polylinker. The cloning strategy generated a polycistronic unit composed of SPP1 genes 11 to 13 and 15 to 16 under the control of a T7 promoter. A DNA fragment containing gene 11 and the beginning of gene 12 was produced by

cleavage of pBT450 with BglI, treated with T4 polymerase to produce blunt ends and digested with Asp718. This fragment was cloned into pBT450 digested with SmaI and KpnI to generate pBT451 in which gene 12 is disrupted by an out-of frame deletion between its internal BglI and SmaI sites (coordinates 7618 to 7646 of the SPP1 sequence). A PstI-XhoI fragment of pBT451 spanning genes 11 to 13 that flank the gene 12 knock-out deletion was then transferred into the *E. coli-B. subtilis* shuttle vector pHP13¹⁹² cut with PstI-SalI to yield pBT452.

The non-permissive strain *B. subtilis* YB886 (pBT452) was infected with a mutant phage carrying conditional lethal mutations in genes 11 and 13 (SPP1*sus7sus31*¹⁹³) forcing a double crossover that leads to integration of the knock-out mutation in gene 12 of the viral genome, as confirmed by DNA sequencing. The resulting phage SPP1*del12* was crossed with SPP1*sus9*, a mutant defective in a tail gene^{80,192}, to yield SPP1*sus9del12*⁸⁰. SPP1*del12* was used to produce SPP1 infectious particles lacking gp12 while SPP1*sus9* and SPP1*sus9del12* were used to produce tailless DNA-filled capsids with (capsids H) and without gp12 (capsids HΔ12), respectively.

4.2 Production and purification of SPP1 virions and DNA-filled capsids

Procapsids, DNA-filled capsids and viral particles were produced and purified as described^{112,194}. Procapsids were kept in buffer R (50 mM potassium glutamate, 10 mM EDTA, 50 mM Hepes-KOH, pH 7.6, 1 mM PMSF freshly added¹⁹⁴) while all other structures were stably stored and manipulated in TBT buffer (100 mM NaCl, 10 mM MgCl₂, 100 mM Tris-Cl pH 7.5). All interactions of tag-gp12/gp12 with viral structures were carried out in TBT.

The concentration of capsid physical particles was estimated based on their DNA content. Ultraviolet absorbance spectra of capsid suspensions were used to assess sample purity and the value at 260 nm to determine DNA concentration. This value was then used to calculate the concentration of capsid physical particles according the

following equation:

$$T = (c \cdot N_A) / (n_{bp} \times 660)$$

where T is the concentration of physical particles/L, c is the DNA concentration in g/L, N_A is the Avogadro constant and n_{bp} is the number of DNA base-pairs. The SPP1 packaged molecules were considered to have an average length of 45.9 kbp⁸⁰.

4.3 Production and purification of tag-gp12

Tag-gp12 was overproduced in *E. coli* BL21 (DE3) (pBT453). Cells were grown at 37°C in LB medium supplemented with 100 µg/mL ampicillin. An overnight culture was diluted 50-fold, grown to an optical density at 600 nm between 0.6 and 0.8, induced with IPTG to a final concentration of 1 mM and shaken for 3 h. Cells were harvested (30,000 g, 30 min, 4°C), resuspended in buffer A (500 mM NaCl, 10 mM imidazole, 50 mM NaH₂PO₄ pH 8.0) supplemented with a protease inhibitors cocktail (Complete™ EDTA-free, Roche Applied Science, Mannheim, Germany) and disrupted by sonication on ice using three cycles of two minutes each, spaced by two minutes pauses (Vibra Cell 72405, Fisher Bioblock, Illkirch, amplitude 60, pulse 3, 30-40 W). The total soluble proteins extract obtained after centrifugation (30,000 g, 1 h, 4°C) was filtered through a 0.22 µm membrane. The filtrate was then loaded on a 5 mL HisTrap™ HP metal affinity column (GE Healthcare Bio-Sciences AB Uppsala, Sweden) coupled to an ÄKTA purification system (GE Healthcare Bio-Sciences AB Uppsala, Sweden). A three steps gradient was applied at 16°C: 2 % buffer B (500 mM NaCl, 500 mM imidazole, 50 mM NaH₂PO₄ pH 8.0) for a first wash, 10 % buffer B for a second wash and 100 % buffer B for elution. The tag-gp12 peak fractions were pooled and run through a preparative size exclusion chromatography column (HiLoad 26/60 Superdex™ 200pg, GE Healthcare Bio-Sciences AB Uppsala, Sweden) pre-equilibrated in buffer C (500 mM NaCl, 50 mM Na₂HPO₄, pH 8.0) at 16°C coupled to an ÄKTA purification system. Aggregates and contaminants were found mostly in the void volume while tag-gp12 eluted as a single peak. Tag-gp12 was obtained at a yield of 3 mg/g wet cell weight and was more than 95% pure as judged from SDS-PAGE

analysis. Purified protein was stored in buffer C and dialysed against other buffers immediately before use. Protein concentration was estimated using the Bio-Rad Protein Assay, following the manufacturer's instructions. Tag-gp12 was used to immunise rabbits following the protocols established for protein (pC)CAT¹⁹⁵ to obtain anti-tag-gp12 polyclonal serum.

TagTEV-gp12 was produced and purified according to the same protocol. The purified protein was then incubated at 16°C for 4 h with TEV protease at a ratio of 1:20 (w/w). In order to remove the tag, the digestion product was loaded to a 1 mL HisTrapTM HP metal affinity column. Gp12 was eluted with 80 mM imidazole. The tag and the TEV protease eluted at 500 mM imidazole. The purified gp12 carries an additional glycine at its amino terminus preceding the initial methionine residue.

4.4 Mass spectrometry

The collagenase digestion of tag-gp12 followed by trypsination was stopped by adding solid guanidine-hydrochloride to a final concentration of 6 M followed by incubation at 90°C for 15 min. Peptides were precipitated at -20°C over weekend by adding 5 volumes of cold acetone. Peptides were recovered by centrifugation, dried and resuspended in ammonium carbonate at 1 µg/µL. They were then analysed by MALDI-TOF and NanoLC-MS/MS.

MALDI Peptide Mass Fingerprinting (PMF): peptides (0.5 µl) were mixed with an equal volume of either α -cyano-4-hydroxycinnamic acid (10 mg/ml, 50% CH₃CN; Sigma-Aldrich) or 2,5-dihydroxybenzoic acid (10 mg/ml, 20 % CH₃CN; Sigma-Aldrich). Peptide mixtures were analyzed by MALDI-TOF (Voyager-DESTR, Applied Biosystems) after external calibration. Crystals were obtained using the dried droplet method, and 500 MALDI mass spectra were averaged per spot. Mass spectrometry measurements were carried out at a maximum accelerating potential of 20 kV, in the positive reflectron mode. Peak lists were generated by the Data Explorer software (Applied Biosystems), and processed data were submitted to the FindPept tool (available on ExPASy portal) using the following parameters: data bank gp12 protein;

mass tolerance, 300 ppm; digest reagents, none.

NanoLC-ESI-MS/MS analyses: the peptide mixture was then analyzed in a Q/TOF Premier mass spectrometer (Waters) coupled to the nanoRSLC chromatography (Dionex) equipped with a trap column (Acclaim PepMap100 C18, 75 μ m I.D. \times 2 cm, 3 μ m, nanoViper) and an analytical column (Acclaim PepMapRSLC C18, 75 μ m I.D. \times 15 cm, 2 μ m, 100 \AA , nanoViper). The loading buffer was H₂O/CH₃CN/TFA (98 % / 2 % / 0.05 %), buffer A and B were H₂O/HCOOH (0.1 %) and CH₃CN/HCOOH (0.1 %), respectively. A 2–50 % B gradient was set for 40 minutes with a flow rate of 300 nL/min. Data-dependent scanning was applied to generate MS/MS spectra with a collision energy ramp of 15 to 40 volts. Standard MS/MS acquisitions were performed on the top of the three most intense parent ions of the previous MS scan. Raw data were processed with ProteinLynx Global Server (Waters). Peptide identification was achieved using the Mascot software with the following parameters: data bank gp12 protein; peptide tolerance 15 ppm; fragment tolerance 0.1 Da; digest reagent none.

4.5 Digestion of tag-gp12 with collagenase

Collagenase VII from *Clostridium histolyticum* (8.8 U/mg) was purchased from Sigma-Aldrich. A stock solution was prepared at 1 mg/mL in buffer D (250 mM NaCl, 10 mM CaCl₂, 10 mM 2-mercaptoethanol, 20 mM HEPES-Na, pH 7.6) and diluted 10-fold before use. Tag-gp12 (50 μ g) was digested with 0.23 μ g of collagenase for 4 h at 16°C. The same result was obtained by digestion for 30 min at 37°C. Digestion products were analysed on SDS-PAGE gel stained with Coomassie Blue and by mass spectrometry as described above.

4.6 Analytical size exclusion chromatography (SEC)

100 μ L of purified tag-gp12 at 2 mg/mL was run at 16°C using a flow of 0.5 mL/min on a SuperdexTM 200 10/300 GL (GE Healthcare Bio-Sciences AB Uppsala, Sweden) column equilibrated in buffer C (500 mM NaCl, 50 mM Na₂HPO₄ pH 8.0) and

coupled to an ÄKTA purification system. Column calibration and Stokes radius estimation were carried out as described¹¹⁵.

4.7 Analytical ultracentrifugation

Analytical ultracentrifugation was carried out on a Beckman Optima XL-A ultracentrifuge (Beckman Coulter, Palo Alto, USA) equipped with 12 mm central cells on an AN-60 Ti rotor. Runs were performed in buffer C at 16°C and monitored by absorption at 280 nm. The tag-gp12 partial specific volume (0.7051 mL/g), buffer C solvent density (1.02647 g/ml), and solvent viscosity (1.1913 cpoise) were calculated using the SEDNTERP software¹⁹⁶.

Sedimentation velocity runs were carried out at a rotor speed of 220,000 *g* using a protein loading concentration of 0.5, 1 and 2 mg/mL. Data points were recorded every 5 min and analysed using the SEDFIT software assuming a non-interacting species model.

Equilibrium sedimentation was performed at 16,300 *g* using protein loading concentrations of 0.3, 0.5 and 0.8 mg/mL. Data were analysed using SEDFAT for average mass determination.

4.8 Circular dichroism (CD) measurements

CD measurements were carried out on a Jasco J810 spectropolarimeter equipped with a Peltier temperature controller. The protein, at 2 mg/mL, was dialysed against either buffer C or TBT buffer and loaded to a 0.1 mm path-length quartz cell. Spectra at fixed temperatures were recorded at equilibrium between 190 and 260 nm every 0.2 nm using a bandwidth of 1 nm and a scanning speed of 20 nm/min. Each spectrum was an accumulation of 5 spectra after baseline correction using the buffer spectrum as blank. Thermal transition profiles were monitored at 200 nm with a 1°C/min heating rate and a protein concentration of 2 mg/mL. One point was recorded for each 1°C. Temperature was raised from 10 to 60°C, kept at 60°C for 30 min and

finally returned back to 10°C using the same rate. Ellipticity was first measured and molar ellipticity was then calculated using the following equation:

$$[\Theta] = (\Theta \times 100M)/(c \times l)$$

where Θ is the ellipticity in degrees, M is the molecular mass, c is the protein concentration in mg/mL and l is the path-length in cm.

T_{ms} were determined from data plots as the transition mid-point.

4.9 Fluorescence-based thermal shift assay (FBTSA)

FBTSA experiments were realised as described previously⁸⁰. In brief, SPP1 virions or capsids (5.7×10^{10} particles) and/or purified tag-gp12 at different concentrations were mixed with diluted SYPRO orange dye (400-fold diluted from stock solution, Invitrogen) in TBT buffer to a final volume of 10 μ L. Experiments were carried out in real-time PCR systems and fluorescence was recorded in real time. Different heating-cooling cycles were applied to the samples as described in figure legends and in the Results Sections. Experiments were carried out in an ABI 7900HT and QuantStudio 12KFlex machines (Applied Biosystems) as detailed in figure legends. The fluorescence profiles, derivatives and T_{ms} were determined using the manufacturer analyses software.

4.10 Gp12 chimerisation experiments

A 2-fold molar excess of purified gp12 was mixed with tag-gp12, and kept either at 16°C or heated for 15 min at 60°C. Mixtures were then loaded to a 1 mL metal affinity column (HisTrapTM HP metal affinity column GE Healthcare Bio-Sciences AB Uppsala, Sweden) and proteins were eluted by applying a step gradient of imidazole concentration as used for tag-gp12 purification (see above).

4.11 Binding of tag-gp12 to SPP1del12 virions

100 μ L of SPP1del12 virions at 5.6×10^{12} pfu/mL were mixed with a 10-fold molar excess of tag-gp12 protein (550 μ L of a 2 mg/mL solution), considering 60 binding sites for tag-gp12 trimers *per* capsid⁸⁰, and incubated overnight at 16°C. The mix was then run through a discontinuous caesium chloride gradient to purify phage particles¹¹² and their protein composition was analysed by western blot with rabbit polyclonal antibodies raised against purified tag-gp12 or against SPP1 purified virions.

4.12 Binding of tag-gp12 to H Δ 12 capsids analysed by a trypsin protection assay or FBTSA

Tag-gp12 was dialysed against TBT. A range of tag-gp12 concentrations and purified SPP1 capsids lacking gp12 (H Δ 12) were incubated separately for 1 h at 16 or 45°C in a PCR machine (Biometra Tprofessional TRIO Thermocycler). A constant number of H Δ 12 was then mixed with variable amounts of tag-gp12 to obtain different molar ratios and incubated at 16 or 45°C for the desired reaction time according to the experimental scheme on the left of Figure 20 (section 5.6). Samples were analysed by FBTSA or incubated with 1 μ g of trypsin at 45°C for 30 min to proteolyse free tag-gp12 (not associated to capsids). Care was taken to avoid any cooling below 45°C for samples whose mixtures were incubated at this temperature before FBTSA or trypsination. This was necessary to prevent rapid refolding/reassociation of free tag-gp12 which would facilitate assembly of free trimers and their binding to capsids. Capsids in trypsinated samples were separated on 0.8 % agarose gels prepared in TAMg buffer (1 mM MgCl₂, 40 mM Tris-acetate, pH 8.3). The running buffer was TAMg and the applied electric field intensity was 70 mA. Gels were stained with ethidium bromide in TAE buffer (40 mM Tris-Acetate supplemented with 1 mM EDTA) in which EDTA leads to disruption of capsids *in situ* rendering viral DNA accessible to ethidium bromide binding, a more sensitive detection method than protein staining with Coomassie Blue.

4.13 Search for gp12 close relatives and sequence based analysis

The gp12 amino acid sequence was submitted to the NCBI online pBLAST protein similarity protein detection tool¹⁹⁷ (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) and run against the non-redundant protein sequences (nr) database (January 2013). A cut-off E-value of 10^{-4} was used to retain specific hits. Those hits constitute the gp12 close relatives. Each hit sequence was then submitted to a pBLAST similarity search using the same criteria as above. The new hits were pooled, constituting the gp12 non-direct relatives. All protein sequences were downloaded on Fasta format and submitted to ClustalW¹⁹⁸ for multiple sequence alignment for detection of conserved segments.

4.14 Genomic context analysis of genes coding gp12 relatives

The protein hits obtained in the NCBI pBLAST searches for gp12 relatives (section 4.13) provide a link to access the protein formatted data of each individual protein. This page provides the access to the complete genome data of the source organism coding for the protein as a GenBank formatted data. The record contains all gene products annotated in the genome sequence including the gene start position, end position, predicted protein sequence and coding strand.

We followed those steps for each gp12 relative and extracted 10 Coding DNA Sequences (CDSs) before the target gp12-like protein gene close and 10 CDSs after. The extracted protein sequences of the CDSs were submitted to NCBI protein to protein BLAST tools which automatically submit it for conserved domain detection (CDD). Based on those results we could annotate putative function(s) for the gene products in the neighbourhood of the gp12-like protein gene to establish its genome context.

4.15 Troglodyte, the CMCPs detection tools

Troglodyte is a software package developed during my PhD thesis for large scale protein database mining in order to search for collagen motif containing proteins. It combines a set of modules that download, parse, index and scan protein sequences for detection of repeated GXY motif.

Those tools were written using the Java programming language. Additional open source frameworks used and are: eUtils, used to access NCBI web services; Apache Derby, an open source relational database; Apache Commons Net, a library that implements the client side of many basic Internet protocols; and PDF Clown, an implementation of the Portable Document Format.

We used the NCBI's Reference Sequence (RefSeq) database as a target for protein data mining. RefSeq is a public available collection of DNA, RNA and protein sequences from distinct life forms. The main characteristic features of the RefSeq database are non-redundancy, rich annotation, diversity, regular updates and curation by NCBI staff and their collaborators. Hence, it constitutes a suitable target for genomic data mining and comparative protein analyses. Archaea, Bacteria and Viral releases were selected and genomic archives were downloaded from the NCBI FTP server (<ftp://ftp.ncbi.nlm.nih.gov/refseq/release/>) in July 15, 2014 (release 66). The Viral release is a complex collection of genomic data from different viruses taxonomic groups that infect different Domains of life. An exhaustive table of viruses and their hosts was downloaded from the NCBI Viral Genome Browser (<http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239>). The genome accession id was used to identify, extract and create two additional taxonomic groups: bacterial viruses and archaeal viruses.

Protein products of CDSs were parsed from genomic GenBank files and saved into a Fasta formatted file. Fasta headers were created following NCBI recommendations but included an additional local auto-generated unique id. CDSs records were indexed into a local database table to allow easy access to their references (unique id, gi, taxon). A unique Fasta file was generated per taxonomic group. In order

to keep only representative sequences, Fasta files were then curated using the CD-HIT clustering program applying a sequence identity threshold below (-c) 0.8 and an alignment coverage for the longer sequence (-aL) of 0.8.

Finally, 5 non-redundant protein sets were generated to cover 5 taxonomic groups: archaea, archaeal viruses, bacteria, bacterial viruses, and viruses.

CMCPs are marked by the presence of a (GXY) n motif. The first amino acid in the triplet is always a glycine (G). X and Y can be any amino acid. The number of successive repeat is defined by the n value that varies from one CMCP to another. We did not impose a limit for the minimum number of (GXY) repeats in the search for CMCPs. Non-redundant protein data sets were searched for a (GXY) n_{\min} motif, where n_{\min} is the minimum repeat number to consider a hit as positive. The searched n_{\min} range was of 1 to 100. At each iteration, the number of total positive records was counted (or saved to an external Fasta file) and the n_{\min} value was incremented by 1. (GXY) n_{\min} was then plotted according to the frequency of the CMCPs in the different data sets to assess the distribution and length of CMCPs in the taxonomic groups analysed.

5 Results

5.1 Gp12 has a collagen-like sequence motif

The SPP1 capsid auxiliary protein gp12 is a 64 amino acid-long polypeptide with a molecular mass of 6613 Da and a theoretical isoelectric point of 8.14. Its carboxyl terminus is predicted to form α -helices while the central part features 8 GXY repeats (Figure 13)¹³¹. The repeated GXY motif is a sequence signature of collagen-like proteins in which three polypeptides are brought together to form an intramolecular left-handed triple helix^{181,183}.

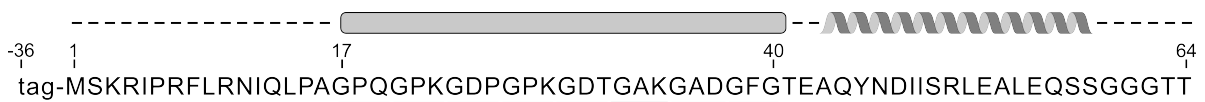


Figure 13: Gp12 amino acid sequence and position of the tag fused to its amino terminus. GXY triplets are underlined. The intramolecular collagen-like triple helix and alpha-helix predicted by bioinformatics are shown above the sequence.

5.2 Gp12 is an elongated trimer in solution

The gp12 amino terminus was fused to a 36 amino acid-long peptide including a hexahistidine tag to enhance protein production and for easy purification. The 10.7 kDa recombinant protein (tag-gp12) eluted from a Superdex 200 analytical SEC column as a single symmetric peak (Figure 14a). Its hydrodynamic radius (R_H) based on a protein calibration data set was 35 Å. The gp12 elongated shape observed in electron microscopy reconstructions of the bacteriophage SPP1 capsid⁸⁰ rendered SEC not suitable to estimate its native mass¹⁹⁹. The shape and oligomerisation state of tag-gp12 were thus investigated by analytical ultracentrifugation at 16 °C. Tag-gp12 behaved as a homogeneous species with a sedimentation coefficient of 1.7 S ($s_{20,w}=2.2 S$) (Figure 14b) at all loading concentrations tested in sedimentation velocity experiments (0.5 to 2

mg/mL). Sedimentation equilibrium centrifugation was then used for shape-independent measurement of the tag-gp12 mass (Figure 14c). The determined molecular mass ($31,270 \pm 590$ Da) was only 3 % lower than the theoretical mass of a tag-gp12 trimer. Using this experimental value and the sedimentation coefficient, we calculated a friction ratio (f/f_0) of 1.79 showing that tag-gp12 is an elongated trimer in solution.

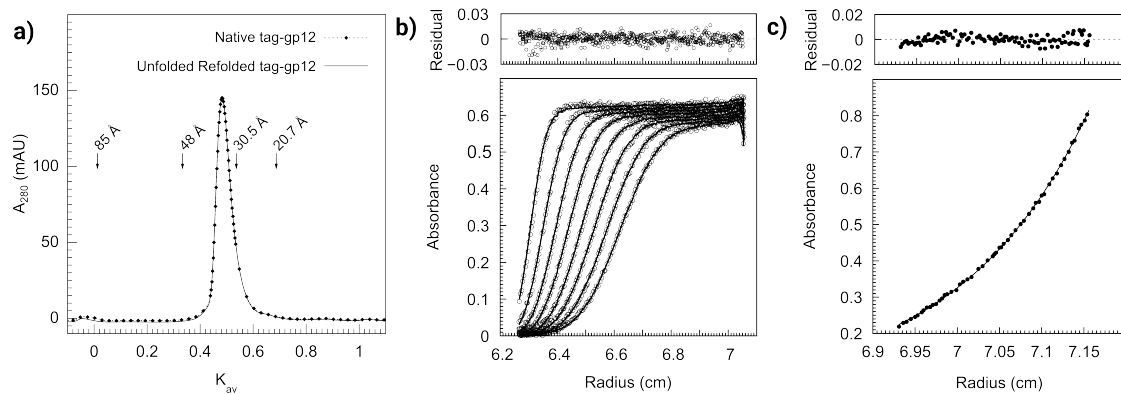


Figure 14: Hydrodynamic properties of tag-gp12. a) R_H determination of native and unfolded-refolded tag-gp12. Native (dotted line) and tag-gp12 heated for 5 min at 90 °C and transferred directly to ice (continuous line) were analyzed by SEC at 16 °C as described under Materials and Methods (section 4). The elution positions of thyroglobulin ($R_H = 85$ Å), γ -globulin ($R_H = 48$ Å), ovalbumin ($R_H = 30,5$ Å), and myoglobin ($R_H = 20,7$ Å) used to calibrate the column are indicated by arrows. K_{av} (partition coefficient) was calculated as described¹¹⁵. mAU, milli absorbance units. b) Sedimentation velocity of tag-gp12 at 220,000 g (loading concentration of 1 mg/ml, 16 °C run). Gp12 has a sedimentation coefficient of 1.7 S ($s_{20,w} = 2.2$ S). c) Sedimentation equilibrium of tag-gp12 at 16,300 g (loading concentration of 0.8 mg/ml, 16 °C run). The data (dots) were fit using a trimer model (continuous line). The best fit was obtained for a single species with an average mass of $31,270 \pm 590$ Da. The top panels in b) and c) show the deviation of experimental points from fitted curves.

5.3 Gp12 has a collagen-like fold

In order to probe that the $(GXY)_8$ repeats of tag-gp12 form a collagen-like triple helix, the protein was challenged with collagenase VII that cuts the triple helix at defined environments²⁰⁰. Control SPP1 proteins gp6 and H16, a tagged form of gp16^{70,190}, were insensitive to proteolysis (not shown) while tag-gp12 was cleaved (inset

in Figure 15b). MALDI-TOF (Figure 15b) and nano LC-MS/MS (Figure 15c) identified the cut between Q₁₉ and G₂₀ of tag-gp12 (arrow in Figure 15a). This site found at the beginning of the GXY repeats region matches one of the expected cutting sites for collagenase VII²⁰⁰.

Collagen right-handed triple helices are also characterized by a CD signature with a deep minimum of negative ellipticity at around 200 nm and a slightly positive ellipticity maximum at around 220 nm^{201,202}. The CD spectrum of native tag-gp12 had a strong minimum at 200 nm and a second minimum at 222 nm where the ellipticity of α -helices masked the positive signal of the collagen helix (Fig. 16a). This profile strongly supports that tag-gp12 combines a collagen-like fold with α -helical regions.

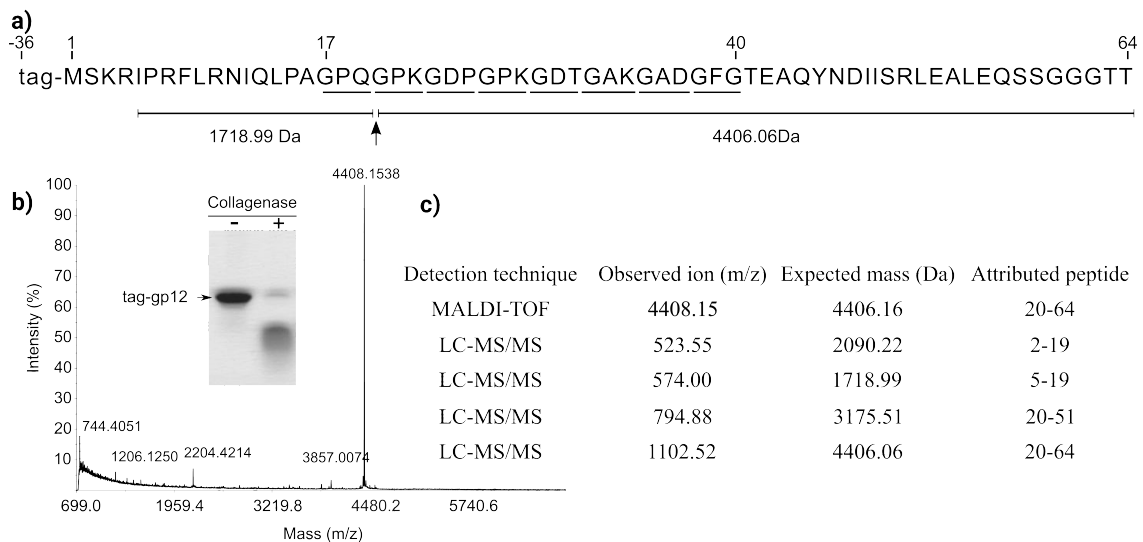


Figure 15: Tag-gp12 sensitivity to collagenase VII. **a)** Collagenase cleavage of tag-gp12 inside the collagen-like sequence motif as indicated by the arrow. Peptides obtained from mass spectrometry analyses (**b**) and **c**) are shown below. **b)** Cleavage of tag-gp12 with collagenase VII analyzed by SDS-PAGE (inset) and mass spectrometry. The observed ion mass (4406.15 Da) in MALDI-TOF (**c**) is attributed to peptide 20 – 64 of the gp12 sequence, identifying the proteolysis site shown in **a**). The same peptide was detected by LC-MS/MS spectrometry, which also showed the presence of three other peptides, resulting from collagenase cleavage at the same position (**c**). The LC-MS/MS analysis had a tag-gp12 sequence coverage of 89%. For clarity, only the peptides in which one end was generated by the collagenase cleavage are listed in **c**). Those peptides were absent in the analysis of tag-gp12 not treated with collagenase.

5.4 Gp12 dissociates and unfolds reversibly at physiological temperature

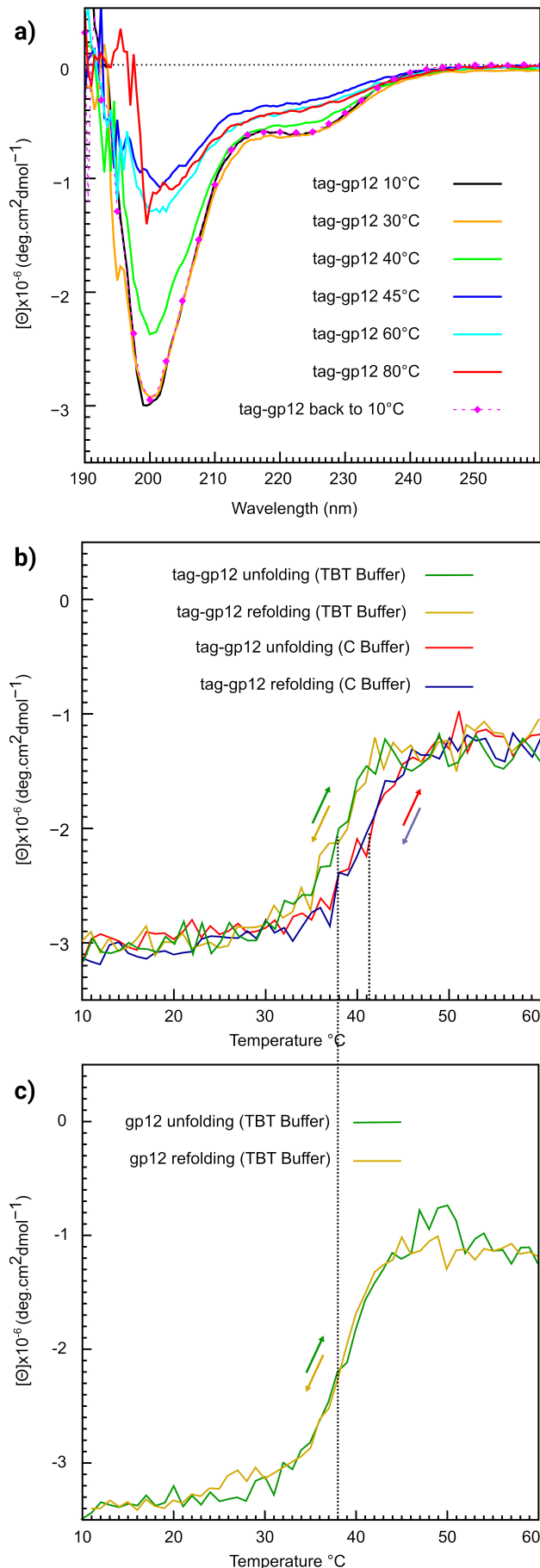


Figure 16 : Reversibility of tag-gp12 trimer unfolding and dissociation. **a)** CD spectra of tag-gp12 (2 mg/ml) in buffer C were recorded using the same sample at different temperatures. Tag-gp12 was then maintained at 80 °C for 30 min and cooled back to 10 °C to record the spectra of the refolded protein (pink dotted line). **b)** Tag-gp12 unfolding and refolding in buffer C and in TBT monitored by CD at 200 nm, the collagen-like triple helix local minimum signal, using a temperature gradient of 1 °C/min. The colored arrows show the direction of the temperature gradient (heating or cooling) for each individual color curve. **c)** Gp12 unfolding and refolding in TBT monitored as in **b)**. The dotted vertical lines in **b)** and **c)** are a visual aid to show the transition midpoints (T_m) in buffer C and TBT. The experiment was repeated twice independently.

CD spectra showed a loss of tag-gp12 structure between 30 and 45°C (Figure 16a). The CD spectra from 45 to 80 °C were characteristic of an unfolded polypeptide chain. Fast (< 1 min) or progressive cooling of the sample back to 10 °C led to complete recovery of the secondary and quaternary structure content with a CD spectrum identical to the one of the native protein (pink dotted line in Figure 16a).

To further analyse the dissociation-unfolding and refolding-reassociation transitions, a CD experiment was monitored at 200 nm (corresponding to the collagen-like helix minimum) by challenging the sample against a heating cycle from 10 to 60 °C (unfolding) and back to 10 °C (refolding) (Figure 16b). Tag-gp12 showed a sharp transition with a T_m of 41°C in the protein high salt buffer and of 38°C in a low monovalent salt solution with magnesium (TBT buffer that stabilizes SPP1 viral particles) (Fig. 16b). Unfolding and refolding followed the same kinetic profile upon heating and cooling (Fig. 16b). The thermal stability study of tag-gp12 by CD revealed a unique transition with complete loss of secondary structure and dissociation of the collagen-like triple helix (Figure 16a,b). The behaviour of tag-free gp12 was identical to tag-gp12 (data not shown; Figure 16c) revealing that the tag influenced neither the protein CD signature nor its dissociation/unfolding – refolding/reassociation properties. The SEC profiles of native and unfolded-refolded tag-gp12 were also indistinguishable with a single symmetric peak of trimers and no detectable intermediate states (Figure 14a). The complete population of refolded tag-gp12 thus retrieved its initial R_H .

In order to define if the tag-gp12 polypeptide chains physically separate upon thermal denaturation we carried out a chimerisation experiment between tag-gp12 (10.7 kDa subunit mass) and tag-free gp12 (6.6 kDa subunit mass). The hexahistidine-tagged tag-gp12 bound strongly to a metal affinity column and eluted only in presence of 500 mM imidazole (Figure 17, top panel). Gp12 adsorbed also to the column matrix but was completely released by a wash with 100 mM imidazole (Figure 17, second panel from top). Loading of a tag-gp12:gp12 mixture kept at 16°C led to differential elution of gp12 at 100 mM imidazole and of tag-gp12 at 500 mM imidazole (Figure 17, third panel from top). When the tag-gp12:gp12 mixture was denatured at 60°C and

reassociated by cooling to 16°C there was a fraction of non-tagged gp12 that co-eluted with tag-gp12 at 500 mM imidazole (Figure 17, bottom panel). This behaviour is explained by presence of heterotrimers in which the tag-gp12 tagged subunit(s) led to retention of the non-tagged gp12 form present in the heterotrimer. The formation of chimeras showed that gp12 and tag-gp12 trimers physically dissociated upon thermal denaturation and that reassociation led to formation of heterotrimers, although homotrimerization appeared to be favoured when comparing the intensity of bands in the bottom panel of Figure 17.

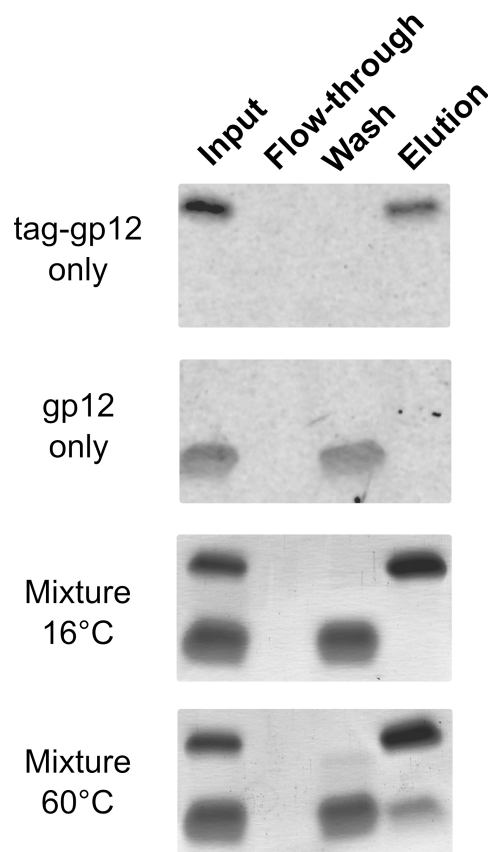


Figure 17: Tag-gp12/gp12 chimerization experiment. Isolated proteins and their mixture incubated at 16 or 60 °C (50 µl of 2 mg/ml in buffer C) were loaded onto a metal affinity column. Aliquots of the input proteins before chromatography, flow-through, washing with 100 mM imidazole, and elution with 500 mM imidazole were analyzed on a 12% Tris-N-[2-hydroxy-1,1-bis(hydroxymethyl)ethyl]glycine gel stained with Coomassie Blue. The experiment was repeated twice independently.

5.5 Binding of gp12 to the capsid lattice is reversible and increases the trimer thermal stability of 20°C

In order to characterize the interaction of gp12 with SPP1 capsids we generated viral particles (SPP1~~12~~) and tailless expanded capsids (H Δ 12) lacking gp12 by genetic engineering (Figure 18a and data not shown) (section 4.1). Gp12 is not present in procapsids but is found in DNA-filled tailless capsids (H in Figure 18a) showing that it attaches to the capsid lattice during DNA packaging (Figure 7)¹²¹. SPP1~~12~~ particles bound tag-gp12 *in vitro* while wild type virions whose capsid carries gp12 did not (Figure 18b). Thus, tag-gp12 interacts strongly and exclusively with specific sites in the SPP1 capsid lattice without any detectable exchange between free (tag-gp12) and capsid-bound (gp12) subunits.

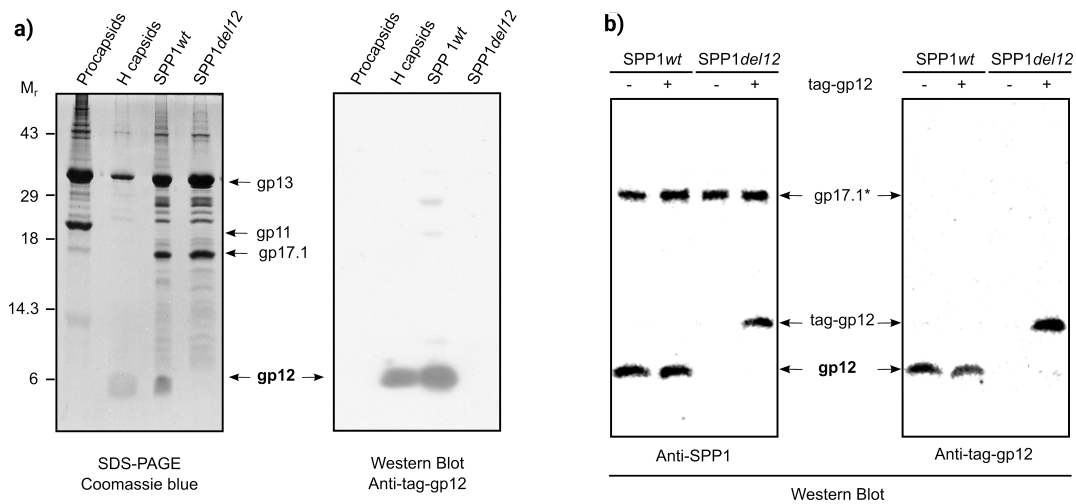
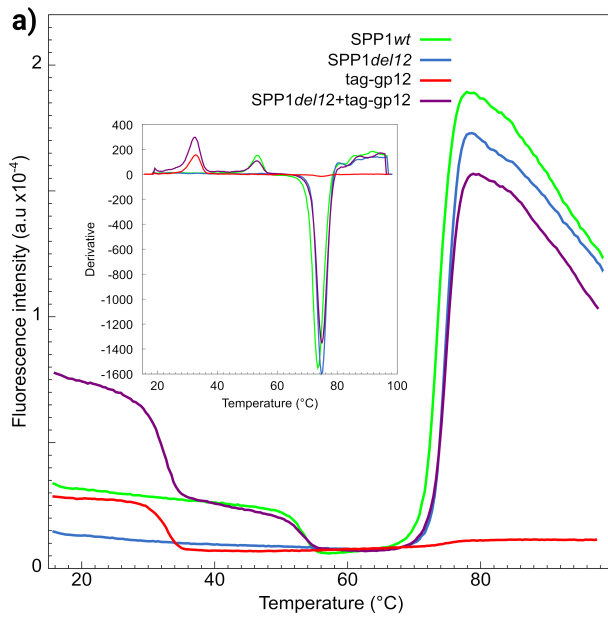


Figure 18: Binding of gp12 to SPP1 particles. a) Composition of SPP1 assembly intermediates (Figure 7) determined by SDS-PAGE gel stained with Coomassie Blue (left panel) and presence of gp12 in the structures detected with anti-tag-gp12 polyclonal antibodies (right panel). The MCP gp13, the MTP gp17.1, and the procapsid internal SFP gp11, present only in procapsids, are identified on the right side of the Coomassie Blue-stained gel. The purified particles are procapsids, tailless DNA-filled capsids (or heads, H), and SPP1 infectious virions wild type (SPP1wt) or lacking gp12 (SPP1~~12~~). b) Binding of tag-gp12 to SPP1wt and to SPP1~~12~~ particles. Virions incubated overnight at 16 °C with tag-gp12, as indicated above the Western blots, were separated from free protein by isopycnic centrifugation in caesium chloride gradients. The composition of particles was analyzed by Western blot with polyclonal antibodies raised against purified SPP1 virions (left panel) and anti-tag-gp12 antibodies (right panel). Note that gp12 and the tail protein gp17.1* are the most immunogenic proteins of the SPP1 particle despite of the fact that they are not the most abundant components of the virion (a) left panel)^{124,131}.



b)

	Isolated tag-gp12	Capsid-bound tag-gp12	Major capsid protein
15°C	Folded	Folded	Folded
40°C	Unfolded	Folded	Folded
60°C	Unfolded	Unfolded	Folded

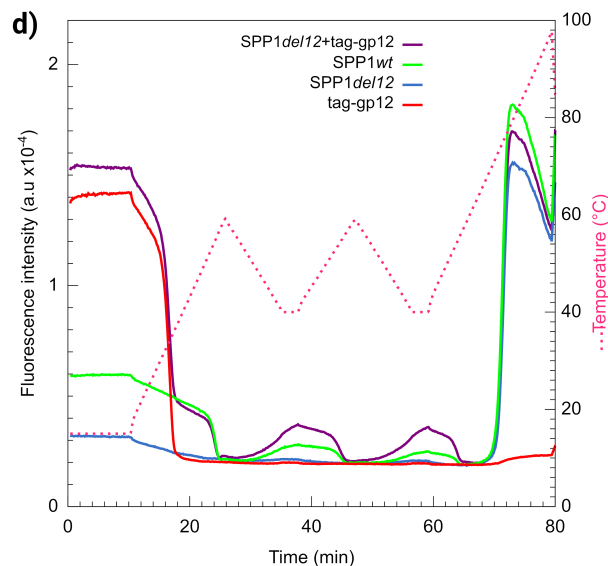
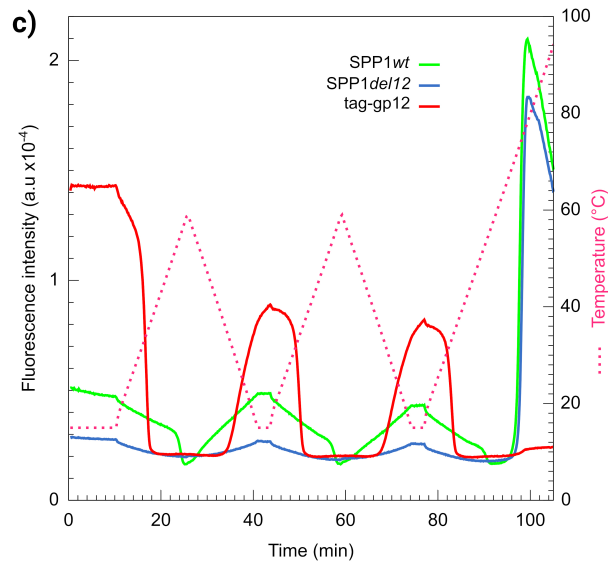


Figure 19: Cyclical gp12 association and dissociation from viral capsids. Sypro Orange was added to protein and purified viral particles samples that were submitted to different heating and cooling regimens at a rate of 3 °C/min and monitored by FBTSA in an ABI 7900HT machine. **a)** FBTSA of isolated tag-gp12 (red curve), SPP1wt virions (green curve), SPP1del12 virions that lack gp12 (blue curve), and SPP1del12 mixed with an excess of tag-gp12 (violet curve). The inset shows the opposite of the first derivative of the fluorescence signal. **b)** Summary of the tag-gp12/gp12 and gp13 states at different temperatures. **c)** Isolated tag-gp12, SPP1wt, and SPP1del12 virions submitted to two cycles of heating to 60 °C and cooling to 15 °C. The experiment was finished with a denaturation step to 99 °C. The pink discontinuous line shows the temperature variation (coordinates are shown on the right). **d)** The same samples and a mix of SPP1del12 virions with a 5.5 molar excess of tag-gp12 (violet curve) challenged with two cycles of heating to 60 °C and cooling to 40 °C. Experiments were repeated at least twice independently.

The FB TSA method allows monitoring independently the thermal denaturation of gp12 and of the major capsid protein gp13⁸⁰. The assay quantifies binding of the Sypro Orange dye to exposed hydrophobic regions of proteins challenged to a temperature gradient²⁰³. In an aqueous environment, Sypro Orange has a low quantum yield and in protein solutions the protein fold normally shields the dye access to non-polar environments. Protein thermal denaturation exposes hydrophobic regions where the dye binds resulting in strong fluorescence emission. Isolated tag-gp12 exhibited the opposite behaviour: an increase of temperature led to progressive loss of the fluorescence followed by a sharp transition at a T_m of $33.4 \pm 0.7^\circ\text{C}$ in TBT buffer (red curve in Figure 19a) which is 4.6°C lower than the unfolding T_m determined by CD (Figure 16b). Gp12 without a tag showed the same behaviour. The profile of this transition revealed that Sypro-Orange has one or several binding sites in native tag-gp12 that were destroyed when the protein starts to lose its secondary and quaternary structure. Such rare property provided a specific signature for tag-gp12 unfolding.

The FB TSA profile of infectious SPP1 phage particles (green curve in Figure 19a) was marked by two transitions similar to the ones found for the tailless capsids (data not shown)⁸⁰, revealing that only the capsid proteins of phage particles gave a detectable signal under our experimental conditions. Both structures carry gp12. The first transition, at $53.6 \pm 0.2^\circ\text{C}$, displayed the tag-gp12 signature and was absent from particles lacking gp12 (SPP1~~12~~; blue curve in Figure 19a). Mixing of SPP1~~12~~ phages with tag-gp12 *in vitro* restored the signal at 53.6°C , while the excess of free protein led to the typical T_m transition at 33.4°C (violet curve in Figure 19a). The identical T_m of gp12 and tag-gp12 was 20.2°C higher than the one observed for isolated tag-gp12 showing that binding to the capsid lattice led to a major stabilisation of the gp12 trimer. The second signal transition of viral particles with or without gp12 was characterized by a strong increase of fluorescence at $75 \pm 0.3^\circ\text{C}$ due to cooperative denaturation of gp13.

The distinct melting temperatures of isolated tag-gp12 (33.4°C), of capsid-bound tag-gp12 or gp12 (53.6°C), and of major capsid protein (75°C) allowed to follow the behaviour of the three species in gp12-capsid binding experiments. At 40°C capsid-

bound gp12 was easily discriminated from isolated tag-gp12, since it was the only folded gp12 form at this temperature while at 60°C only the capsid protein was stable (Figure 19b). Isolated tag-gp12, SPP1 wild type and SPP1~~12~~ particles were submitted to two heating-cooling cycles between 15 and 60°C followed by a final heating step from 15 to 99°C (15-60-15-60-15-99°C, 3°C/min heating/cooling rate). Free tag-gp12 exhibited a loss of signal upon heating and its partial reacquisition when cooling to 15°C showing that the fluorophore binding site(s) was/were not completely restored in the tag-gp12 population (red curve in Figure 19c) in spite that the protein fully reacquired its quaternary structure CD signature (Figure 16). The temperatures of transition were remarkably reproducible revealing that tag-gp12 undergone dissociation/unfolding and folding/reassociation cycles. Gp12 bound to SPP1 capsids exhibited a transition corresponding to a T_m of 53.6°C (continuous green curve in Figure 19c). The process was reversible upon cooling and re-heating apart from a slight loss of fluorescence from one cycle to another. Gp12 thus dissociated/unfolded reversibly from wild type capsids and maintained its binding activity to the capsid.

In order to assess if the capsid lattice influences gp12 refolding/reassociation the cycling experiment was repeated with cooling steps to 40 °C (15-60-40-60-40-99 °C program, Figure 19d), a temperature at which free tag-gp12 remained unfolded after the first heating step. Gp12 bound to phage capsids kept its signature (T_m of 53.6°C) in heating cycles to 60°C. Cooling to 40°C led to recovery of some fluorescence signal (green curve in Figure 19d) but significantly less than when the temperature was reduced to 15°C (Figure 19c). Therefore, at 40°C a sub-population of gp12 rebound to phage capsids yielding folded trimers that fix Sypro Orange. Addition of a 5.5-fold molar excess of exogenous tag-gp12 to wild type capsids restored most of the gp12 signal associated to capsids after each 60-40°C cycle (violet curve in Figure 19d) showing that tag-gp12 had efficiently replaced gp12 which left its capsid sites upon denaturation. Restoring of the tag-gp12 signal at 40°C occurred exclusively in presence of the capsid lattice showing that this structure promoted tag-gp12 refolding and reassociation.

5.6 Native and unfolded gp12 bind to SPP1 capsids in a distinct way

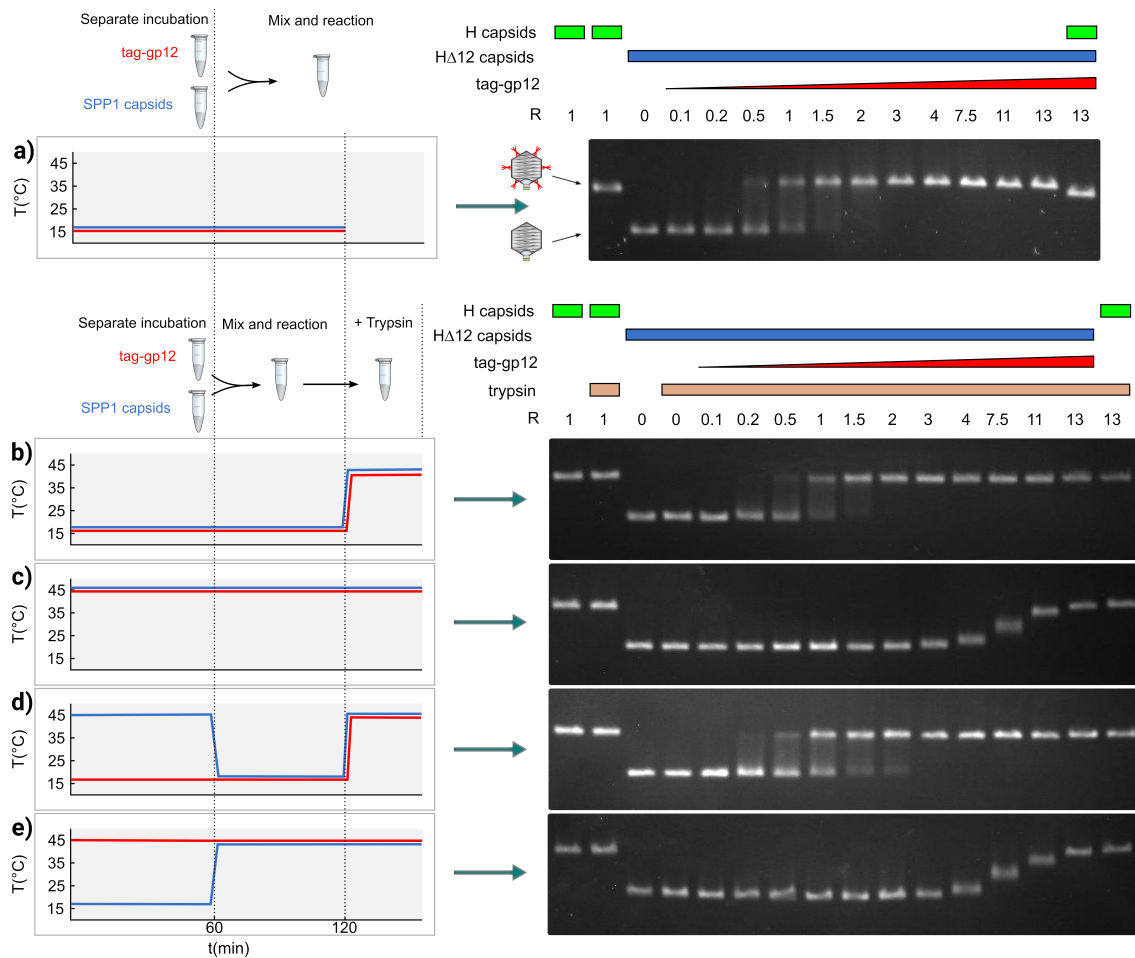


Figure 20: Capsid binding behavior of native and unfolded tag-gp12. Purified SPP1 tailless capsids lacking gp12 (capsid H Δ 12) (blue characters and blue curves on the left and blue rectangles above the gels on the right) and tag-gp12 (red) were preincubated separately, mixed, and treated with trypsin (except in **a**) according to the different combinations of incubation conditions used in the experiments in **a**-**e**) (see text for details), as outlined on the left of each panel. Samples treated with trypsin in **b**-**e**) are identified by salmon rectangles above the gel lanes on the right. Capsids were then resolved by agarose gel electrophoresis to assess their occupancy with tag-gp12. Wild-type SPP1 capsids with gp12 (H capsids) (green rectangles above the gels on the right) and H Δ 12 were used as controls. The schematics in the center of **a**) show the electrophoretic mobility of capsid H (with gp12 represented in red) and H Δ 12. The experiment was repeated four times independently.

The finding that both native and unfolded tag-gp12 bound to SPP1 phage capsids (Figure 19c,d) suggested two distinct types of interaction prompting their

characterization. Tailless capsids without gp12 (H Δ 12) and purified tag-gp12 were pre-incubated separately at 16 or 45°C followed by mixing at different ratios for interaction at the two temperatures (Figure 20). Reactions were then incubated at 45°C with trypsin that degrades free gp12/tag-gp12 (Figure 21a). The tag of capsid-bound tag-gp12 was prone to trypsin attack but the gp12 moiety attached to the capsid remained intact (Figure 21b) explaining the lower electrophoretic mobility of capsids loaded with tag-gp12 not treated with trypsin when compared to those that were trypsinated (compare the two leftmost lanes in the gels of Figure 20). This proteolysis step prevented subsequent interactions of free tag-gp12 with capsids during downstream sample manipulation at room temperature and separation by gel agarose electrophoresis. SPP1 capsids with gp12 (H capsids) or H Δ 12 loaded with tag-gp12 had a slower electrophoretic mobility than capsids lacking gp12 (Figure 20) most likely because gp12/tag-gp12 reduces the capsid surface electronegative charge. In contrast, gp12 does not have a major effect on the capsid diameter that is almost identical in H and H Δ 12 (~610 Å⁸⁰).

When H Δ 12 capsids were mixed at 16°C with increasing amounts of tag-gp12 native trimers the capsid species shifted from tag-gp12-free to capsids fully loaded with tag-gp12 (Figure 20a,b,d). At a ratio (R = 60 tag-gp12 trimers/capsid) of 0.5 most capsids lacked tag-gp12 but a minority was already saturated with tag-gp12 while at R=1.5 almost all capsids were decorated with tag-gp12. Species with intermediate electrophoretic mobility were poorly detected revealing that capsids partially occupied with tag-gp12 were a minor population even at limiting amounts of tag-gp12 (e.g. R=0.5). We attribute this behaviour to high cooperative binding of tag-gp12 trimers to its 60 sites in the SPP1 capsid.

To characterise the interaction of unfolded tag-gp12 with the capsid, the two species were preheated individually at 45°C and mixed at the same temperature (Figure 20c). A significant excess of tag-gp12 *per* capsid (R between 4 and 13) was needed to promote a change of capsids electrophoretic mobility. Their discrete bands showed a migration pattern that progressed from the capsids lacking gp12 band behaviour (R<3) to the full tag-gp12-loaded capsid band (R \geq 13) (Figure 20c). Similar results were

obtained when the tag-gp12 - H Δ 12 interaction reaction was prolonged overnight at 45°C (not shown).

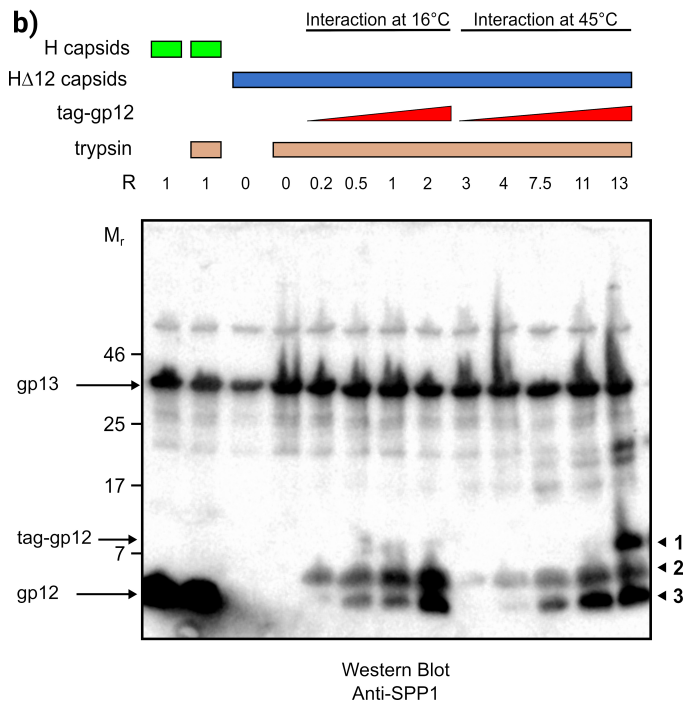
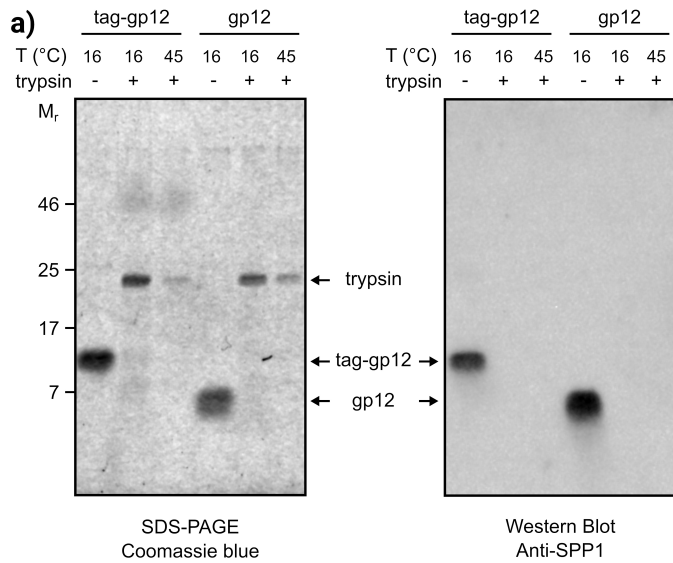


Figure 21 : Trypsin sensitivity of free and capsid-bound gp12. a) Purified tag-gp12 and gp12 were incubated with trypsin either at 16 or at 45 °C. Both proteins were completely digested by the protease at the tested temperatures, as assessed by Coomassie-stained SDS-PAGE (left panel) and Western blot analysis) with polyclonal anti-SPP1 antibodies that recognize gp12 (Figure 18b) (right panel). The position of migration of gp12, tag-gp12 and trypsin is shown between the panels. b) Trypsin digestion of the binding reaction between capsids and tag-gp12 (labeled band 1 on the right of the figure) under the same conditions as in Figure 20b,c. Note that gp12 bound to H capsids is not sensitive to trypsin, whereas the tag of tag-gp12 is partially (band 2) or fully (band 3) digested by trypsin. The Western blot was developed with anti-SPP1 antibodies that recognize gp12 but also, although giving a comparatively weak signal, the MCP gp13. The gp13 band was used to control the normalized input of capsids in the binding reaction.

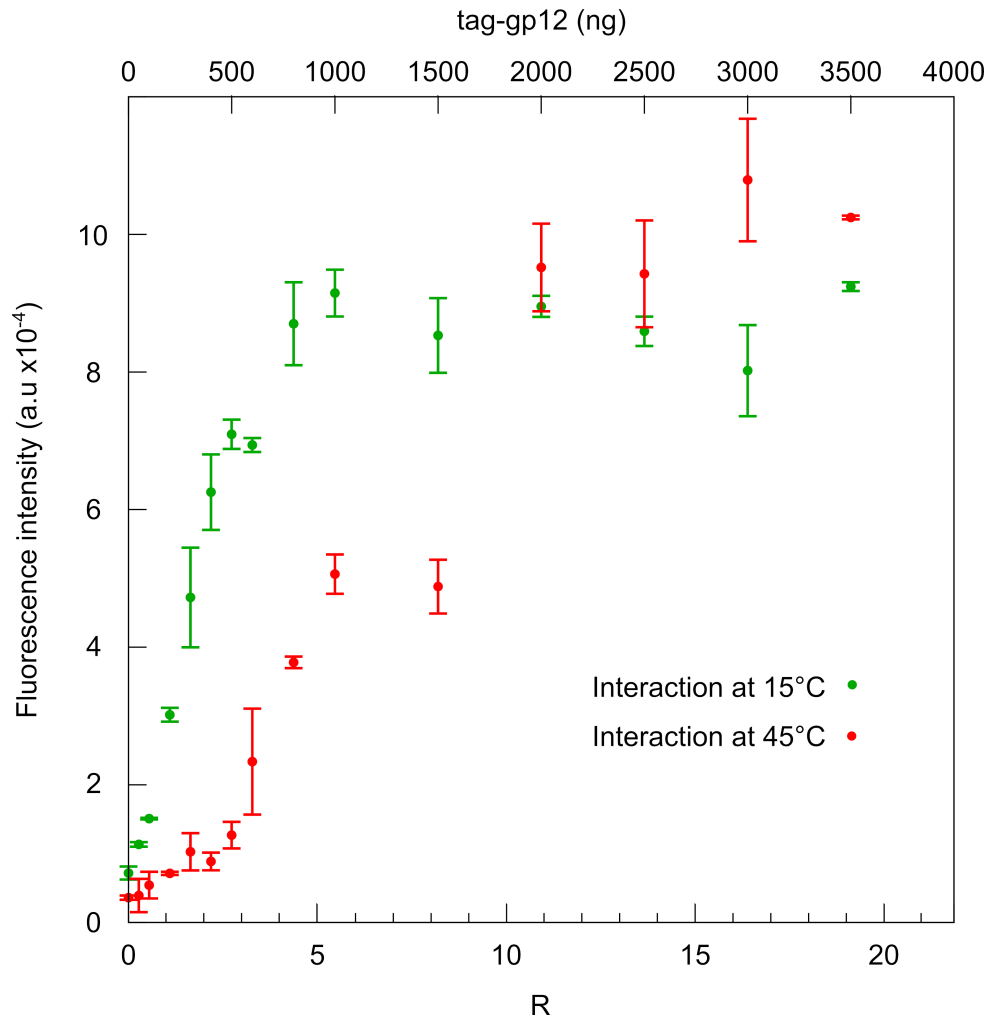


Figure 22: FB-TSA of HΔ12 incubated with increasing amounts of tag-gp12. Capsids and tag-gp12 were preincubated separately at 15 (green) or 45 °C (red) and then mixed at the same temperature according to the experimental setup shown in the left panels of Figure 20a and 20c (not trypsinated), respectively. After co-incubation, the samples were transferred to a QuantStudio 12Kflex machine for thermal denaturation at a heating rate of 3 °C/min in the presence of Sypro Orange. The amplitude of the gp12 signal with its characteristic transition at 53.6 °C (cf. Figure 19a) was plotted against the R ratio of tag-gp12 relative to the input of HΔ12 capsids. The experimental points are averages of triplicates in two independent experiments.

Furthermore, pre-heating of capsids at 16°C or 45°C showed that temperature did not affect their binding properties (Figure 20). Stable interaction of unfolded tag-gp12 with HΔ12 capsids thus required an excess of tag-gp12 that bound in an inefficient manner leading to a population of capsids whose binding sites are only partially occupied by tag-gp12 at molar ratios as high as R=11 (Figure 20c,e). The increase of occupancy with the rise of R correlated with an augmentation of the tag-

gp12 signal in FBTSAs experiments consistent with the formation of tag-gp12 trimers in the capsid lattice (Figure 22).

5.7 Identification and modular organization of gp12-related proteins

Our studies showed a novel structural organization for a phage capsid auxiliary protein (section 1.6). The SPP1 gp12 modular organization with a central collagen-like fold (Figure 13) provides very versatile association properties to this highly specific binder of the phage DNA-filled capsid. We thus aimed to investigate if other phages code gp12-like proteins.

A pBLAST search for proteins with sequence similarity to gp12 in the NCBI non-redundant protein sequence databank (January 2013) delivered five hits when a cut-off E-value $<10^{-4}$ was applied (pBlast 1 in Figure 23a,b). These gp12-related proteins, abbreviated B,C,D,E,F (A being gp12), were then submitted individually to a new pBLAST search (pBlast 2). Only proteins C and D gave new similarity hits (Figure 23a,c). Ten different additional proteins were identified in each search (C1-C10 (Figure 23c) and D1-D10). The same hits were obtained with the two template proteins due to the similarity between C and D (Figure 23b). The multiple sequence alignments of sequences obtained in pBlast 1 (template gp12) and in pBlast 2 (template proteins C or D) showed that all have the modular organization of gp12: an N-terminal domain, a central stretch of GXY repeats and a C-terminal domain (Figures 23b,c and 24). Gp12 has the shortest number of repeats that varies among the proteins analysed to a maximum of (GXY)₂₄ in protein C8 (Figure 24).

Two groups of proteins can be distinguished. The first group with strong similarity to gp12 (Blast 1) has an homologous C-terminus and a N-terminus part that varies both in size and sequence (Figures 23b and 24b). They have a short number of GXY repeats apart from proteins C and D. The second group typified by proteins C and D (Blast 2) shares a common N-terminus part and is more divergent at the C-terminus end (Figure 23c). Their central collagen-like sequence varies in length from 10 to 24

GXY repeats (Figure 24c).

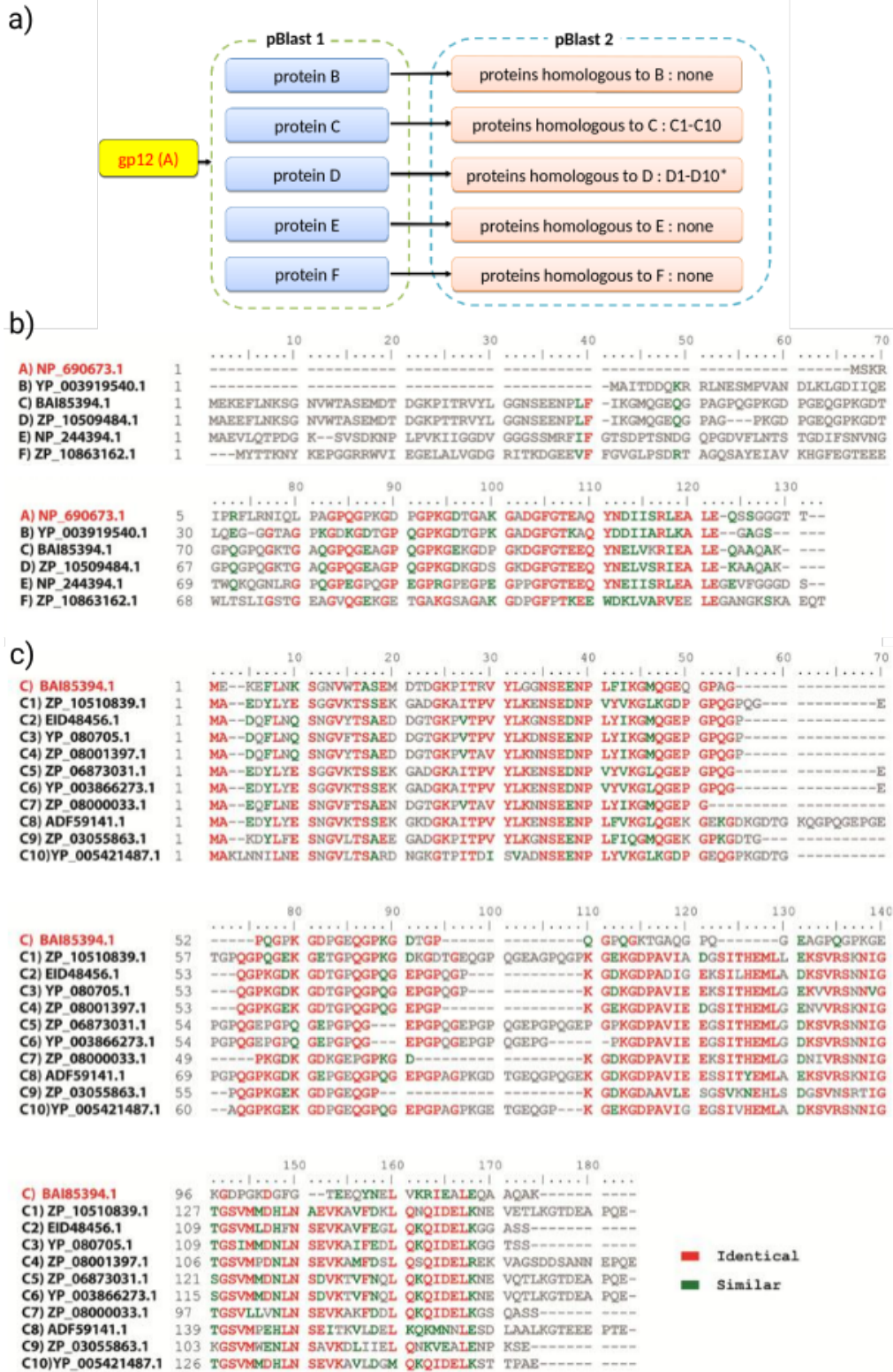
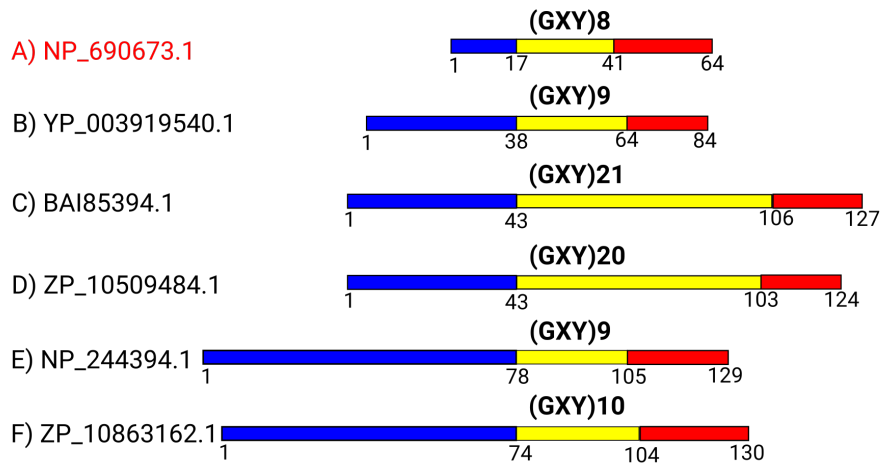


Figure 23 :Similarity search of gp12 relatives by using pBLAST. **a)** Search protocol and hits obtained. Proteins are named A (gp12), B to F (hits of pBlast 1 using the gp12 sequence as template for similarity search at a threshold E-value $<10^{-4}$), and C1 to C10 or D1 to D10 (hits of Blast 2 using the C and D proteins as search templates, respectively). B,E,F template searches gave no new hits in pBlast 2. *- the D1 to D10 hits obtained are identical to C1 to C10. **b)** Multiple sequence alignment of proteins identified in Blast 1 using gp12 as search template. GenBank protein names are shown on the left with the search template protein name (gp12) on top in red. **c)** Multiple sequence alignment of proteins identified in pBlast 2 using protein C as search template.

We have then analysed the genome context of the open reading frames (ORFs) coding for the proteins identified in the pBLAST analyses. They are all found in genomes of *Bacilli* spp. In the genomes functionally annotated (proteins C,D,E,C3,C5,C6,C8,C9), they are encoded by an ORF localized in a cluster of prophage genes. This ORF is located adjacent to a gene coding for a putative phage major capsid protein. This is also the case for SPP1 gp12 that is encoded by the gene that precedes the MCP gp13 gene. The proteins identified by our similarity search are thus most likely capsid auxiliary proteins. An interesting case is protein C8 (GenBank: ADF59141.1) of temperate phage phi105 of *B. subtilis*. This is the largest protein identified here. It has a (GXY)₂₄ repeat that is 3-fold longer than the one from gp12 suggesting that it uses a collagen-like triple helix to build filaments that irradiate from the phi105 capsid. A pBLAST similarity search carried out in April 2019 using the stringency criteria used in Blast 1 increased the number of gp12 directly related proteins from five (Figure 23a,b) to 55 hits. The proteins identified are encoded by genomes of *Bacilli* spp. or by closely related Firmicutes confirming the widespread presence of collagen-like proteins in prophages of this taxonomic group.

a)



b)

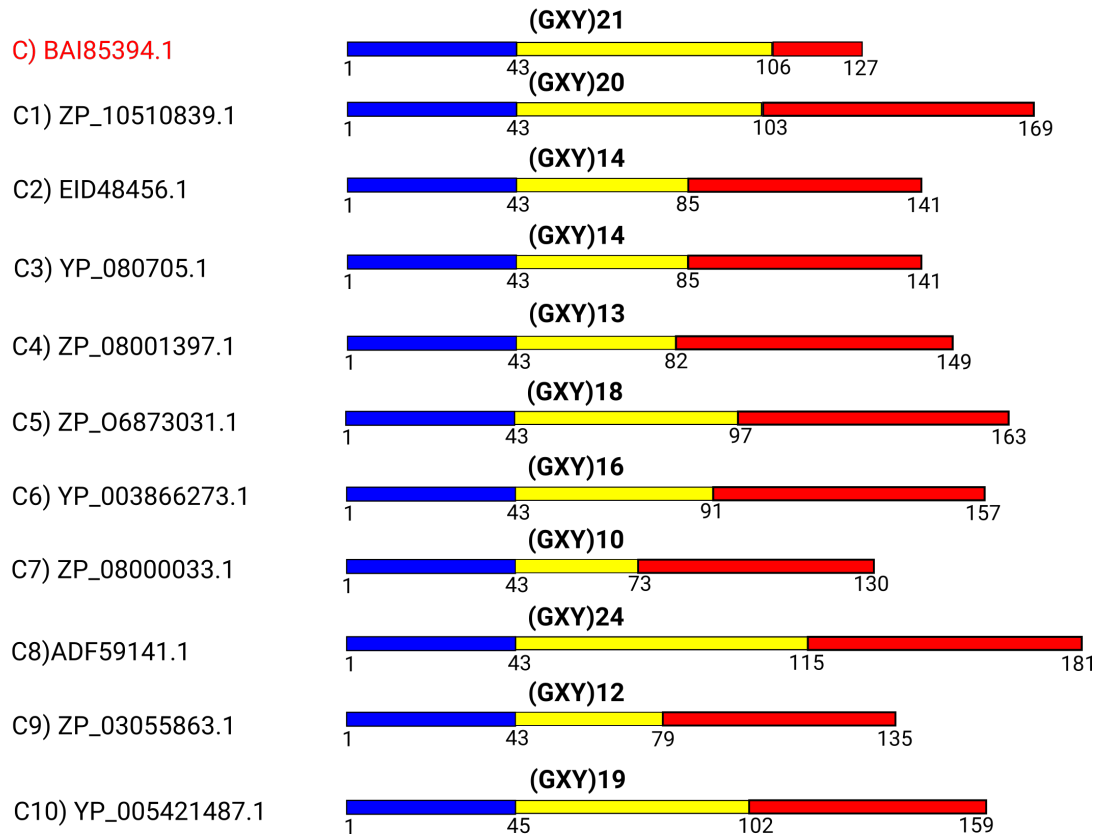


Figure 24: Modular organization of gp12-related proteins. **a)** Schematics of the proteins identified in pBlast 1 (Figure 23) aligned through the beginning of the GXY repeats segment. The N-terminus domain is displayed in blue, the (GXY)_n repeats in yellow, and the C-terminus in red. Numbers identify the residues at the boundaries of the protein domains. **b)** Schematics of the proteins identified in Blast 2 (Figure 23) using protein C as template. The alignment and figure organization is as in **a)**.

5.8 Distribution of CMCP proteins in prokaryotes and their viruses

Proteins with collagen-like properties were found in bacteriophage tail fibers¹⁵⁸ and capsids (this PhD work) as well as at the surface of bacteria^{173-176,180,161}. This observation prompted us to carry out a quantitative analysis of the distribution of proteins with GXY repeats (CMCPs) in viruses and prokaryotes.

I developed the Troglodyte software suite of tools, including available free programs, to download, index and scan protein sequences for detection of GXY repeats (section 4.15). Protein products of CDSs were download from the RefSeq non-redundant database (July 2014) for five different taxonomic groups: viruses, bacterial viruses, archaeal viruses, bacteria and archaea. Bacterial and archaeal viruses are subsets of the viruses group to allow a direct correlation between these viruses and the Domains of Life they infect. A unique FASTA file was created for each taxonomic group and curated with CD-HIT to eliminate redundant sequences (Figure 25a). All proteins in the datasets were then searched for a $(\text{GXY})_{n_{\min}}$ motif in which n_{\min} is the minimum number of sequential GXY repeats found in a protein. The number of protein positive hits for n_{\min} ranging between 1 and 100 was plotted as a percentage of the total number of proteins in the dataset. For interpretation of the resulting graphics (Figure 25b), note that a protein with, for example, 21 successive GXY repeats will be counted as positive hit for $\text{GXY } n_{\min}=1$ to $\text{GXY } n_{\min}=21$ and not counted for $\text{GXY } n_{\min}\geq 22$.

Proteins with GXY repeats are found in all taxonomic groups but their frequency and length varies significantly. When $n_{\min}=5$, the frequency of CMCPs is lower than 0.6 % of the total proteins in the dataset. Viruses shows the highest level of CMCPs (0.2 % of total proteins for $n_{\min}=10$). Almost half of those proteins are from bacterial viruses (phages). The length of their collageneous segments reduces steadily to $n_{\min}\approx 65$ when it reaches a drastic cut-off. Eukaryotic viruses CMCPs account for the viral proteins with very long GXY repeats ($n_{\min}>70$ on the red line in Figure 25b), most probably overrepresented in Mimiviruses and its related giant viruses^{171,172}. Bacteria have less CMCPs with short GXY repeats than its viruses but they feature proteins with

longer repeats (magenta line in Figure 25b). CMCPs are rare and the length of their GXY repeats is short in archaea. They are also virtually not found in archaeal viruses showing that the collagen fold is not a major element on archaeal proteins architecture.

This bioinformatics survey highlights the widespread distribution of CMCPs in bacteria and in their phages showing that the study of non-eukaryotic collagen is of significant biological importance.

a)

	Total proteins				
	Non curated		Curated		
	Proteins	Genomes	Proteins	Genomes	Gly %
Archaea	855 021	352	429 687	352	7.67
Archaeal viruses	3 923	64	3 086	64	6.85
Bacteria	49 809 020	14 927	10 015 887	14 698	7.54
Bacterial viruses	125 889	1 301	74 367	1 295	7.07
Viruses	187 375	3 959	121 614	3 732	6.26

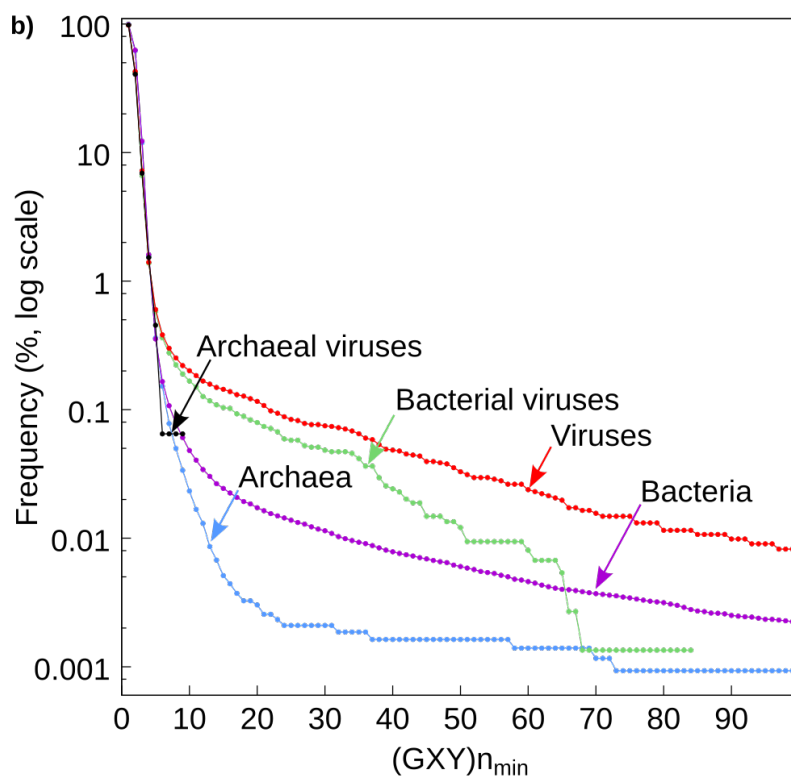


Figure 25 : CMCPs in prokaryotes and viruses. **a)** Protein sequences were extracted from GenBank Refseq files according to the five taxonomic groups shown and then linked to their corresponding genomes. The table compares the total number of proteins and genomes, before and after the sequence datasets curation process with CD-HIT to generate the non-redundant datasets. The glycine frequency is reported for the non-redundant protein datasets. Despite the reduction in proteins number, glycine frequencies are extremely close between non curated and curated protein data sets (not shown). **b)** Frequency of CMCPs within prokaryotes (archaea and bacteria) and

viruses (all viruses, bacterial viruses or phages, and archaeal viruses). The complete sets of non-redundant proteins were searched for a minimum number of successive (GXY) repeats (n_{min}). Each n_{min} value (x-axis) corresponds to an independent CMCPs search. The frequency of CMCPs (y-axis) is relative to the total number of non-redundant proteins of corresponding dataset and is reported in a log scale.

6 General Discussion

Collagen is the most abundant protein in vertebrates where it is the major component of the extracellular matrix. Eukaryotic proteins with triple helix segments are also involved in host defense processes such as complement factor C1q, mammalian lectins and macrophage scavenger receptors^{136,158}. The collagen super helix is made of three left-handed elongated helices that twist together around a central axis to form a right-handed super helix. Intramolecular hydrogen bonding perpendicular to that axis stabilizes the structure. The quaternary structure forms an elongated rod that imposes a glycine at every third position because its small side chain can be accommodated in the crowded interior of the super-helix. Consequently, the repetitive GXY motif is the sequence signature of collagen. Post-translational modification of collagen such hydroxylation of proline at position Y within the repetitive GXY motif and glycosylation of the polypeptide stabilize the structure of eukaryotic collagen¹⁸².

In prokaryotes, the presence of hydroxylated proline was not observed although presence of a prolyl-4-hydroxylase activity was reported in *B. anthracis*²⁰⁴. Incorporation of hydroxyproline in recombinant collagens synthesized in *Streptococcus pyogenes* was also observed when this bacterium was grown in medium with hydroxyproline²⁰⁵. However, CMCPs were identified in bacteria and were subsequently shown to feature biochemical properties of a stable collagen protein in absence of post-translational modifications¹⁵⁹⁻¹⁶⁴ (section 2.2). Studies with synthetic peptides used as model systems of the collagen fold also confirmed that hydroxyproline could be substituted by other amino acids leading to a stable collagen triple helix structure^{175,206} (his work). Genes coding for CMCPs were identified in all Domains of Life and in their viruses^{160,161} (this work). The study of prokaryotic and viral CMCPs structural organization, of the properties conferred by their GXY repeats and how they combine with other domain modules to achieve diverse biological functions is a field of intensive research at present.

In this thesis I studied the protein gp12 of bacteriophage SPP1. Gp12 binds to

the SPP1 capsid surface at the center of each hexamer of the MCP gp13⁸⁰. The gp12 6.6 kDa polypeptide is characterized by a stretch of 8 GXY repeats that occupy its central region which led to the initial hypothesis that gp12 is a CMCP. Our results showed that gp12 is a trimer in solution whose hydrodynamic properties reveal an elongated or oblate shape (Figure 14). This shape fits to the observation that gp12 is an elongated rod that emerges from the center of gp13 hexamers as found from comparison of cryo-electron microscopy structures of the SPP1 capsid with and without gp12 (Figure 8d)⁸⁰. A collagen triple helix could be docked in the gp12 density of the wild type capsid but the remaining non-assigned density in the cryoEM reconstruction could not account for the complete gp12 trimer (H White and EV Orlova, personal communication). This observation suggests that gp12 has some flexible parts that are smeared out in the icosahedral reconstruction of the SPP1 capsid.

Secondary structure prediction based on amino acid sequence and experimental CD data revealed the modular organization of gp12 (Figures 13 and 16A). The CD spectrum of the native gp12 trimer shows two minima. The first one at 200nm corresponds to the deep minimum of the polyproline II fold, characteristic of the collagen triple helix quaternary structure. The second minimum around 222nm is the signature of alpha-helices. Its intensity is likely reduced by the positive CD signal of the polyproline II fold at 222 nm²⁰² but the spectrum clearly revealed that gp12 has an alpha-helical content. The central position of the (GXY)₈ repeat in the gp12 sequence and the sensitivity of the gp12 trimer to collagenase (Figure 15) show that the collagen super helix forms an elongated spacer that separates the two protein end segments. Secondary structure prediction suggests that the gp12 C-terminus forms an alpha helix while the N-terminus is predicted to be unstructured. Protein disordered regions were found to participate in interactions that lead to their subsequent folding during assembly of large macromolecular assemblies, including phage particles^{71,207}. We thus propose that the gp12 N-terminus binds to the capsid while the C-terminus is exposed at the capsid surface. The same strategy might be used by the gp12-related CMCPs that have a similar modular organization (Figure 24). In such case, the 10 proteins with a highly conserved N-terminus domain identified by similarity to protein C might bind to the same capsid sites from different phages (Figures 23c and 24b). Alternatively, their few

amino acid sequence differences still ensure specific binding to a single phage capsid type. The experimental validation of these hypotheses and the understanding how the gp12 trimer positions relative to the gp13 hexamer structure in the SPP1 capsid remain to be established.

The combination of a collagen segment with alpha helices was found in other phage CMCPs. Ghosh et al.¹⁵⁹ studied collagen-like proteins from prophages of pathogenic *E. coli* strains. Those proteins feature various structural organizations. In all cases, one or two collagenous segments are combined to other domains such as phage fiber domains and alpha coiled coil domains that were found to be trimerisation domains¹⁸⁹. This protein building plan leads to formation of long flexible rods connecting globular domains that were proposed to form side fibers in the tails of bacteriophages¹⁵⁹. They probably support phage adhesion to bacterial surfaces. CMCPs are also found at the surface of bacteria where they play roles in bacterial pathogenesis and biofilm formation^{167,168}. These proteins feature collagen rods of different lengths that connect non-collagenous amino and carboxyl terminus domains which vary significantly in size¹⁶¹. One of these domains can include transmembrane segments that anchor the protein at the cytoplasmic membrane¹⁶⁴. Taken together these results show that the stable flexible rod structure provided by the collagen-like triple helix is combined with a variety of other domains to build extracellular proteins of different biological functions. The CMCPs in phage tails and capsids might play also a role to provide adhesion properties to the phage particle.

Despite the short length of the gp12 single collagenous segment, the purified polypeptide forms a stable trimer with a melting temperature around 40°C (Figure 16). Compared to the known T_m values of model peptides with short GXY repeats, the thermal stability of gp12 is higher than expected if provided exclusively by the collagen triple helix. Persikov et al. had shown the implication of the length (n repeats) and the nature of amino acid in position X and Y of the (GXY)_n motif in the triple helix stability¹⁸⁵. For example, the melting temperature of the (GPP)₉ collagen-like peptide is around 15°C against 60°C for the (GPP)₂₀ peptide. Replacement of the proline at position Y with a hydroxyproline (O) increased the thermal stability of the peptide. The

(GPO)₁₂ peptide featured a melting temperature around 70°C. The effect of the sequence length levels when n reaches 14 to 16 perfect repeats in model peptides¹⁸⁵. However, in collagens and CMCPs other structural features intervene to yield collagenous domains whose thermal stability is around 38°C¹⁶¹, as in case of collagen type I that has more than 1000 GXY repeats. Gp12 has only 8 GXY repeats and no post-translational modifications that could be detected by mass spectrometry of tryptic gp12 (Figure 15). The collagen-like domain sequence is marked by the presence of both positively and negatively charged residues at the X and Y positions. It had been observed in bacteria that collagen-like domain sequences are also characterized by the use of charged residues as an efficient replacement for hydroxyproline^{160,187}. The intermolecular contacts between those charged residues could be either direct or mediated by water molecules, stabilizing the triple helix via side chain interactions and organization of the water shell around the protein. Non-collagenous domains also play a major role in the assembly and stabilization of the triple helix. The C-terminus domain of collagens was shown to play an essential role in the triple helix nucleation rate, in thermal stability and in correct alignment of the polypeptide chains. In order to investigate this function, Frank et al. reported that the fusion of a (GPP)₁₀ polypeptide with the T4 foldon domain increases the thermal stability of the engineered protein, the (GPP)₁₀-foldon protein, by about 42°C and increases the rate of the collagen triple helix formation^{176,206}. In case of SPP1 gp12 it is likely that its amino and/or carboxyl domains contribute to the protein stability. Our working model is that the carboxyl terminus alpha-helical region plays a central role in gp12 trimerization promoting assembly of the collagen-like region and that the predicted unfolded amino terminus mediates interaction with the phage capsid hexamers. The latter interaction strongly stabilizes the gp12 structure, increasing its thermal stability by ~20°C (Figure 19).

Upon heating, gp12 trimers unfold and lose their secondary structure. The same sharp transition is seen on CD spectra and in fluorescence-based thermal shift assays (FBTSA) using the fluorescent probe Sypro Orange. The transition monitored by FBTSA occurs at a T_m 4.6°C lower than in the CD experiments (Figures 16 and 19A) revealing that the site(s) for binding of the Sypro Orange probe in the gp12 trimer is/are

destroyed prior to gp12 dissociation and unfolding. This behavior shows that the trimer structure is disturbed, probably within defined locations of the protein, before losing its three-dimensional structure. Dissociation of the trimer by disruption of the triple collagen helix and eventually of a putative trimerization domain is likely concomitant with unfolding of the gp12 polypeptide chain as assessed by the parallel loss of the left-handed triple helix and α -helical signals in CD spectra between 40° and 45°C (Figure 16A). Upon cooling, unfolded gp12 regains its trimer structure. When gp12 is mixed with a tagged gp12 protein and their trimers are unfolded together at 60°C followed by refolding at 16°C hetero-trimers are formed (Figure 17). This experiment showed that unfolded gp12 is in a monomeric state competent for hetero-trimerization. The refolded and reassociated gp12 appears to fully recover its secondary and quaternary structures as seen in the CD experiment. However, in the FBTSA experiment it does not recover all its initial Sypro Orange binding capability (Figure 19C). Thus, most probably, the gp12 protein recovers its quaternary and secondary structure but the exact atomic arrangement is not completely restored. The presence of the dye during the refolding process could also affect the monomers refolding/reassociation behavior. Exogenous probe-free techniques such NMR could be used to investigate this hypothesis.

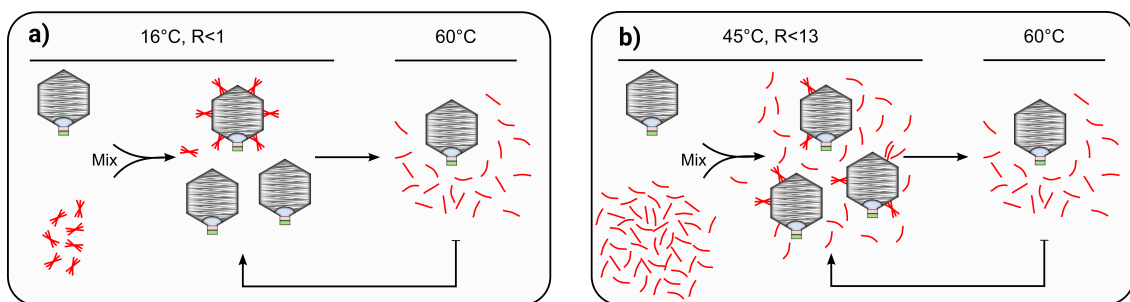


Figure 26: Models of native gp12 trimers (in limiting amounts) **(a)** and unfolded gp12 polypeptides **(b)** binding to capsids and their dissociation from the capsid lattice in a temperature-dependent fashion.

The temperature-dependent reversibility of gp12 refolding/reassociation was instrumental to study the pathway to assemble the trimer state but also to investigate its interaction with the SPP1 viral capsid. Both native and unfolded gp12 are able to bind to the capsid but following different pathways. The native trimeric gp12 binds cooperatively to the 60 gp13 hexamers of DNA-filled capsids lacking gp12. We

hypothesize that binding of a gp12 trimer to the center of an hexamer produces a tectonic effect that spreads through the SPP1 capsid lattice favoring subsequent attachment of trimers to other gp12 binding sites. Under gp12 limiting amounts this mechanism generates a mixed population in which capsids fully decorated with gp12 or that stay naked are predominant (Figures 20a,b,d and 26a). In contrast, binding of unfolded gp12 to capsids leads to partial occupancy of gp12 binding sites over a broad range of gp12-capsid ratios (Figures 20c,e and 26b). This interaction correlates with formation of Sypro Orange binding site(s) showing that gp12 reacquires its quaternary structure upon interaction with the capsid (Figure 22). The results suggest that the capsid acts as a nucleation platform where three unfolded monomers meet at a gp13 hexamer and then twist together to rebuild the gp12 capsid bound trimer. Interestingly, the concentration-dependence of unfolded gp12 binding to capsids shows that a relatively homogeneous population of capsids with similar amounts of gp12 is present for a defined concentration of gp12, as assessed by their electrophoretic mobility (Figure 20c,e). In spite of the limited resolution of this technique, the result reveals that the binding reaction lacks significant cooperativity and calls for structural analysis of these capsid intermediates to investigate if there is a defined pattern of gp12 distribution in the icosahedral capsid lattice associated to partial occupancy. Cryo-electron microscopy and extensive processing to sort the capsid particles images dataset will be necessary to deal with gp12 flexibility⁸⁰, with some heterogeneity of gp12 occupancy, and that the occupancy pattern might not follow an icosahedral distribution: a (very) challenging task!

This thesis work provided a comprehensive description how the bacteriophage SPP1 auxiliary protein gp12 folds and associates to build a trimer with a collagen fold. The isolated protein exhibits common biophysical properties with eukaryotic collagens and prokaryotic CMCPs while association with viral capsids strongly stabilizes its fold rendering it temperature resistant. The gp12 reversible temperature unfolding/dissociation behavior renders it a versatile biotechnological tool to engineer SPP1 capsids *in vitro* with trimeric hybrid proteins that change the capsid surface properties. Gp12 thus appears as a very promising minimal model system to study the behavior of collagen-like proteins, as a module for protein engineering, and to

investigate viral capsids behavior.

Bioinformatics studies of gp12, of its protein relatives that share a similar modular organization and a large-scale survey of prokaryotic and viral CMCPs showed that gp12 is not a singular case in Nature. CMCPs appears as a frequent protein structural module that was conserved during evolution, although rarely used in archaea and their viruses (Figure 25). The number of CMCPs identified in our bioinformatics analyses is very high, which renders their study a challenging task in order to understand how and why bacterial and phages need those proteins.

7 Perspectives

This thesis work provided insights on the gp12 folding/association, on its structural organization and on its interaction with the SPP1 DNA-filled capsid. It also opened the way for future studies while more functional questions remain to be elucidated.

A future direction of study is to define experimentally the role of the gp12 amino and carboxyl terminus domains, namely their contribution for gp12 trimerization and for binding to the SPP1 capsid. Determination of the gp12 atomic structure by X-ray crystallography or NMR would be a major step to address this question but also to rationalize the molecular behavior of gp12 addressed in this work. I have obtained diffracting crystals of gp12 trimers but was unable to solve the phase to determine the structure in spite of numerous efforts with Dr S Bressanelli (I2BC). NMR spectra were also recorded with Dr S Zinn-Justin (I2BC). These initial studies provide a good basis for future structural studies. Their combination with cryo-EM reconstructions of the SPP1 icosahedral capsid determined in the group of Prof EV Orlova (Birkbeck College) aims to unravel the fold of the isolated gp12 trimer in solution and how it changes during attachment at the quasi-equivalent environment of the capsid hexamer. Such analysis would hopefully provide a clue how the interaction raises the thermal stability of the gp12 trimer by 20°C. This information will provide a molecular framework of the gp12 building plan as a model CMCP and open the way for its rational engineering as a trimeric scaffold at the SPP1 capsid surface.

A second research direction will be to engineer the gp12 protein. We compared the SPP1 gp12 sequence with homologous proteins from close SPP1 relative viruses (SF6, rho15 and 41c). The sequence of the proteins is conserved with the exception that the SPP1 gp12 central region counts 8 GXY triplets while 10 triplets are found in the other phages. The two additional triplets are the duplication of the GPQGPD motif (data not shown; K Djacem, unpublished). This duplication shows the extendable nature of the GXY motif. The central region of gp12 can thus be engineered by

extension/duplication of the GXY triplets to make longer collagen-like rod offering a longer spacer between the amino and carboxyl terminus domains while maintaining gp12 functional for capsid binding. Longer GXY repeats and more variable N and C-termini were identified in the search for gp12-related proteins (Figures 23b and 24a). It will be also interesting to construct hybrid proteins in which the gp12 amino or carboxyl non-collagenous terminus is swapped with domains from those putative phage proteins with central collagen-like segments. For that purpose we identified the protein ADF59141.1 from the *B. subtilis* temperate phage ϕ 105 that has a modular organization similar to gp12 (Figures 23c and 24b) and is encoded within the phage capsid genes cluster of an infectious phage. Its protein sequence features a central (GXY)₂₄ repeat and the carboxyl terminus has some similarity with gp12. An actualized extensive search for CMCPs associated to phage capsids will be useful to provide modules for assembling gp12 hybrid proteins carrying amino or carboxyl terminus from other phages putative auxiliary proteins. Their study aims to identify which domain intervenes for gp12 trimerization and for binding to the capsid. They might also provide information on the function of this group of proteins in the phage particle, e.g. in adhesion to bacteria or surfaces, a question that remains presently unanswered.

8 References

1. Levine, A.J. The origins of Virology. in *Fields Virology* (Fourth edition), vol. 1. (Lippincott Williams & Wilkins, 2001).
2. Dimmock, N., Easton, A. & Leppard, K. *Introduction to Modern Virology*. (Wiley, 2007).
3. Kung, S. & Yang, S.-F. *Discoveries in Plant Biology*. (World Scientific, 1998).
4. Creager, A. N. H., Scholthof, K.-B. G., Citovsky, V. & Scholthof, H. B. Tobacco Mosaic Virus: Pioneering Research for a Century. *Plant Cell Online* **11**, 301–308 (1999).
5. Douglas, Harper. Virus. *The Online Etymology Dictionary* Available at: http://www.etymonline.com/index.php?term=virus&allowed_in_frame=0.
6. Kausche, G. A., Pfankuch, E. & Ruska, H. Die Sichtbarmachung von pflanzlichem Virus im Übermikroskop. *Naturwissenschaften* **27**, 292–299 (1939).
7. Borries, B. von, Ruska, E. & Ruska, H. Bakterien und Virus in Übermikroskopischer Aufnahme. *Klin. Wochenschr.* **17**, 921–925 (1938).
8. Twort, F. W. An investigation on the nature of ultra-microscopic viruses. *The Lancet* **186**, 1241–1243 (1915).
9. D’Herelle F. Sur un microbe invisible antagoniste des bacillus dysentériques. *Acad Sci Paris.* **165**, 373–375 (1917).
10. Abedon, S. T., Thomas-Abedon, C., Thomas, A. & Mazure, H. Bacteriophage prehistory : Is or is not Hankin, 1896, a phage reference? *Bacteriophage* **1**, 174–178 (2011).
11. Ackermann, H.-W. Ruska H. Visualization of bacteriophage lysis in the hypermicroscope. *Naturwissenschaften* 1940 28:45-6. *Bacteriophage* **1**, 183–185 (2011).

12. Cann, A. *Principles of Molecular Virology*. (Academic Press, 2012).
13. Bamford, D. H., Grimes, J. M. & Stuart, D. I. What does structure tell us about virus evolution? *Curr. Opin. Struct. Biol.* **15**, 655–663 (2005).
14. 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).
15. Koonin, E. V. Virology: Gulliver among the Lilliputians. *Curr. Biol.* **15**, R167–R169 (2005).
16. Yutin, N., Wolf, Y. I. & Koonin, E. V. Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. *Virology* **466–467**, 38–52 (2014).
17. Scola, B. L. *et al.* A Giant Virus in Amoebae. *Science* **299**, 2033–2033 (2003).
18. Raoult, D. *et al.* The 1.2-megabase genome sequence of Mimivirus. *Science* **306**, 1344–1350 (2004).
19. Svenstrup, H. F., Fedder, J., Abraham-Peskir, J., Birkelund, S. & Christiansen, G. Mycoplasma genitalium attaches to human spermatozoa. *Hum. Reprod.* **18**, 2103–2109 (2003).
20. Taylor-Robinson, D. & Jensen, J. S. Mycoplasma genitalium: from Chrysalis to Multicolored Butterfly. *Clin. Microbiol. Rev.* **24**, 498–514 (2011).
21. Walker, D. H. Rickettsiae. in *Medical Microbiology* (ed. Baron, S.) (University of Texas Medical Branch at Galveston, 1996).
22. Becker, Y. Chlamydia. in *Medical Microbiology* (ed. Baron, S.) (University of Texas Medical Branch at Galveston, 1996).
23. Davey, N. E., Travé, G. & Gibson, T. J. How viruses hijack cell regulation. *Trends Biochem. Sci.* **36**, 159–169 (2011).
24. Alonso, J. M., Górzny, M. Ł. & Bittner, A. M. The physics of tobacco mosaic virus and virus-based devices in biotechnology. *Trends Biotechnol.* **31**, 530–538 (2013).
25. Richardson, J. S., Dekker, J. D., Croyle, M. A. & Kobinger, G. P. Recent advances

- in Ebolavirus vaccine development. *Hum. Vaccin.* **6**, 439–449 (2010).
26. Poranen, M. M. & Tuma, R. Self-assembly of double-stranded RNA bacteriophages. *Virus Res.* **101**, 93–100 (2004).
 27. Gronenborn, B. Nanoviruses: genome organisation and protein function. *Vet. Microbiol.* **98**, 103–109 (2004).
 28. Baker, T. S., Olson, N. H. & Fuller, S. D. Adding the third dimension to virus life cycles: three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs. *Microbiol. Mol. Biol. Rev.* **63**, 862–922 (1999).
 29. Caspar, D. L. & Klug, A. Physical principles in the construction of regular viruses. *Cold Spring Harb. Symp. Quant. Biol.* **27**, 1–24 (1962).
 30. Prasad, B. V. V. & Schmid, M. F. Principles of Virus Structural Organization. *Adv. Exp. Med. Biol.* **726**, 17–47 (2012).
 31. Zhou Z.H. & Chiou J. Protein chainmail variants in dsDNA viruses. *AIMS Biophys.* **2**, 200–218 (2015).
 32. Dai, X. & Zhou, Z.H. Structure of the herpes simplex virus 1 capsid with associated tegument protein complexes. *Science* **360**, pii: eaao7298 (2018).
 33. Tavares P. The bacteriophage head-to-tail interface. *Subcell. Biochem.* **88**, 305–328 (2018).
 34. Murphy, F. A. Virus Taxonomy and Nomenclature. in *Laboratory Diagnosis of Infectious Diseases Principles and Practice* 153–176 (Springer New York, 1988).
 35. Gibbs, A. J. Viral taxonomy needs a spring clean; its exploration era is over. *Viol. J.* **10**, 254 (2013).
 36. Baltimore, D. Expression of animal virus genomes. *Bacteriol. Rev.* **35**, 235–241 (1971).
 37. Navaratnarajah, C. K., Warriar, R. & Kuhn, R. J. Assembly of Viruses: Enveloped Particles. in *Encyclopedia of Virology* (Third Edition) (ed. Regenmortel, B. W. J.

- M. H. V. V.) 193–200 (Academic Press, 2008).
38. Huang, H.-H. *et al.* Global comparison of multiple-segmented viruses in 12-dimensional genome space. *Mol. Phylogenet. Evol.* **81**, 29–36 (2014).
 39. Lopes, A., Tavares, P., Petit, M.-A., Guérois, R. & Zinn-Justin, S. Automated classification of tailed bacteriophages according to their neck organization. *BMC Genomics* **15**, 1027 (2014).
 40. Xu, F. *et al.* Exploring virus relationships based on virus-host protein-protein interaction network. *BMC Syst. Biol.* **5 Suppl 3**, S11 (2011).
 41. Adriaenssens, E. M. *et al.* Integration of genomic and proteomic analyses in the classification of the Siphoviridae family. *Virology* **477**, 144-154 (2015).
 42. van Regenmortel, M. H. V. & Mahy, B. W. J. Emerging Issues in Virus Taxonomy. *Emerg. Infect. Dis.* **10**, 8–13 (2004).
 43. Simmonds, P. Methods for virus classification and the challenge of incorporating metagenomic sequence data. *J. Gen. Virol.* **96**, 1193-1206 (2015).
 44. Ackermann, H.-W. 5500 Phages examined in the electron microscope. *Arch. Virol.* **152**, 227–243 (2007).
 45. 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).
 46. Ackermann, H.-W. Phage classification and characterization. *Methods Mol. Biol.* **501**, 127–140 (2009).
 47. Krupovic, M., Prangishvili, D., Hendrix, R. W. & Bamford, D. H. Genomics of Bacterial and Archaeal Viruses: Dynamics within the Prokaryotic Virosphere. *Microbiol. Mol. Biol. Rev.* **75**, 610–635 (2011).
 48. Van Twest, R. & Kropinski, A. M. Bacteriophage enrichment from water and soil. *Methods Mol. Biol.* **501**, 15–21 (2009).

49. Ackermann, H. W. Tailed bacteriophages: the order caudovirales. *Adv. Virus Res.* **51**, 135–201 (1998).
50. Rixon, F. J. & Schmid, M. F. Structural similarities in DNA packaging and delivery apparatuses in Herpesvirus and dsDNA bacteriophages. *Curr. Opin. Virol.* **5**, 105–110 (2014).
51. Baker, M. L., Jiang, W., Rixon, F. J. & Chiu, W. Common Ancestry of Herpesviruses and Tailed DNA Bacteriophages. *J. Virol.* **79**, 14967–14970 (2005).
52. Casjens, S. & Hendrix, R. Control mechanisms in dsDNA bacteriophage assembly. in *The Bacteriophages*, vol 1, Calendar R. (ed). (Plenum Press, New York, 1988).
53. Veesler, D. & Cambillau, C. A common evolutionary origin for tailed bacteriophage functional modules and bacterial machineries. *Microbiol Mol. Biol. Rev.* **75**, 423-433 (2011).
54. Böck, D. *et al.* In situ architecture, function, and evolution of a contractile injection system. *Science* **357**, 713-717 (2017).
55. Cascales, E. Microbiology: And Amoebophilus Invented the Machine Gun! *Curr. Biol.* **27**, R1170-R1173 (2017).
56. Prevelige, P.E. & Fane, B.A. Building the machines: scaffolding protein functions during bacteriophage morphogenesis. *Adv. Exp. Med. Biol.* **726**, 325-350 (2012).
57. Oh, B., Moyer, C.L., Hendrix, R.W. & Duda, R.L. The delta domain of the HK97 major capsid protein is essential for assembly. *Virology* **456-457**, 171-178 (2014).
58. 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).
59. Bazinet, C. & King, J. The DNA translocating vertex of dsDNA bacteriophage. *Annu. Rev. Microbiol.* **39**, 109–129 (1985).
60. 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).
61. Newcomb, W.W., Homa, F.L. & Brown, J.C. Involvement of the portal at an early step in herpes simplex virus capsid assembly. *J. Virol.* **79**, 10540–10546 (2005).
 62. Motwani, T. *et al.* A viral scaffolding protein triggers portal ring oligomerization and incorporation during procapsid assembly. *Sci. Adv.* **3**, e1700423 (2017).
 63. Rao, V.B. & Feiss, M. Mechanisms of DNA packaging by large double-stranded DNA viruses. *Annu. Rev. Virol.* **2**, 351-378 (2015).
 64. Smith, D.E. *et al.* The bacteriophage ϕ 29 portal motor can package DNA against a large internal force. *Nature* **413**, 748–752 (2001).
 65. Cornilleau, C. *et al.* The nuclease domain of the SPP1 packaging motor coordinates DNA cleavage and encapsidation. *Nucleic Acids Res.* **41**, 340-354 (2013).
 66. Oliveira, L., Tavares, P. & Alonso, J.C. Headful DNA packaging: bacteriophage SPP1 as a model system. *Virus Res.* **173**, 247-259 (2013).
 67. Oliveira, L., Cuervo, A. & Tavares, P. Direct interaction of the bacteriophage SPP1 packaging ATPase with the portal protein. *J. Biol. Chem.* **285**, 7366-7373 (2010).
 68. Chemla, Y.R. & Smith, D.E. Single-molecule studies of viral DNA packaging. *Adv. Exp. Med. Biol.* **726**, 549–584 (2012).
 69. Jardine, J.P. Slow and steady wins the race: physical limits on the rate of viral DNA packaging. *Curr. Opin. Virol.* **36**, 32-37 (2019).
 70. Orlova, E.V. *et al.* Structure of a viral DNA gatekeeper at 10 Å resolution by cryo-electron microscopy. *EMBO J.* **22**, 1255-1262 (2003).
 71. Tavares, P., Zinn-Justin, S. & Orlova, E. V. Genome gating in tailed bacteriophage capsids. *Adv. Exp. Med. Biol.* **726**, 585–600 (2012).
 72. Kalejta, R.F. Tegument proteins of human cytomegalovirus. *Microbiol. Mol. Biol. Rev.* **72**, 249-265 (2008).

73. Katsura, I. Determination of bacteriophage lambda tail length by a protein ruler. *Nature* **327**, 73–75 (1987).
74. Leiman, P.G. & Shneider, M.M. Contractile tail machines of bacteriophages. *Adv. Exp. Med. Biol.* **726**, 93–114 (2012).
75. Davidson, A.R., Cardarelli, L., Pell, L.G., Radford, D.R. & Maxwell, K.L. Long noncontractile tail machines of bacteriophages. *Adv. Exp. Med. Biol.* **726**, 115–142 (2012).
76. 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.* **91**, 1164–1178 (2014).
77. Lander, G. C. *et al.* Bacteriophage lambda stabilization by auxiliary protein gpD: timing, location, and mechanism of attachment determined by cryo-EM. *Structure* **16**, 1399–1406 (2008).
78. Rao, V. B. & Black, L. W. Structure and assembly of bacteriophage T4 head. *Viol. J.* **7**, 356 (2010).
79. Effantin, G., Boulanger, P., Neumann, E., Letellier, L. & Conway, J. F. Bacteriophage T5 structure reveals similarities with HK97 and T4 suggesting evolutionary relationships. *J. Mol. Biol.* **361**, 993–1002 (2006).
80. White, H. E. *et al.* Capsid structure and its stability at the late stages of bacteriophage SPP1 assembly. *J. Virol.* **86**, 6768–6777 (2012).
81. 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).
82. Hernando-Pérez, M., Lambert, S., Nakatani-Webster, E., Catalano, C. E. & de Pablo, P. J. Cementing proteins provide extra mechanical stabilization to viral cages. *Nat. Commun.* **5**, 4520 (2014).

83. Sae-Ueng, U. *et al.* Major capsid reinforcement by a minor protein in herpesviruses and phage. *Nucleic Acids Res.* **42**, 9096–9107 (2014).
84. Mateu, M. G. Assembly, stability and dynamics of virus capsids. *Arch. Biochem. Biophys.* **531**, 65–79 (2013).
85. Singh, P., Nakatani, E., Goodlett, D. R. & Catalano, C. E. A Pseudo-Atomic Model for the Capsid Shell of Bacteriophage Lambda Using Chemical Cross-Linking/Mass Spectrometry and Molecular Modeling. *J. Mol. Biol.* **425**, 3378–3388 (2013).
86. Yang, Q., Maluf, N. K. & Catalano, C. E. Packaging of a unit-length viral genome: the role of nucleotides and the gpD decoration protein in stable nucleocapsid assembly in bacteriophage lambda. *J. Mol. Biol.* **383**, 1037–1048 (2008).
87. Yang, F. *et al.* Novel fold and capsid-binding properties of the lambda-phage display platform protein gpD. *Nat. Struct. Biol.* **7**, 230–237 (2000).
88. Forrer, P., Chang, C., Ott, D., Wlodawer, A. & Plückthun, A. Kinetic stability and crystal structure of the viral capsid protein SHP. *J. Mol. Biol.* **344**, 179–193 (2004).
89. Stone, N.P. *et al.* A hyperthermophilic phage decoration protein suggests common evolutionary origin with herpesvirus triplex proteins and an Anti-CRISPR protein. *Structure* **26**, 936–947 (2018).
90. Wang, Z. *et al.* Structure of the Marine Siphovirus TW1: Evolution of Capsid-Stabilizing Proteins and Tail Spikes. *Structure* **26**, 238-248 (2018).
91. Newcomer, R.L. *et al.* The phage L capsid decoration protein has a novel OB-fold and an unusual capsid binding strategy. *Elife* **8**, pii: e45345 (2019).
92. Gilcrease, E.B., Winn-Stapley, D.A., Hewitt, F.C., Joss, L. & Casjens, S.R. Nucleotide sequence of the head assembly gene cluster of bacteriophage L and decoration protein characterization. *J. Bacteriol.* **187**, 2050–2057 (2005).
93. 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. U. S. A.* **114**, E8184-E8193 (2017).
94. Qin, L., Fokine, A., O'Donnell, E., Rao, V. B. & Rossmann, M. G. Structure of the small outer capsid protein, Soc: a clamp for stabilizing capsids of T4-like phages. *J. Mol. Biol.* **395**, 728–741 (2010).
95. Robertson, K., Furukawa, Y., Underwood, A., Black, L. & Liu, J. L. Deletion of the Hoc and Soc capsid proteins affects the surface and cellular uptake properties of bacteriophage T4 derived nanoparticles. *Biochem. Biophys. Res. Commun.* **418**, 537–540 (2012).
96. Wikoff, W.R. *et al.* Topologically linked protein rings in the bacteriophage HK97 capsid. *Science* **289**, 2129-2133 (2000).
97. Fokine, A. *et al.* Structure of the three N-terminal immunoglobulin domains of the highly immunogenic outer capsid protein from a T4-like bacteriophage. *J. Virol.* **85**, 8141-8148 (2011).
98. Vernhes, E. *et al.* High affinity anchoring of the decoration protein pb10 onto the bacteriophage T5 capsid. *Sci. Rep.* **7**, 43977 (2017).
99. Lander, G.C. *et al.* Capsomer dynamics and stabilization in the T=12 marine bacteriophage SIO-2 and its procapsid studied by CryoEM. *Structure* **20**, 498-503 (2012).
100. Mikawa, Y. G., Maruyama, I. N. & Brenner, S. Surface display of proteins on bacteriophage lambda heads. *J. Mol. Biol.* **262**, 21–30 (1996).
101. Li, Q., Shivachandra, S. B., Zhang, Z. & Rao, V. B. Assembly of the small outer capsid protein, Soc, on bacteriophage T4: a novel system for high density display of multiple large anthrax toxins and foreign proteins on phage capsid. *J. Mol. Biol.* **370**, 1006–1019 (2007).
102. Gamkrelidze, M. & Dąbrowska, K. T4 bacteriophage as a phage display platform.

- Arch. Microbiol.* **196**, 473–479 (2014).
103. Tao, P. *et al.* A Bacteriophage T4 Nanoparticle-Based Dual Vaccine against Anthrax and Plague. *MBio* **9**, pii: e01926-18 (2018).
104. Schwarz, B. *et al.* Symmetry Controlled, Genetic Presentation of Bioactive Proteins on the P22 Virus-like Particle Using an External Decoration Protein. *ACS Nano* **9**, 9134–9147 (2015).
105. 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).
106. Alonso, J.C., Tavares, P., Lurz, R., and Trautner, T.A. Bacteriophage SPP1. In *The Bacteriophages* (Calendar, R., ed.), pp. 331-349, Oxford University Press, New York, USA (2006).
107. Plisson, C. *et al.* Structure of bacteriophage SPP1 tail reveals trigger for DNA ejection. *EMBO J.* **26**, 3720–3728 (2007).
108. 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).
109. Godinho L.M. *et al.* The revisited genome of *Bacillus subtilis* bacteriophage SPP1. *Viruses* **10**, pii: E705 (2018).
110. 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.* **190**, 4989-4996 (2008).
111. São-José, C. *et al.* The ectodomain of the viral receptor YueB forms a fiber that triggers DNA ejection of bacteriophage SPP1 DNA. *J. Biol. Chem.* **281**, 11464-11470 (2006).
112. Jakutyte 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).

113. 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* **8**, e67840 (2013).
114. Jakutyte, L. *et al.* First steps of bacteriophage SPP1 entry into *Bacillus subtilis*. *Virology* **422**, 425–434 (2012).
115. Poh, S.L. *et al.* Oligomerization of the SPP1 scaffolding protein. *J. Mol. Biol.* **378**, 551-564 (2008).
116. Stiege, A., Isidro, A., Dröge, A. & Tavares, P. Specific and stoichiometric targeting of a DNA-binding protein to the SPP1 procapsid by interaction with the portal oligomer. *Mol. Microbiol.* **49**, 1201-1212 (2003).
117. Vinga, I. *et al.* The minor capsid protein gp7 of bacteriophage SPP1 is required for efficient infection of *Bacillus subtilis*. *Mol. Microbiol.* **61**, 1609-1621 (2006).
118. 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.* **252**, 386-398 (1995).
119. Djacem K., Tavares P. & Oliveira, L. Bacteriophage SPP1 pac cleavage: a precise cut without sequence specificity requirement. *J. Mol. Biol.* **429**, 1381-1395 (2017).
120. 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).
121. 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).
122. Lhuillier, S. *et al.* Structure of bacteriophage SPP1 head-to-tail connection reveals mechanism for viral DNA gating. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 8507–8512 (2009).
123. 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).

124. Auzat, I., Dröge, A., Weise, F., Lurz, R. & Tavares, P. Origin and function of the two major tail proteins of bacteriophage SPP1. *Mol. Microbiol.* **70**, 557–569 (2008).
125. Langlois, C. *et al.* Bacteriophage SPP1 Tail Tube Protein self-assembles into β -structure rich tubes. *J. Biol. Chem.* **290**, 3836–3849 (2015).
126. Goulet, A. *et al.* The opening of the SPP1 bacteriophage tail, a prevalent mechanism in Gram-positive-infecting siphophages. *J. Biol. Chem.* **286**, 25397–25405 (2011).
127. Vinga, I. *et al.* Role of bacteriophage SPP1 tail spike protein gp21 on host cell receptor binding and trigger of phage DNA ejection. *Mol. Microbiol.* **83**, 289–303 (2012).
128. Nováček, J. *et al.* Structure and genome release of Twort-like Myoviridae phage with a double-layered baseplate. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 9351–9356 (2016).
129. Arnaud, C.A. *et al.* Bacteriophage T5 tail tube structure suggests a trigger mechanism for Siphoviridae DNA ejection. *Nat. Commun.* **8**, 1953 (2017).
130. 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).
131. Dröge, A. Capsidmorphogenese des Bakteriophagen SPP1. *Doctoral Thesis*. Technische Universität Berlin, Germany (1998).
132. Gay S. & Miller EJ. What is collagen, what is not. *Ultrastruct Pathol.* **4**, 365–377 (1983).
133. Bella J. Collagen structure: new tricks from a very old dog. *Biochem J.* **473**, 1001–1025 (2016).
134. Bella J & Hulmes DJ. Fibrillar Collagens. *Subcell Biochem.* **82**, 457–490 (2017).

135. Kadler, K. E., Baldock, C., Bella, J. & Boot-Handford, R. P. Collagens at a glance. *J. Cell Sci.* **120**, 1955–1958 (2007).
136. Ricard-Blum, S. The collagen family. *Cold Spring Harb Perspect Biol.* **3**, a004978. (2011).
137. Gordon M.K. & Hahn R.A. Collagens. *Cell Tissue Res.* **339**, 247-257 (2010).
138. Hulmes, D. J. S. Collagen Diversity, Synthesis and Assembly. in *Collagen* (ed. Fratzl, P.) 15–47 (Springer US, 2008).
139. Shoulders M.D. & Raines R.T. Collagen structure and stability. *Annu Rev Biochem.* **78**, 929-958 (2009).
140. Gelse, K., Pöschl, E. & Aigner, T. Collagens-structure, function, and biosynthesis. *Adv. Drug Deliv. Rev.* **55**, 1531–1546 (2003).
141. Garnero, P. *et al.* Extracellular post-translational modifications of collagen are major determinants of biomechanical properties of fetal bovine cortical bone. *Bone* **38**, 300–309 (2006).
142. Myllyharju, J. & Kivirikko, K. I. Collagens, modifying enzymes and their mutations in humans, flies and worms. *Trends Genet.* **20**, 33–43 (2004).
143. Ricard-Blum, S., Baffet, G. & Théret N. Molecular and tissue alterations of collagens in fibrosis. *Matrix Biol.* **68-69**, 122-149 (2018).
144. Hicks, D *et al.* Mutations in the collagen XII gene define a new form of extracellular matrix-related myopathy. *Hum. Mol. Genet.* **23**, 2353-63 (2014).
145. Punetha J. *et al.* Novel Col12A1 variant expands the clinical picture of congenital myopathies with extracellular matrix defects. *Muscle Nerve* **55**, 277-281 (2017).
146. Logan, C.V. *et al.* Congenital Myasthenic Syndrome Type 19 Is Caused by Mutations in *COL13A1*, Encoding the Atypical Non-fibrillar Collagen Type XIII α 1 Chain. *Am. J. Hum. Genet.* **97**, 878–885 (2015).
147. Dusl, M. *et al.* Congenital myasthenic syndrome caused by novel COL13A1

- mutations. *J. Neurol.* **266**, 1107-1112 (2019).
148. Su, J. *et al.* Collagen XIX is expressed by interneurons and contributes to the formation of hippocampal synapses. *J. Comp. Neurol.* **518**, 229-253 (2010).
149. Khan, M.I. *et al.* Whole-exome sequencing analysis reveals co-segregation of a COL20A1 missense mutation in a Pakistani family with striate palmoplantar keratoderma. *Genes Genomics* **40**, 789-795 (2018).
150. Ton, Q.V. *et al.* Collagen COL22A1 maintains vascular stability and mutations in COL22A1 are potentially associated with intracranial aneurysms. *Dis. Model Mech.* **11**, dmm033654 (2018).
151. Spivey, K.A. *et al.* A role for collagen XXIII in cancer cell adhesion, anchorage-independence, and metastasis. *Oncogene* **31**, 2362–2372 (2012).
152. Shinwari J.M.A. *et al.* Recessive Mutations in COL25A1 Are a Cause of Congenital Cranial Dysinnervation Disorder. *Am. J. Hum. Genet.* **96**, 147–152 (2015).
153. Khan, A.O. & Al-Mesfer, S. Recessive COL25A1 mutations cause isolated congenital ptosis or exotropic Duane syndrome with synergistic divergence. *J. AAPOS* **19**, 463-465 (2015).
154. Gonzaga-Jauregui, C. *et al.* Mutations in COL27A1 cause Steel syndrome and suggest a founder mutation effect in the Puerto Rican population. *Eur. J. Hum. Genet.* **23**, 342-346 (2015).
155. Gariballa, N. *et al.* A novel aberrant splice site mutation in COL27A1 is responsible for Steel syndrome and extension of the phenotype to include hearing loss. *Am J Med Genet A.* **173**, 1257-1263 (2017).
156. Chen, A. *et al.* The Molecular Basis of Genetic Collagen Disorders and Its Clinical Relevance. *J. Bone Joint Surg. Am.* **100**, 976-986 (2018).
157. Lin, C.J., Lin, C.Y. & Stitzel, N.O. Genetics of the extracellular matrix in aortic

- aneurysmal diseases. *Matrix Biol.* **71-72**, 128-143 (2018).
158. Plüddemann, A., Neyen, C. & Gordon, S. Macrophage scavenger receptors and host-derived ligands. *Methods* **43**, 207–217 (2007).
159. Ghosh, N. *et al.* Collagen-Like Proteins in Pathogenic E. coli Strains. *PLoS One* **7**, e37872 (2012).
160. Rasmussen, M., Jacobsson, M. & Björck, L. Genome-based identification and analysis of collagen-related structural motifs in bacterial and viral proteins. *J. Biol. Chem.* **278**, 32313–32316 (2003).
161. Yu, Z., An, B., Ramshaw, J. A. M. & Brodsky, B. Bacterial collagen-like proteins that form triple-helical structures. *J. Struct. Biol.* **186**, 451–461 (2014).
162. Lukomski, S. *et al.* Identification and characterization of the scl gene encoding a group A Streptococcus extracellular protein virulence factor with similarity to human collagen. *Infect. Immun.* **68**, 6542–6553 (2000).
163. Lukomski S *et al.* Collagen-like proteins of pathogenic streptococci. *Mol Microbiol.* **103**,919-930 (2017)
164. Xu, Y., Keene, D. R., Bujnicki, J. M., Höök, M. & Lukomski, S. Streptococcal Scl1 and Scl2 proteins form collagen-like triple helices. *J. Biol. Chem.* **277**, 27312–27318 (2002).
165. Leikina, E., Merts, M. V., Kuznetsova, N. & Leikin, S. Type I collagen is thermally unstable at body temperature. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 1314–1318 (2002).
166. Cao, H. & Xu, S.-Y. Purification and characterization of type II collagen from chick sternal cartilage. *Food Chem.* **108**, 439–445 (2008).
167. Oliver-Kozup, H. A. *et al.* The streptococcal collagen-like protein-1 (Scl1) is a significant determinant for biofilm formation by group A Streptococcus. *BMC Microbiol.* **11**, 262 (2011).

168. Boydston, J. A., Chen, P., Steichen, C. T. & Turnbough, C. L. Orientation within the Exosporium and Structural Stability of the Collagen-Like Glycoprotein BclA of *Bacillus anthracis*. *J. Bacteriol.* **187**, 5310–5317 (2005).
169. Caldentey, J., Tuma, R. & Bamford, D. H. Assembly of bacteriophage PRD1 spike complex: role of the multidomain protein P5. *Biochemistry* **39**, 10566–10573 (2000).
170. Merckel, M. C., Huiskonen, J. T., Bamford, D. H., Goldman, A. & Tuma, R. The structure of the bacteriophage PRD1 spike sheds light on the evolution of viral capsid architecture. *Mol. Cell* **18**, 161–170 (2005).
171. Shah, N. *et al.* Exposure to mimivirus collagen promotes arthritis. *J. Virol.* **88**, 838–845 (2014).
172. Luther, K. B. *et al.* Mimivirus collagen is modified by bifunctional lysyl hydroxylase and glycosyltransferase enzyme. *J. Biol. Chem.* **286**, 43701–43709 (2011).
173. Kramer, R. Z., Bella, J., Brodsky, B. & Berman, H. M. The crystal and molecular structure of a collagen-like peptide with a biologically relevant sequence. *J. Mol. Biol.* **311**, 131–147 (2001).
174. Kramer, R. Z. *et al.* X-ray crystallographic determination of a collagen-like peptide with the repeating sequence (Pro-Pro-Gly). *J. Mol. Biol.* **280**, 623–638 (1998).
175. Brodsky, B. & Persikov, A. V. Molecular structure of the collagen triple helix. *Adv. Protein Chem.* **70**, 301–339 (2005).
176. Stetefeld, J. *et al.* Collagen stabilization at atomic level: crystal structure of designed (GlyProPro)₁₀foldon. *Structure* **11**, 339–346 (2003).
177. Gjaltema R.A. & Bank R.A. Molecular insights into prolyl and lysyl hydroxylation of fibrillar collagens in health and disease. *Crit Rev Biochem Mol*

- Biol.* **52**, 74-95 (2017).
178. Bella J. & Hulmes D.J. Fibrillar Collagens. *Subcell. Biochem.* **82**, 457-490 (2017).
179. Engel, J. & Bächinger, H. P. Cooperative equilibrium transitions coupled with a slow annealing step explain the sharpness and hysteresis of collagen folding. *Matrix Biol.* **19**, 235–244 (2000).
180. Ackerman, M. S. *et al.* Sequence dependence of the folding of collagen-like peptides. Single amino acids affect the rate of triple-helix nucleation. *J. Biol. Chem.* **274**, 7668–7673 (1999).
181. Bella, J., Brodsky, B. & Berman, H. M. Hydration structure of a collagen peptide. *Structure* **3**, 893–906 (1995).
182. Shoulders, M. D. & Raines, R. T. Collagen structure and stability. *Annu. Rev. Biochem.* **78**, 929–958 (2009).
183. Bella J., Eaton, M., Brodsky, B. & Berman, H.M. Crystal and molecular structure of a collagen-like peptide at 1.9 Å resolution. *Science* **266**, 75-81 (1994).
184. Raman, S. S., Parthasarathi, R., Subramanian, V. & Ramasami, T. Role of length-dependent stability of collagen-like peptides. *J. Phys. Chem. B* **112**, 1533–1539 (2008).
185. Persikov, A. V., Ramshaw, J. A. M. & Brodsky, B. Prediction of collagen stability from amino acid sequence. *J. Biol. Chem.* **280**, 19343–19349 (2005).
186. Bächinger, H. P. & Davis, J. M. Sequence specific thermal stability of the collagen triple helix. *Int. J. Biol. Macromol.* **13**, 152–156 (1991).
187. Mohs, A. *et al.* Mechanism of stabilization of a bacterial collagen triple helix in the absence of hydroxyproline. *J. Biol. Chem.* **282**, 29757–29765 (2007).
188. McAlinden, A. *et al.* α -Helical Coiled-coil Oligomerization Domains Are Almost Ubiquitous in the Collagen Superfamily. *J. Biol. Chem.* **278**, 42200–42207 (2003).
189. Beck, K. & Brodsky, B. Supercoiled Protein Motifs: The Collagen Triple-Helix

- and the α -Helical Coiled Coil. *J. Struct. Biol.* **122**, 17–29 (1998).
190. Lurz, R. *et al.* Structural organisation of the head-to-tail interface of a bacterial virus. *J. Mol. Biol.* **310**, 1027-1037 (2001).
191. Alonso, J.C. *et al.* The complete nucleotide sequence and functional organization of Bacillus subtilis bacteriophage SPP1. *Gene* **204**, 201-212 (1997).
192. Haima, P., Bron, S. & Venema, G. The effect of restriction on shotgun cloning and plasmid stability in Bacillus subtilis Marburg. *Mol. Gen. Genet.* **209**, 335-342 (1987).
193. Dröge, A. & Tavares, P. *In vitro* packaging of DNA of the Bacillus subtilis bacteriophage SPP1. *J. Mol. Biol.* **296**, 103-115 (2000).
194. Becker, B. *et al.* Head morphogenesis genes of the Bacillus subtilis bacteriophage SPP1. *J. Mol. Biol.* **268**, 822-839 (1997).
195. Isidro, A., Santos, M.A., Henriques, A.O. & Tavares, P. The high-resolution functional map of bacteriophage SPP1 portal protein. *Mol. Microbiol.* **51**, 949-962 (2004).
196. Philo, J.S. An improved function for fitting sedimentation velocity data for low-molecular-weight solutes. *Biophys. J.* **72**, 435–444 (1997).
197. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).
198. Chenna, R. *et al.* Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31**, 3497-3500 (2003).
199. Cabré, F., Canela, E.I. & Canela, M.A. Accuracy and precision in the determination of Stokes radii and molecular masses of proteins by gel filtration chromatography. *J. Chromatogr.* **472**, 347-356 (1989).
200. Seifter, S. & Gallop, P.M. Collagenase from Clostridium histolyticum: Collagen + H₂O → Peptides Gelatin + H₂O → Peptides. *Methods Enzymol.* **5**, 659-665 (1962).

201. Wallace, B.A. & Janes, R.W. Synchrotron radiation circular dichroism spectroscopy of proteins: secondary structure, fold recognition and structural genomics. *Curr. Opin. Chem. Biol.* **5**, 567-571 (2001).
202. Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nat. Protoc.* **1**, 2876–2890 (2006).
203. Ericsson, U.B., Hallberg, B.M., Detitta, G.T., Dekker, N. & Nordlund, P. Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Anal. Biochem.* **357**, 289-298 (2006).
204. Miller, M. A., Scott, E. E. & Limburg, J. Expression, purification, crystallization and preliminary X-ray studies of a prolyl-4-hydroxylase protein from *Bacillus anthracis*. *Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun.* **64**, 788–791 (2008).
205. Peng, Y.Y., Nebl, T., Glattauer, V. & Ramshaw, J.A.M. Incorporation of hydroxyproline in bacterial collagen from *Streptococcus pyogenes*. *Acta Biomater.* **80**, 169-175 (2018).
206. Frank, S. *et al.* Stabilization of short collagen-like triple helices by protein engineering. *J. Mol. Biol.* **308**, 1081–1089 (2001).
207. Pell, L. G., Kanelis, V., Donaldson, L. W., Howell, P. L. & Davidson, A. R. The phage lambda major tail protein structure reveals a common evolution for long-tailed phages and the type VI bacterial secretion system. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 4160–4165 (2009).

The Collagen-like Protein gp12 Is a Temperature-dependent Reversible Binder of SPP1 Viral Capsids*

Received for publication, June 23, 2014, and in revised form, July 27, 2014. Published, JBC Papers in Press, July 29, 2014, DOI 10.1074/jbc.M114.590877

Mohamed Zairi^{†1}, Asita C. Stiege[§], Naima Nhiri[¶], Eric Jacquet^{¶||}, and Paulo Tavares^{‡2}

From the [†]Unité de Virologie Moléculaire et Structurale, UPR 3296 CNRS, Centre de Recherche de Gif, 91190 Gif-sur-Yvette, France, the [§]Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany, the [¶]Institut de Chimie des Substances Naturelles, UPR 2301 CNRS, Centre de Recherche de Gif, Gif-sur-Yvette, France, and the ^{||}IMAGIF CTPF and qPCR Platform, Centre de Recherche de Gif, 91190 Gif-sur-Yvette, France

Background: Auxiliary proteins bind to viral capsid surfaces, forming symmetric arrays of polypeptides.

Results: Collagen-like gp12 binds cooperatively to multiple sites of the bacteriophage SPP1 capsid in a reversible fashion.

Conclusion: The collagen fold and interaction with the capsid determine gp12 thermostability and folding/association properties.

Significance: Gp12 represents a novel type of viral capsid binders characterized by thermoswitchable properties.

Icosahedral capsids of viruses are lattices of defined geometry and homogeneous size. The (quasi-)equivalent organization of their protein building blocks provides, in numerous systems, the binding sites to assemble arrays of viral polypeptides organized with nanometer precision that protrude from the capsid surface. The capsid of bacterial virus (bacteriophage) SPP1 exposes, at its surface, the 6.6-kDa viral polypeptide gp12 that binds to the center of hexamers of the major capsid protein. Gp12 forms an elongated trimer with collagen-like properties. This is consistent with the fold of eight internal GXY repeats of gp12 to build a stable intersubunit triple helix in a prokaryotic setting. The trimer dissociates and unfolds at near physiological temperatures, as reported for eukaryotic collagen. Its structural organization is reacquired within seconds upon cooling. Interaction with the SPP1 capsid hexamers strongly stabilizes gp12, increasing its T_m to 54 °C. Above this temperature, gp12 dissociates from its binding sites and unfolds reversibly. Multivalent binding of gp12 trimers to the capsid is highly cooperative. The capsid lattice also provides a platform to assist folding and association of unfolded gp12 polypeptides. The original physicochemical properties of gp12 offer a thermoswitchable system for multivalent binding of the polypeptide to the SPP1 capsid surface.

Viruses are infectious agents characterized by an extracellular state, the virus particle or virion, which protects the viral genome from environmental aggression and ensures its highly efficient delivery to host cells for virus multiplication. The viral particle is a protein nanocage, sometimes combined with a lipid membrane, surrounding the nucleic acid molecule(s) that code(s) for the hereditary genetic information of the virus. A large number of prokaryotic and eukaryotic virions have an icosahedral protein shell of homogeneous size, termed the cap-

sid. Its self-assembly exploits (quasi-)equivalent interactions between a large number of identical protein subunits (1–3). Viruses with long dsDNA genomes, like tailed bacterial viruses (bacteriophages or phages) and the eukaryotic pathogen herpesvirus, first assemble an icosahedral protein lattice, the procapsid (4, 5) (Fig. 1). This structure is formed by major capsid protein subunit hexamers found at the planar faces of the icosahedron and by pentamers that define its angular vertices (2, 3). Viral DNA is then translocated to the procapsid interior through a specialized portal vertex by a powerful nanomotor, leading to tight packing of dsDNA in the capsid interior. During DNA packaging, the capsid undergoes a major conformational change called expansion. It leads to a gain in volume, stability, and, in numerous viral systems, to the creation of capsid auxiliary protein binding sites (3, 6–9). Those proteins cement structurally weak capsid points by establishing additional inter-hexamer bonding or attach to the center of hexamers (9–16). In both cases, they establish a symmetrically organized array of polypeptides at the capsid surface. In contrast to the conserved fold of the major capsid protein of the tailed bacteriophage-herpesviruses lineage (3, 5), their auxiliary proteins can have diverse length, structure, and biochemical properties (9–16).

The high fidelity of viral capsids assembly yields a population of homogeneous, robust particles. Their symmetric elements are arranged accurately with nanometer precision, offering excellent systems to engineer versatile enzymatic or bioactive nanoparticles (17–21). This can be efficiently achieved by molecular biology and chemical approaches (22, 23) that rely on detailed knowledge of the molecular structure of the virion and of the biochemical behavior of its components.

SPP1, like all other tailed bacterial viruses and herpesviruses, assembles a procapsid that serves as a container for subsequent viral DNA packaging (Fig. 1). The structure is composed of 415 subunits of the major capsid protein gp13, organized following quasi-equivalent interaction rules to build an icosahedron with a triangulation number (T) of 7 (16, 24). DNA pumping to the procapsid interior through a specialized portal vertex is accompanied by a major rearrangement of the capsid lattice that acquires a clear icosahedral outfit, increasing in diameter more than 50 Å (24, 25). This expansion process creates the binding

* This work was supported by institutional funding from the CNRS and from the Max Planck Institute for Molecular Genetics.

¹ Supported by a doctoral fellowship from Ministère de l'Éducation Nationale, de l'Enseignement supérieur et de la Recherche (MNERT).

² To whom correspondence should be addressed: Unité de Virologie Moléculaire et Structurale, UPR 3296 CNRS, Centre de Recherche de Gif, Bâtiment 14B, 1 ave. de la Terrasse, 91190 Gif-sur-Yvette, France. Tel.: +331-6982-3860; Fax: +331-6982-4308; E-mail: tavares@vms.cnrs-gif.fr.

A Collagen-like Binder of the SPP1 Viral Capsid

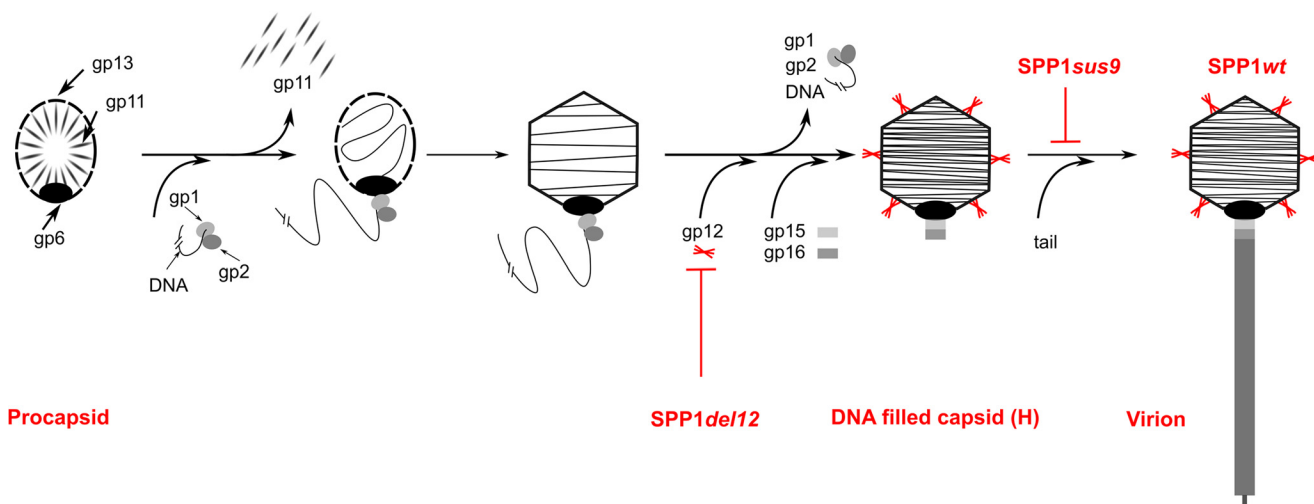


FIGURE 1. **Schematic of the SPP1 virion assembly pathway.** Phage proteins involved in capsid assembly are labeled as follows: *gp1-gp2*, terminase complex; *gp6*, portal protein; *gp11*, procapsid scaffolding protein; *gp12*, capsid auxiliary protein; *gp13*, major capsid protein; *gp15*, connector adaptor protein; *gp16*, connector stopper protein (25). Virus strains impaired in production of gp12 (SPP1*del12*) and of tail structures (SPP1*sus9*) are shown in red, and the step of assembly they affect is identified. SPP1*del12* was used to produce SPP1 infectious particles lacking gp12, whereas SPP1*sus9* and SPP1*sus9del12* were used to produce tailless DNA-filled capsids with (capsid H) and without gp12 (capsid HΔ12). Gp12 trimers are also highlighted in red.

site for gp12 (Fig. 1) (26) at the center of each of the 60 gp13 hexamers in the icosahedral lattice (16). Viral DNA packaging is followed by binding of a tail to the DNA-filled capsid, yielding the infectious virion (Fig. 1) (27, 28). To uncover the molecular principles of how auxiliary proteins interact with the surfaces of viral particles, we investigated the properties of gp12. This 6.6-kDa polypeptide is shown to adopt a collagen-like fold with the remarkable property of binding reversibly in a temperature-dependent fashion to its 60 sites at the SPP1 capsid surface.

EXPERIMENTAL PROCEDURES

Cloning Procedures and Creation of the Gene 12 Knockout SPP1 Strain

Gene 12 was cloned into plasmid pRSET A (Invitrogen) for protein overproduction in *Escherichia coli* using the strategy described by Lurz *et al.* (29). The resulting plasmid, pBT453, codes for tag-gp12, in which the gp12 amino terminus is fused in-frame to a 36-amino acid-long tag that includes a hexahistidine sequence. To engineer a cleavage site for tobacco etch virus (TEV)³ protease (ENLYFQG) between the tag and the gp12 amino acid sequence, gene 12 was amplified from pBT453 DNA with oligonucleotides TGS (TAAGGTACCGGATCCGAGAATCTGTACTTCCAGGGCATGTCTAAGCGTATACCGCGTTTCTTGC; the BamHI site is underlined, the sequence coding for a TEV protease cleavage site is in italicized, and the beginning of gene 12 is shown in boldface) and TGA (ATACTCGAGTACCAGCTGCAGTTATTAAGTCGTTCC; the PstI site is underlined, the gene 12 complementary coding sequence is shown in boldface, and stop codons are double-underlined). The PCR fragment was then cleaved with BamHI-PstI and cloned into pRSET A, generating pMZ1.

Plasmid pBT450 was constructed in two steps. First a SfcI fragment bearing genes 15 and 16 of SPP1 (coordinates 8830–9787 of the SPP1 genome sequence, GenBankTM accession

number X97918 (30)) was treated with a Klenow fragment to produce blunt ends and cloned in the SmaI site of pBluescript SK- (Stratagene). Secondly, the resulting plasmid was used to clone a NruI-EarI blunt-ended fragment (coordinates 6699–8778 of the SPP1 sequence), bearing genes 11–13, in the HincII site of the pBluescript SK- polylinker. The cloning strategy generated a polycistronic unit composed of SPP1 genes 11–13 and 15 and 16 under the control of a T7 promoter. A DNA fragment containing gene 11 and the beginning of gene 12 was produced by cleavage of pBT450 with BglII, treated with T4 polymerase to produce blunt ends, and digested with Asp718. This fragment was cloned into pBT450 previously digested with SmaI and KpnI to generate pBT451. In pBT451, gene 12 is disrupted by an out-of-frame deletion between its internal BglII and SmaI sites (coordinates 7618–7646 of the SPP1 sequence). The *E. coli-Bacillus subtilis* shuttle vector pHP13 (31) cut with PstI-Sall was used for cloning a PstI-XhoI fragment of pBT451 spanning genes 11 to 13 that flank the gene 12 knock-out deletion. The resulting plasmid was named pBT452.

The non-permissive strain *B. subtilis* YB886 (pBT452) was infected with a mutant phage carrying conditional lethal mutations in genes 11 and 13 (SPP1*sus7sus31* (32)), forcing a double crossover that led to integration of the knockout mutation in gene 12 of the viral genome, as confirmed by DNA sequencing. The resulting phage, SPP1*del12*, was crossed with SPP1*sus9*, a mutant defective in a tail gene (33), to yield SPP1*sus9del12* (16). SPP1*del12* was used to produce SPP1 infectious particles lacking gp12, whereas SPP1*sus9* and SPP1*sus9del12* were used to produce tailless DNA-filled capsids with (capsid H) and without gp12 (capsid HΔ12), respectively.

Production and Purification of SPP1 Virions and DNA-filled Capsids

Procapsids, DNA-filled capsids, and viral particles were produced and purified as described previously (16, 33). Procapsids were kept in buffer R (50 mM potassium glutamate, 10 mM EDTA, 50 mM Hepes-KOH (pH 7.6), and 1 mM PMSF, added

³ The abbreviations used are: TEV, tobacco etch virus; SEC, size exclusion chromatography; FBTSa, fluorescence-based thermal shift assay.

freshly (33)), whereas all other structures were stored stably and manipulated in TBT buffer (100 mM NaCl, 10 mM MgCl₂, and 100 mM Tris-Cl (pH 7.5)). All interactions of tag-gp12/gp12 with viral structures were carried out in TBT buffer.

The concentration of capsid physical particles was estimated on the basis of their DNA content. Ultraviolet absorbance spectra of capsid suspensions were used to assess sample purity and the value at 260 nm to determine DNA concentration. This value was then used to calculate the concentration of capsid physical particles according the following equation:

$$T = (c \cdot N_A) / (n_{bp} \times 660) \quad (\text{Eq. 1})$$

where T is the concentration of physical particles/liter, c is the DNA concentration in grams/liter, N_A is the Avogadro constant, and n_{bp} is the average number of base pairs per SPP1 DNA molecule. The SPP1-packaged molecules were considered to have an average length of 45.9 kbp (16).

Production and Purification of Tag-gp12

Tag-gp12 was overproduced in *E. coli* BL21 (DE3) (pBT453). Cells were grown at 37 °C in Luria broth medium supplemented with 100 μg/ml ampicillin. An overnight culture was diluted 50-fold, grown to an optical density at 600 nm between 0.6 and 0.8, induced with isopropyl 1-thio-β-D-galactopyranoside to a final concentration of 1 mM and shaken for 3 h. Cells were harvested (30,000 × g, 30 min, 4 °C), resuspended in buffer A (500 mM NaCl, 10 mM imidazole, and 50 mM NaH₂PO₄ (pH 8.0)) supplemented with protease inhibitor mixture (Complete™ EDTA-free, Roche Applied Science) and disrupted by sonication on ice using three cycles of 2 min each spaced by 2-min pauses (Vibra Cell 72405, Fisher Bioblock, Illkirch, amplitude 60, pulse 3, 30–40 watt). The total soluble proteins extract obtained after centrifugation (30,000 × g, 1 h, 4 °C) was filtered through a 0.22-μm membrane. The filtrate was then loaded on a 5-ml HisTrap™ HP metal affinity column (GE Healthcare) coupled to an ÄKTA purification system (GE Healthcare). A three-step gradient was applied at 16 °C: 2% buffer B (500 mM NaCl, 500 mM imidazole, and 50 mM NaH₂PO₄ (pH 8.0)) for a first wash, 10% buffer B for a second wash, and 100% buffer B for elution. The tag-gp12 peak fractions were pooled and run through a preparative size exclusion chromatography column (HiLoad 26/60 Superdex™ 200pg, GE Healthcare) pre-equilibrated in buffer C (500 mM NaCl and 50 mM Na₂HPO₄ (pH 8.0)) at 16 °C coupled to an ÄKTA purification system. Aggregates and contaminants were found mostly in the void volume, whereas tag-gp12 eluted as a single peak. Tag-gp12 was obtained at a yield of 3 mg/g wet cell weight and was more than 95% pure, as judged from SDS-PAGE analysis. Purified protein was stored in buffer C and dialyzed against other buffers immediately before use. Protein concentration was estimated using the Bio-Rad protein assay following the instructions of the manufacturer. Tag-gp12 was used to immunize rabbits following the protocols established for protein (pC)CAT (34) to obtain anti-tag-gp12 polyclonal serum.

TagTEV-gp12 was produced and purified according to the same protocol. The purified protein was then incubated at 16 °C for 4 h with TEV protease at a ratio of 1:20 (w/w). To

remove the tag, the digestion product was loaded onto a 1-ml HisTrap™ HP metal affinity column. Gp12 was eluted with 80 mM imidazole. The tag and the TEV protease eluted at 500 mM imidazole. The purified gp12 carries an additional glycine at its amino terminus, preceding the initial methionine residue.

Mass Spectrometry

The collagenase digestion of tag-gp12 followed by trypsinolysis was stopped by adding solid guanidine hydrochloride to a final concentration of 6 M, followed by incubation at 90 °C for 15 min. Peptides were precipitated at –20 °C over a weekend by adding 5 volumes of cold acetone. Peptides were recovered by centrifugation, dried, and resuspended in ammonium carbonate at 1 μg/μl. They were then analyzed by MALDI-TOF and nano-LC-MS/MS.

MALDI Peptide Mass Fingerprinting—Peptides (0.5 μl) were mixed with an equal volume of either α-cyano-4-hydroxycinnamic acid (10 mg/ml and 50% CH₃CN, Sigma-Aldrich) or 2,5-dihydroxybenzoic acid (10 mg/ml and 20% CH₃CN, Sigma-Aldrich). Peptide mixtures were analyzed by MALDI-TOF (Voyager-DESTR, Applied Biosystems) after external calibration. Crystals were obtained using the dried droplet method, and 500 MALDI mass spectra were averaged per spot. Mass spectrometry measurements were carried out at a maximum accelerating potential of 20 kV in positive reflectron mode. Peak lists were generated by Data Explorer software (Applied Biosystems), and processed data were submitted to the FindPept tool (available on the ExPASy portal) using the following parameters: data bank gp12 protein; mass tolerance, 300 ppm; digest reagents, none.

Nano-LC-ESI-MS/MS Analyses—The peptide mixture was then analyzed with the Q/TOF Premier mass spectrometer (Waters) coupled to a nanoRSLC chromatography unit (Dionex) equipped with a trap column (Acclaim PepMap100 C18, 75 μm inner diameter × 2 cm, 3 μm, nanoViper) and an analytical column (Acclaim PepMapRSLC C18, 75 μm inner diameter × 15 cm, 2 μm, 100 Å, nanoViper). The loading buffer was H₂O/CH₃CN/TFA (98/2/0.05%). Buffer A and B were H₂O/HCOOH (0.1%) and CH₃CN/HCOOH (0.1%), respectively. A 2–50% B gradient was set for 40 min with a flow rate of 300 nl/min. Data-dependent scanning was applied to generate MS/MS spectra with a collision energy ramp of 15–40 volts. Standard MS/MS acquisitions were performed on the top of the three most intense parent ions of the previous MS scan. Raw data were processed with ProteinLynx Global Server (Waters). Peptide identification was achieved using the Mascot software with the following parameters: data bank gp12 protein; peptide tolerance, 15 ppm; fragment tolerance, 0.1 Da; digest reagent, none.

Digestion of Tag-gp12 with Collagenase

Collagenase VII from *Clostridium histolyticum* (8.8 units/mg) was purchased from Sigma-Aldrich. A stock solution was prepared at 1 mg/ml in buffer D (250 mM NaCl, 10 mM CaCl₂, 10 mM 2-mercaptoethanol, and 20 mM HEPES-Na (pH 7.6)) and diluted 10-fold before use. Tag-gp12 (50 μg) was digested with 0.23 μg of collagenase for 4 h at 16 °C. The same result was obtained by digestion for 30 min at 37 °C. Digestion products

A Collagen-like Binder of the SPP1 Viral Capsid

were analyzed on SDS-PAGE gel stained with Coomassie Blue and by mass spectrometry as described above.

Analytical Size Exclusion Chromatography (SEC)

100 μ l of purified tag-gp12 at 2 mg/ml was run at 16 °C using a flow of 0.5 ml/min on a SuperdexTM 200 10/300 GL (GE Healthcare) column equilibrated in buffer C (500 mM NaCl and 50 mM Na₂HPO₄ (pH 8.0)) and coupled to an ÄKTA purification system. Column calibration and Stokes radius estimation were carried out as described previously (35).

Analytical Ultracentrifugation

Analytical ultracentrifugation was carried out on a Beckman Optima XL-A ultracentrifuge (Beckman Coulter, Palo Alto, CA) equipped with 12-mm central cells on an ANTi-60 rotor. Runs were performed in buffer C at 16 °C and monitored by absorption at 280 nm. The tag-gp12 partial specific volume (0.7051 ml/g), buffer C solvent density (1.02647 g/ml), and solvent viscosity (1.1913 cP) were calculated using the SEDNTERP software (36).

Sedimentation velocity runs were carried out at a rotor speed of 220,000 \times *g* using protein loading concentrations of 0.5, 1, and 2 mg/ml. Data points were recorded every 5 min and analyzed using the SEDFIT software, assuming a non-interacting species model.

Equilibrium sedimentation was performed at 16,300 \times *g* using protein loading concentrations of 0.3, 0.5, and 0.8 mg/ml. Data were analyzed using SEDFAT for average mass determination.

CD Measurements

CD measurements were carried out on a Jasco J810 spectropolarimeter equipped with a Peltier temperature controller. The protein, at 2 mg/ml, was dialyzed against either buffer C or TBT buffer and loaded to a 0.1-mm path length quartz cell. Spectra at fixed temperatures were recorded at an equilibrium of between 190–260 nm every 0.2 nm using a bandwidth of 1 nm and a scanning speed of 20 nm/min. Each spectrum was an accumulation of five spectra after baseline correction using the buffer spectrum as blank. Thermal transition profiles were monitored at 200 nm with a 1 °C/min heating rate and a protein concentration of 2 mg/ml. One point was recorded for each 1 °C. The temperature was raised from 10 to 60 °C, kept at 60 °C for 30 min, and finally returned back to 10 °C using the same rate. Ellipticity was measured first and molar ellipticity was then calculated using the following equation:

$$[\theta] = (\theta \times 100M)/(c \times l) \quad (\text{Eq. 2})$$

where θ is the ellipticity in degrees, *M* is the molecular mass, *c* is the protein concentration in mg/ml, and *l* is the path length in centimeters. *T_m*s was determined from data plots as the transition midpoint.

Fluorescence-based Thermal Shift Assay (FBTSA)

FBTSA experiments were performed as described previously (16). In brief, SPP1 virions or capsids (5.7×10^{10} particles) and/or purified tag-gp12 at different concentrations were mixed with diluted Sypro Orange dye (400-fold diluted from

stock solution, Invitrogen) in TBT buffer to a final volume of 10 μ l. Experiments were carried out in real-time PCR systems, and fluorescence was recorded in real time. Different heating-cooling cycles were applied to the samples as described in the figure legends and under “Results.” Experiments were carried out in ABI 7900HT and QuantStudio 12KFlex machines (Applied Biosystems) as detailed in the figure legends. The fluorescence profiles, derivatives, and *T_m*s were determined using the analysis software of the manufacturer.

Gp12 Chimerization Experiments

A 2-fold molar excess of purified gp12 was mixed with tag-gp12 and kept either at 16 °C or heated for 15 min at 60 °C. Mixtures were then loaded onto a 1-ml metal affinity column (HisTrapTM HP metal affinity column, GE Healthcare), and proteins were eluted by applying a step gradient of imidazole concentration as used for tag-gp12 purification (see above).

Binding of Tag-gp12 to SPP1del12 Virions

100 μ l of SPP1del12 virions at 5.6×10^{12} pfu/ml were mixed with a 10-fold molar excess of tag-gp12 protein (550 μ l of a 2 mg/ml solution), considering 60 binding sites for tag-gp12 trimers per capsid, and incubated overnight at 16 °C. The mixture was then run through a discontinuous cesium chloride gradient to purify phage particles (37), and their protein composition was analyzed by Western blotting with rabbit polyclonal antibodies raised against purified tag-gp12 or against purified SPP1 virions.

Binding of Tag-gp12 to H Δ 12 Capsids Analyzed by a Trypsin Protection Assay or FBTSA

Tag-gp12 was dialyzed against TBT. A range of tag-gp12 concentrations and purified SPP1 capsids lacking gp12 (H Δ 12) were incubated separately for 1 h at 16 or 45 °C in a PCR machine (Biometra Tprofessional Trio thermocycler). A constant number of H Δ 12 was then mixed with variable amounts of tag-gp12 to obtain different molar ratios and incubated at 16 or 45 °C for the desired reaction time according to the experimental schematic in Fig. 7, *left panels*. Samples were analyzed by FBTSA or incubated with 1 μ g of trypsin at 45 °C for 30 min to proteolyse free tag-gp12 (not associated with capsids). Care was taken to avoid any cooling below 45 °C for samples whose mixtures were incubated at this temperature before FBTSA or trypsination. This was necessary to prevent rapid refolding/reassociation of free tag-gp12, which would facilitate assembly of free trimers and their binding to capsids. Capsids in trypsinated samples were separated on 0.8% agarose gels prepared in TAMg buffer (1 mM MgCl₂ and 40 mM Tris acetate (pH 8.3)). The running buffer was TAMg, and the applied electric field intensity was 70 mA. Gels were stained with ethidium bromide in TAE buffer (40 mM Tris-acetate supplemented with 1 mM EDTA) in which EDTA led to disruption of capsids *in situ*, rendering viral DNA accessible to ethidium bromide binding, a more sensitive detection method than protein staining with Coomassie Blue.

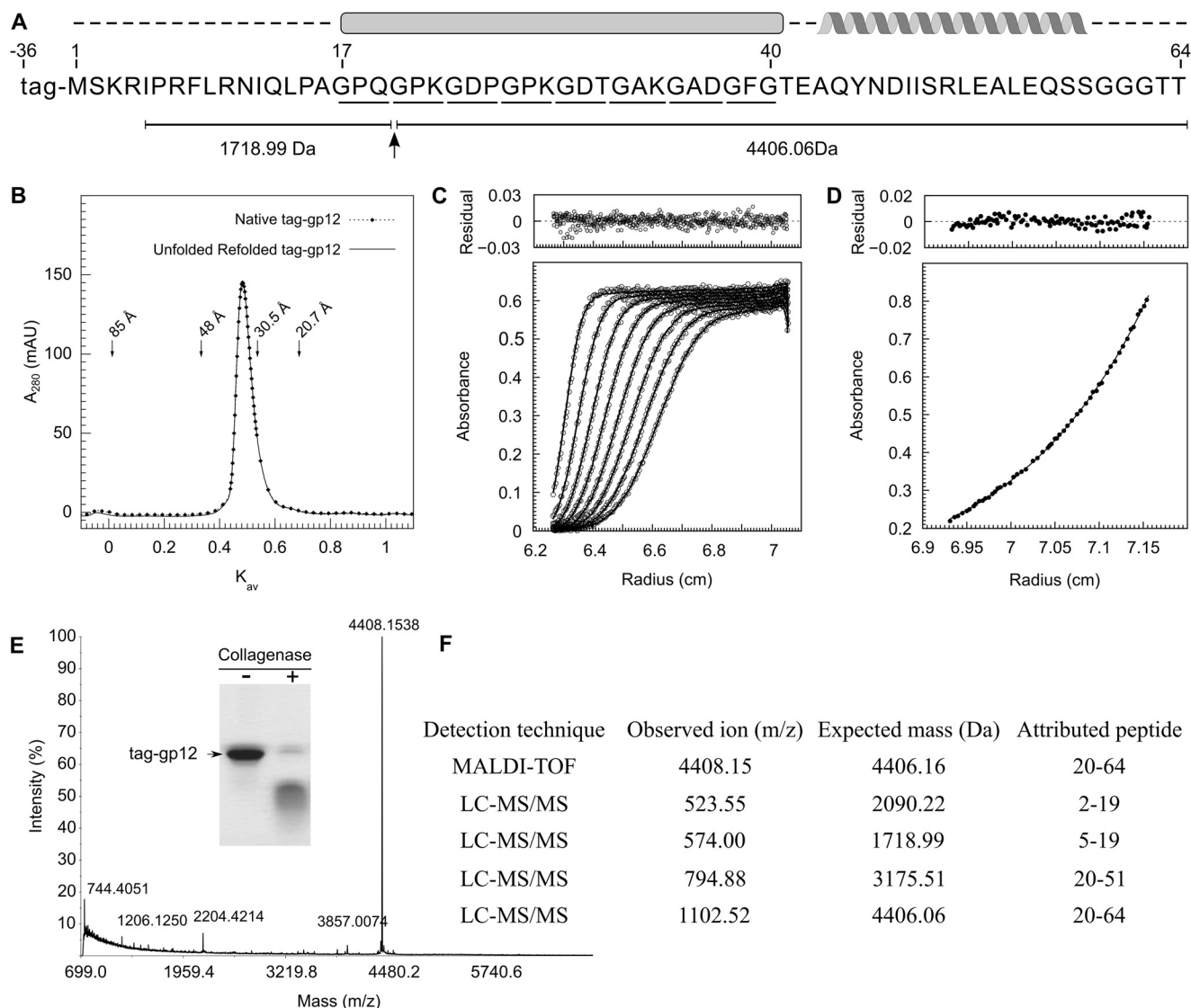


FIGURE 2. Properties of the SPP1 auxiliary protein gp12. A, Gp12 amino acid sequence showing position of the tag fused to its amino terminus. GXY triplets are *underlined*. The intermolecular collagen-like triple helix and α -helix predicted by bioinformatics are shown above the sequence. The collagenase cut of tag-gp12 inside the collagen-like sequence motif and peptides obtained from mass spectrometry analysis (E and F) are shown below. B, R_H determination of native and unfolded-refolded tag-gp12. Native (*dotted line*) and tag-gp12 heated for 5 min at 90 °C and transferred directly to ice (*continuous line*) were analyzed by SEC at 16 °C as described under "Experimental Procedures." The elution positions of thyroglobulin ($R_H = 85$ Å), γ -globulin ($R_H = 48$ Å), ovalbumin ($R_H = 30.5$ Å), and myoglobin ($R_H = 20.7$ Å) used to calibrate the column are indicated by *arrows*. K_{av} (partition coefficient) was calculated as described (35). *mAU*, milli absorbance units. C, sedimentation velocity of tag-gp12 at 220,000 $\times g$ (loading concentration of 1 mg/ml, 16 °C run). Gp12 has a sedimentation coefficient of 1.7 S ($s_{20,w} = 2.2$ S). D, sedimentation equilibrium of tag-gp12 at 16,300 $\times g$ (loading concentration of 0.8 mg/ml, 16 °C run). The data (dots) were fit using a trimer model (*continuous line*). The best fit was obtained for a single species with an average mass of $31,270 \pm 590$ Da. The *top panels* in C and D show the deviation of experimental points from fitted curves. E and F, cleavage of tag-gp12 with collagenase VII analyzed by SDS-PAGE (*inset* in E) and mass spectrometry. The observed ion mass (4406.15 Da) in MALDI-TOF (E) is attributed to peptide 20–64 of the gp12 sequence, identifying the proteolysis site shown in A. The same peptide was detected by LC-MS/MS spectrometry, which also showed the presence of three other peptides, resulting from collagenase cleavage at the same position (F). The LC-MS/MS analysis had a tag-gp12 sequence coverage of 89%. For clarity, only the peptides in which one end was generated by the collagenase cleavage are listed in F. Those peptides were absent in the analysis of tag-gp12 not treated with collagenase.

Bioinformatics

Protein secondary structure predictions were carried out using Jpred (38), and the three-dimensional structure was predicted using HHPred (39).

RESULTS

Gp12 Has a Collagen-like Sequence Motif—The SPP1 capsid auxiliary protein gp12 is a 64-amino acid-long polypeptide with a molecular mass of 6613 Da and a theoretical isoelectric point of 8.14. Its carboxyl terminus is predicted to form α -helices,

whereas the central part features eight GXY repeats (Fig. 2A) (40). The repeated GXY motif is a sequence signature of collagen-like proteins in which three polypeptides are brought together to form an intermolecular, left-handed triple helix (41, 42).

Gp12 Is an Elongated Trimer in Solution—The gp12 amino terminus was fused to a 36-amino acid-long peptide including a hexahistidine tag to enhance protein production and allow easy purification. The 10.7-kDa recombinant protein (tag-gp12) eluted from a Superdex 200 analytical SEC column as a single

A Collagen-like Binder of the SPP1 Viral Capsid

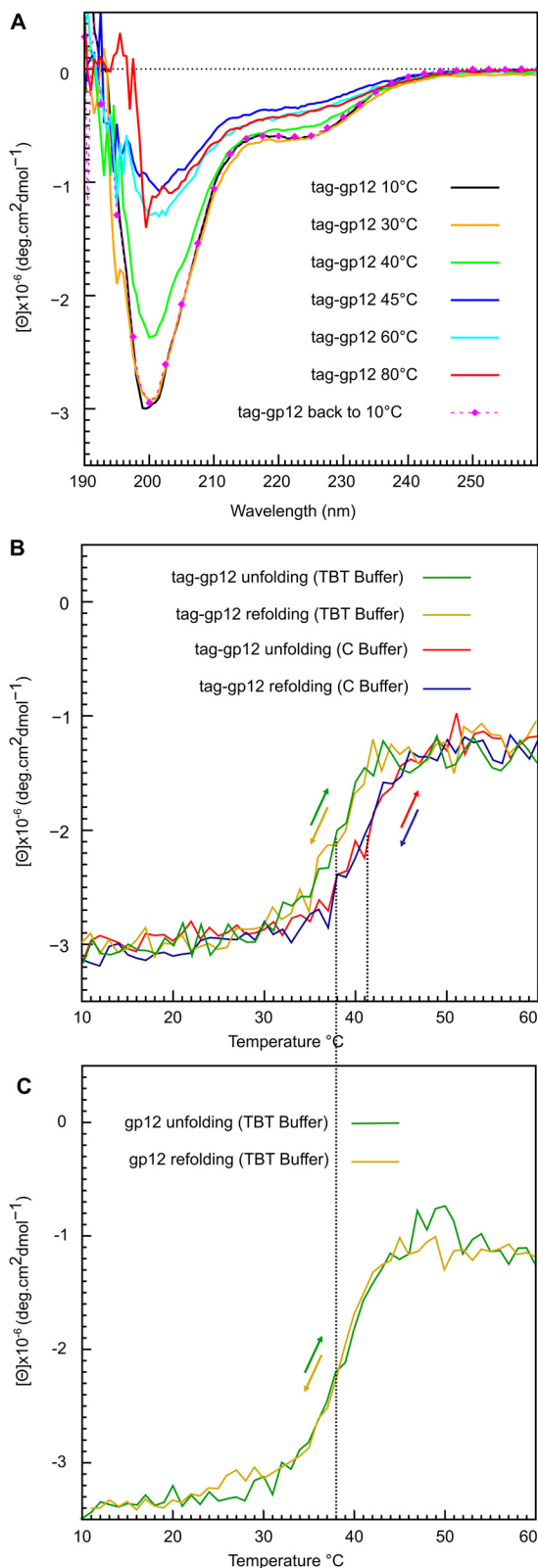


FIGURE 3. Reversibility of tag-gp12 trimer unfolding and dissociation. A, CD spectra of tag-gp12 (2 mg/ml) in buffer C were recorded using the same sample at different temperatures. Tag-gp12 was then maintained at 80 °C for 30 min and cooled back to 10 °C to record the spectra of the refolded protein (pink dotted line). B, tag-gp12 unfolding and refolding in buffer C and in TBT monitored by CD at 200 nm, the collagen-like triple helix local minimum signal, using a temperature gradient of 1 °C/min. The colored arrows show the direction of the temperature gradient (heating or cooling) for each individual color curve. C, gp12 unfolding and refolding in

symmetric peak (Fig. 2B). Its hydrodynamic radius (R_H) on the basis of a protein calibration data set was 35 Å. The gp12 elongated shape observed in electron microscopy reconstructions of the bacteriophage SPP1 capsid (16) rendered SEC not suitable to estimate its native mass (43). The shape and oligomerization state of tag-gp12 were therefore investigated by analytical ultracentrifugation at 16 °C. Tag-gp12 behaved as a homogeneous species with a sedimentation coefficient of 1.7 S ($s_{20,w} = 2.2$ S) (Fig. 2C) at all loading concentrations tested in sedimentation velocity experiments (0.5–2 mg/ml). Sedimentation equilibrium centrifugation was then used for shape-independent measurement of the tag-gp12 mass (Fig. 2D). The determined molecular mass ($31,270 \pm 590$ Da) was only 3% lower than the theoretical mass of a tag-gp12 trimer. Using this experimental value and the sedimentation coefficient, we calculated a friction ratio (f/f_0) of 1.79, showing that tag-gp12 is an elongated trimer in solution.

Gp12 Has a Collagen-like Fold—To probe that the (GXY)₈ repeats of tag-gp12 form a collagen-like triple helix, the protein was challenged with collagenase VII, which cuts the triple helix at defined environments (44). The control SPP1 proteins gp6 and H16, a tagged form of gp16 (45), were insensitive to proteolysis (not shown), whereas tag-gp12 was cleaved (Fig. 2E, inset). MALDI-TOF (Fig. 2E) and nano-LC-MS/MS (Fig. 2F) identified the cut between Gln-19 and Gly-20 of tag-gp12 (Fig. 2A, arrow). This site, found at the beginning of the GXY repeat region, matches one of the expected cutting sites for collagenase VII (44).

Collagen left-handed triple helices are also characterized by a CD signature with a deep minimum of negative ellipticity at around 200 nm and a slightly positive ellipticity maximum at around 220 nm (46). The CD spectrum of native tag-gp12 had a strong minimum at 200 nm and a second minimum at 222 nm, where the ellipticity of α -helices masked the positive signal of the collagen helix (Fig. 3A). This profile strongly supports that tag-gp12 combines a collagen-like fold with α -helical regions.

Gp12 Dissociates and Unfolds Reversibly at Physiological Temperature—CD spectra showed a loss of tag-gp12 structure between 30 and 45 °C (Fig. 3A). The CD spectra from 45–80 °C were characteristic of an unfolded polypeptide chain. Fast (<1 min) or progressive cooling of the sample back to 10 °C led to complete recovery of the secondary and quaternary structure content, with a CD spectrum identical to the one of the native protein (Fig. 3A, pink dotted line).

To further analyze the dissociation-unfolding and refolding-reassociation transitions, a CD experiment was monitored at 200 nm (corresponding to the collagen-like helix minimum) by challenging the sample against a heating cycle from 10–60 °C (unfolding) and back to 10 °C (refolding) (Fig. 3B). Tag-gp12 showed a sharp transition with a T_m of 41 °C in the protein-high salt buffer and of 38 °C in a low monovalent salt solution with magnesium (TBT buffer that stabilizes SPP1 viral particles) (Fig. 3B). Unfolding and refolding followed the same kinetic profile upon heating and cooling (Fig. 3B). The thermal stability

TBT monitored as in B. The dotted vertical lines in B and C are a visual aid to show the transition midpoints (T_m) in buffer C and TBT. The experiment was repeated twice independently.

study of tag-gp12 by CD revealed a unique transition with complete loss of secondary structure and dissociation of the collagen-like triple helix (Fig. 3, A and B). The behavior of tag-free gp12 was identical to tag-gp12 (data not shown and Fig. 3C), revealing that the tag influenced neither the protein CD signature nor its dissociation/unfolding and refolding/reassociation properties. The SEC profiles of native and unfolded-refolded tag-gp12 were also indistinguishable, with a single symmetric

peak of trimers and no detectable intermediate states (Fig. 2B). The complete population of refolded tag-gp12, therefore, retrieved its initial R_H .

To define whether the tag-gp12 polypeptide chains physically separate upon thermal denaturation, we carried out a chimerization experiment between tag-gp12 (10.7 kDa subunit mass) and tag-free gp12 (6.6 kDa subunit mass). The hexahistidine-tagged tag-gp12 bound strongly to a metal affinity column and eluted only in the presence of 500 mM imidazole (Fig. 4, *first panel*). Gp12 also adsorbed to the column matrix but was completely released by a wash with 100 mM imidazole (Fig. 4, *second panel*). Loading of a tag-gp12:gp12 mixture kept at 16 °C led to differential elution of gp12 at 100 mM imidazole and of tag-gp12 at 500 mM imidazole (Fig. 4, *third panel*). When the tag-gp12:gp12 mixture was denatured at 60 °C and reassociated by cooling to 16 °C, there was a fraction of non-tagged gp12 that coeluted with tag-gp12 at 500 mM imidazole (Fig. 4, *fourth panel*). This behavior is explained by the presence of heterotrimers in which the tag-gp12-tagged subunit(s) led to retention of the non-tagged gp12 form present in the heterotrimer. The formation of chimeras showed that the gp12 and tag-gp12 trimers physically dissociated upon thermal denaturation and that reassociation led to the formation of heterotrimers, although homotrimerization appeared to be favored when comparing the intensity of bands in Fig. 4, *fourth panel*.

Binding of gp12 to the Capsid Lattice Is Reversible and Increases the Trimer Thermal Stability of 20 °C—To characterize the interaction of gp12 with SPP1 capsids, we generated viral particles (SPP1 Δ el12) and tailless expanded capsids (H Δ 12) lacking gp12 by genetic engineering (Figs. 1 and 5A). These particles bound tag-gp12 *in vitro*, whereas wild-type virions whose capsid carries gp12 did not (Fig. 5B). Therefore, tag-gp12 interacts strongly and exclusively with specific sites in the SPP1 capsid lattice without any detectable exchange between free (tag-gp12) and capsid-bound (gp12) subunits.

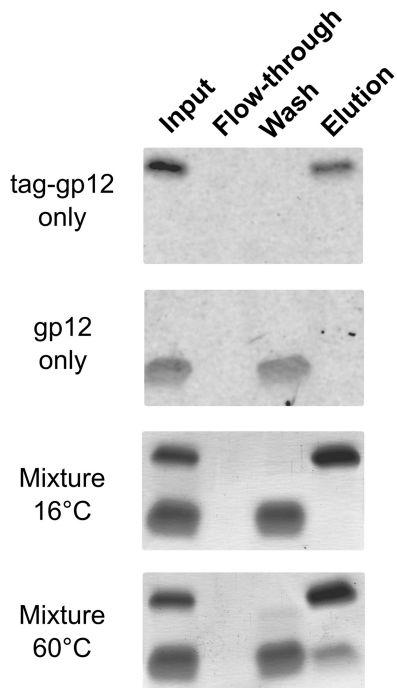


FIGURE 4. **Tag-gp12/gp12 chimerization experiment.** Isolated proteins and their mixture incubated at 16 or 60 °C (50 μ l of 2 mg/ml in buffer C) were loaded onto a metal affinity column. Aliquots of the input proteins before chromatography, flow-through, washing with 100 mM imidazole, and elution with 500 mM imidazole were analyzed on a 12% Tris-*N*-[2-hydroxy-1,1-bis(hydroxymethyl)ethyl]glycine gel stained with Coomassie Blue. The experiment was repeated twice independently.

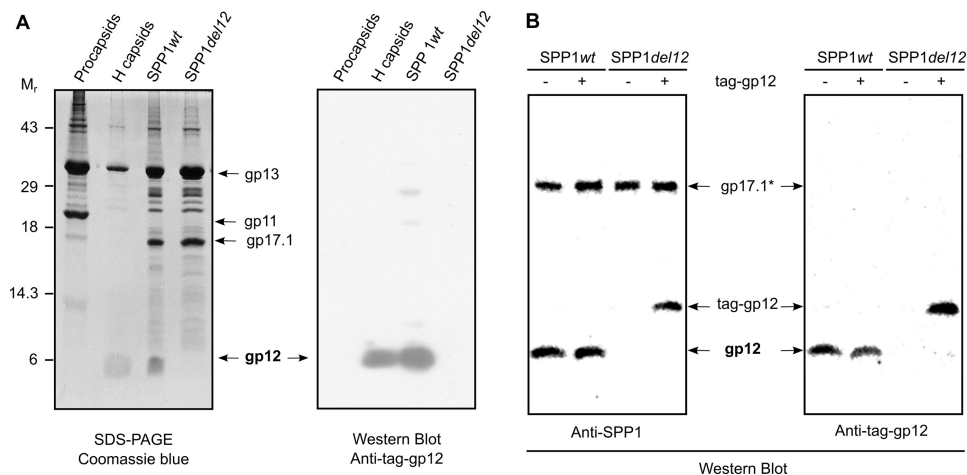


FIGURE 5. **Binding of gp12 to SPP1 capsids.** A, composition of SPP1 assembly intermediates (*cf.* Fig. 1) determined by SDS-PAGE gel stained with Coomassie Blue (*left panel*) and presence of gp12 in the structures detected with anti-tag-gp12 polyclonal antibodies (*right panel*). The major capsid protein gp13, the major tail tube protein gp17.1, and the procapsid internal scaffolding protein gp11 are also identified in the Coomassie Blue-stained gel. B, binding of tag-gp12 to wild-type SPP1 and to SPP1 Δ el12 particles. Virions incubated overnight at 16 °C with tag-gp12, as indicated above the Western blot analyses, were separated from free protein by isopycnic centrifugation in cesium chloride gradients, and the composition of particles was analyzed by Western blotting with polyclonal antibodies raised against purified SPP1 virions (*left panel*) and anti-tag-gp12 antibodies (*right panel*). Note that gp12 and the tail protein gp17.1* (28) are the most immunogenic proteins of the SPP1 particle despite of the fact that they are not the most abundant components of the virion (A, *left panel*) (Refs. 28, 40 and this work).

A Collagen-like Binder of the SPP1 Viral Capsid

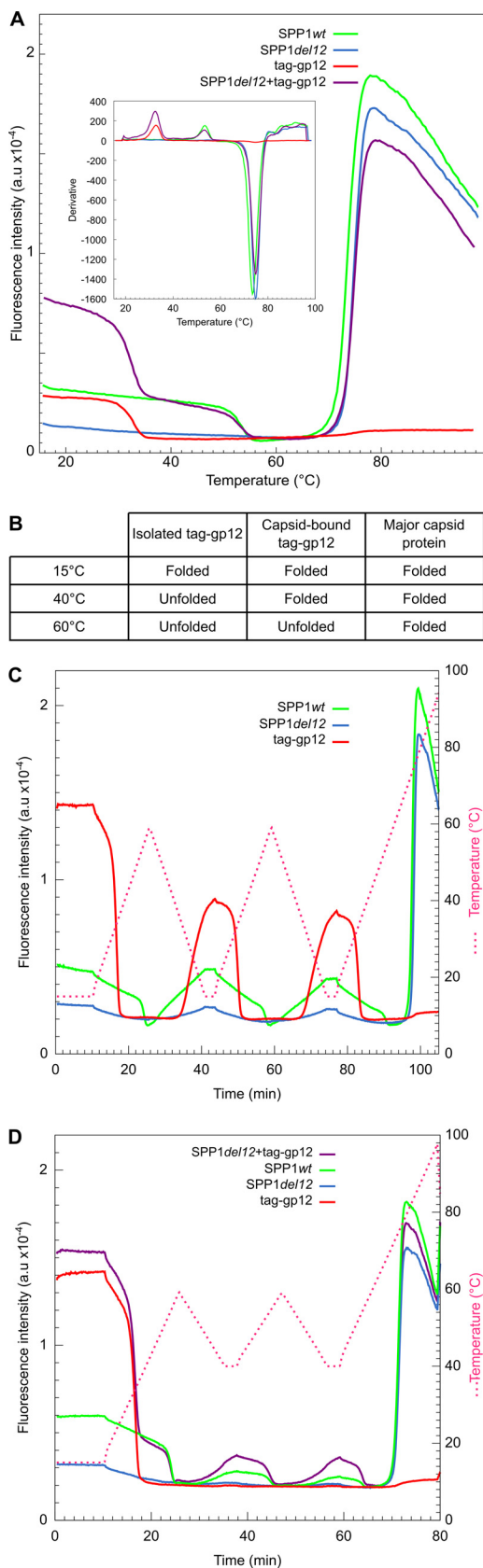


FIGURE 6. Cyclical gp12 association and dissociation from viral capsids. Sypro Orange was added to protein and viral particle samples that were submitted to different heating and cooling regimens at a rate of 3 $^{\circ}\text{C}/\text{min}$ and monitored by FBTSA in an ABI 7900HT machine. *A*, FBTSA of isolated tag-gp12 (red curve), wild-type SPP1 virions (green curve),

The FBTSA method allows monitoring independently of the thermal denaturation of gp12 and of the major capsid protein gp13 (16). The assay quantifies binding of the Sypro Orange dye to exposed hydrophobic regions of proteins challenged to a temperature gradient (47). In an aqueous environment, Sypro Orange has a low quantum yield, and in protein solutions, the dye access to non-polar environments is normally shielded by the protein fold. Protein thermal denaturation exposes hydrophobic regions where the dye binds, resulting in strong fluorescence emission. Isolated tag-gp12 exhibited the opposite behavior. An increase of temperature led to progressive loss of fluorescence, followed by a sharp transition at a T_m of 33.4 ± 0.7 $^{\circ}\text{C}$ in TBT buffer (Fig. 6A, red curve), which is 4.6 $^{\circ}\text{C}$ lower than the unfolding T_m determined by CD (Fig. 3B). Gp12 without a tag showed the same behavior. The profile of this transition revealed that Sypro Orange has one or several binding sites in native tag-gp12 that were destroyed when the protein started to lose its secondary and quaternary structure. Such a rare property provided a specific signature for tag-gp12 unfolding.

The FBTSA profile of infectious SPP1 phage particles (Fig. 6A, green curve) was marked by two transitions similar to the ones found for the tailless capsids (data not shown) (16), showing that only the capsid proteins of phage particles gave a detectable signal under our experimental conditions. Both structures carry gp12 (Fig. 5A). The first transition, at 53.6 ± 0.2 $^{\circ}\text{C}$, displayed the tag-gp12 signature and was absent from particles lacking gp12 (SPP1del12, Fig. 6A, blue curve). Mixing of SPP1del12 phages with tag-gp12 *in vitro* restored the signal at 53.6 $^{\circ}\text{C}$, whereas the excess of free protein led to the typical T_m transition at 33.4 $^{\circ}\text{C}$ (Fig. 6A, violet curve). The identical T_m of gp12 and tag-gp12 was 20.2 $^{\circ}\text{C}$ higher than the one observed for isolated tag-gp12, showing that binding to the capsid lattice led to a major stabilization of the gp12 trimer. The second signal transition of viral particles with or without gp12 was characterized by a strong increase of fluorescence at 75 ± 0.3 $^{\circ}\text{C}$ because of cooperative denaturation of gp13.

The distinct melting temperatures of isolated tag-gp12 (33.4 $^{\circ}\text{C}$), of capsid-bound tag-gp12 or gp12 (53.6 $^{\circ}\text{C}$), and of major capsid protein (75 $^{\circ}\text{C}$) allowed us to follow the behavior of the three species in gp12-capsid binding experiments. At 40 $^{\circ}\text{C}$, capsid-bound gp12 was easily distinguished from isolated tag-gp12 because it was the only folded gp12 form at this temperature, whereas, at 60 $^{\circ}\text{C}$, only the capsid protein was stable (Fig. 6B). Isolated tag-gp12, wild-type SPP1, and SPP1del12 particles were submitted to two heating-cooling cycles between 15 and 60 $^{\circ}\text{C}$, followed by a final heating step from 15–99 $^{\circ}\text{C}$ (15–60–15–60–15–99 $^{\circ}\text{C}$, 3 $^{\circ}\text{C}/\text{min}$ heating/cooling rate). Free

SPP1del12 virions that lacked gp12 (blue curve), and SPP1del12 mixed with an excess of tag-gp12 (violet curve). The inset shows the opposite of the first derivative of the fluorescence signal. *B*, summary of the tag-gp12/gp12 and gp13 states at different temperatures. *C*, isolated tag-gp12, wild-type SPP1, and SPP1del12 virions submitted to cycles of heating to 60 $^{\circ}\text{C}$ and cooling to 15 $^{\circ}\text{C}$. The experiment was finished with a denaturation step to 99 $^{\circ}\text{C}$. The pink discontinuous line shows the temperature variation (coordinates are shown on the right). *D*, the same samples and a mix of SPP1del12 virions with a 5.5 molar excess of tag-gp12 (violet curve) challenged with two cycles of heating to 60 $^{\circ}\text{C}$ and cooling to 40 $^{\circ}\text{C}$. Experiments were repeated at least twice independently.

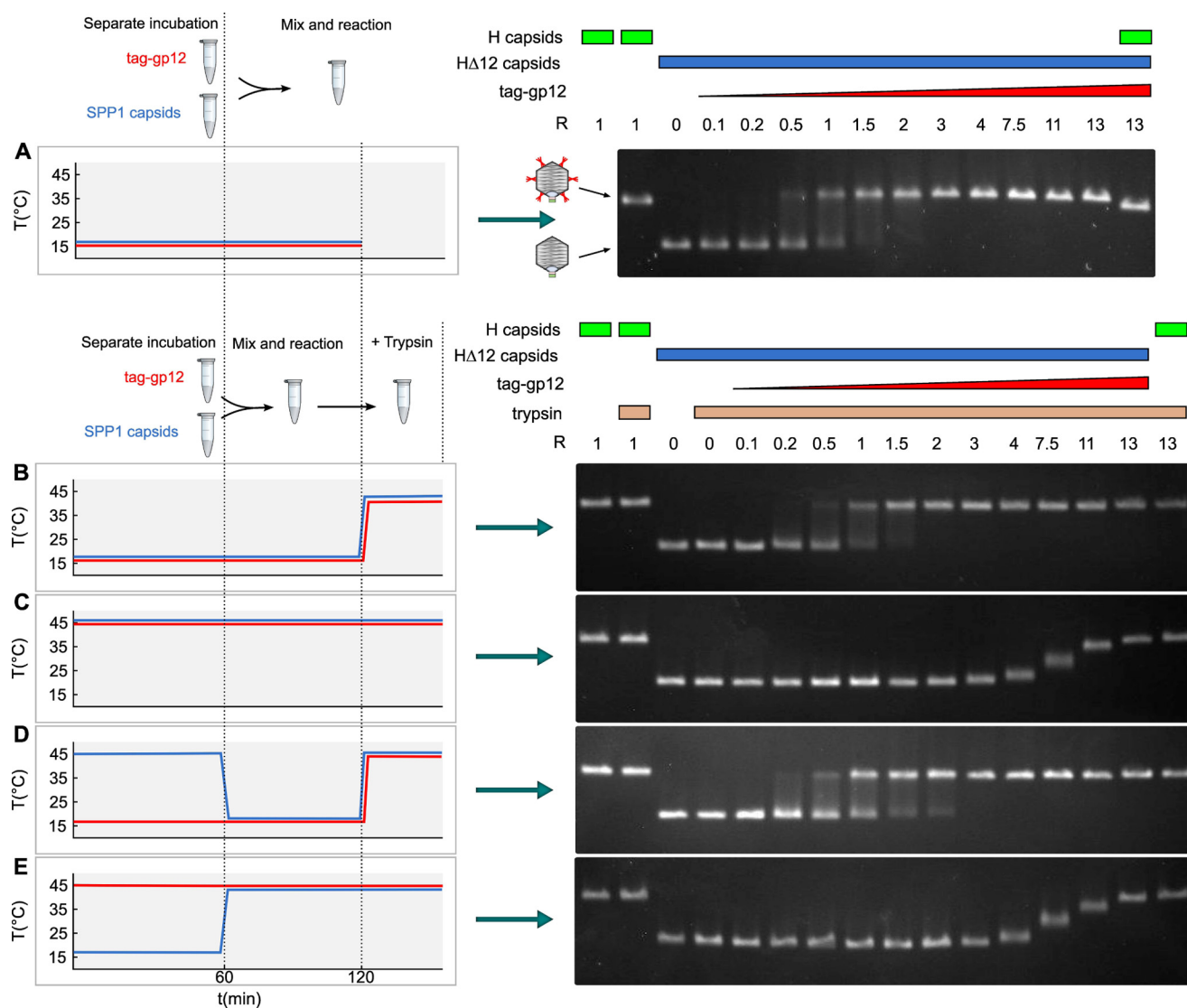


FIGURE 7. **Capsid binding behavior of native and unfolded tag-gp12.** Purified SPP1 tailless capsids lacking gp12 (capsid HA Δ 12) (blue characters and blue curves on the left and blue rectangles above the gels on the right) and tag-gp12 (red) were preincubated separately, mixed, and treated with trypsin (except in A) according to the different combinations of incubation conditions used in the experiments in A–E (see “Results” for details), as outlined on the left of each panel. Samples treated with trypsin in B–E are identified by yellow rectangles above the gel lanes on the right. Capsids were then resolved by agarose gel electrophoresis to assess their occupancy with tag-gp12. Wild-type SPP1 capsids with gp12 (H capsids) (green rectangles above the gels on the right) and HA Δ 12 were used as controls. The schematics in the center of A show the electrophoretic mobility of capsid H (with gp12 represented in red) and HA Δ 12. The experiment was repeated four times independently.

tag-gp12 exhibited a loss of signal upon heating, and its partial reacquisition when cooling to 15 °C showed that the fluorophore binding site(s) was/were not completely restored in the tag-gp12 population (Fig. 6C, red curve) in spite of the fact that the protein fully reacquired its quaternary structure CD signature (Fig. 3). The temperatures of transition were remarkably reproducible, revealing that tag-gp12 underwent dissociation/unfolding and folding/reassociation cycles. Gp12 bound to SPP1 capsids exhibited a transition corresponding to a T_m of 53.6 °C (Fig. 6C, continuous green curve). The process was reversible upon cooling and reheating, apart from a slight loss of fluorescence from one cycle to another. Therefore, gp12 dissociated/unfolded reversibly from wild-type capsids and maintained its binding activity to the capsid.

To assess whether the capsid lattice influences gp12 refolding/reassociation, the cycling experiment was repeated with cooling steps to 40 °C (15–60–40–60–40–99 °C program, Fig. 6D), a temperature at which free tag-gp12 remained unfolded after the first heating step (Fig. 3D, red curve). Gp12 bound to phage capsids kept its signature (T_m of 53.6 °C) in heating cycles to 60 °C. Cooling to 40 °C led to recovery of some fluorescence signals (Fig. 6D, green curve) but significantly less than when the temperature was reduced to 15 °C (Fig. 6C). Therefore, at 40 °C, a subpopulation of gp12 rebound to phage capsids, yielding folded trimers that fixed Sypro Orange. Addition of a 5.5-fold molar excess of exogenous tag-gp12 to wild-type capsids restored most of the gp12 signal associated with capsids after each 60–40 °C cycle (Fig. 6D, violet curve), showing that tag-gp12 had efficiently replaced gp12, which left its capsid

A Collagen-like Binder of the SPP1 Viral Capsid

sites upon denaturation. Restoration of the tag-gp12 signal at 40 °C occurred exclusively in presence of the capsid lattice, showing that this structure promoted tag-gp12 refolding and reassociation.

Native and Unfolded gp12 Binds to SPP1 Capsids in a Distinct Way—The finding that both native and unfolded tag-gp12 bound to SPP1 phage capsids (Fig. 6, C and D) suggested two distinct types of interaction, prompting their characterization. Tailless capsids without gp12 (HΔ12) and purified tag-gp12 were preincubated separately at 16 or 45 °C, followed by mixing at different ratios for interaction at the two temperatures (Fig. 7). Reactions were then incubated at 45 °C with trypsin, which degraded free gp12/tag-gp12 (Fig. 8A). The tag of capsid-bound tag-gp12 was prone to trypsin attack, but the gp12 moiety attached to the capsid remained intact (Fig. 8B), explaining the lower electrophoretic mobility of capsids loaded with tag-gp12 not treated with trypsin when compared with those that were trypsinated (Fig. 7, A and B). This step prevented subsequent interactions of free tag-gp12 with capsids during downstream sample manipulation at room temperature and separation by gel agarose electrophoresis. SPP1 capsids with gp12 (H capsids) or HΔ12 loaded with tag-gp12 had a slower electrophoretic mobility than capsids lacking gp12 (Fig. 7), most likely because gp12/tag-gp12 reduces the capsid surface electronegative charge. In contrast, gp12 does not have a major effect on capsid diameter, which is almost identical in H and HΔ12 (~610 Å (16)).

When HΔ12 capsids were mixed at 16 °C with increasing amounts of tag-gp12 native trimers, the capsid species shifted from tag-gp12-free to capsids fully loaded with tag-gp12 (Fig. 7, A, B, and D). At a ratio (R = 60 tag-gp12 trimers/capsid) of 0.5, most capsids lacked tag-gp12, but a minority was already saturated with tag-gp12, whereas, at R = 1.5, almost all capsids were decorated with tag-gp12. Species with intermediate electrophoretic mobility were poorly detected, revealing that capsids partially occupied with tag-gp12 were a minor population, even at limiting amounts of tag-gp12 (e.g. R = 0.5). We attribute this behavior to high cooperative binding of tag-gp12 trimers to its 60 sites in the SPP1 capsid.

To characterize the interaction of unfolded tag-gp12 with the capsid, the two species were preheated individually at 45 °C and mixed at the same temperature (Fig. 7C). A significant excess of tag-gp12 per capsid (R between 4 and 13) was needed to promote a change of capsid electrophoretic mobility. Their discrete bands showed a migration pattern that progressed from the capsids lacking gp12 band behavior (R < 3) to the full tag-gp12-loaded capsid band (R ≥ 13) (Fig. 7C). Similar results were obtained when the interaction reaction was prolonged overnight at 45 °C (not shown). Furthermore, preheating of capsids at 16 °C or 45 °C showed that temperature did not affect their binding properties (Fig. 7). Stable interaction of unfolded tag-gp12 with HΔ12 capsids therefore required an excess of tag-gp12 that bound in an inefficient manner, leading to a population of capsids whose binding sites are only partially occupied by tag-gp12 at molar ratios as high as R = 11 (Fig. 7, C and E). The increase of occupancy with the rise of R correlated with an augmentation of the tag-gp12 signal in FB-TSA experiments,

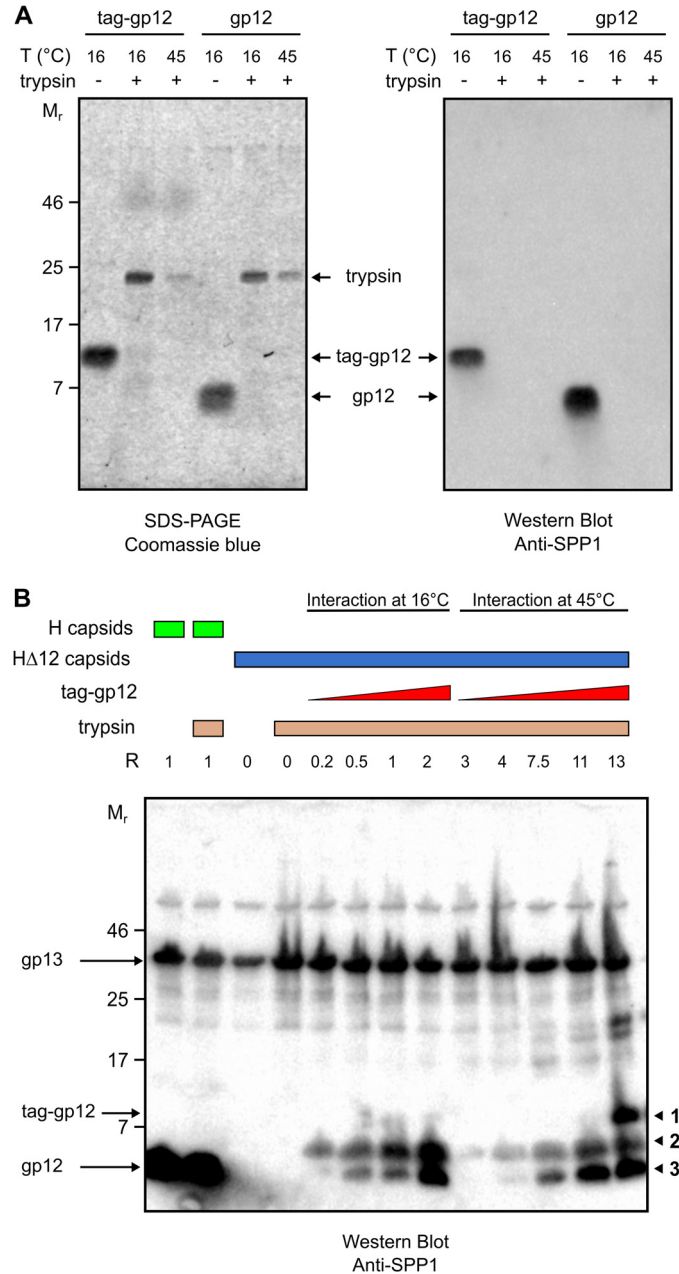


FIGURE 8. Trypsin sensitivity of free and capsid-bound gp12. A, purified tag-gp12 and gp12 were incubated with trypsin either at 16 or at 45 °C. Both proteins were completely digested by the protease at the tested temperatures, as assessed by Coomassie-stained SDS-PAGE (left panel) and Western blot analysis (right panel) with polyclonal anti-SPP1 antibodies that recognize gp12 (Fig. 5B). B, trypsin digestion of the binding reaction between capsids and tag-gp12 (labeled band 1 on the right of the figure) under the same conditions as in Fig. 7, B and C. Note that gp12 bound to H capsids is not sensitive to trypsin, whereas the tag of tag-gp12 is partially (band 2) or fully (band 3) digested by trypsin. The Western blot analysis was developed with anti-SPP1 antibodies that recognize gp12 but also, although giving a comparatively weak signal, the major capsid protein gp13, whose band was used to control the normalized input of capsids in the binding reaction.

consistent with the formation of tag-gp12 trimers in the capsid lattice (Fig. 9).

DISCUSSION

The 6.6-kDa gp12 polypeptide of bacterial virus SPP1 was shown here to build an elongated trimer. Its properties indicate

the presence of an intermolecular collagen-like triple helix that correlates with presence of eight GXY repeats at the center of the gp12 sequence, revealing a modular organization in which the collagen-like elongated segment connects two short amino and carboxyl terminus domains. Collagens are well studied protein components of the extracellular matrix of animals. They are characterized by the presence of 4-hydroxyproline at position Y of the GXY triplet, which is considered a major determinant of collagen stability (42). However, stable triple helices are built in synthetic model peptides by repeats longer than $(GXY)_8$, showing that their association requires no amino acid posttranslational modifications (48). This property is consis-

tent with the presence of collagen-like segments in prokaryote systems reported for streptococcal surface proteins (49, 50) and tail fibers of bacterial viruses (51), where several dozens of GXY triplets assemble long, flexible filaments. The remarkable feature of gp12 is that its short $(GXY)_8$ stretch confers to the overall polypeptide a collagen-like behavior with a characteristic loss of quaternary structure, corresponding to a sharp transition at temperatures around 40 °C, like animal collagen (52, 53), that is rapidly and fully reversible upon cooling (Fig. 3). The process is accompanied by physical separation of the polypeptide chains that can reassociate into heterotrimers (Fig. 4). The unusual binding of Sypro Orange dye to folded gp12 resulting in fluorescence emission reveals the presence of an accessible hydrophobic binding environment in the trimer that is lost at the beginning of denaturation, correlating with a fast drop of fluorescence (Fig. 6). The gp12 native structure is, therefore, mainly stabilized by its intermolecular collagen-like triple helix rather than by a buried hydrophobic core that would become exposed for high-affinity binding of Sypro Orange upon unfolding, in contrast to the usual behavior of proteins (54).

The interaction of gp12 with SPP1 capsids does not change its capacity to bind Sypro Orange (Fig. 6A). However, it increases the protein gp12 thermal stability by 20.2 °C, to 53.6 °C. Such stability is not limited anymore by the collagen fold intrinsic stability, being strongly enhanced by gp12 binding to the capsid lattice. This stabilization mechanism ensures the perennial association of gp12 to viral particles that are liberated to the environment when infected cells lyse. Native trimers bind cooperatively to their 60 sites in the capsid, as best appraised when gp12 is provided in limiting concentrations to interact with HΔ12 capsids. A mixed population of capsids whose majority is either fully loaded with gp12 trimers or devoided of this auxiliary protein is observed under such conditions (Figs. 7, A, B, and D, and 10A). We hypothesize that initial binding of one trimer to a capsid hexamer creates a tectonic effect that spreads across the overall icosahedral shell, promoting a conformational change of other hexamers that strongly favors interaction with gp12 trimers. Such a maturation event uncovers a novel dynamic role of the expanded capsid surface that has previously been viewed as a rather passive lattice of independent binding sites for auxiliary viral polypeptides. The rearrangement resulting from the cross-talk between the 60 gp12 attachment sites is subtle, leading to no detectable

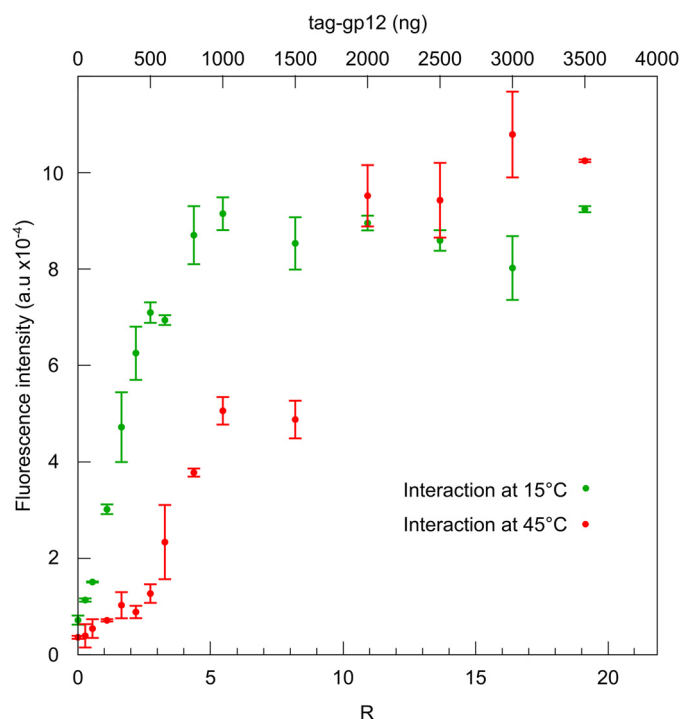


FIGURE 9. FBTSA of HΔ12 incubated with increasing amounts of tag-gp12. Capsids and tag-gp12 were preincubated separately at 15 (green) or 45 °C (red) and then mixed at the same temperature according to the experimental setup shown in the left panels of Fig. 7, A and C (not trypsinated), respectively. After coincubation, the samples were transferred to a QuantStudio 12Kflex machine for thermal denaturation at a heating rate of 3 °C/min in the presence of Sypro Orange. The amplitude of the gp12 signal with its characteristic transition at 53.6 °C (cf. Fig. 6A) was plotted against the R ratio of tag-gp12 relative to the input of HΔ12 capsids. The experimental points are averages of triplicates in two independent experiments.

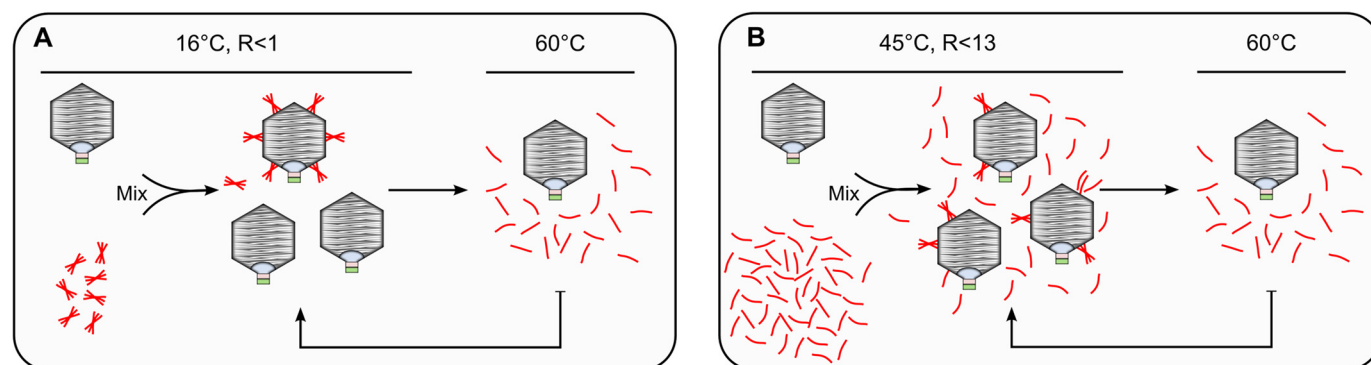


FIGURE 10. Models of native gp12 trimers (A) and unfolded gp12 polypeptides (B) binding to capsids and their dissociation from the capsid lattice in a temperature-dependent fashion.

A Collagen-like Binder of the SPP1 Viral Capsid

difference when the structure of capsids before and after gp12 binding is compared at nanometer resolution (16).

Denatured gp12 also binds to capsid lattices (Fig. 7, C and E), leading to assembly of folded trimers when present in molar excess, as assessed in Sypro Orange binding experiments (Fig. 9). Therefore, the SPP1 capsid provides a platform for attachment of unfolded gp12 polypeptide chains. When three chains meet at a gp13 hexamer interaction site, their physical proximity likely provides a window of opportunity to twist together to trimerize at a temperature (45 °C) at which free gp12 chains remain fully unstructured (Fig. 10B). If unfolded gp12 is provided at less than a 13-fold excess relative to the number of its capsid binding sites, the reaction yields a relatively homogeneous population of capsids but those are only partially filled with gp12 (Fig. 7, C and E). Such behavior, resulting from the complexity of the interaction, contrasts with the very efficient cooperative binding of folded trimers.

CONCLUSIONS

The precise architecture of viral particles achieved by tightly regulated assembly of a few different polypeptides is an excellent system to understand how the polypeptides fold and how their physicochemical properties are exploited to build megadalton biomolecular assemblies of precise architecture with exquisite efficiency. The small capsid auxiliary protein gp12 of bacterial virus SPP1 exhibits novel and noteworthy properties. It uses a collagen-like fold to assemble an elongated trimer whose thermal stability properties render it a temperature dependent binder to the capsid multivalent icosahedral platform. Cooperative binding ensures very efficient full occupancy of gp12 sites in the capsid (Figs. 7, A, B, and D, and 10A), providing experimental evidence that an initial interaction of the viral auxiliary protein exerts long-range effects in the capsid lattice, favoring attachment to its other sites in the capsid. These properties of gp12, combined with its capacity to undergo fast reversible cycles of dissociation-unfolding and refolding-reassociation to capsids, offer a versatile system to engineer the SPP1 viral particle.

Acknowledgments—We thank Anja Dröge for generously communicating the initial identification of GXY repeats in the gp12 sequence. We also thank Manuela Argentini and David Cornu for mass spectrometry analyses (mass spectrometry platform of IMAGIF); Christophe Velours and Karine Madiona for CD and analytical ultracentrifugation analyses (biophysics platform of IMAGIF); and Isabelle Auzat, Sandrine Brasilès, Charlene Cornilleau, Stéphane Roche, and Stéphane Bressanelli for advice and experimental training (Unité de Virologie Moléculaire et Structurale).

REFERENCES

1. Caspar, D. L., and Klug, A. (1962) Physical principles in the construction of regular viruses. *Cold Spring Harb. Symp. Quant. Biol.* **27**, 1–24
2. Baker, T. S., Olson, N. H., and Fuller, S. D. (1999) Adding the third dimension to virus life cycles: three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs. *Microbiol. Mol. Biol. Rev.* **63**, 862–922
3. Prasad, B. V., and Schmid, M. F. (2012) Principles of virus structural organization. *Adv. Exp. Med. Biol.* **726**, 17–47
4. Casjens, S., and King, J. (1975) Virus assembly. *Annu. Rev. Biochem.* **44**, 555–611
5. Prevelige, P. E., and Fane, B. A. (2012) Building the machines: scaffolding protein functions during bacteriophage morphogenesis. *Adv. Exp. Med. Biol.* **726**, 325–350
6. Ren, Z. J., Lewis, G. K., Wingfield, P. T., Locke, E. G., Steven, A. C., and Black, L. W. (1996) Phage display of intact domains at high copy number: a system based on SOC, the small outer capsid protein of bacteriophage T4. *Protein Sci.* **5**, 1833–1843
7. Yang, F., Forrer, P., Dauter, Z., Conway, J. F., Cheng, N., Cerritelli, M. E., Steven, A. C., Plückthun, A., and Wlodawer, A. (2000) Novel fold and capsid-binding properties of the λ -phage display platform protein gpD. *Nat. Struct. Biol.* **7**, 230–237
8. Roos, W. H., Radtke, K., Kniesmeijer, E., Geertsema, H., Sodeik, B., and Wuite, G. J. (2009) Scaffold expulsion and genome packaging trigger stabilization of herpes simplex virus capsids. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 9673–9678
9. Parent, K. N., Khayat, R., Tu, L. H., Suhanovsky, M. M., Cortines, J. R., Teschke, C. M., Johnson, J. E., and Baker, T. S. (2010) P22 coat protein structures reveal a novel mechanism for capsid maturation: stability without auxiliary proteins or chemical crosslinks. *Structure* **18**, 390–401
10. Tang, L., Gilcrease, E. B., Casjens, S. R., and Johnson, J. E. (2006) Highly discriminatory binding of capsid-cementing proteins in bacteriophage λ . *Structure* **14**, 837–845
11. Effantin, G., Boulanger, P., Neumann, E., Letellier, L., and Conway, J. F. (2006) Bacteriophage T5 structure reveals similarities with HK97 and T4 suggesting evolutionary relationships. *J. Mol. Biol.* **361**, 993–1002
12. Lander, G. C., Evilevitch, A., Jeembaveva, M., Potter, C. S., Carragher, B., and Johnson, J. E. (2008) Bacteriophage λ stabilization by auxiliary protein gpD: timing, location, and mechanism of attachment determined by cryo-EM. *Structure* **16**, 1399–1406
13. Yang, Q., Maluf, N. K., and Catalano, C. E. (2008) Packaging of a unit-length viral genome: the role of nucleotides and the gpD decoration protein in stable nucleocapsid assembly in bacteriophage λ . *J. Mol. Biol.* **383**, 1037–1048
14. Li, Q., Shivachandra, S. B., Zhang, Z., and Rao, V. B. (2007) Assembly of the small outer capsid protein, Soc, on bacteriophage T4: a novel system for high density display of multiple large anthrax toxins and foreign proteins on phage capsid. *J. Mol. Biol.* **370**, 1006–1019
15. Qin, L., Fokine, A., O'Donnell, E., Rao, V. B., and Rossmann, M. G. (2010) Structure of the small outer capsid protein, Soc: a clamp for stabilizing capsids of T4-like phages. *J. Mol. Biol.* **395**, 728–741
16. White, H. E., Sherman, M. B., Brasilès, S., Jacquet, E., Seavers, P., Tavares, P., Orlova, E. V. (2012) Capsid structure and its stability at the late stages of bacteriophage SPP1 assembly. *J. Virol.* **86**, 6768–6777
17. Lucon, J., Qazi, S., Uchida, M., Bedwell, G. J., LaFrance, B., Prevelige, P. E. Jr., and Douglas, T. (2012) Use of the interior cavity of the P22 capsid for site-specific initiation of atom-transfer radical polymerization with high-density cargo loading. *Nat. Chem.* **4**, 781–788
18. O'Neil, A., Prevelige, P. E., Basu, G., and Douglas T. (2012) Coconfinement of fluorescent proteins: spatially enforced communication of GFP and mCherry encapsulated within the P22 capsid. *Biomacromolecules* **13**, 3902–3907
19. Clark, J. R., and March, J. B. (2006) Bacteriophages and biotechnology: vaccines, gene therapy and antibacterials. *Trends Biotechnol.* **24**, 212–218
20. Fischlechner, M., and Donath, E. (2007) Viruses as building blocks for materials and devices. *Angew. Chem. Int. Ed. Engl.* **46**, 3184–3193
21. Tao, P., Mahalingam, M., Marasa, B. S., Zhang, Z., Chopra, A. K., and Rao, V. B. (2013) *In vitro* and *in vivo* delivery of genes and proteins using the bacteriophage T4 DNA packaging machine. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 5846–5851
22. Steinmetz, N. F., Lin, T., Lomonosoff, G. P., and Johnson, J. E. (2009) Structure-based engineering of an icosahedral virus for nanomedicine and nanotechnology. *Curr. Top. Microbiol. Immunol.* **327**, 23–58
23. Strable, E., and Finn, M. G. (2009) Chemical modification of viruses and virus-like particles. *Curr. Top. Microbiol. Immunol.* **327**, 1–21
24. Dröge, A., Santos, M. A., Stiege, A. C., Alonso, J. C., Lurz, R., Trautner, T. A., and Tavares, P. (2000) Shape and DNA packaging activity of bacteriophage SPP1 procapsid: protein components and interactions during

- assembly. *J. Mol. Biol.* **296**, 117–132
25. Oliveira, L., Tavares, P., and Alonso, J. C. (2013) Headful DNA packaging: bacteriophage SPP1 as a model system. *Virus Res.* **173**, 247–259
 26. Isidro, A., Henriques, A. O., and Tavares, P. (2004) The portal protein plays essential roles at different steps of the SPP1 DNA packaging process. *Virology* **322**, 253–263
 27. Plisson, C., White, H. E., Auzat, I., Zafarani, A., São-José, C., Lhuillier, S., Tavares, P., and Orlova, E. V. (2007) Structure of bacteriophage SPP1 tail reveals trigger for DNA ejection. *EMBO J.* **26**, 3720–3728
 28. Auzat, I., Dröge, A., Weise, F., Lurz, R., and Tavares, P. (2008) Origin and function of the two major tail proteins of bacteriophage SPP1. *Mol. Microbiol.* **70**, 557–569
 29. Lurz, R., Orlova, E. V., Günther, D., Dube, P., Dröge, A., Weise, F., van Heel, M., and Tavares, P. (2001) Structural organisation of the head-to-tail interface of a bacterial virus. *J. Mol. Biol.* **310**, 1027–1037
 30. Alonso, J. C., Lüder, G., Stiege, A. C., Chai, S., Weise, F., and Trautner, T. A. (1997) The complete nucleotide sequence and functional organization of *Bacillus subtilis* bacteriophage SPP1. *Gene*. **204**, 201–212
 31. 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
 32. Dröge, A., and Tavares, P. (2000) *In vitro* packaging of DNA of the *Bacillus subtilis* bacteriophage SPP1. *J. Mol. Biol.* **296**, 103–115
 33. Becker, B., de la Fuente, N., Gassel, M., Günther, D., Tavares, P., Lurz, R., Trautner, T. A., and Alonso, J. C. (1997) Head morphogenesis genes of the *Bacillus subtilis* bacteriophage SPP1. *J. Mol. Biol.* **268**, 822–839
 34. Isidro, A., Santos, M. A., Henriques, A. O., and Tavares, P. (2004) The high-resolution functional map of bacteriophage SPP1 portal protein. *Mol. Microbiol.* **51**, 949–962
 35. Poh, S. L., el Khadali, F., Berrier, C., Lurz, R., Melki, R., and Tavares, P. (2008) Oligomerization of the SPP1 scaffolding protein. *J. Mol. Biol.* **378**, 551–564
 36. Philo, J. S. (1997) An improved function for fitting sedimentation velocity data for low-molecular-weight solutes. *Biophys. J.* **72**, 435–444
 37. Jakutyte, L., Baptista, C., São-José, C., Daugelavičius, R., Carballido-López, R., and Tavares, P. (2011) 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
 38. Cole, C., Barber, J. D., and Barton, G. J. (2008) The Jpred3 secondary structure prediction server. *Nucleic Acids Res.* **36**, W197–W201
 39. Söding, J., Biegert, A., and Lupas, A. N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244–W248
 40. Dröge, A. (1998) *Capsidmorphogenese des Bakteriophagen SPP1*. Doctoral thesis, Technische Universität Berlin, Berlin, Germany
 41. Bella, J., Eaton, M., Brodsky, B., and Berman, H. M. (1994) Crystal and molecular structure of a collagen-like peptide at 1.9 Å resolution. *Science* **266**, 75–81
 42. Bella, J., Brodsky, B., and Berman, H. M. (1995) Hydration structure of a collagen peptide. *Structure* **3**, 893–906
 43. Cabré, F., Canela, E. I., and Canela, M. A. (1989) Accuracy and precision in the determination of Stokes radii and molecular masses of proteins by gel filtration chromatography. *J. Chromatogr.* **472**, 347–356
 44. Seifter, S., and Gallop, P. M. (1962) Collagenase from *Clostridium histolyticum*: collagen + H₂O → peptides gelatin + H₂O → peptides. *Methods Enzymol.* **5**, 659–665
 45. Orlova, E. V., Gowen, B., Dröge, A., Stiege, A., Weise, F., Lurz, R., van Heel, M., and Tavares, P. (2003) Structure of a viral DNA gatekeeper at 10 Å resolution by cryo-electron microscopy. *EMBO J.* **22**, 1255–1262
 46. Wallace, B. A., and Janes, R. W. (2001) Synchrotron radiation circular dichroism spectroscopy of proteins: secondary structure, fold recognition and structural genomics. *Curr. Opin. Chem. Biol.* **5**, 567–571
 47. Ericsson, U. B., Hallberg, B. M., Detitta, G. T., Dekker, N., and Nordlund, P. (2006) Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Anal. Biochem.* **357**, 289–298
 48. Persikov, A. V., Ramshaw, J. A., and Brodsky, B. (2005) Prediction of collagen stability from amino acid sequence. *J. Biol. Chem.* **280**, 19343–19349
 49. Xu, Y., Keene, D. R., Bujnicki, J. M., Höök, M., and Lukomski, S. (2002) Streptococcal Scl1 and Scl2 proteins form collagen-like triple helices. *J. Biol. Chem.* **277**, 27312–27318
 50. Mohs, A., Silva, T., Yoshida, T., Amin, R., Lukomski, S., Inouye, M., and Brodsky, B. (2007) Mechanism of stabilization of a bacterial collagen triple helix in the absence of hydroxyproline. *J. Biol. Chem.* **282**, 29757–29765
 51. Ghosh, N., McKillop, T. J., Jowitt, T. A., Howard, M., Davies, H., Holmes, D. F., Roberts, I. S., and Bella, J. (Jun 6, 2012) Collagen-like proteins in pathogenic *E. coli* strains. *PLoS ONE* 10.1371/journal.pone.0037872
 52. Leikina, E., Mertts, M. V., Kuznetsova, N., and Leikin, S. (2002) Type I collagen is thermally unstable at body temperature. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 1314–1318
 53. Cao, H., and Xu, S.-H. (2008) Purification and characterization of type II collagen from chick sternal cartilage. *Food Chem.* **108**, 439–445
 54. Kranz, J. K., and Schalk-Hihi, C. (2011) Protein thermal shifts to identify low molecular weight fragments. *Methods Enzymol.* **493**, 277–298

Titre : GP12 : une protéine de type collagène qui se fixe à la capsid du bactériophage SPP1

Mots clés : collagen-like, protéines auxiliaires de la capsid, gp12

Résumé : Gp12 est une protéine qui se fixe symétriquement au centre de chacun des 60 hexamères de la capsid icosaédrique du bactériophage SPP1. La protéine produite dans un système d'expression hétérologue se lie à la capsid de particules virales dont le gène codant gp12 a été inactivé. Cette interaction a lieu spécifiquement avec des capsids qui ont subi le processus d'expansion et encapsulé l'ADN viral. L'analyse de la séquence de gp12 montre la présence d'un motif (GXY)_n retrouvé dans des protéines de type collagène.

Nous avons démontré que gp12 est un trimère allongé en solution. Ce trimère s'avère sensible à la collagénase VII qui coupe la protéine gp12 dans un site spécifique du motif (GXY)₈. Le profil de dichroïsme circulaire de gp12 porte aussi la signature d'une protéine de type collagène. La fixation de gp12 sur la capsid virale conduit à une augmentation de 20°C de sa stabilité thermique. Gp12 peut être dénaturée-dissociée et puis renaturée-reassociée sous l'effet de la température.

Le trimère de gp12 et sa forme dénaturée se fixent à la capsid de SPP1 mais avec des profils d'interaction différents. Ces propriétés permettent d'utiliser gp12 comme un ligand réversible de la capsid phagique en fonction de la température. Gp12 a une organisation modulaire avec un motif collagène qui sépare les modules amino et carboxyl-terminaux.

Des protéines avec une organisation similaire sont codées par des gènes adjacents à celui codant pour la protéine majoritaire de la capsid dans des prophages de *Bacilli*, suggérant une fonction similaire à gp12. Leurs modules ont une taille variable. Une recherche de protéines procaryotes et virales avec des segments collagène a montré qu'elles sont abondantes parmi les bactéries et les virus. Le motif est rare parmi les archées et leurs virus.

Ces résultats montrent l'importance des protéines avec des séquences de type collagène dans le monde non-eucaryote et du développement de leur étude biochimique et fonctionnelle.

Title : Gp12 : a collagen-like protein that binds to the bacteriophage SPP1 capsid

Keywords : collagen-like, capsid auxiliary protein, gp12

Abstract : Gp12 is a protein found distributed symmetrically at the surface of the icosahedral capsid from bacteriophage SPP1. Recombinant gp12 binds to phage particles whose gene coding for gp12 was disrupted. This interaction occurs specifically with capsids that undergone expansion and packaged DNA.

The gp12 protein sequence is marked by the presence of a stretch of 8 repeats of a GXY motif, which is the sequence signature of collagen. Our results showed that gp12 is an elongated trimer in solution. The trimer is sensitive to collagenase VII that cuts the gp12 protein inside the collagen motif. Its circular dichroism profile has also the signature of a collagen-like protein. Binding of gp12 to SPP1 capsids increases its thermal stability by 20°C.

Gp12 is denatured and dissociated reversibly by temperature shift.

The gp12 trimer and its denatured form bind to SPP1 capsids but with a different interaction behavior. These properties allow to use gp12 as thermo-switchable SPP1 capsid binder. Gp12 has a modular organization with a central collagen motif that connects the amino and carboxyl termini. Proteins with a similar organization that are encoded by genes adjacent to the gene coding for the major capsid protein were identified in prophages of *Bacilli*, suggesting a function similar to gp12. Their modules have a variable length.

A pangenome-wide search for collagen-like proteins in prokaryotes and viruses shows that they are abundant among bacteria and viruses. In contrast, this motif is rare in archaea and their viruses. Our analysis highlights the importance of collagen-like proteins in the non-eukaryotic world and supports the interest to develop their biochemical and structural study.

