École Doctorale Biologie Santé de Lille

THÈSE
Pour l'obtention du grade de
DOCTEUR DE L'UNIVERSITE DE LILLE

Discipline
Aspects moléculaires et cellulaires de la biologie

Structural and functional study of the HEV ORF3 protein by solution NMR

Étude structurale et fonctionnelle de la protéine ORF3 du VHE par RMN en solution

Présentée et soutenue publiquement par
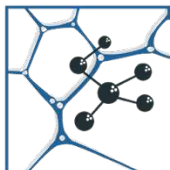
Danai Kalliopi MOSCHIDI

Le 25 janvier 2023

Directeur de thèse : Docteur Xavier HANOULLE

JURY

Docteur Anja BOCKMANN   Rapporteur
Docteur Jérôme GOUTTENOIRE   Rapporteur
Docteur Laurence COCQUEREL   Examinateur
Docteur Xavier HANOULLE   Directeur de thèse

Integrative Structural Biology
EMR 9002 CNRS U1167

Acknowledgements

These three years of my PhD journey have been incredibly educational and have helped me improve my knowledge in various fields and shaped me personally. This work could not have been successful without the contribution of many people in the laboratory, but also people in my personal life who supported me during this difficult period also due to the Covid19 pandemic.

My greatest acknowledgements are addressed to my supervisor Dr. Xavier Hanoulle for his guidance and support throughout these years. First, he gave me the opportunity to come to Lille and follow my dreams. Our discussions and his help in any question I had, helped me expand my scientific, management and critical thinking skills during this work. I am also grateful for correcting, suggesting and discussing all my writing attempts and especially this thesis. I cannot forget his help whenever I had to manage something in French from the very beginning.

I would like to sincerely thank Dr. Anja Böckmann and Dr. Jérôme Gouttenoire, who firstly accepted to be the members of *Comité de Suivi Individuel* (CSI) jury, and their questions and suggestions at each of our annual meetings provide me with more ideas to continue this project. Their contribution in reviewing this thesis, their insightful comments and suggestions were incredibly helpful. I have to also thank Dr. Laurence Cocquerel who is a valuable member of my thesis committee and her knowledge helped me to improve this dissertation.

My esteemed appreciations also go to ANRS agency which funded this research (research grant ECTZ103422 obtained during the AAP 2019-2 of the CSS12, ECTZ101316 project) and supported me through these three years of my research.

I am also grateful to Dr. Isabelle Landrieu who gave me the opportunity to come two months before my PhD contract started and work in the lab, meet the other lab members and familiarize myself with the -new for me- lab environment and city. I want to thank all the members of Integrative Structural Biology group who contributed to this work. Especially, Dr. François-Xavier Cantrelle and Emmanuelle Boll who recorded and analyzed some of the NMR data presented in this work, and Dr. Elian Dupré and Dr. Zoe Lens who conducted the initial experiments that were

<div style="text-align: right">

Thank you all,

Danai-Kalliopi Moschidi

</div>

Abstract

Hepatitis E Virus (HEV) is the most common cause of acute viral hepatitis worldwide with over 20 million infections and around 44,000 deaths recorded annually. Until today, 8 HEV genotypes are reported with only the first four genotypes infecting humans. In developing countries, HEV is transmitted via the fecal-oral route through contaminated drinking water, whereas in developed countries, the transmission occurs by consumption of uncooked or undercooked meat from infected animals. There is no specific treatment for HEV infection, apart from a vaccine which is available only in China. Therefore, HEV represents a public health problem which is constantly growing around the world. HEV is small, icosahedral virus with 27 to 40 nm diameter and classified in the *Hepeviridae* family. Virions can be found as quasi-enveloped, by host-cell-derived membranes, in bloodstream or as non-enveloped in the faeces and bile of infected individuals. The virus contains a ~7.2 kb positive-sense, single-stranded RNA genome which comprises three open reading frames: ORF1, ORF2 and ORF3. ORF1 encodes a non-structural polyprotein that includes multiple functional domains responsible for the replication of the viral genome. ORF2 encodes the viral capsid protein that assembles to make the viral particles. ORF3 encodes a small regulatory multifunctional protein which is poorly characterized. Previous studies have shown that ORF3 is mainly involved in the release of the infectious viral particles and interacts with other viral and host proteins inside the cell. This small protein is thought to interact with the human Ubiquitin E2 variant (UEV) domain of Tsg101 protein (Tsg101 UEV), a member of endosomal sorting complex required for transport-I (ESCRT-I), which has been involved in the release of HIV, Ebola and other viruses. Previous studies have also shown that ORF3 protein is associated with the cellular membranes either via its oligomerization and transmembrane insertion, or via palmitoylation of its N-terminal Cysteine-rich region.

In this study, a detailed molecular characterization of the ORF3 protein in order to decipher its functional role(s) during the HEV life cycle is achieved using NMR spectroscopy and other biophysical techniques. Firstly, we performed HEV ORF3 recombinant expression in. *E. coli* and purified the protein. Its structural characterization by solution-state NMR spectroscopy shown that it is an intrinsically disordered protein devoid of any stable 3D structure organization.

Secondly, for the further determination of the ORF3 protein membrane association, the Nanodisc (ND) technology is used which mimics a bilayer membrane and thus, is closed to the native environment of membrane proteins. Although the attachment of the protein into a ND assembly was successful, a reliable conclusion about the membrane attachment of the protein cannot be drawn based on our experimental results. Finally, we in-depth characterized the molecular interaction between HEV ORF3 and Tsg101 UEV proteins. We identified the binding sites in both ORF3 and Tsg101 UEV proteins by NMR spectroscopy, we measured the affinity of the interaction using Isothermal Titration Calorimetry (ITC) and finally a high-resolution atomic structure of the complex between Tsg101 UEV protein and a 10-residues ORF3-derived peptide was solved by X-ray crystallography. In addition, we used our biochemical and structural data to start the setup of an experimental assay that could be used to detect potential antiviral compounds targeting the ORF3-Tsg101 interaction.

Keywords: Protein biochemistry, NMR spectroscopy, Structural biology, HEV

Résumé

Le virus de l'hépatite E (VHE) est la cause la plus fréquente d'hépatite virale aiguë dans le monde avec plus de 20 millions d'infections et environ 44 000 décès enregistrés chaque année. A ce jour, 8 génotypes de VHE ont été identifiés, seuls les quatre premiers génotypes infectent les humains. Dans les pays en développement, le VHE se transmet par voie fécale-orale principalement via de l'eau contaminée, alors que dans les pays développés, la transmission se fait par l'intermédiaire de la consommation de viande crue ou insuffisamment cuite provenant d'animaux infectés. Il n'existe pas de traitement spécifique pour l'infection par le VHE, à part un vaccin qui n'est disponible qu'en Chine. Le VHE représente par conséquent un problème de santé publique qui ne cesse de croître dans le monde. Le VHE est un petit virus icosaédrique de 27 à 40 nm de diamètre qui est classé dans la famille des *Hepeviridae*. Les virions peuvent être trouvés quasi-enveloppés, par des membranes dérivées de la cellule hôte, dans la circulation sanguine ou non enveloppés dans les selles et la bile des individus infectés. Le virus contient un génome à ARN simple brin positif d'environ 7.2 kb qui comprend trois cadres de lecture ouverts : ORF1, ORF2 et ORF3. ORF1 code pour une polyprotéine non structurale qui comprend plusieurs domaines fonctionnels responsables de la réplication du génome viral. ORF2 code la protéine de capside virale qui s'assemble pour former les particules virales. ORF3 code pour une petite protéine multifonctionnelle régulatrice qui est mal caractérisée. Des études antérieures ont montré que l'ORF3 est principalement impliquée dans la libération des particules virales infectieuses et qu'elle interagit avec d'autres protéines virales et hôtes à l'intérieur de la cellule. Cette petite protéine pourrait interagir avec le domaine de la variante humaine de l'ubiquitine E2 (UEV) de la protéine Tsg101 (Tsg101 UEV), un membre du 'endosomal sorting complex required for transport-I' (ESCRT-I), qui a été impliqué dans la libération d'autres virus comme le VIH et Ebola. Des études antérieures ont également montré que la protéine ORF3 est associée aux membranes cellulaires soit via son oligomérisation et son insertion transmembranaire, soit via la palmitoylation de sa région N-terminale riche en cystéine.

Dans cette étude, une caractérisation moléculaire détaillée de la protéine ORF3 afin de déchiffrer son ou ses rôles fonctionnels au cours du cycle de vie du VHE est réalisée à l'aide de la

spectroscopie RMN et d'autres techniques biophysiques. Tout d'abord, nous avons produit l'ORF3 du VHE dans *E. coli* puis purifié cette protéine recombinante. Sa caractérisation structurale par spectroscopie RMN en solution a montré qu'il s'agit d'une protéine intrinsèquement désordonnée dépourvue de structure 3D stable. Ensuite, pour un examen plus poussé de l'association membranaire de la protéine ORF3, la technologie Nanodisque (ND) qui imite une membrane bicouche physiologique a été utilisée. Bien que l'attachement de la protéine dans un ND ait réussi, une conclusion fiable sur le mode de fixation membranaire de ORF3 n'a pu être tirée sur la base de nos résultats expérimentaux. Enfin, nous avons caractérisé en détail l'interaction moléculaire entre les protéines HEV ORF3 et Tsg101 UEV. Nous avons identifié les sites de liaison dans les protéines ORF3 et Tsg101 UEV par spectroscopie RMN, nous avons mesuré l'affinité de l'interaction à l'aide de la calorimétrie de titrage isotherme (ITC) et enfin une structure atomique à haute résolution du complexe entre la protéine Tsg101 UEV et un peptide de 10 résidus dérivé de ORF3 a été résolue par cristallographie aux rayons X. Finalement, nous avons utilisé nos données biochimiques et structurales pour mettre au point un test expérimental qui pourrait être utilisé pour détecter des composés antiviraux potentiels ciblant l'interaction ORF3-Tsg101.

Mots clés : Biochimie des protéines, Spectroscopie RMN, Biologie structurale, VHE

Table of Contents

Table of Figures

15

21

Table of Tables

Abbreviations

aa	amino acid
ALG-2	Apoptosis-linked gene 2
Alix	Apoptosis-linked gene-2 interacting protein X
App-1"-p	ADP-ribose-1"-monophsosphate
Appr-1"-pase	ADP-ribose-1"-monophsosphatase
Asn	Asparagine
BMRB	Biological Magnetic Resonance Data Bank
BNYVV	Beet necrotic yellow vein virus
Bro1	BCK1-like resistance to osmotic shock protein-1
CBV	coxsackie B virus
CHMP	charged multivesicular body proteins
CMV	cytomegalovirus
CPMG	Carr-Purcell Meiboom-Gill relaxation dispersion
CREs	cis-reactive elements
CSP	chemical shift perturbations
CV	calicivirus
CV	Column Volume
DGS-NTA(Ni)	1,2-dioleoyl-sn-glycero-3-[(N-(5-amino-1-carboxypentyl)iminodiacetic acid)succinyl] (nickel salt) (18:1 DGS-NTA(Ni))
DLS	Dynamic Light Scattering
DMPC	1,2-dimyristoyl-sn-glycero-3-phosphocholine (14:0 PC – DMPC)
DOPC	1,2-dioleoyl-sn-glycero-3-phosphocholine (18:1 (Δ9-Cis) PC – DOPC)
DOPS	1,2-dioleoyl-sn-glycero-3-phospho-L-serine (sodium salt) (18:1 PS – DOPS)
DSF	Differential Scanning Fluorimetry
eEF1a1	eukaryotic elongation factor 1 isoform-1
EGFRs	epidermal growth factor receptors
eHEV	quasi-enveloped virions
eIF4F	eukaryotic initiation factor 4F complex
EMCV	encephalomyocarditis virus
ER	endoplasmic reticulum
ERK	extracellularly regulated kinase
ESCRT	endosomal sorting complex required for transport
FID	Free Induction Decay
FP	Fluorescence Polarization
FRET	fluorescence resonance energy transfer
GDD	Mg2+ binding sequence highly conserved motif
Gln	Glutamine
GLUE	GRAM-Like Ubiquitin-binding in EAP45
HCCA	α-Cyano-4-hydroxycinnamic acid
HCV	Hepatitis C Virus
Hel	helicase

HEV	Hepatitis E Virus
HEV 239	Hecolin® vaccine
HIF-1	hypoxia-inducible factor 1
HIV-1	human immunodeficiency virus-1
HRS	hepatocyte growth factor-regulated tyrosine kinase substrate
HSPGs	heparan sulphate proteoglycans
HSQC	Heteronuclear Single Quantum Coherence spectrum
HSV 1/2	Herpes Simplex Virus 1/2
HVR	hypervariable region
IAV	influenza A virus
IDP	internally displaced persons
IDP(s)	intrinsically disordered protein(s)
IDR(s)	intrinsically disordered protein region(s)
IEM	Immune Electron Microscopy
IFV	influenza virus
IgG	Immunoglobulin G
IgM	Immunoglobulin M
IPTG	isopropyl-β-D-thiogalactopyranoside
ITC	Isothermal Titration Calorimetry
KSHV	Kaposi's sarcoma-associated herpesvirus
LB	Lysogeny Broth
LDDT	Local Distance Difference Test
LPR	Lipid-to-Protein Ratio
LuMPIS	luminescence-based mammalian interactome mapping pull-down assays
M	ORF2 middle domain
m^{7}G	7-methylguanosine cap
MALDI-TOF	Matrix Assisted Laser Desorption Ionization-Time Of Flight mass spectrometry
MARK	mitogen-activated protein kinase
MARV	Marburg virus
MeT	methyltransferase
Met	Methionine
MIT	microtubule-interacting and trafficking) domains
MMseqs2	Many-against-Many searching) software
ms	milliseconds
MSA	Multiple Sequence Alignments
MSPs	Membrane Scaffold Proteins
MVB	multivesicular bodies
ND	Nanodisc
NF-jB	nuclear factor kappa B
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser effect
NPC1	Niemann-Pick disease type C1
ns	nanoseconds

NSP	Nuclear spin relaxation
nt	nucleotides
NTPase	nucleoside-triphosphatase
NZF	Npl4-type zinc finger
ORF(s)	open reading frame(s)
P or E2s	ORF2 protruding domain
PAE	Predicted Aligned Error
PAMPs	pathogen-associated motif patterns
PCP	papain-like cysteine protease
PDB	Protein Data Bank
PE	phosphatidylethanolamine
PE-MCC	1,2-dioleoyl-sn-glycero-3-phosphoethanolamine-N-[4-(p-maleimidomethyl)cyclohexane-carboxamide] (sodium salt) (18:1 PE-MCC)
PEF	penta-EF-hand
PG	1,2-dimyristoyl-sn-glycero-3-phospho-(1'-rac-glycerol) (sodium salt) (14:0 PG)
PG	phosphatidylglycerol
PI3P	phosphatidylinositol 3-phosphate
polyA	polyadenylated
PRE	Paramagnetic relaxation enhancement
Pro	proline-rich region
PRR	proline-rich region
ps	picoseconds
RCI-S^2	Random Coil Index S^2
RDC	Residual dipolar coupling
RdRp	RNA-dependent RNA polymerase
rHEV	recombinant HEV protein vaccine
ribavirin	1-β-D-ribofuranosyl-1,2,4-triazole
RIG-I	retinoic acid-inducible gene I
RIP1	receptor-interacting protein kinase 1
RMSD	root-mean-square deviation
RP	Reverse Phase Chromatography
RT-PCR	Reverse transcriptase polymerase chain reaction
RubV	rubella virus
S	ORF2 shell domain
SAMSA Fluorescein	(5-((2-(and-3)-S-(acetylmercapto)succinoyl)amino)fluorescein)
SB	α-helical/steadiness box domain
sc	side chains
SDS-PAGE	Sodium dodecyl-sulfate polyacrylamide gel electrophoresis
SEC	Size Exclusion Chromatography
SH3	Src homology 3
SOF	sofosbuvir
ss	single-stranded
SSP	Secondary Structure Propensities

STAM1/2	signal transducing adaptor molecule1/2
SV40	simian virus 40
SW	spectral width
TEV	Tobacco Etch Virus
THP	Tris(hydroxypropyl)phosphine)
TLR-3	toll-like receptor 3
Tm	melting temperature
TMSP	3-(Trimethylsilyl)propanoic acid
TRADD	tumor necrosis factor receptor 1-associated death domain protein
Trp	Tryptophan
TSA	Thermal Shift Assay
Tsg101	tumor susceptibility gene 101 protein
UBAP1	ubiquitin-associated protein 1
UEV	ubiquitin E2 variant domain
VLP	virus-like particle
WT	Wild Type
Y2H	yeast two-hybrid
β-OG	octyl-β-D-glucopyranoside

Publications

Published in peer-review journal:

Cantrelle FX[#], Boll E[#], Brier L[#], **Moschidi D**, Belouzard S, Landry V, Leroux F, Dewitte F, Landrieu I, Dubuisson J, Deprez B, Charton J, Hanoulle X. "NMR spectroscopy of the main protease of SARS-CoV-2 and fragment-based screening identify three protein hotspots and an antiviral fragment." Angewandte Chemie International Edition. 2021 Nov 22;60(48):25428-35.

[#] Authors contributed equally.

Moschidi D, Cantrelle FX, Boll E, Hanoulle X. "Backbone NMR resonance assignment of the apo human Tsg101-UEV domain." Biomolecular NMR Assignments. 2023 Feb 6:1-6.

Brier L[#], Hassan H[#], Hanoulle X[#], Landry V, **Moschidi D**, Desmarets L, Rouillé Y, Dumont J, Herledan A, Warenghem S, Piveteau C, Carré P, Ikherbane S, Cantrelle FX, Dupré E, Dubuisson J, Belouzard S, Leroux F, Deprez B, Charton J. "Novel dithiocarbamates selectively inhibit 3CL protease of SARS-CoV-2 and other coronaviruses". European Journal of Medicinal Chemistry. 2023 Feb 6:115186.

[#] Authors contributed equally.

In preparation:

The work in this study about HEV ORF3 protein is included in a journal paper which is in preparation.

Communications

<u>Oral presentations</u>:

"PhD project on the ORF3 protein from the Hepatitis E virus", Presented at Réunion de labo, February 3rd 2020, Pasteur Institute of Lille.

MT180 – 3 min thesis presentation: September 14th 2020 and September 20th 2021, Pasteur Institute of Lille.

"Structural and functional characterization of HEV ORF3 protein by NMR Spectroscopy" on 1ère conférence virtuelle du GERM2021 on 6th, <u>8th</u> and 9th April 2021 (online conference).

"Structural and functional characterization of HEV ORF3 protein by NMR Spectroscopy" on 9es Rencontres RMN-RPE-IRM on 10th June 2021 (online seminar).

"Structural and functional characterization of HEV ORF3 protein by NMR Spectroscopy" on Réunion Annuelle AC42 de l'ANRS Réseau National Hépatites on 15th – <u>16th</u> June 2021 (online conference) and on <u>7th</u> – 8th June 2022, Espace Saint-Martin, Paris, France.

"Structural and functional characterization of HEV ORF3 protein by NMR Spectroscopy" on 20ème Journée André Verbert 2021 (online course/seminar).

<u>Poster presentations</u>:

<u>Moschidi D</u>, Dupré E, Boll E, Cantrelle FX, Lens Z, Hanoulle X. *Structural and functional characterization of HEV ORF3 protein by NMR Spectroscopy*; Poster presented at Recent advances in structural biology of membrane proteins, EMBO virtual Workshop, 29th November – 1st December 2021.

<u>Moschidi D</u>, Dupré E, Boll E, Cantrelle FX, Lens Z, Hanoulle X. *Structural and functional characterization of HEV ORF3 protein by NMR Spectroscopy*; Poster presented at Workshop on Structural Biophysics, 5th – 10th December 2021, Bordeaux, France.

Introduction

1. Hepatitis E Virus

1.1 General

Hepatitis E Virus (HEV) is the most common cause of acute viral hepatitis worldwide with over 20 million infections and around 44,000 deaths recorded annually[1]. The first documented HEV infection was in New Delhi, India in 1955 when the scientists could not categorize the infection to an existed hepatitis virus and therefore it was referred to as non-A, non-B hepatitis virus[2]. After 28 years, in 1983, there was another outbreak of viral hepatitis among Soviet soldiers in Afghanistan. The Russian virologist Mikhail Balayan voluntary ingested pooled fecal extracts from infected soldiers causing himself to become ill. Then, collecting during the incubation period and analyzing his own stool by Immune Electron Microscopy (IEM), he proved that the non-A, non-B hepatitis virus transmitted via the fecal-oral route in human, identified 27- to 30-nm spherical virus-like particles in his stool could cause the infection and then attempted to inoculate the virus into *Macacus cynomolgus* monkeys[3]. In 1990, Reyes *et al.* successfully isolated, cloned and sequenced the first partial cDNA clone of the non-A, non-B hepatitis virus and thus proposed to name hepatitis E virus (HEV)[4,5]. By the next year, the cloning and sequencing of the full-length viral genome of HEV were achieved[6].

HEV infection is acute, self-limiting and mainly asymptomatic which lasts usually few weeks, 5 to 6 weeks on average[7]. The symptoms are nonspecific and indistinguishable from acute hepatitis A, B and C infection[1]. The HEV illness have two phases. The first one starts with mild fever, anorexia, nausea and vomiting for up to 10 days[8]. In the next phase, the icteric phase, the patients with acute hepatitis E present also jaundice, darkened urine, pale stools and in some cases abdominal pain, generalized itching and skin rash, which lasts 15-40 days[1,8]. After the incubation period, the virus is released from the liver cells (hepatocytes) to the bile and then excreted into the stools[9]. The majority of patients fully recover, but rarely acute hepatitis E could be fatal as results in fulminant hepatitis and acute liver failure[1]. HEV infection in pregnant women in the second or third trimester could cause acute liver failure, fetal loss and mortality up to 30%[2,10]. In

immunocompromised patients, patients after organ transplantation, chemotherapy or with HIV infection, HEV leads to chronicity and these patients are at a higher risk of rapid progression to cirrhosis[2].

The early detection of a HEV infection is difficult because of the nonspecific and indistinguishable from other acute hepatitis infections symptoms[1]. The diagnosis of HEV infection is made either serologically or by molecular techniques. The detection of specific antibodies, the anti-HEV immunoglobulin M (IgM) and G (IgG) antibodies, during the incubation period of the infection in the patient's blood is achieved by commercial tests that are not still approved by Food and Drug Administration (FDA)[11]. However, these tests are not reliable because of the significant variation in their sensitivity and their specificity. For chronic HEV infections, where the levels of the anti-HEV antibodies are limited, molecular techniques are used to detect the virus directly[5]. By the reverse transcriptase polymerase chain reaction (RT-PCR) technique, the HEV RNA is detected in blood approximately 20 days after the onset of symptoms while in stool remains for further 2 weeks[12]. This assay has higher specificity than the serologic diagnosis, but it is performed in specialized research laboratories and thus it is usually impractical for quick diagnosis[1,12].

1.2 Epidemiology – Transmission of the virus

Hepatitis E virus is classified in the *Hepeviridae* family which contains two genera, the genus *Orthohepevirus* and the genus *Piscihepevirus*[13,14]. Until today, there are reported 8 HEV genotypes (HEV1 – HEV8) and only one serotype[2]. Only the HEV genotypes 1, 2, 3, 4 and 7 of species *Orthohepevirus A* infect humans. The three other HEV species of genus *Orthohepevirus* infect animals, but not humans and especially the *Orthohepevirus B* found in chicken, the *Orthohepevirus C* isolated from rat, greater bandicoot, Asian musk shrew, ferret, and mink and the *Orthohepevirus D* found in bat[2]. The genus *Piscihepevirus* has one species *Piscihepevirus A* found in trout and also is not contagious to humans[15]. Genotypes 1 and 2 infect only humans and are found mainly in developing countries where the transmission occurs via contaminated water. The illness occurred by these genotypes is reported to be self-limiting without progression to chronic disease in epidemic cases in Asia, Africa and South America for genotype 1 and in Mexico, Nigeria and Chad for genotype 2[11,16]. Genotype 3 and 4 are found mainly in developed countries and are reported in sporadic cases causing zoonotic infections in humans who consumed uncooked or undercooked meat from infected animals[2]. The main host animals reported are domestic swine, bats, ferrets, rabbits, mongooses, deer and wild boars with swine to be the primary reservoirs of HEV infection in humans[15]. In addition, these genotypes are associated with chronic HEV cases in immunocompromised patients[17] as well there are reports of HEV infections caused by blood transfusion[14]. Genotypes 5, 6 and 8 are reported only in animals, with HEV5 and HEV6 found only in Japanese wild boar[16] and HEV8 only in Bactrian camels[18]. Regarding genotype 7, it is found in dromedary camels and in 2016 one case of human infection reported in a patient with liver transplant receptor consumed camel milk and meat[19]. The natural hosts and the transmission route of all HEV genotypes are presented in Figure 1[2].
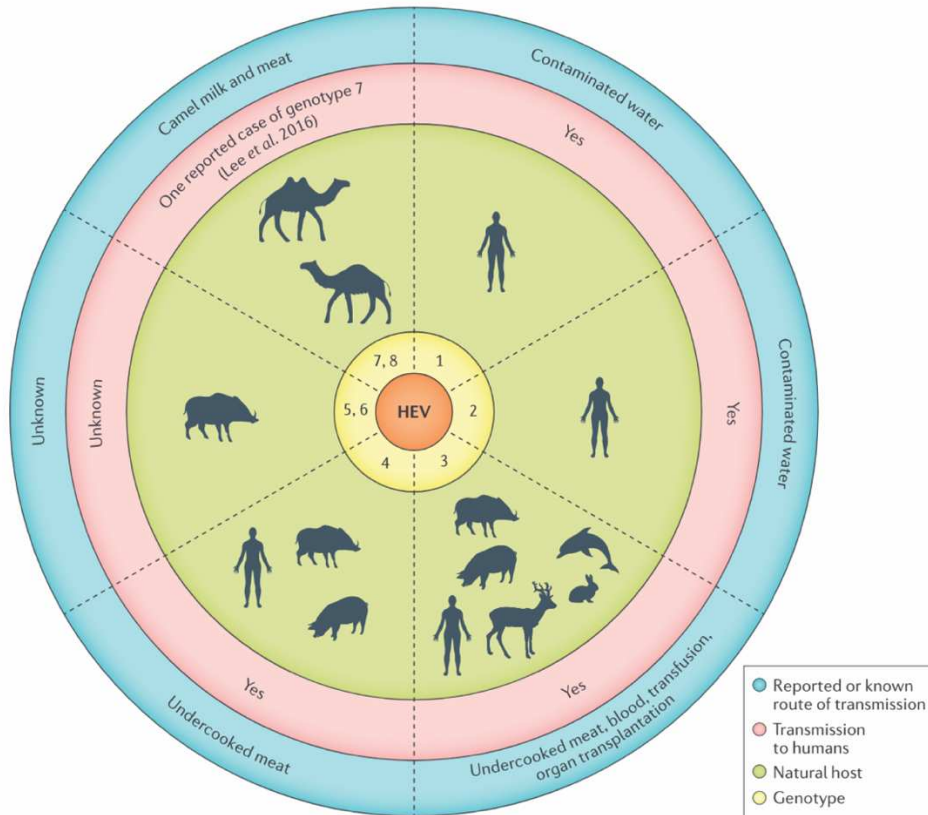
Figure 1. HEV Genotypes (HEV1-HEV8) classification with natural hosts and transmission route reported for each genotype. From Nimgaonkar et al.[2] with permission from Nature Reviews Gastroenterology & Hepatology, Copyright 2018.

Based on nucleotide sequences and phylogenetic analysis, the first 4 HEV genotypes are further divided into sub-genotypes, the assigned ones presented in Figure 2 combined with the remaining unassigned demonstrate the genetic diversity of the virus[14,20]. The HEV3c, HEV3e, HEV3f and HEV3g sub-genotypes are mainly found in Europe[21,22].

Figure 2. Sub-genotypes of the first 4 HEV genotypes. HEV1 divided into 6 (a-f), HEV2 into 2 (a-b), HEV3 into 11 (a-j, ra) and HEV4 into 9 (a-i) sub-genotypes. From Raji at al.[14] available under a Creative Commons Attribution-NonCommercial-No Derivatives License (CC BY NC ND).

HEV represents a major public health problem which is constantly growing around the world. The last decades and especially from 2005 to 2015, there is a 10-fold increase of incidents recorded in Europe[23], but also the number of the silent infections could not be precisely determined[21]. The worldwide distribution of the four HEV genotypes (HEV1-HEV4) that infect humans, directly and indirectly, is shown in Figure 3[24].

Figure 3. Worldwide distribution of the four HEV genotypes (HEV1-HEV4) that infect humans, directly and indirectly. From Kamar et al.[24] with permission from Nature Reviews Disease Primers, Copyright 2017.

1.3 Genome Organization

Hepatitis E virus is a small, icosahedral virus, member of the alphavirus-like supergroup III of RNA single-stranded (ss) + virus and sole member of *Hepeviridae* family[25,26]. It is either non-enveloped virions found in the faeces and bile with 27 to 34 nm diameter or quasi-enveloped virions (eHEV) by host-cell-derived membranes found in bloodstream with ~40 nm diameter[2,27,28]. However, the eHEV particles do not contain glycoproteins in their surface as the other enveloped alpha-like viruses[26]. It contains a ~7.2 kb positive-sense, single-stranded RNA genome which comprises short non-coding region in both 5' and 3' ends and three open reading frames (ORFs), ORF1, ORF2 and ORF3[2]. The viral genome organization is shown in Figure 4[2].
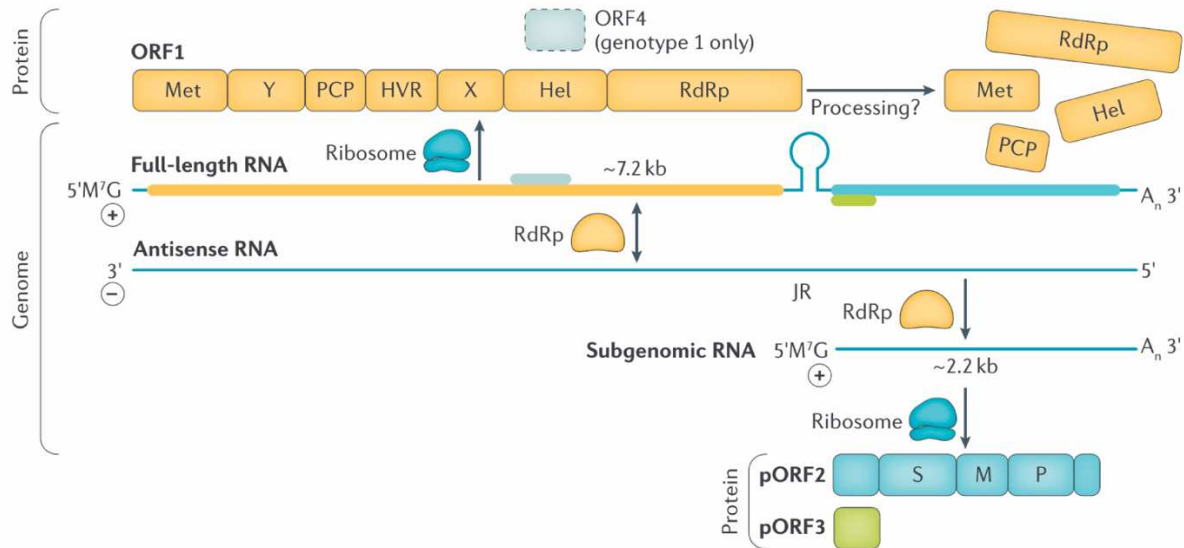
Figure 4. Genome organization of Hepatitis E virus. The non-coding and coding regions of the full-length ~7.2 kb and the sub-genomic ~2.2 kb RNA and the translated HEV proteins are shown. From Nimgaonkar et al.[2] with permission from Nature Reviews Gastroenterology & Hepatology, Copyright 2018.

The 5' untranslated region contains a 7-methylguanosine cap (m^7G) structure with length of 27 nucleotides (nt) which plays an essential role in the initiation of the viral replication and infectivity[29,30]. The 3' untranslated region is a polyadenylated (polyA) tail with 68 nt length and its U-rich region plays important role in the induction of the interferon response in the retinoic acid-inducible gene I (RIG-I) pathway through the pathogen-associated motif patterns (PAMPs) found in this non-coding region[30,31].

Regarding the three open reading frames (ORFs), ORF1 protein is expressed from the full-length ~7.2 kb RNA whereas the ORF2 and ORF3 proteins from the ~2.2 kb bicistronic subgenomic RNA[32]. ORF1 encodes a non-structural polyprotein of 1693 amino acid (aa) in total that includes multiple functional domains responsible for the replication of the viral genome[33]. Until today, it is not fully clear if the polyprotein functions as a single polyprotein or is cleaved into these domains with more favorable its proteolysis with recent study showed the detection of cleaved products in different cellular compartments[30,34]. The putative functional domains are the methyltransferase (MeT), the Y domain, the papain-like cysteine protease (PCP), the hypervariable region (HVR) that includes also a proline-rich region (Pro), the X domain, the helicase (Hel) and the RNA-dependent RNA polymerase (RdRp)[2,30,35–38]. The expression of the methyltransferase domain (MeT) in insect cells has two products, the 110 kDa protein (P110) and a 80 kDa proteolytic product, which both have guanine-7-methyltransferase and guanylyltransferase activities crucial for the viral infectivity[30]. The Y domain contains 227 amino acids (residues 216 to 442) with highly sequence similarity to rubella virus (RubV), but its function is still unclear assuming that it is an extension of the MeT domain and plays role in the viral replication and infectivity[25,35]. Next, the papain-like cysteine protease (PCP) is similar to the rubella virus (RubV) protease, a crystal structure has been recently solved (PDB ID: 6nu9), but the current data about its function(s) are controversial[30,37,39]. Regarding the hypervariable region (HVR) and the proline-rich region (Pro), there is not clear discrimination of these regions as in some studies they are considered as only HVR and in other ones as two different domains. Both regions are highly variable among the HEV genotypes in terms of the sequence and length, it seems that they play a role in the viral replication efficiency, but their function(s) have to be further investigated[29,30]. The X domain known as macro-domain belongs to the ADP-ribose-1"-monophsosphatase (Appr-1"-pase) family which catalyzes the ADP-ribose-1"-monophsosphate (App-1"-p) to ADP-ribose[26]. Previous studies have shown that it is important for the viral replication at the post-translational stage, identified as putative interferon antagonist and interacts with both MeT domain and ORF3 protein forming a viral replication complex[26,29]. The helicase (Hel) protein is a member of the helicase superfamily SF1 with motifs I and III to be crucial compared to the non-essential I, IV and VI motifs possessing NTPase (nucleoside-triphosphatase) activity and unwind RNA strands using ATP hydrolysis energy

function[30]. The RNA-dependent RNA polymerase (RdRp) is the last functional domain of ORF1 polyprotein and member of the supergroup III with high similarity to the corresponding ones of the rubella virus (RubV) and the beet necrotic yellow vein virus (BNYVV). Apart from the conserved RdRp motifs, it contains an extra Mg^{2+} binding sequence highly conserved motif (GDD) which is crucial for its activity. RdRp is vital for the viral replication as plays important role in the catalytic activity and the coordination of metal ions, synthesizes the complementary RNA strand and studies have shown that it is present in the endoplasmic reticulum (ER), a potential replication site[26,29,30].

The second open reading frame ORF2 encodes the viral capsid protein, major component of the HEV virions, that enables the entry in the cell and contains the main targets for neutralizing antibodies. ORF2 encoded protein contains 660 amino acids with a predicted molecular weight at 72 kDa[30]. At least three ORF2 capsid protein forms are detected, the infectious ORF2i, the glycosylated ORF2g and the cleaved ORF3c in the efficient HEV cell culture system described by Montpellier *et al*[40]. While only the ORF2i form is associated with infectious particles, the other two forms, ORF2g and ORF2c, are secreted glycoproteins and the most detected in infected patient sera[40]. In 2009, Yamashita *et al*. solved a 3.56 Å crystal structure of HEV virus-like particle (VLP) (PDB ID: 2ztn) which consists of 60 subunits of the truncated ORF2 protein and contains three domains, the shell (S) domain, 129-319 aa, the middle (M) domain, 320-455 aa, and the protruding (P or E2s) domain, 456-606 aa, forming icosahedral 2, 3, and 5-fold axes[41] (Figure 5). Between the M and P domain, a long proline-rich hinge (445-462 aa) is detected which provides protease resistance in the capsid and also conduces to dimer formation on the surface[41]. The capsid protein plays important role in the cell signaling while interacts with cellular proteins and is detected in various organelles, such as the ER, Golgi and nucleus[26,29,30]. Based on the structural data which are valuable for vaccine development, many efforts for a specific capsid-detected vaccine were made targeting either the ORF2 protein unit or the VLP resulting in the development of the Hecolin® (HEV 239) vaccine which is available only in China[26].
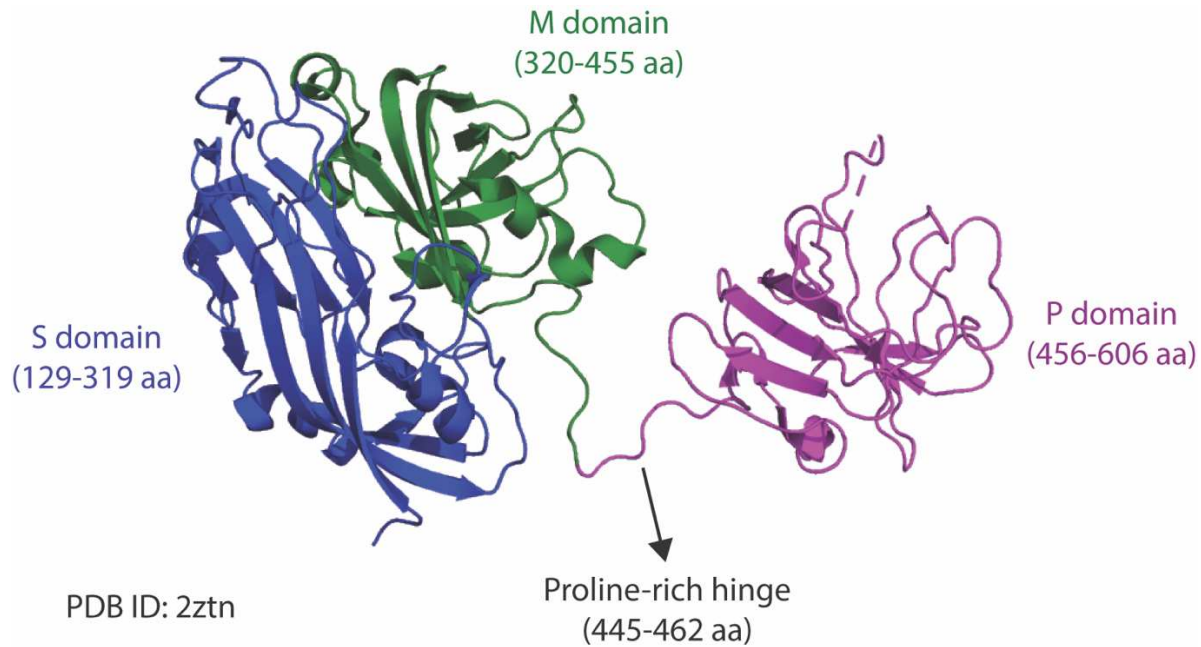
Figure 5. Crystal structure of HEV virus-like particle (VPL) (PDB ID: 2ztn)[41] at 3.56 Å resolution which consists of 60 subunits of the truncated ORF2 protein and contains three domains, the shell (S) domain (129-319 aa) in blue, the middle (M) domain (320-455 aa) in green and the protruding (P or E2s) domain (456-606 aa) in magenta. The proline-rich hinge is between the M and P domains and provides protease resistance in the viral capsid.

ORF3 encodes a small regulatory multifunctional protein which is poorly characterized. As shown in Figure 4, ORF3 is expressed from the subgenomic RNA resulting in a 113-115 amino acids protein with a molecular weight at 13 kDa[42]. Its coding sequence overlaps almost entirely with the ORF2 coding sequence (5145-5475 nucleotides) in different reading frame and therefore it constitutes the most conserved protein among the HEV genotypes[30]. Previous studies have shown that ORF3 is involved in the release of the infectious viral particles and interacts with other virus and host proteins inside the cell[43]. It is also reported ORF3 to be localized in the intracellular membranes and the plasma membrane[44,45]. More details about its function(s) are described in chapter 1.6.

Only in HEV Genotype 1, a fourth open reading frame is reported, ORF4, and encodes a ~20 kDa short-lived protein expressed only under ER stress conditions. It interacts with host proteins such as eukaryotic elongation factor 1 isoform-1 (eEF1a1) and tubulin β, but also with the Hel, X and RdRp viral proteins assembling a replication complex that increases the viral RdRp activity and thus promotes the viral replication[46].

Moreover, there are four highly-conserved across the HEV genotypes *cis*-reactive elements (CREs) important for the viral replication. The first one is located at the start of ORF1 coding region, the second one between the end of ORF2 and 3' non-coding region which binds to the RNA-dependent RNA polymerase, the third one in the junction region between ORF1 and ORF3 regions forming a stem-loop and it is the starting point of the subgenomic RNA synthesis and the last one at the end of ORF2 coding region[26,29].

1.4 Life cycle of the virus

The life cycle of Hepatitis E virus is not fully characterized until today. The entry of the virus into the cell is the starting point and the mechanism followed is distinct and needs more investigation for both non-enveloped (naked) and quasi-enveloped (eHEV) virions[47]. The information about the receptors involved in the entry of naked virions is limited with the heparan sulphate proteoglycans (HSPGs) and integrin α3 as cofactor to be the main candidates for the cellular binding[18]. Previous studies have shown that eHEV virions enters the cell through the dynamin- and clathrin-dependent, receptor-mediated endocytosis and the GTPases Ras-related proteins Rab5 and Rab7 involved in the process[29]. The lysosomal protein Niemann-Pick disease type C1 (NPC1) is then responsible for the degradation of the lipid membrane[29]. In addition, the role of the ORF3 protein molecules in the lipid-delivered membrane of eHEV virions during the entry and the uncoating process has to be further studied[18]. After the dynamin- and clathrin-mediated endocytosis, the positive-strand RNA genome is released into the cytoplasm. The translation of ORF1 protein from the full-length RNA is occurred by the host translational machinery attached in the 5' non-coding region with the eukaryotic initiation factor 4F (eIF4F) complex to be essential for the viral replication as described also in other viruses, such as cytomegalovirus (CMV), influenza virus (IFV), encephalomyocarditis virus (EMCV), Kaposi's sarcoma-associated herpesvirus (KSHV) and calicivirus (CV)[18,30,48]. The ORF1 domains are expressed and then the RNA-dependent RNA polymerase (RdRp) transcribes a complementary negative-sense RNA which is a template for the transcription of both full-length and subgenomic RNA[29]. Previous studies have shown that ORF1 protein is identified in endoplasmic reticulum membranes which is possible the viral replication site[49,50]. The subgenomic RNA is thus used for the expression of the capsid ORF2 protein and the small regulatory ORF3 protein. The final step is the release of newly formed virus in enveloped form[18]. ORF3 protein is shown that plays essential role in the secretion step while it probably binds to tumor susceptibility gene 101 protein (Tsg101), a member of the endosomal sorting complex required for transport-I (ESCRT-I), as shown in other viruses, such as human immunodeficiency virus (HIV)-1 or Ebola through a P(S/T)AP motif, as called "late domain"[51]. The eHEV virions that exit the cells contain a lipid membrane derived probably from the Golgi network which is then degraded by bile salts and therefore the HEV virions detected in the stool are

naked[2]. The eHEV virions that enters the bloodstream in the enveloped form, where it is protected from neutralizing antibodies against ORF2 and ORF3 proteins[2]. The life cycle of Hepatitis E virus with all steps described above is shown in Figure 6[2].
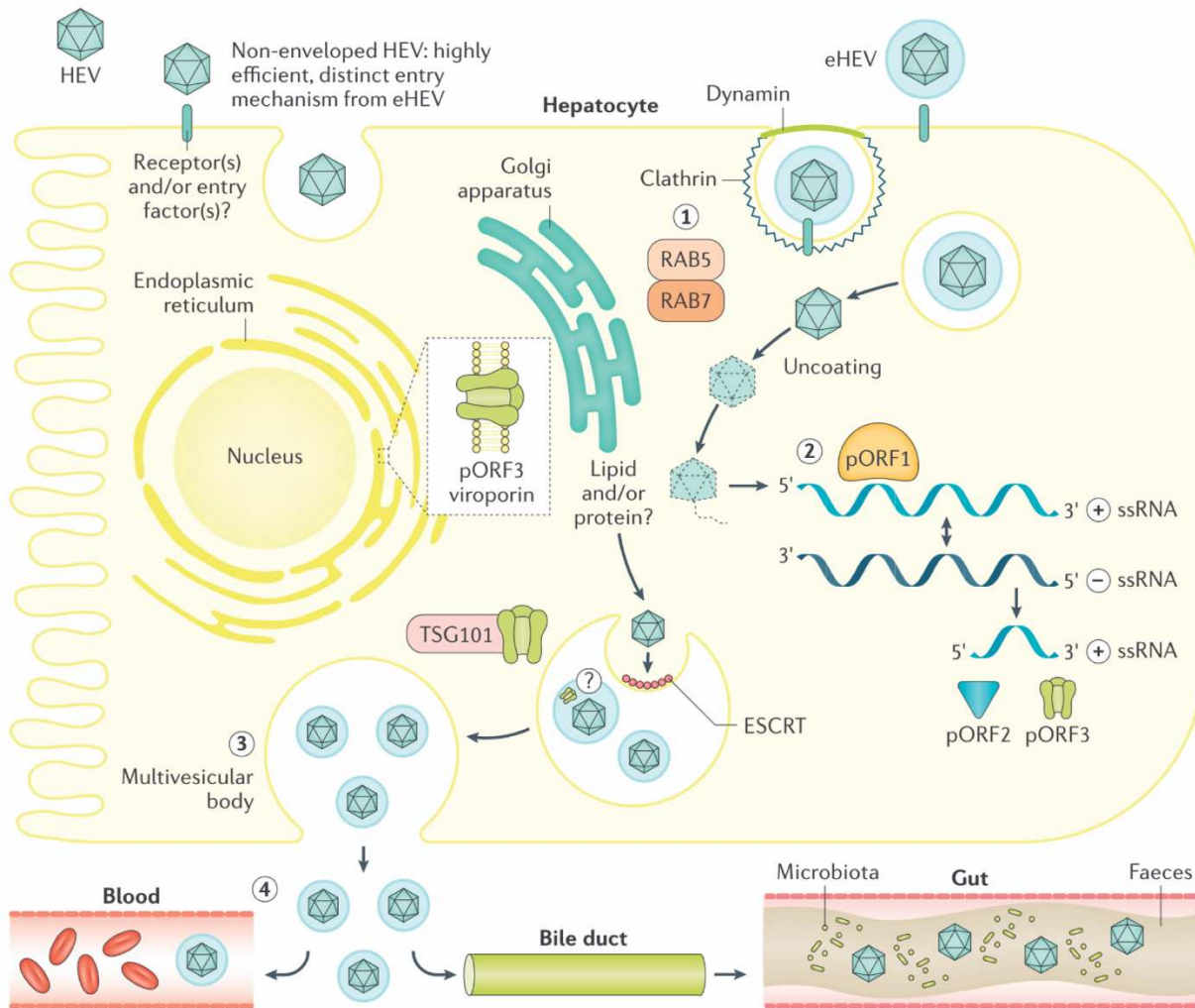
Figure 6. Life cycle of Hepatitis E virus. The entry (1) of naked and eHEV virions, the viral replication (2), the release (3) of the newly infectious virions to the bile and faeces in naked form and to the blood in quasi-enveloped form (4) are depicted. From Nimgaonkar et al.[2] with permission from Nature Reviews Gastroenterology & Hepatology, Copyright 2018.

1.5 Treatment of HEV infection

The HEV infection is mostly self-limiting, but the mortality rate, especially in pregnant women, and the endemic character of the virus mainly in developing countries has been forced to develop HEV-specific vaccines. For chronic HEV cases, the ribavirin (1-β-D-ribofuranosyl-1,2,4-triazole) treatment is approved in which the synthetic guanosine/adenosine analog with a broad antiviral spectrum targets the RdRp hindering the viral replication[26]. Recent case study on a liver transplant male patient who received a combination of ribavirin with sofosbuvir (SOF, GS-331007), an inhibitor of hepatitis C virus (HCV) polymerase, therapy, has as a result the elimination of HEV RNA in plasma during the treatment, but not the viral clearance[52]. Although more than 11 experimental preclinical studies on HEV vaccines are conducted since the virus discovery, only three studies continued to clinical trials in humans[53].

The recombinant Hecolin® p239 vaccine, developed and manufactured by Xiamen Innovax Biotech Co., Ltd., China, is licensed only in China since December 2011[54]. It is based on a 239 amino-acid peptide of the ORF2 capsid protein (residues 368-606 of ORF2 protein sequence) derived from HEV genotype 1 found in China expressed in *E.coli* system[1,55]. A large-scale phase III clinical trial in China showed a 100% efficacy and safety in 16-65 years-old participants after 3 completed doses and 96% efficacy and safety in ones after at least one dose of the Hecolin® p239 vaccine[56]. In addition, two clinical trials were carried out to assess the safety, immunogenicity and effectiveness of the vaccine, a phase I clinical trial in USA (NCT03827395)[57] and a phase IV clinical trial in rural areas of Bangladesh in pregnant women resulting in the prevention of maternal and neonatal deaths due to HEV infection (NCT02759991)[58,59]. Moreover, it is reported by World Health Organization that in March 2022, Hecolin® p239 vaccine is used for the first time in a vaccination outbreak at Bentiu IDP (internally displaced persons) camp in South Sudan's Unity state initiated by Médecins Sans Frontières and South Sudan's Ministry of Health[1].

A second recombinant HEV protein (rHEV) vaccine, developed and manufactured by GlaxoSmithKline Biologicals, Rixensart, Belgium, is based on the 56 kDa ORF2 capsid protein[53]. A phase II clinical trial in Nepal shows a 95.5% efficacy after 3 completed doses and 88.5-89.9% efficacy after the first dose (NCT00287469)[60,61].

The last HEV VLP vaccine (p179) is developed and manufactured by Changchun Institute of Biological Products Co., Ltd, China, based on genotype 4 found in China[53]. A phase I clinical trial in China showed that after 3 completed doses the HEV VLP p179 vaccine is safe and well-tolerated in 16-65 years-old participants without serious adverse side effects[53,62].

Because the recombinant Hecolin® p239 vaccine is not commercially available worldwide, the prevention against the HEV infection is important. The overall good hygiene practices on individual level and improvement in the quality of public water supplies and disposal systems are the main practices to limit the outbreak in endemic areas[63]. Especially, travelers to these areas have to be cautious, drinking bottled water and avoiding to consume raw shellfish which could be contaminated through the water[1,63].

1.6 ORF3 protein

ORF3 is a small multifunctional protein, highly conserved among the HEV strains due to the fact that its nucleotide RNA region is almost entirely overlapped with the ORF2 coding sequence by 330 nucleotides in different reading frame[30]. It is shown that ORF3 protein plays important role in the secretion step of newly formed infectious virions, but the full function(s) during the HEV life cycle remains to be elucidated.

Previous studies have shown that ORF3 protein is phosphorylated at serine residue in position 70 (Ser70) for genotype 3 and in position 71 (Ser71) for genotypes 1, 2 and 4. This is the first proline-rich domain identified that can be phosphorylated by the extracellularly regulated kinase (ERK), a member of the mitogen-activated protein kinase (MAPK) family. The serine phosphorylation probably triggers the interaction with the non-glycosylated capsid ORF2 protein which therefore plays a role in the virions assembly[64,65]. Due to the controversial results about the importance of the phosphorylation state of ORF3 protein in the viral replication, further investigation is needed[29]. ORF3 protein is also shown that interacts with microtubules, directly or through another proteins, resulting in the facilitation of HEV infection as promoting the capsid transport during egress[30,66]. Moreover, there are various studies that shown that ORF3 protein upregulates the expression of glycolytic pathway enzymes through the stabilization of hypoxia-inducible factor 1 (HIF-1) resulting the inhibition of the mitochondrial depolarization and death[67,68], but also downregulates the toll-like receptor 3 (TLR-3) mediated nuclear factor kappa B (NF-jB) signaling via tumor necrosis factor receptor 1-associated death domain protein (TRADD) and receptor-interacting protein kinase 1 (RIP1)[69].

The ORF3 protein sequence from genotype 3 contains two PXXP motifs, called also late domains, in the C-terminal region at 86-89 aa and 95-98 aa, whereas the genotypes 1, 2 and 4 contains only the second one[70]. The second PSAP motif detected in all four genotypes is known to interact with many Src homology 3 (SH3) class I domains and plays essential role in the HEV virion release[71]. Generally, proteins containing SH3 domains are important in signaling pathways involved in cell growth, differentiation and other regulatory functions[72]. Previous studies have shown that ORF3 protein interacts via [95]PSAP[98] motif with CIN85, a multidomain adaptor protein

implicated in the Cbl-mediated downregulation of receptor tyrosine kinases, resulting in the delayed degradation of the endomembrane growth factor and therefore the prolongation of the cell survival[73]. Moreover, the highly conserved "late domain" of ORF3 protein functionally interacts with the human tumor susceptibility gene 101 (Tsg101) protein, a member of the endosomal sorting complex required for transport-I (ESCRT-I), essential for viral budding[71,74]. The same kind of interaction is also shown between Tsg101 protein with "late domains" of other viruses, such as human immunodeficiency virus (HIV)-1 or Ebola, involved in the viral secretion[51,75]. The ORF3-Tsg101 interaction leads to the synthesis of the quasi-enveloped (eHEV) particles with ORF3 protein to be present in the outer lipid membrane of the infectious virions[28,76]. More details about Tsg101 protein and its interaction with ORF3 protein are presented in the chapter 2.4. Genotype 1 ORF3 protein contains a second PXXP motif at 66-75 proline-rich region and also a PMSPLR (PXXPX+, which + is either arginine or lysine) motif which is characteristic for class II SH3 domains, but with not yet known function[77–79].

Besides the host cellular proteins, based on high-throughput yeast two-hybrid (Y2H) and modified luminescence-based mammalian interactome mapping pull-down assays (LuMPIS) screening, ORF3 has found to interact with viral proteins, the ORF1 domains, methyltransferase (MeT), papain-like cysteine protease (PCP), X domain, helicase (Hel) and the RNA-dependent RNA polymerase (RdRp), suggesting a regulatory function in the formation of the replication complex[80].

As mentioned above, ORF3 protein has been localized in the intracellular membranes and the plasma membrane[44,45]. While two different groups were trying to identify its function and the involved regions, both identified ORF3 as associated to the cellular membrane, but two different modes of binding have been proposed[81,82]. In early 2017, Ding *et al.* proposed a transmembrane insertion of ORF3 protein with N-terminal and C-terminal localized in the lumen of ER and the cytoplasm, respectively, and its oligomerization that forms an ion channel and correlated the ORF3 function with a viroporin function[81]. In general, viroporin plays a vital role in promoting viral pathogenesis and especially affecting cell entry, viral replication and mainly viral release, but the mechanism is not yet known[81]. Previous studies have shown that viroporins have also found

in both enveloped (Hepatitis C virus (HCV), HIV-1, influenza A virus (IAV), rotavirus, alphavirus/Sindbis virus, coronaviruses) and non-enveloped viruses (simian virus 40 (SV40), coxsackie B virus (CBV), polio virus)[83,84]. However, this study did not include biophysical and structural data to support this model. Almost two years later, in December 2018, Gouttenoire *et al.* published a paper showing the association of ORF3 with the membrane via a post-translational modification, the palmitoylation of cysteine residues in the N-terminal region at position 1-28 aa and also the presence of both N- and C-terminus in the cytoplasm without a transmembrane insertion[82]. The palmitoylation of ORF3 protein is essential for the membrane anchoring and stability of the protein as well the subcellular localization and they proved that mutation of involved cysteines affects the viral secretion[82]. However, the cysteine region that is responsible for the anchoring of ORF3 in the membrane has to be further investigated. The results obtained by these two groups are apparently divergent regarding the topology of ORF3 protein. However, influenza virus (M2 protein) is an example which combines both modes while it has a viroporin function and it is anchored to the membrane by a palmitoylated cysteine in the C-terminal region[85]. Figure 7 illustrates the proposed modes of membrane anchoring of ORF3 protein by Ding *et al.*[81] (a) and Gouttenoire *et al.*[82] (b), but also the potential modes combining the information obtained by these studies (c).
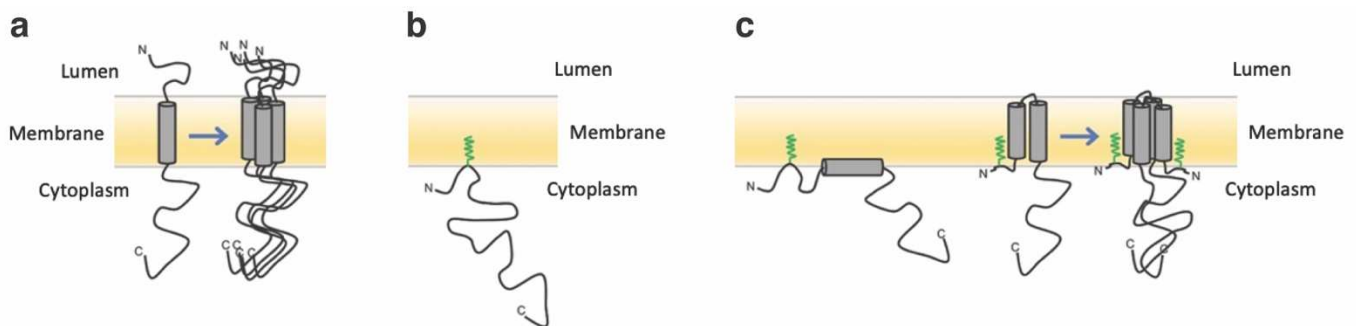
Figure 7. Membrane anchoring modes of ORF3 protein. (a) Transmembrane insertion of ORF3 protein with N-terminal and C-terminal localized in the lumen of ER and the cytoplasm, respectively, forming oligomers to constitute a viroporin, as proposed by Ding et al.[81]. (b) Membrane anchoring via a post-translational modification, the palmitoylation of cysteine residues in the N-terminal region and the presence of both N- and C-terminus in the cytoplasm without a transmembrane insertion, as proposed by Gouttenoire et al.[82]. (c) potential modes of membrane anchoring combing the information obtained by the two studies (a) and (b). Grey barrel: transmembrane region and green line: palmitoylation of cysteine residues.

In order to better clarify the function(s) of ORF3 protein, its topology, the mode of membrane anchoring and the mechanism of the viral release, further studies are required and therefore used to design antivirals to combat HEV infection.

2. Tumor susceptibility gene 101 (Tsg101) protein

2.1 ESCRT machinery

The endosomal sorting complex for transport (ESCRT) machinery was initially identified as ubiquitin-dependent protein sorting pathway of membrane proteins in yeast *Saccharomyces cerevisiae* within vacuole which is equivalent to lysosome in eukaryotes[86,87]. Until today, its functions in many biological processes have been studied and Figure 8 depicts the up-to-date ones as described by Christ *et al.*[88]. ESCRT pathway is mainly involved in the sorting, trafficking and lysosomal degradation of ubiquitinated proteins though the multivesicular bodies (MVB), as well the membrane recycling, cytokinesis, autophagy or exosome secretion[87,89]. ESCRT machinery mediates inverse membrane remodeling with the formation of vesicles which contain cytosol and bud away from it, either at cell surface or inside cellular organelles[90]. The consequences of failure to tightly regulate of cell receptors signaling by the ESCRT machinery are associated with many human diseases, such as cancers and neurodegenerative diseases[91]. The mechanism for ESCRT pathway functions have to be further investigated.
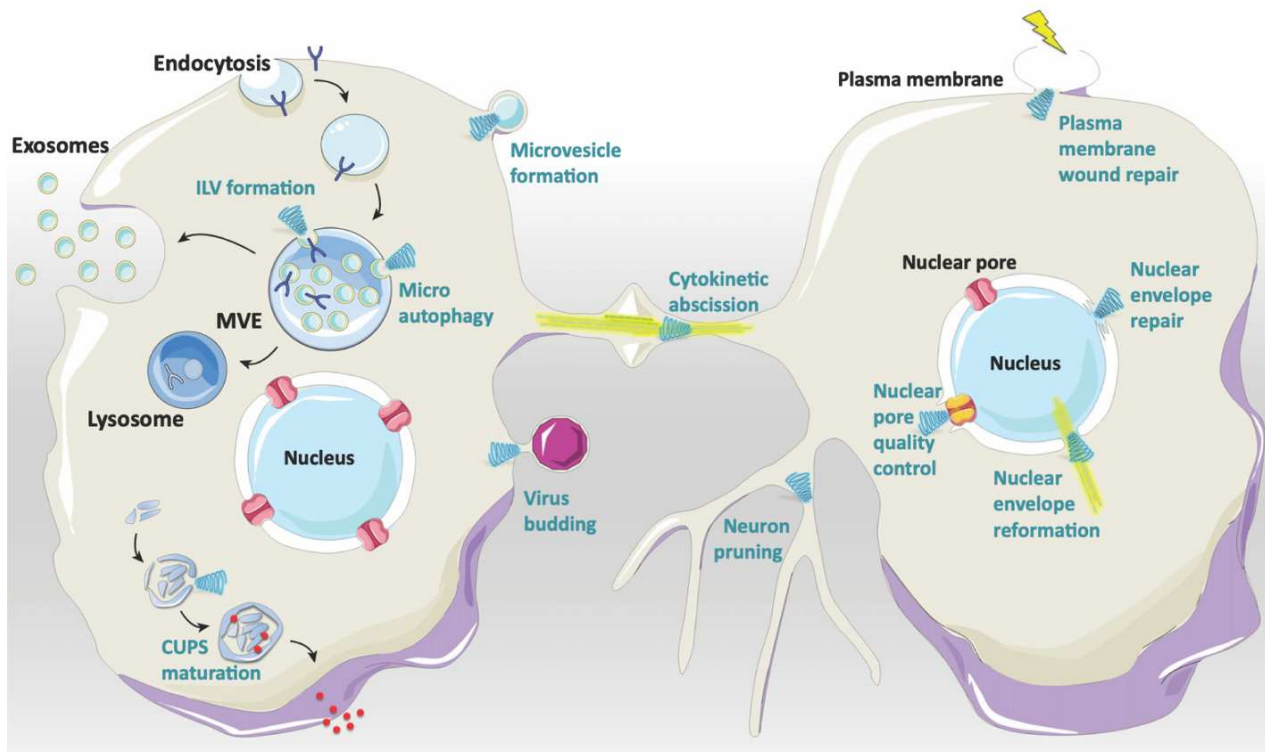
Figure 8. ESCRT pathway involved in many biological processes. From Christ et al.[88] with permission from Trends in Biochemical Sciences, Copyright 2017.

While many previous studies have focused on the HIV-1 release from the plasma membrane utilizing the ESCRT pathway, the study of other enveloped viruses' budding based on this machinery was performed[51,75,92,93].

The ESCRT machinery consists of five distinct complexes, the ESCRT-0, -I, -II, -III and Vps4 complexes, which have different functions required for MVB formation and are sequentially recruited from the cytoplasm to late endosomes[94]. ESCRT-0 is the first complex of the machinery that localizes to endosomes where it binds to ubiquitin moieties of membrane proteins intended to degrade and phosphatidylinositol 3-phosphate (PI3P) lipids initiating the MVB pathway[94]. It was not classified as a ESCRT complex, but because it was shown to recruit the ESCRT-I complex, it was finally included in the pathway[95]. ESCRT-0 comprises of two subunits, the hepatocyte growth factor-regulated tyrosine kinase substrate (HRS) (Vps27 in yeast) and the signal transducing adaptor molecule1/2 (STAM1/2) (Hse1 in yeast)[87]. These two proteins assemble a 1:1 heterodimer interacting via long coiled-coil GAT domains as shown in a 2.3 Å crystal structure solved by Ren *et al.* (Figure 9, PDB ID: 3f1i)[96]. They contain a VHS domain in N-terminal region with not yet known function and ubiquitin- and clathrin-binding domains, one found in STAM and two in HRS protein[94]. The latter one contains an extra FYVE zinc-finger domain which binds endosomal PI3P lipids and is responsible for locating the ESCRT-0 complex in endosomes[88,94]. It is also shown that the ubiquitin-binding protein, Eps15b, binds to HRS subunit of ESCRT-0 in human cells with proposed involvement in sorting epidermal growth factor receptors (EGFRs) for degradation[95]. Moreover, HRS recruits the ESCRT-I complex interacting directly with the N-terminal region of Tsg101 (Vps23 in yeast) protein, a subunit of ESCRT-I, via a P(T/S)AP motif in its C-terminal region[94]. ESCRT-0 complex is not conserved, while it is only found in fungi and animals and is not present in plants and protists in which a yet unknown system recruits ESCRT-I to initiate the MVB pathway[95].
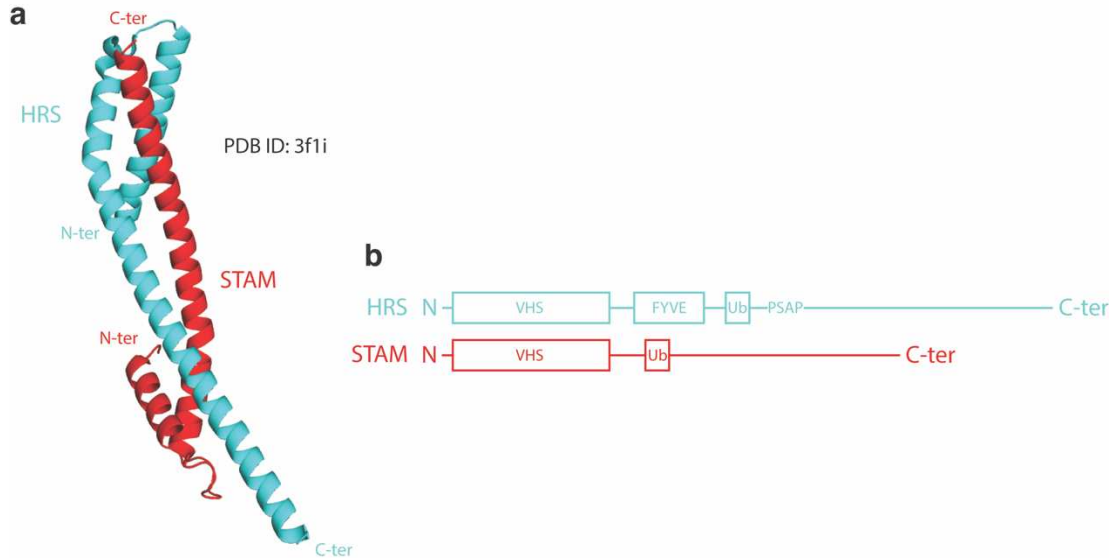
Figure 9. (a) Crystal structure of ESCRT-0 complex at 2.3 Å resolution solved by Ren at al. (PBD ID: 3f1i)[96]. (b) Main domains of HRS and STAM subunits of ESCRT-0 complex. HRS subunit in cyan and STAM subunit in red.

ESCRT-I complex is a soluble hetero-tetrameric complex containing the Tsg101 (Vps23 in yeast) protein, the Vps28 (same in yeast), the Vps37(A, B, C, D) (Vps37 in yeast) and the Mvb12(A, B) (Mvb12 in yeast) or the ubiquitin-associated protein 1 (UBAP1)[94]. More details about the subunits and the known structure of the ESCRT-I core complex are presented in the following chapter (chapter 2.2).

Then, ESCRT-I recruits the ESCRT-II Y-shaped hetero-tetrameric complex which consists of one Vps22 (same in yeast and EAP30 in mammalian) molecule, one Vps36 (same in yeast and EAP45 in mammalian) molecule and two Vps25 (same in yeast and EAP20 in mammalian) molecules[97,98]. The Vps22 and Vps36 subunits tightly interact forming the base of the Y-shaped complex while each one interacts with one Vps22 molecule corresponding to the arms of the Y[97,98]. In addition, Vps36 subunit contains a GLUE (GRAM-Like Ubiquitin-binding in EAP45) domain in the N-terminal regions which is a split pleckstrin homology domain and binds ubiquitin, PI3P lipids found in endosome membrane and also to the C-terminal domain of Vps28 of ESCRT-I complex with nanomolecular affinity in yeast[99,100]. Moreover, the Vps36 GLUE domain in yeast contains two NZF (Npl4-type zinc finger) domains, one interacts with ubiquitin and the second one with Vps28 of ESCRT-I complex, whereas the human Vps36 GLUE domain does not contain NZF domains, but still binds ubiquitin and the mammalian lacking NZF domains needs further investigation to

elucidate the interaction between ESCRT-I and ESCRT-II complexes[99]. In 2008, Im and Hurley solved the crystal structure of the human ESCRT-II complex at 2.6 Å resolution (PDB ID: 3cuq) (and also a second one at 2.9 Å resolution, PDB ID: 2zme)[101] as shown in Figure 10.
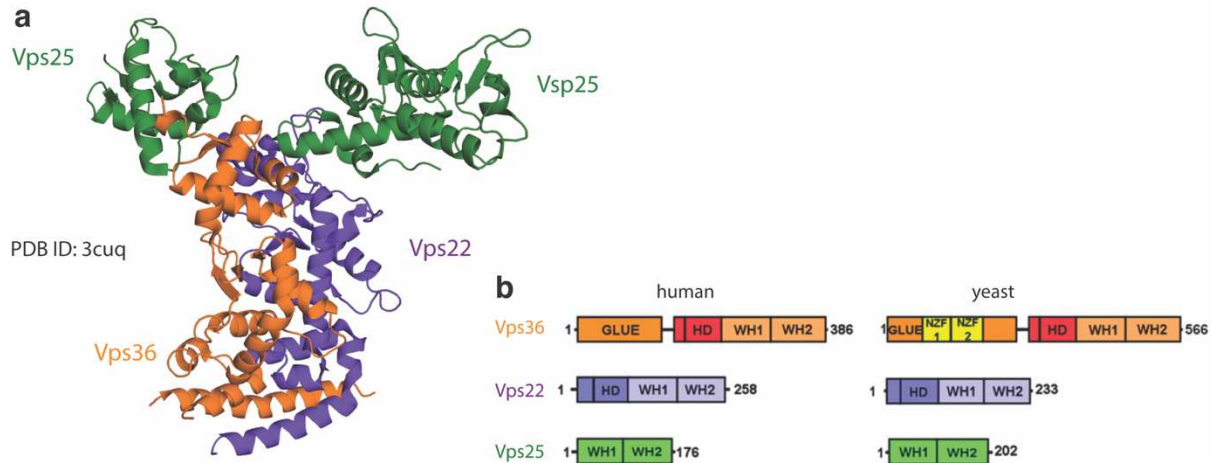
Figure 10. (a) Crystal structure of ESCRT-II complex at 2.6 Å resolution solved by Im and Hurley (PBD ID: 3cuq)[101]. (b) Main domains of Vps36, Vps22 and Vps25 subunits of ESCRT-II complex in human and in yeast. Vps36 subunit in orange, Vps36 subunit in purple and Vps36 subunit in green. Modified from Im and Hurley[101] with permission from Developmental Cell, Copyright 2008.

Next step is the recruitment of ESCRT-III complex which only transiently assembles on endosomes[102]. It is a multimeric complex consisting of four core subunits, the highly charged multivesicular body proteins (CHMP) in mammals, particularly the Vps20 (same in yeast and CHMP6 in mammalian), the Vps32(A, B, C) (Vps32 or Snf7 in yeast and CHMP4(A, B, C) in mammalian), the Vps24 (same in yeast and CHMP3 in mammalian) and the Vps2(A, B) (Vps2 in yeast and CHMP2(A, B) in mammalian) proteins, and several accessory proteins, the most studied to be the Ist1 (same in yeast), the Vps46(A, B) (Vps46 or Did2 in yeast and CHMP1(A, B) in mammalian), the Vps60 (Vps60 or Mos10 in yeast and CHMP5 in mammalian) and the Vta1 (same in yeast and LIP5 in mammalian) proteins[94,95]. The four core subunits are small proteins containing approximately ~230 aa, with similar structure, but distinct function[87,94]. Although the crystallization of entire ESCRT-III complex cannot be obtained because of its non-stable nature, crystal structures of some subunits were solved. The crystal structures of the core human CHMP3 subunit (PDB IDs: 2gd5[103], 3frt and 3frv[104]) and the human Ist1 (PDB IDs: 3frr and 3frs[104]) display that the ESCRT-III subunits share structural features as the helical fold[88]. The inactive form and autoinhibition of ESCRT-III complex is stabilized when the two helices of the negative charged C-terminal region fold back to the positively charged hairpin structure formed by two helices of the

N-terminal region in cytoplasm with the mechanism to be yet unknown[105,106]. The high-affinity binding of the Vps25 subunit of ESCRT-II to Vps20 subunit of ESCRT-III complex results to the activation of the latter to the endosomes and therefore the further binding of all its subunits[98]. Then, the Vps20 protein recruits the homo-oligomerized Vps32 (Snf7 in yeast) on the complex and the latter in turn recruits the ESCRT-III adaptor protein Alix (Apoptosis-linked gene-2 interacting protein X) (Bro1, BCK1-like resistance to osmotic shock protein-1, in yeast) that stabilizes the Vps32 into filamentous oligomers capped by Vps24 subunit and recruits the deubiquitinating enzyme Doa4[107,108]. The assembly of ESCRT-III complex is completed on the endosomes when the Vps24 recruits the Vps2 subunit which mediates recruitment of the Vps4 through the MIT (microtubule-interacting and trafficking) domains located in the C-terminal region[94,107]. Finally, the Vps4 complex consists of the type I AAA-ATPase Vps4 and its co-factor Vta1 and assembles into a stable dodecamer of two stacked hexameric rings with a central pore[94]. The Vta1 protein, which also be an accessory protein of ESCRT-III complex and interacts with Vps60, could be a potential adaptor for Vps4 and ESCRT-III interaction and upon binding the ATPase activity of Vps4 is enhanced[109,110]. Vps4 is a mechanoenzyme which is inactive promoter in monomer or dimer form in the nucleotide-free or ADP-bound state in cytoplasm and necessary for the dissociation of ESCRT-III from the membrane in the end of a MVB cargo sorting and vesicle formation[111–113]. Once ESCRT-III disassembly, the Vps4 complex is also dissociates into its inactive protomers until the next assembly process[87]. Figure 11 illustrates the assembly of ESCRT machinery with the components of each complex (ESCRT-0, -I, -II, -III and Vps4) and depicted interactions between the subunits[88].
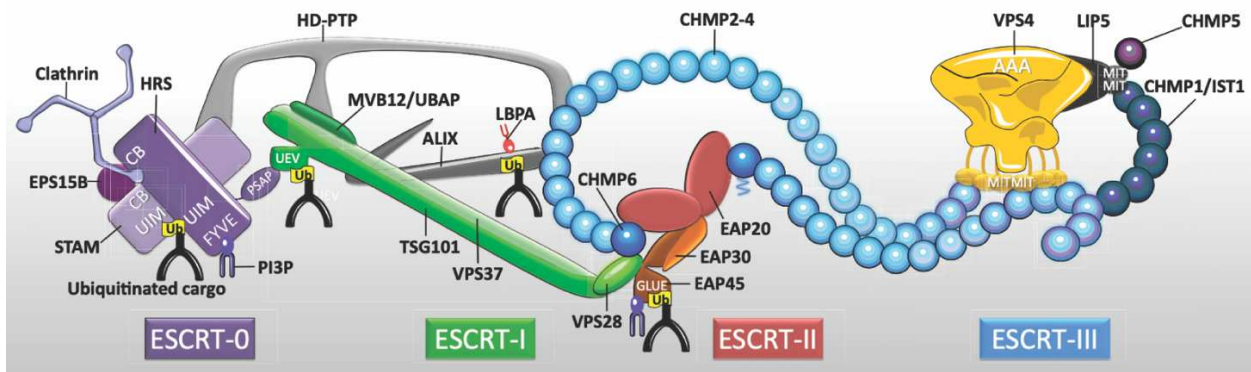
Figure 11. Assembly of ESCRT machinery with the components of each complex (ESCRT-0, -I, -II, -III and Vps4) and depicted interactions between the subunits. From Christ et al.[88] with permission from Trends in Biochemical Sciences, Copyright 2017.

2.2 ESCRT-I complex

As mentioned in the previous chapter, ESCRT-I is a hetero-tetrameric complex recruited by ESCRT-0 essential for the cargo sorting. It is the first ESCRT complex identified in yeast initially consisting of only Vps23, Vps28 and Vps37 subunits while the Mvb12 protein was later determined to stabilize the complex[114,115]. In mammalian cells, the existence of multiple isoforms of Vps37 and Mvb12 subunits could be related to the tissue in which the ESCRT machinery is utilized[87,116]. The crystal structure of yeast ESCRT-I complex reveals that one copy of each subunit assembled in an approximately 18 nm long complex containing a headpiece necessary for binding with ESCRT-II, a rigid 13-nm long coiled-coil stalk required for ESCRT-I cargo sorting and an endpiece in the other end of the complex necessary for binding with ESCRT-0[117]. Figure 12 depicts the 2.7 Å crystal structure of yeast ESCRT-I complex solved by Kostelansky *et al.* (PBD ID: 2p22)[117].
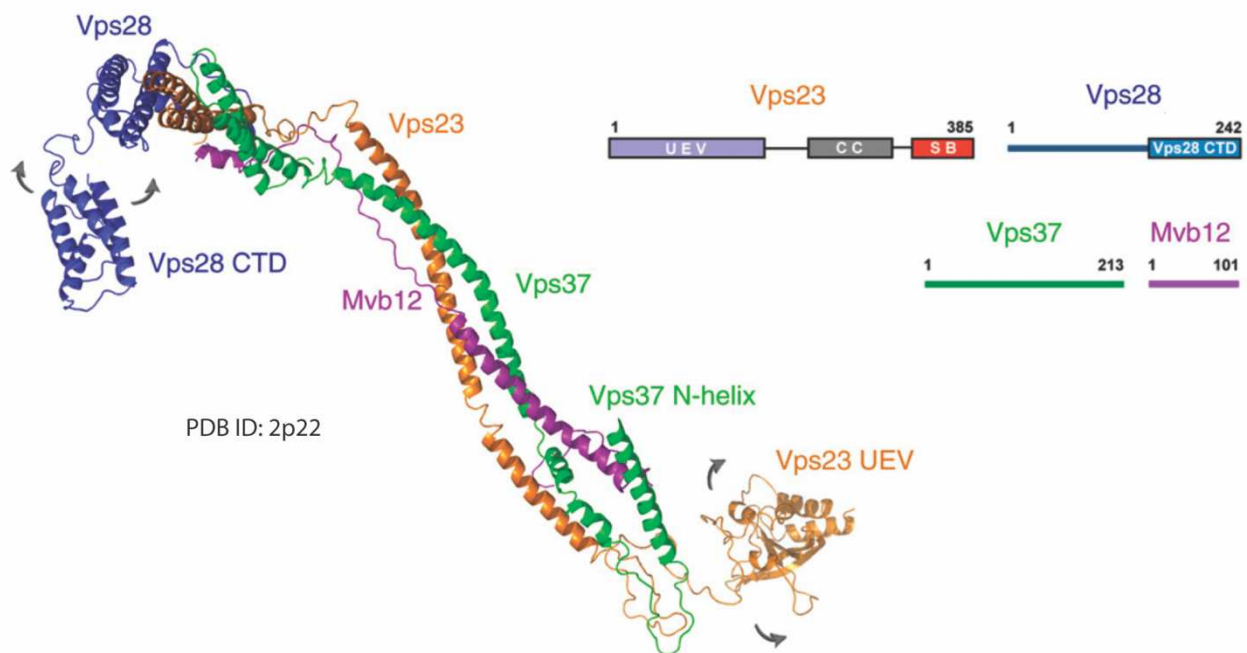
Figure 12. Crystal structure of yeast ESCRT-I complex at 2.7 Å resolution solved by Kostelansky et al. (PBD ID: 2p22)[117] and the main domains of the four subunits, Vps23 in orange, Vps28 in blue, Vps37 in green and Mvb12 in magenta. The grey arrows indicate the dynamic motions of the domains which binds to other ESCRTs. Modified from Kostelansky et al.[117] with permission from Cell, Copyright 2007.

Based on the solved structure of the yeast ESCRT-I, the Vps23 (the C-terminal region), Vps37 and Mvb12 subunits interact to each other in order to form the Stalk. All these subunits are important in the complex formation, but the two first proteins to be needed for the assembly of the

headpiece[117]. In addition, the Vps23 (Tsg101 in human) contains a ubiquitin E2 variant domain (UEV) in the N-terminal region which binds the P(T/S)AP motif in the C-terminal region of the HRS protein, a subunit of ESCRT-0 complex[87,117]. The absence of ESCRT-0 complex results the failing in the recruitment of ESCRT-I complex on the endosomal membranes[95]. The Vps23-HRS interaction is similar to the interaction of Tsg101 with the Gag protein of human immunodeficiency virus-1 as described previously, important for virus budding[118]. In the other end of the complex, in the headpiece, the Vps28 subunit interacts with the GLUE domain of Vps36 subunit of ESCRT-II complex resulting the recruiting of the latter one[87,117]. The UEV domain of Vps23 and the C-terminal domain of Vps28 showed to be dynamic and transit from inactive to active open conformation in which the binding with other subunits is occurred[117]. Regarding the UBAP1 protein, it binds ubiquitin as the Mvb12 subunit on endosomes, and is more specific for MVB sorting while the ESCRT-I with Mvb12 subunit is involved in the HIV virus budding[94]. The ESCRT-I and ESCRT-II complexes create a stable rigid structure responsible for budding membranes into the lumen of unilamellar vesicles, but also stabilizing the bud neck of a growing vesicle[94]. The ubiquitination of the ESCRT-I substrate Cps1, the binding of GLUE domain of Vps36 of ESCRT-II complex to the membrane embedded PI3P lipids and the interaction of Vps37 of ESCRT-I complex with acidic phospholipids have shown that the ESCRT-I located parallel to the membranes with the three positions for the membrane attachment[117]. In Figure 13, the structural model (a) and schematic representation (b) of the yeast ESCRT-I complex located on an endosomal membrane and the interaction with other ESCRT complexes are shown[117].
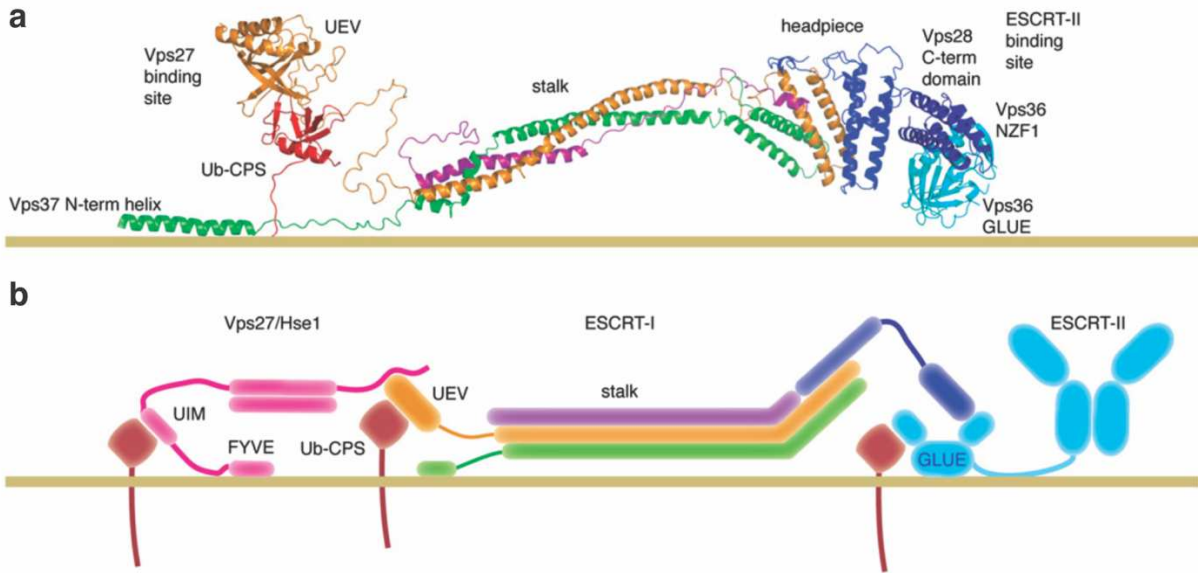
Figure 13. (a) Structural model of the yeast ESCRT-I complex located on an endosomal membrane and interaction with subunits of ESCRT-0 and ESCRT-II complexes. The ESCRT-I subunits Vps23 in orange, Vps28 in blue, Vps37 in green and Mvb12 in magenta (PDB ID: 2p22)[117] as well the GLUE domain (PDB ID: 2cay)[99] and NZF1 domain (PDB ID: 2j9u)[100] of ESCRT-II in cyan are shown. (b) Schematic representation of yeast ESCRT-I complex in interaction with ESCRT complexes and with the membrane in the three previously described positions in red. Modified from Kostelansky et al.[117] with permission from Cell, Copyright 2007.

Previous studies have shown that all ESCRT complexes are essential, but ESCRT-I and ESCRT-III complexes are involved in the viral maturation and release. Many enveloped RNA viruses, such as HIV-1[75,93], Ebola[51], human T-lymphotropic virus[119] or Hepatitis E virus (HEV)[70], hijack the ESCRT pathway to bud away from the plasma membrane of infected cells, a process that is similar to vesicles budding into intracellular organelles from the cytosol. This study focuses in the mechanism of the virus release and specifically the key role of human Tsg101 protein, the ESCRT-I subunit, in this process.

2.3 Tsg101 protein – structure

Regarding the ESCRT-I complex, the tumor susceptibility gene 101 (Tsg101) protein (Vps23 in yeast) constitutes a central subunit which binds ubiquitinated cargo proteins and then manages their sorting into endosomes. Apart from the role in the ESCRT machinery, Tsg101 is reported to play an important role in the formation and release of extracellular vesicles as it is found in exosomes where their purification process is occured[120,121]. In addition, previous studies proposed that Tsg101 could be an oncoprotein in certain cell types, but experimental data showing that its abnormal function results in genesis and progression of cancer are missing[91]. In general, Tsg101 protein is involved in many intracellular processes with the role in endosomal trafficking of cargo proteins to be the most studied.

In 1995, Tsg101 protein was initially identified as a stathmin downstream regulator through coiled-coil interactions using a yeast two-hybrid screen assay[122]. A year later, Li and Cohen cloned, expressed and proposed a potential tumor suppressor function of the full-length Tsg101 protein using a controlled homozygous functional knockout assay in mammalian cells[123].

The human Tsg101 protein contains 390 amino acids with a predicted molecular weight approximately at 44 kDa and expressed in all tissues and cell types[124,125]. Comparing the human protein sequence with the one derived from rat and the mouse one, the amino-acid sequence is predicted to have 94% and 99% similarity, respectively, using the multiple sequence alignment tool Clustal Omega[126,127]. Tsg101 protein is consists of four conserved domains with distinct function and structure, the N-terminal ubiquitin E2 variant (UEV) domain, followed by a proline-rich region (PRR), a Stalk domain (coiled-coil, CC) and a C-terminal Head domain (α-helical/steadiness box, SB) (Figure 14)[90,124].
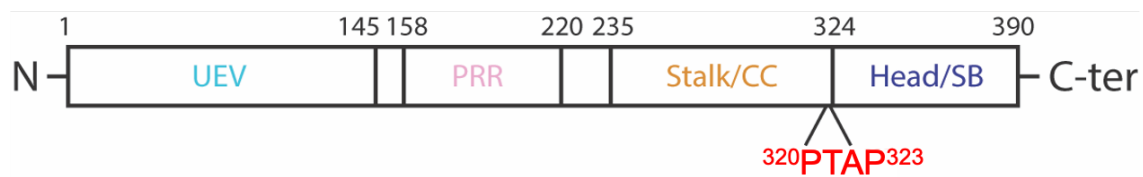
Figure 14. Domains of the human Tsg101 protein. The N-terminal ubiquitin E2 variant (UEV) domain, the proline-rich region (PRR), the Stalk/coiled-coil (Stalk/CC) domain and the Head/α-helical-steadiness box (Head/SB) domain are presented.

Figure 15 depicts the predicted AlphaFold2 structure for full-length human Tsg101 protein in which apart from the disordered PRR domain, the other domains have high prediction scores[128].
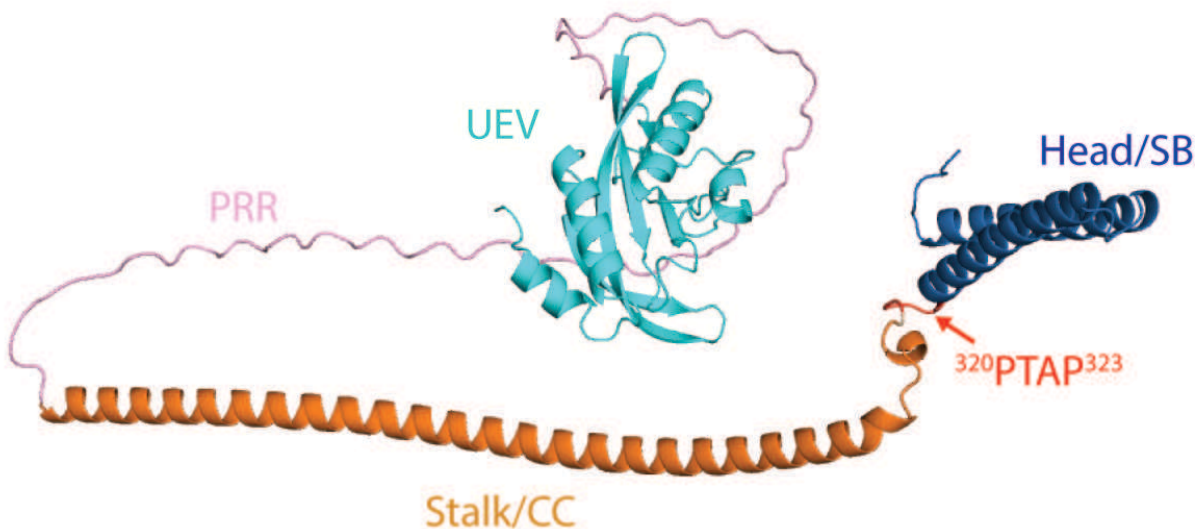
Figure 15. Predicted AlphaFold2 structure for full-length human Tsg101 protein[128]. The N-terminal UEV domain is in cyan, the disordered proline-rich (PRR) region in pink, the Stalk (Coiled-coil) (Stalk/CC) domain in orange, the PTAP motif in red and the Head (α-helical-steadiness box) (Head/SB) domain in blue.

The functions and the structure of N-terminal UEV domain will be described in detail in the following chapter (chapter 2.4).

Next to the UEV region, there is a proline-rich region of approximately 70 amino acids long in which 30% of its residues are prolines and usually found in surface proteins and transcription factors[91,124]. Due to the high percentage of proline residues, it is predicted to be disordered as also shown in the AlphaFold2 structure in Figure 15. Previous studies have shown that the [154]QATGPPNTSYMPG[166] sequence of the PRR region competes with the corresponding proline-rich sequence of ESCRT-III adaptor protein Alix (Apoptosis-linked gene-2 interacting protein X) for binding to the mid-body protein CEP55A hinge region required for cell abscission during cytokinesis[129–132]. Moreover, the PRR region also interacts directly with the ALG-2 (apoptosis-linked gene 2) protein, which is a dimeric Ca^{2+} binding penta-EF-hand (PEF) protein, but the function of this interaction remains unclear[133,134].

As previously mentioned, the crystal structure of ESCRT-I complex has shown that the C-terminal region of Tsg101 protein, the Stalk and Head domains, interacts with the ESCRT-I Vps37 and

Mvb12 subunits to form the a triple-coiled-helices core structure important for the stability of the complex[117,135]. The Stalk or coiled-coil domain was the initial region identified to bind the cytosolic phosphoprotein stathmin by using a yeast two-hybrid screen assay[122]. It is also shown that it acts as transcriptional suppressor for estrogen receptor and other nuclear hormone receptors[136]. Moreover, Tsg101 protein forms an efficient repressive transcription complex with Daxx, a Fas interacting protein and transcription regulator, through its coiled-coil domain, that is co-localized in the nucleus. Studies have associated the deletion of the coiled-coil domain with Burkitt lymphomas and non-Hodgkin's lymphomas diseases[137,138]. Between the Stalk and Head domain, a PTAP motif in position 320-323 aa is proposed that modulates the binding of various proteins to the Tsg101 UEV domain[91,139].

Finally, the C-terminal domain, the Head or α-helical/steadiness box (SB) domain, in combination with the N-terminal region of Vps28 and the C-terminal region of Vps37 subunits form the headpiece of ESCRT-I complex[117]. Looking closer to the structure of SB domain, two long antiparallel α helices form a hairpin in the end of the protein[124]. Although the three subunits that comprise the ESCRT-I headpiece do not share sequence similarity, they have similar structure features[140]. Not only the SB domain is important for the stability of ESCRT-I complex together with the coiled-coil domain, but also previous studies have shown that it plays essential role for post-translational autoregulation of steady-state levels of Tsg101 protein in murine and human cells[141].

2.4 Tsg101 UEV domain

In the N-terminal region of Tsg101 protein, the ubiquitin E2 variant (UEV) domain contains the first 145 amino acid of the protein sequence with a predicted molecular weight approximately at 16.6 kDa[142]. Although it is homologous to ubiquitin E2 ligases family and binds ubiquitin, the catalytic cysteine is replaced by tyrosine residue and therefore it is lacking the enzymatic activity required for the transfer of ubiquitin[142]. The UEV-ubiquitin interaction plays important role in the sorting ubiquitinated cargo in the MVB pathway as well the virus release process[124]. In addition, previous studies have shown that Tsg101 has an autoregulate function by interaction of this UEV domain with its PTAP motif located between the Stalk and Head domains[139,143].

The Tsg101 UEV domain is the most structurally characterized among Tsg101 domains as many NMR and crystal structures are solved in apo- and bound-state[142,144]. In 2002, Pornillos *et al.* revealed the first NMR structure of Tsg101 UEV domain in apo-state[142]. Three years later, the first crystal structure in free state confirmed the E2 fold structure of the domain consisting of four α helices packed against one side of four stranded antiparallel β-sheet (PDB ID: 2f0r) (Figure 16)[144].
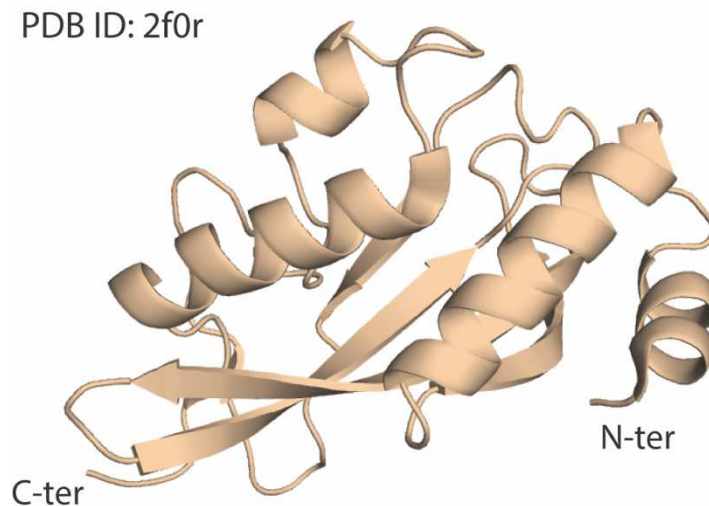
Figure 16. Crystal structure of Tsg101 UEV domain in apo-state in 2.26 Å resolution consisting of four α helices packed against one side of four stranded antiparallel β-sheet (PDB ID: 2f0r)[144].

The high resolution (at 2 Å) crystal structure of UEV domain in complex with ubiquitin (PDB ID: 1s1q) reveals the concave binding site located opposite to the β-sheet as shown in Figure 17[145].

Comparing the crystal structure of UEV domain in apo-state and bound to a ubiquitin molecule, the root-mean-square deviation (RMSD) at 2 Å proves that they are extremely similar[144,145].
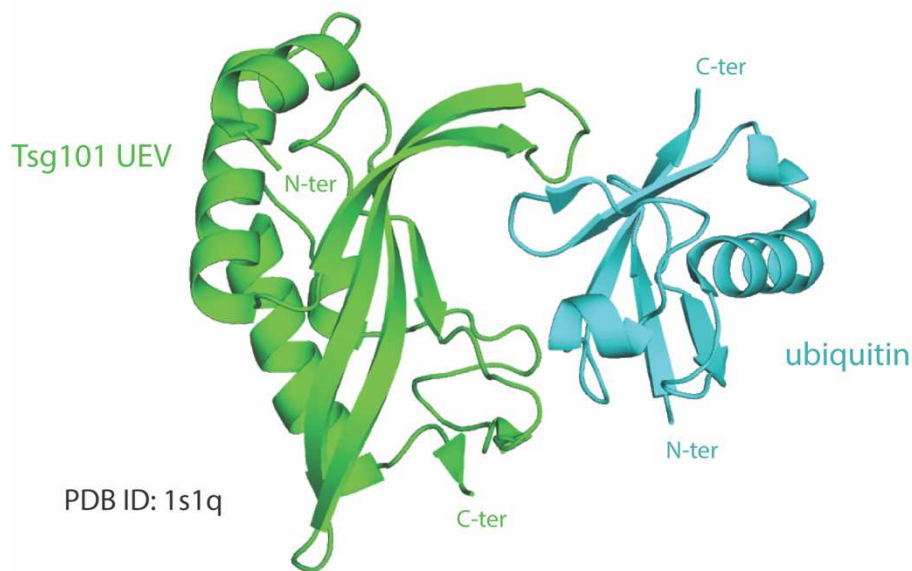
Figure 17. Crystal structure of Tsg101 UEV domain (in green) in complex with ubiquitin (in cyan) at 2 Å resolution (PDB ID: 1s1q)[145].

Next to the ubiquitin binding pocket, there is a hydrophobic groove at the protein surface through which it interacts with P(T/S)AP motifs found in cellular and viral proteins, such as regulatory proteins for intracellular trafficking and retroviral proteins[118,142,145,146]. NMR and crystal structures of Tsg101 UEV domain in complex with P(T/S)AP peptide sequences are also solved[146,147]. The Tsg101 UEV domain interacts with HRS subunit of ESCRT-0 complex via its PSAP motif and therefore the ESCRT-I recruitment occurs[96,147]. A crystal structure of Tsg101 UEV domain with a nine-residue human HRS PSAP peptide at high resolution (at 1.4 Å) provides more insight into the interaction[147]. In addition, previous studies have shown that the ESCRT-I complex is involved in the virus budding and especially the Tsg101 UEV domain interacts with the viral "late domains", P(T/S)AP, to promote their release from the cell[146,148]. Specifically, peptides containing the late motifs from viral HIV-1 Gag, Ebola VP40 and MARV NP proteins derived from HIV-1, Ebola and Marburg virus (MARV) viruses, respectively, bind to the Tsg101 resulting the efficient transfer and budding of infectious viral particles binding[146,147,149]. NMR (DYANA and CNS ensembles, PDB IDs: 1m4p and 1m4q, respectively)[146] and crystal structures (PDB ID: 3obu, at 1.6 Å)[147] of Tsg101 UEV domain in complex with HIV-1 PTAP peptides and structurally-modified-

PTAP-derived peptides (FA459 and FA258, PDB IDs: 3p9g at 1.8 Å and 3p9h at 1.8 Å, respectively)[150] as well with Ebola PTAP peptide (PDB ID: 4eje, 2.2 Å) are available (Figure 18).
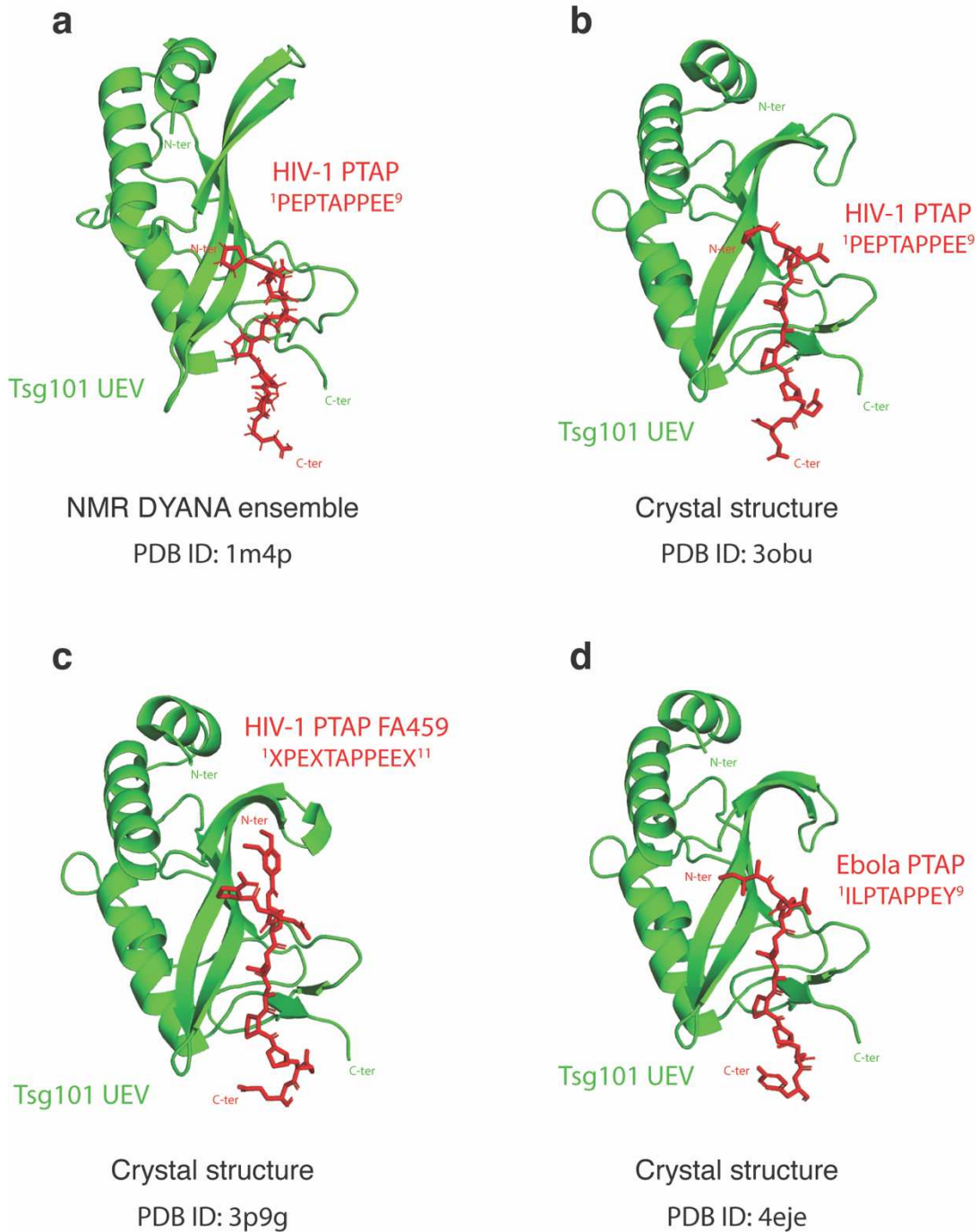
Figure 18. NMR and crystal structures of Tsg101 UEV domain (in green) in complex with viral late domain PTAP peptides (in red). (a) NMR DYANA ensemble with HIV-1 PTAP peptide (PDB ID: 1m4p)[146]. (b) Crystal structure with HIV-1 PTAP peptide at 1.6 Å resolution (PDB ID: 3obu)[147]. (c) Crystal structure with HIV-1 structurally-modified-PTAP-derived peptide FA459 at 1.8 Å resolution (PDB ID: 3p9g)[150]. (d) Crystal structure with Ebola PTAP peptide at 2.2 Å resolution (PDB ID: 4eje).

In these solved structures, the peptides induce conformational changes on Tsg101 structure and also comparing the crystal structure of Tsg101 UEV domain in complex with ubiquitin molecule, it seems that the two binding sites are independent and therefore Tsg101 UEV could simultaneously interact with both ubiquitin and PTAP motif from the same or different protein partners (Figure 19).
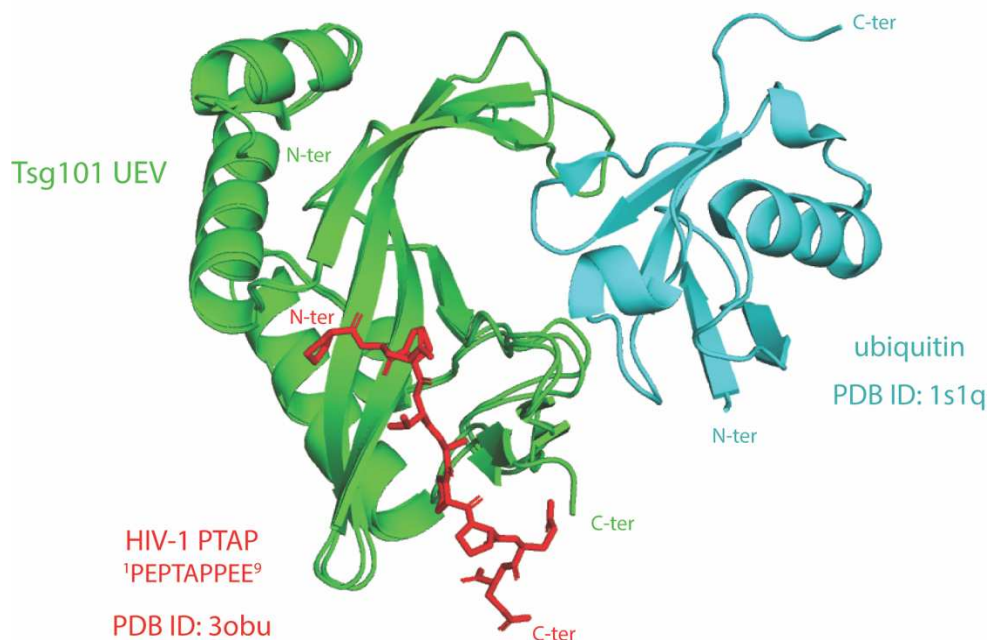
Figure 19. Alignment of crystal structures of human Tsg101 UEV domain (in green) in complex with ubiquitin (in cyan) (PDB ID: 1s1q)[145] and with HIV-1 PTAP peptide (in red) (PDB ID: 3obu)[147] showing the two binding sites.

This second binding site in the protein surface has already constituted an attractive drug target to develop antivirals against enveloped viruses. Studies based on peptidomimetics have shown that designed cyclic peptides and small molecules could inhibit the interaction between the human Tsg101 UEV domain and HIV-1 Gag protein and thus, impede the HIV particles budding from the cell[151–153].

In 2017, Strickland et al. published that two prazole-based compounds, esomeprazole (referred as F15) and tenatoprazole (referred as N16), bind to the Tsg101 UEV domain in the ubiquitin binding pocket disrupting the interaction with ubiquitin and its function, but not the PTAP binding, resulting in interference with the early HIV-1 assembly[154]. They also reported an NMR

structure of Tsg101 UEV domain in complex with tenatoprazole which could be used for further antiviral target compound development (Figure 20, PDB ID: 5vkg)[154].
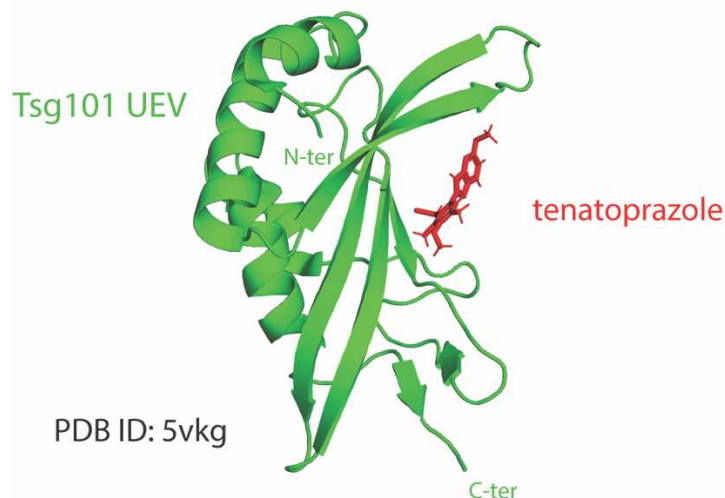
Figure 20. NMR structure of Tsg101 UEV domain (in green) in complex with tenatoprazole (in red) (PDB ID: 5vkg)[154].

Moreover, a recent study showed that tenatoprazole and the more potent ilaprazole, a related prazole compound, could not only block the HIV-1 particles release, but also the Herpes Simplex Virus (HSV) 1/2 from infected Vero cells in culture suggesting that these prazole-based compounds could be good candidates for further development of potential drugs[155].

Therefore, both binding sites of Tsg101 UEV domain, the ubiquitin pocket and the surface groove, could be used for antiviral drug discovery targeting the enveloped virus budding and which could be used for multiple viral infections. This study focuses on the interaction of Tsg101 UEV domain with HEV ORF3 protein important for the release of infectious viral particles.

3. Nuclear Magnetic Resonance (NMR) Spectroscopy

3.1 General

In 1938, the nuclear magnetic resonance phenomenon was initially reported by Rabi *et al.* while they were measuring the nuclear magnetic moment in molecular beams and thus, he received the Nobel Prize in Physics in 1944 *"for his resonance method for recording the magnetic properties of atomic nuclei"*[156–158]. In early 1946, two research groups working independently, the group of Edward M. Purcell, Howard C. Torrey and Richard V. Pound in Massachusetts Institute of Technology in USA and the group of Felix Bloch, William W. Hansen and Martin Packard in Stanford University in USA, were able to further develop and use the nuclear magnetic resonance on liquids and solids[159,160]. Edward M. Purcell and Felix Bloch were awarded and shared the Nobel Prize in Physics in 1952 *"for their development of new methods for nuclear magnetic precision measurements and discoveries in connection therewith"*[161]. After mid-1950s the nuclear magnetic resonance spectroscopy was also used in chemistry and commercial instruments, which were very primitive compared to today's instruments, were available[158]. Since then, the development and the application in chemical and biological studies have been expeditious with milestones, the development of the Fourier transform pulse sequence by Richard R. Ernst and Weston A. Anderson in 1966 and then the conception of multidimensional NMR spectroscopy originated by Jean L. C. Jeener introducing the 2D NMR spectroscopy in 1971 and further developed by Richard R. Ernst during the mid-1970s[158,162]. Three more Nobel Prizes have been awarded related to the development of NMR Spectroscopy in distinct fields until today. Richard R. Ernst received the Nobel Prize in Chemistry in 1991 *"for his contributions to the development of the methodology of high resolution nuclear magnetic resonance (NMR) spectroscopy"*[163], Kurt Wuthrich the Nobel Prize in Chemistry in 2002 *"for his development of nuclear magnetic resonance spectroscopy for determining the three-dimensional structure of biological macromolecules in solution"*[164] and Paul C. Lauterbur and Sir Peter Mansfield the Nobel Prize in Physiology or Medicine in 2003 *"for their discoveries concerning magnetic resonance imaging"*[165].

Nowadays, NMR Spectroscopy consists a powerful tool in multiple fields and especially in structural biology while it is used for structural characterization from small organic and inorganic compounds to biological macromolecules. Focusing on the determination of the structure of the biomolecules, NMR Spectroscopy, X-ray crystallography and 3D electron microscopy are the available techniques to provide atomic resolution structures and they have different advantages and disadvantages. NMR Spectroscopy provides the structural and dynamical parameters needed to calculate the three-dimensional structure of the protein. In this study, the solution-state NMR Spectroscopy is used for the characterization of the proteins of interest. The main limitation of this technique is the size of the studied macromolecule which cannot be more than 30 kDa. In some cases, this can be overcome due to the development of the labeling methods, for example the segmental labeling in which only specific domains of the protein are labeled, and the recorded advanced experiments. In addition, it does not require the crystallization of the protein as happened in X-ray crystallography and thus, it can be also used to study highly dynamic systems. The 3D electron microscopy is more favorable used for large complexes, more than one protein system or proteins with higher than 150 kDa molecular weight, and requires small amount of sample (about 0.1 mg) compared to solution-state NMR spectroscopy in which the concentration of the sample depends on the protein size and the available magnetic field. Moreover, using the solution-state NMR Spectrometry, the interaction of the biomolecule of interest with small compounds/ligands or other macromolecule can be detected directly and therefore the binding interface as well in many cases the binding affinity can be easily determined. The latter could not be determined with X-ray crystallography and 3D electron microscopy techniques, although the interaction can be observed if the crystallization of the protein complex and high signal-to-noise ratio images, respectively, are obtained.

NMR spectroscopy like other forms of spectroscopy observes the effect of energy transfer occurred in the atomic nuclei by an external magnetic field in radiofrequency. The transition from the ground to the excited state causes a change in the nuclear spin (I), an intrinsic property of all electrically charged nuclei. Not all the nuclei could be detected in NMR Spectrometry, but only the ones possess net spin ($I \neq 0$), such as hydrogen (^1H) which has spin equal to ½ and constitutes the most natural abundant isotope (99.98%). Regarding the proteins and their atomic

composition, the most abundant isotopes of carbon (^{12}C, 98.90%), nitrogen (^{14}N, 99.64%) and oxygen (^{16}O, 99.76%) are not detectable by NMR[166]. The NMR-active isotopes for carbon and nitrogen, which have spin equal to ½, are not naturally abundant with ^{15}N isotope to be found about 0.37% and ^{13}C isotope about 1.11%[166]. Therefore, the studied proteins have to be enriched with these isotopes during their production in the selected expression system, which could be *Escherichia coli* as used in this study, eukaryotic cells, cell-free systems or synthetic systems.

The active nuclei in the protein sample behave as magnetic dipoles that are arranged in random manner in the ground state without the application of a static magnetic field B_0 while the spins get aligned during magnetization process, parallel and antiparallel to the direction of the magnetic field. When the spin returns to the ground state, the energy emitted as radiofrequency produces a characteristic signal named free induction decay (FID) which is subsequently Fourier transformed into an NMR signal of the corresponding nucleus called chemical shift[167].

Figure 21 depicts the main components of a liquid-state NMR spectrometer, the superconducting magnet which generates the strong magnetic field and nowadays could be up to 1.2 GHz field, a probe in which the sample is placed and is mainly a cryoprobe for enhancing the sensitivity, and a complex electronic system controlled by a workstation where the experiments' set up is occurred[168]. In this study, two different NMR spectrometers, both equipped with distinct cryoprobe, are used, a 900 MHz spectrometer with 21.1 tesla field strength and a 600 MHz spectrometer with 14 tesla field strength.
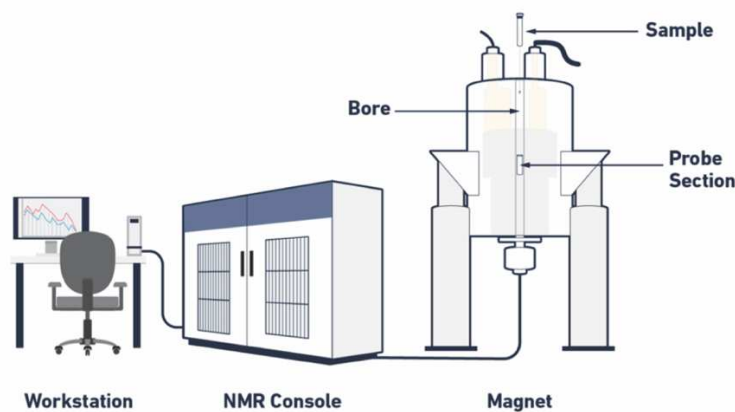
Figure 21. Main components of a liquid-state NMR spectrometer including the superconducting magnet, a probe, an NMR console and the workstation. (Available at http://www.technologynetworks.com/analysis/articles/nmr-spectroscopy-principles-interpreting-an-nmr-spectrum-and-common-problems-355891)[168].

3.2 Solution-state NMR of disordered proteins vs folded proteins

Biomolecular solution-state NMR Spectroscopy as the other biophysical techniques mentioned above could be used in order to determine the three-dimensional structure of the proteins. Based on the central dogma of structural biology, the proteins are required to adapt a stable structure which determines their function(s). In 1999, this theory is overturned by Wright and Dyson who first presented the existence of disordered proteins and therefore the structural characterization of an intrinsically unstructured protein which folds upon the further interaction with another protein partner by NMR spectroscopy[169]. Since then, the study of disordered protein or protein regions, named as intrinsically disordered proteins (IDPs) or intrinsically disordered protein regions (IDRs), respectively, has increased dramatically. The further understanding was essential while they play vital role in cellular signaling and regulatory networks as well many studies have shown that their abnormal regulation is associated with diverse diseases. Previous studies have shown that in human proteome the estimated percentage of intrinsically disordered proteins are close to 32% and the corresponding one of intrinsically disordered protein regions close to 19% and therefore together gives an overall percentage of 51%, half of the human proteins[170]. Last year, Kumar *et al.* performed an extended bioinformatics analysis on 6,108 viral proteomes which showed that the broad variability of the content of IDPs/IDRs in these proteomes. They categorized the studied 283,000 viral proteins based on their IDRs content and physicochemical properties and observed that DNA virus-encoded proteins contain more IDRs than RNA ones[171]. Because of their biological significance in various steps of cellular life and the ongoing discovery of more of their functions, the intrinsically disordered proteins constitute potential drug targets, while different designed small molecules have shown to interact with them[172].

Therefore, apart from the well-folded proteins with the stable three-dimensional structure over time, the intrinsically disordered proteins or protein regions exist as dynamic ensemble of conformations lacking of stable secondary or tertiary structure over time. Figure 22 illustrates the difference of a well-folded (globular) protein in panel (a) with an intrinsically disordered protein in panel (b).
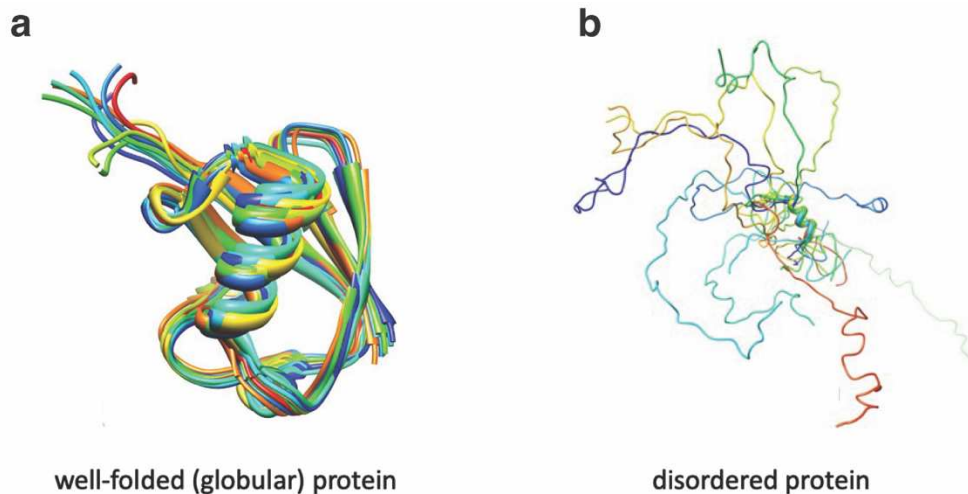
Figure 22. Three-dimensional structure of a well-folded (globular) protein over time (a) compared to dynamic ensemble of conformations of an intrinsically disordered protein (b).

The simplest NMR spectrum recorded first is the 1D ^{1}H spectrum in which all the protons are detectable. Because of the protein is in aqueous sample, the pulse sequence of the experiment included a water suppression step to eliminate the proton signals from the water molecules (proton chemical shift at 4.7 ppm). Although the 1D spectrum is not used for the structural determination, it contains useful qualitative information about the nature of the protein, if it is well-folded or disordered, an estimation of the concentration of the sample depending the signal heights and the buffer composition, but also it can be used for detection of any changes on the protein during the recording time. Figure 23 depicts an 1D ^{1}H NMR spectrum of globular ubiquitin molecule and the specific regions of different types of protons found in the protein[173]. The methyl groups (-CH$_3$) and the aliphatic (-CH$_2$) protons are the most shielded with chemical shifts around 1 ppm and up to ~3.5 ppm, respectively. From 3.5 to 5.5 ppm of 1D spectrum, the Hα protons are detected and also the water protons resonate at 4.7 ppm which could cover the Hα peaks in case of insufficient water suppression. The region of 6 to 11 ppm is the most important for protein samples because the protons of side chains of glutamine, asparagine and tryptophan residues, the aromatic protons and the backbone amide protons (H^N) are located in this region.
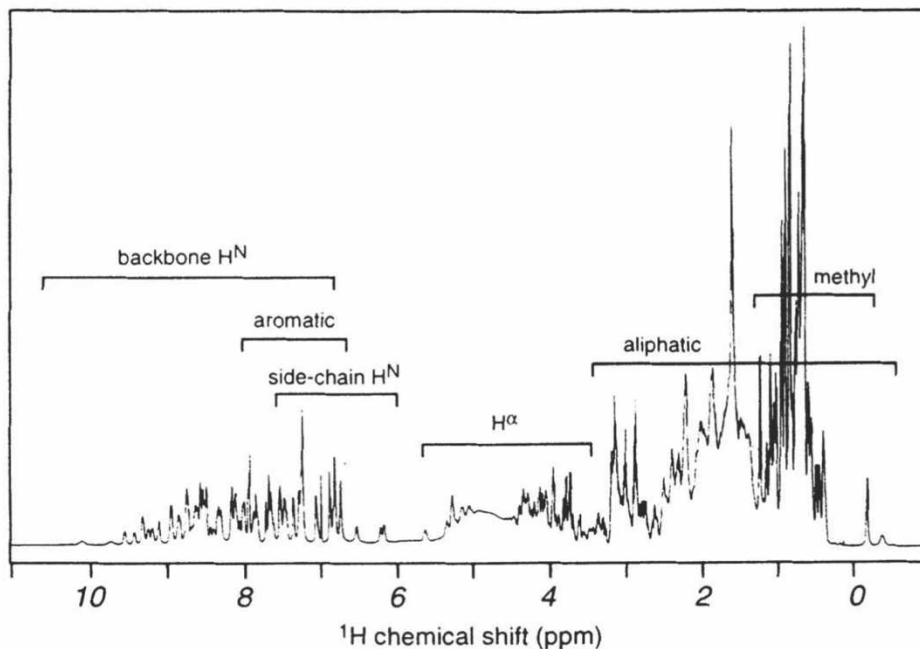
Figure 23. 1D ¹H NMR spectrum of globular ubiquitin with notated regions of different types of protons found in protein. From Schanda Paul, Copyright 2007[173].

Looking in the 6 to 11 ppm on the 1D ¹H NMR spectrum of the studied protein, the nature of the protein could be determined. The well-folded proteins have dispersed peaks within this region whereas the intrinsically disordered proteins have narrow dispersion and all the protons are overlapped in ~6.5 to 8.5 ppm region as shown in Figure 24[174].
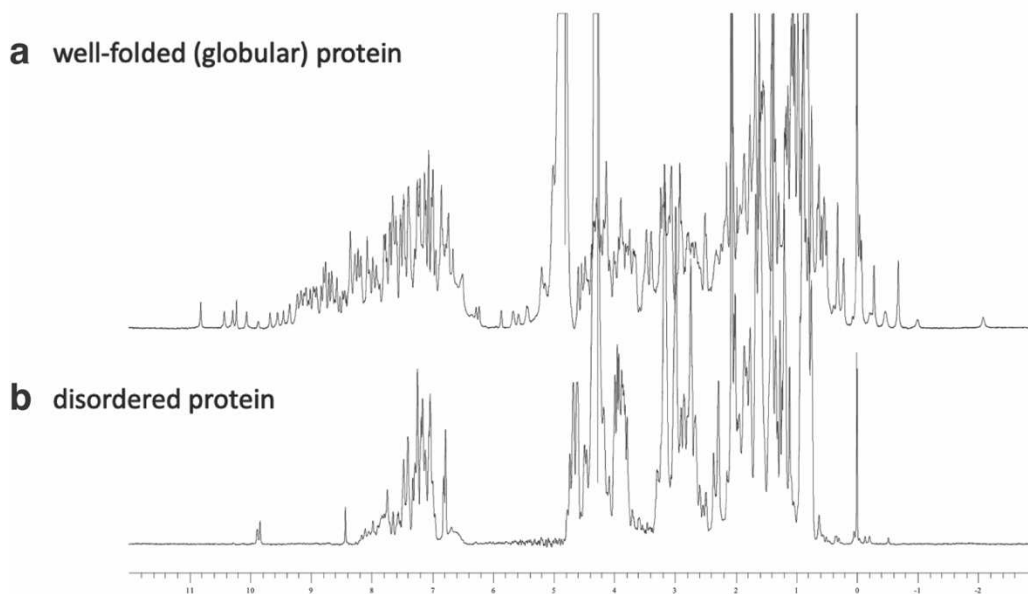
Figure 24. Comparison of (a) 1D ¹H NMR spectrum of well-folded (globular) protein with (b) 1D ¹H NMR spectrum of intrinsically disordered protein. (Available at https://www.slideserve.com/garren/nmr-spectroscopy)[174]

A 1D ^{1}H NMR spectrum can be recorded for any protein sample without enrichment with NMR-active isotopes. The protein sample volume and concentration vary depending the NMR tube used and the size of the protein, respectively. The sample volume varies from 250 μL to 500 μL while the protein concentration limit for small proteins (~10 kDa) could be ~70 μM which is increasing with the increasing of the protein size. For only recording 1D ^{1}H NMR spectrum, the protein does not have to be highly concentrated due to the sensitivity of the spectrum and even lower concentration could be reasonable.

For recording the multi-dimensional NMR experiments and gain structural information, the protein sample has to be labeled with nitrogen (^{15}N) and carbon (^{13}C) active isotopes.

The most common and basic 2D heteronuclear NMR spectrum is the 2D ^{1}H, ^{15}N HSQC (Heteronuclear Single Quantum Coherence) spectrum. It is considered the "fingerprint" of the protein because it is unique. It provides the correlation between the nitrogen (^{15}N) and the amide proton (^{1}H) of the protein backbone as well the H-N signals from the side chains of glutamine, asparagine, tryptophan and less often arginine are also detected. Each resonance or peak of the HSQC spectrum corresponds to a residue of the protein sequence except of prolines. As in 1D ^{1}H NMR spectrum, the 2D ^{1}H, ^{15}N HSQC spectrum have characteristics regions where the glycine residues can be found on the top of the spectrum, the side chains of glutamine and asparagine residues in pairs, the side chains of tryptophan on the left down corner and also the last residue of protein has the highest intensity on the spectrum (Figure 25)[175].
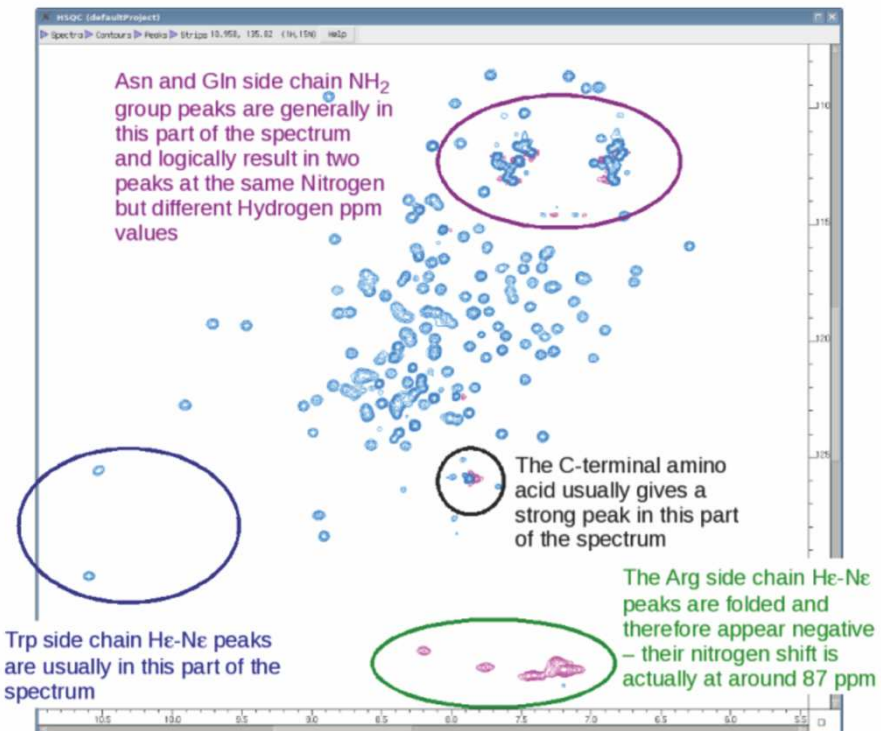
Figure 25. 2D ^1H, ^{15}N HSQC spectrum of folded protein with notated regions for side chains of glutamine, asparagine, tryptophan and arginine and the region where the last residue is located. From Protein NMR[175], available at https://www.protein-nmr.org.uk/solution-nmr/spectrum-descriptions/1h-15n-hsqc/.

The HSQC spectrum in Figure 25 is a typical spectrum obtained for a well-folded protein as the peaks are dispersed and the overlap is limited, a factor that also depends on the size of the protein and the frequency of the spectrometer. The corresponding spectrum for an intrinsically disordered protein is narrow dispersed in the ~6.5 - 8.5 ppm proton dimension as also observed in the 1D spectrum and the peaks are overlapped and it is difficult to clearly distinguish the resonances. The comparison of the 2D HSQC spectra of the two kind of proteins is shown in Figure 26[176].
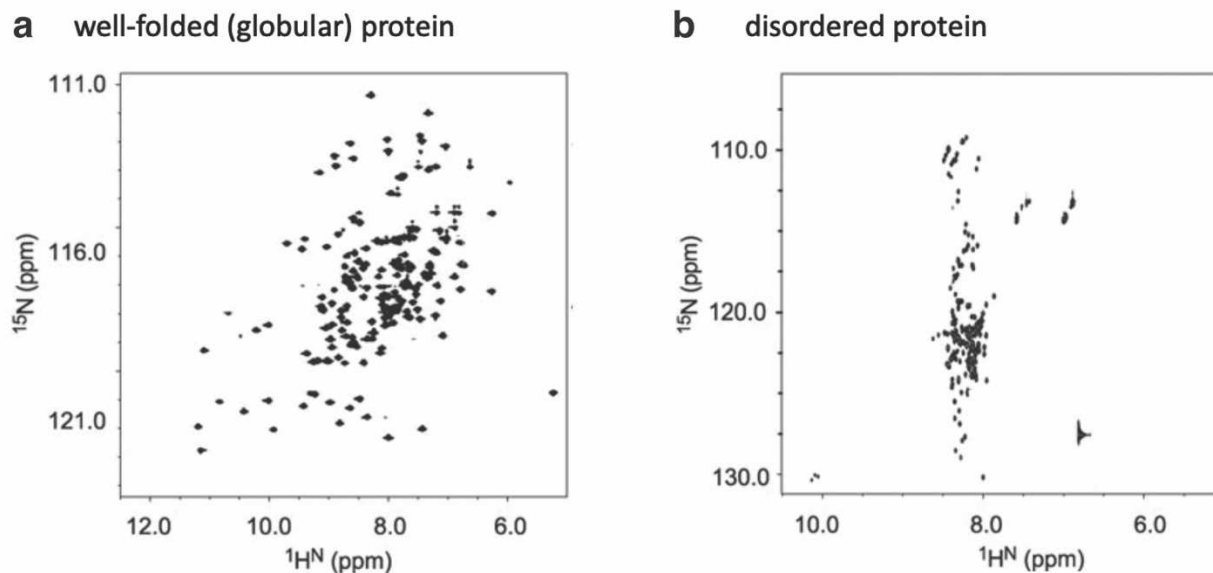
Figure 26. Comparison of (a) 2D ^1H, ^{15}N HSQC spectrum of well-folded (globular) protein with (b) 2D ^1H, ^{15}N HSQC spectrum of intrinsically disordered protein. From Breukels et al.[176], with permission from Current Protocols in Protein Science, Copyright 2011 John Wiley & Sons, Inc.

In order to acquire a high quality 2D HSQC spectrum, the protein has to be concentrated (for 10 kDa proteins to be around ~100 μM while for bigger proteins has to be higher) and pure from other proteins enriched with only ^{15}N isotope as well its size is decisive for the experimental time, while bigger proteins require more scans and thus, longer acquisition time.

This spectrum is used in combination with various 3D NMR spectra in order to make the link between every observed peak with an amino acid of the protein sequence, a procedure called "assignment" of the protein which will be further described in the next chapter. In addition, 2D HSQC spectra are acquired between 3D experiments in order to observe any changes and check the stability and the quality of the sample throughout the recording time.

3.3 NMR assignments and NMR titration protein – protein interaction

The assignment procedure is the first step for the structural characterization of the protein using NMR spectroscopy. In order to record 3D ^{1}H, ^{15}N, ^{13}C NMR spectra, the protein has to be enriched with both nitrogen (^{15}N) and carbon (^{13}C) isotopes. Combining the information for backbone atoms from the different triple resonance spectra, the connection of the peaks of 2D HSQC spectrum with the amino acids of the protein sequence is achieved. This sequentially linkage of the backbone atoms results in each peak corresponds to a probe for further structural study. The recording time of the 3D experiments is longer than the time required for acquiring a 2D HSQC spectrum and varies from couple of hours to few days as well the sensitivity of the 3D experiments is overall lower than the 2D ones. Due to the long duration of dataset recording, the protein has to be stable for several days, otherwise the precipitation of the protein has as a result the limited information and consequently the assignment procedure is harder or even impossible.

The main 3D experiments necessary for the backbone assignment procedure are the ^{1}H, ^{15}N, ^{13}C HNCO, HNCACO, HNCACB and HN(CO)CACB. Figure 27 depicts the strategy for the sequential assignment of ^{13}Cα and ^{13}Cβ resonances based on the 3D HNCACB and HN(CO)CACB spectra. The 3D HNCACB spectrum provides the correlation of each NH group with Cα and Cβ chemical shifts of the same residue (more intense peaks) and of the preceding one while the 3D HN(CO)CACB spectrum provides only the correlation of the NH group with Cα and Cβ chemical shifts of the preceding residue. Combining the information of these two spectra, a long chain of connected NH groups based on the Cα and Cβ resonances can be obtained. The next step is their assignment on specific amino acid in the protein sequence achieved based on characteristic chemical shifts of specific residues, such as Glycine has only Cα resonance at ~45 ppm and the Cβ of Alanine, Serine and Threonine residues are at ~20 ppm, ~65 ppm and ~70 ppm, respectively.
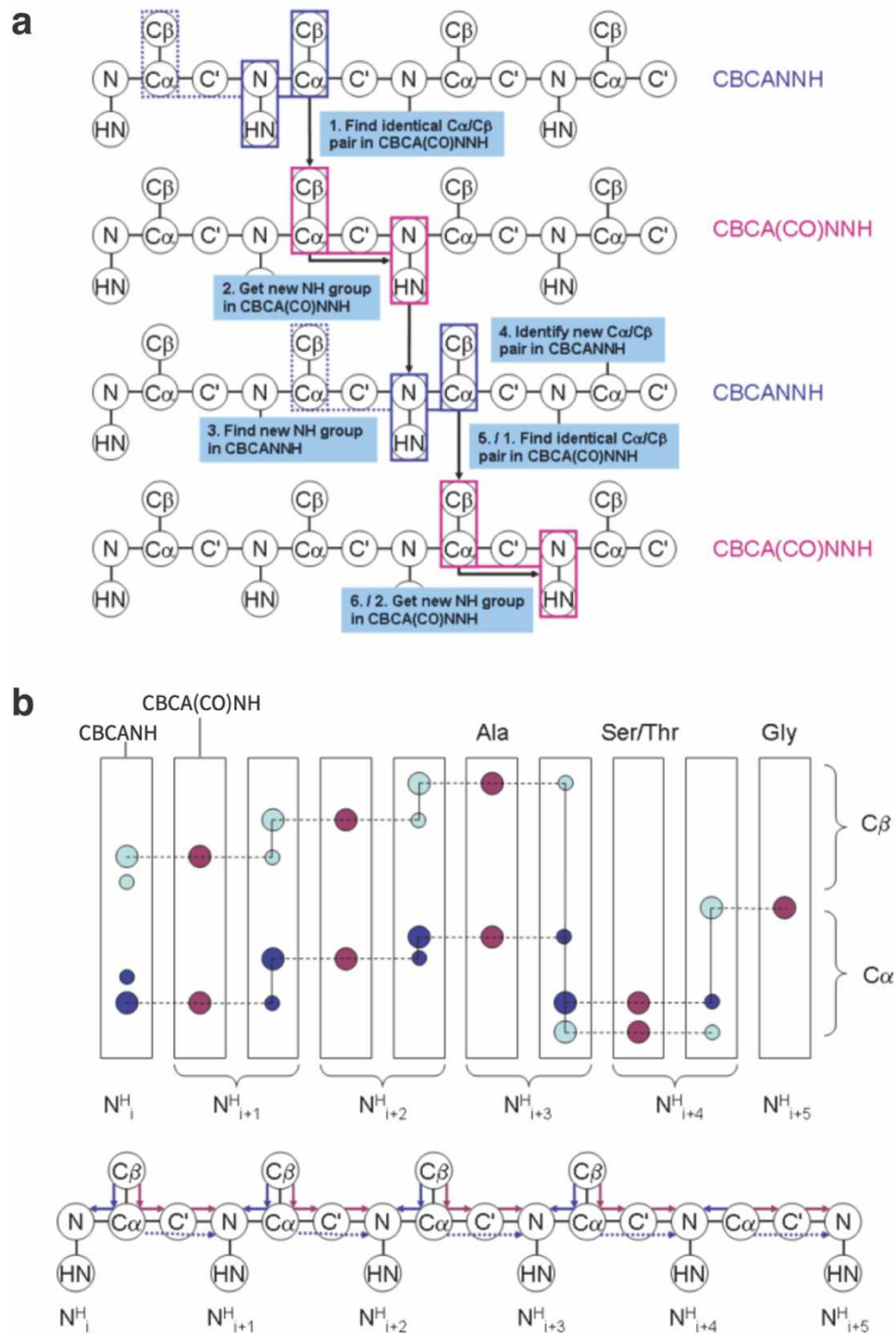
Figure 27. Backbone assignment strategy based on 3D ^1H, ^{15}N, ^{13}C HNCACB and HN(CO)CACB spectra. (a) Steps followed for sequentially linking of Cα and Cβ resonances of NH groups. (b) Representation of the long chain of connected NH groups based on the Cα and Cβ resonances obtained from 3D HNCACB and HN(CO)CACB spectra. The Cα resonances are shown in blue and the Cβ in cyan and the more intense peaks corresponds to the resonances of the same residue on HNCACB spectrum. The Cα and Cβ peaks of the preceding residue of HN(CO)CACB spectrum are shown in magenta. Characteristic chemical shifts of specific residues, such as Glycine, Alanine, Serine and Threonine residues, are shown. From Protein NMR[177], available at https://www.protein-nmr.org.uk/solution-nmr/assignment-theory/triple-resonance-backbone-assignment/.

Analyzing these 3D spectra and connecting every peak of the 2D HSQC spectrum, the $^1H^N$, $^1H\alpha$, ^{15}N, $^{13}C\alpha$, $^{13}C\beta$ and ^{13}CO resonances are successfully assigned for non-Proline residues. As mentioned above, the proline residues are not detectable in the 2D ^1H,^{15}N HSQC spectrum due to the absence of an amide proton. In order to assign the prolines of the protein sequence, specific NMR experiments are recorded based on the carbon detection. The most common 2D carbon detection spectrum is the 2D ^{15}N, ^{13}C NCO spectrum which provides the correlation between the amide (N_i) of the same residue and the backbone carbonyl (CO_{i-1}) of the preceding residue. In this experiment, the proline residues are visible in ^{15}N region ~132-142 ppm and combing with either the 3D HNCO or HNCACO experiment, the proline residues could be successfully assigned (Figure 28).
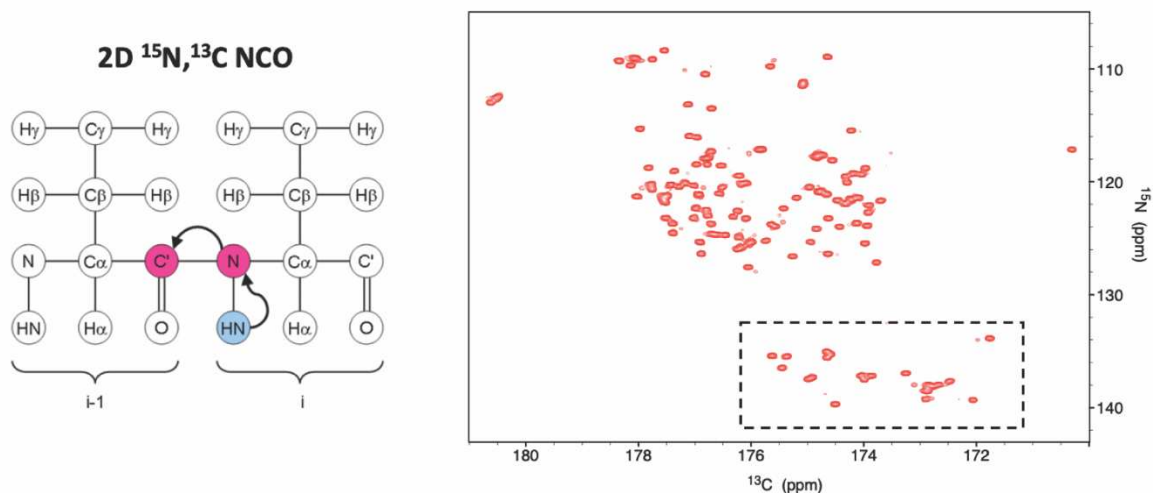
Figure 28. 2D carbon detection ^{15}N, ^{13}C NCO spectrum. (left) The magnetization pathway performed[178] and (right) 2D ^{15}N, ^{13}C NCO spectrum with the proline residues visible in ^{15}N region ~132-142 ppm (shown in dashed box). From Protein NMR[178], available at https://www.protein-nmr.org.uk/solid-state-mas-nmr/spectrum-descriptions/nco/.

The obtained backbone assignments of the protein or in other words the connection of all the peaks on the 2D HSQC spectrum with the amino acids of the protein sequence has as a result each peak corresponds to a probe for further study.

The interaction of an assigned protein with a protein partner or a small compound/molecule or any other interacting partner, such as RNA, can be studied by NMR spectroscopy and especially with NMR titration experiments. The protein of interest labeled either only with ^{15}N precursor or with both ^{15}N and ^{13}C precursors is in constant concentration and the protein partner or the small

compound is unlabeled and added in increasing concentration while 2D HSQC spectra and 2D carbon detection NCO spectra, in case of the double labeled sample for the proline residues, are recorded for each concentration point. The two parameters monitored during the titration experiments are the peak shifting and the peak broadening. The interaction of the two proteins induces chemical shift perturbations (CSP). Analyzing the CSP of all residues along the protein sequence, the peaks with the highest values correspond to the residues involved in the interaction or the ones that undergo conformational changes upon binding. Having the 3D structure of the studied protein and combining the CSP data, the binding site and/or the conformational changes aside the binding site can be determined.

The exchange rate between the two protein states, the free and the bound state, could be slow, intermediate or fast (Figure 29). In the slow exchange rate, two peaks corresponding to the free and bound state of a residue are observed during the titration, the peak of the free state gradually disappears, its intensity is decreasing, while the bound one appears (Figure 29a). In the fast exchange rate, the affected peak is shifted smoothly from the free to bound state position, which corresponds to the weighted average of the chemical shifts of two states in each titration point (Figure 29c). In the intermediate exchange rate, the peak signal broadens and shifts simultaneously (Figure 29b). In NMR experiments, the slow exchange usually corresponds to tight-binding interactions due to the long-lasting interaction that does not dissociate during the recording time, whereas the fast exchange to weak-binding interactions due to the rapid interconversion between the two states.
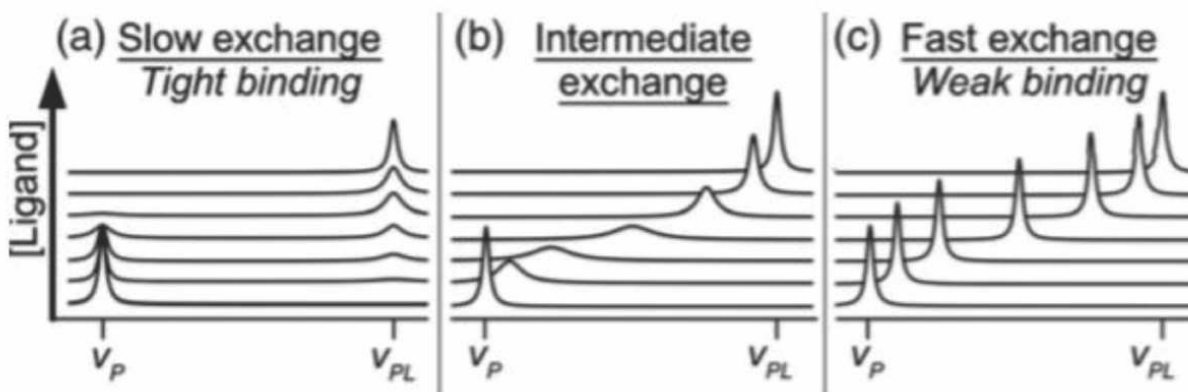
Figure 29. Exchange rates between two proteins could be (a) slow, (b) intermediate or (c) fast. The peaks for the free (v_P) and bound (v_{PL}) state in each case are observed. Modified from Kleckner and Foster[179], with permission from Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, Copyright 2011.

Moreover, the dissociation constant could be also determined analyzing the CSP values of the affected peaks. In Figure 30, the affected residue L83 is shifted from the free (black) to the bound (green) state position and by plotting the CSP values against the molar ratio of ligand to protein concentration, the dissociation constant can be calculated using the appropriate binding equation[180].
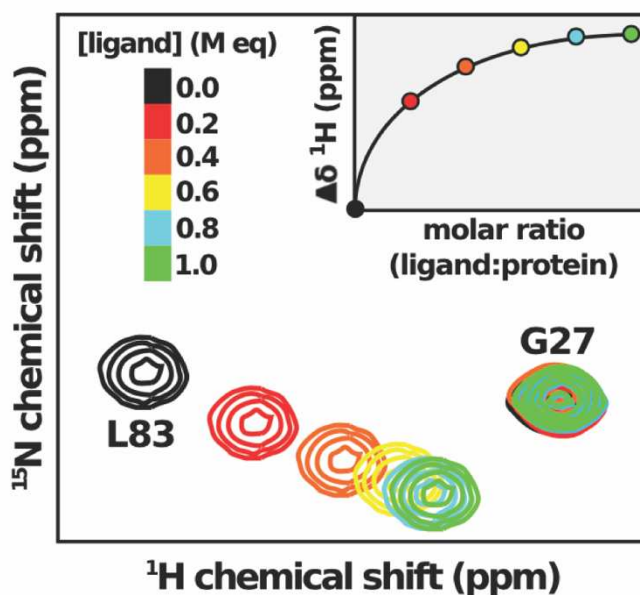
Figure 30. NMR titration experiment of a labeled protein with increasing concentration of a ligand. The overlay of 2D ^{1}H, ^{15}N HSQC spectra shows the affected peak L83 to move from free state in black to the bound state in green. Analyzing the CSP of the L83 peak during the titration by plotting the CSP values against the molar ratio of ligand to protein, the dissociation constant can be calculated using the appropriate binding equation. From Ziarek, Baptista and Wagner[180], with permission from Journal of Molecular Medicine, Copyright 2018.

Although the binding site on the protein of interest can be easily determined analyzing the CSP values, the dissociation constant of the interaction cannot be always calculated using the NMR titration data because of quick broadening of the affected peaks. Therefore, other biophysical techniques are used for determination of the kinetics of the interaction.

3.4 NMR relaxation

NMR spectroscopy provides information also for the dynamics of the protein on large timescale range, from picoseconds (ps) to seconds (s) timescale motions. There are various NMR methods for studying the protein dynamics, such as Carr-Purcell Meiboom-Gill (CPMG) relaxation dispersion, Residual dipolar coupling (RDC), Paramagnetic relaxation enhancement (PRE) and Nuclear spin relaxation (NSP) and especially ^{15}N relaxation measurements. The latter measures the fast correlated motions from picoseconds (ps) to nanoseconds (ns) timescales, but also slower motions up to milliseconds (ms) timescales and thus, provides information for the structure, the rigidity and any existed chemical exchange phenomena. In the ^{15}N relaxation measurements, three parameters are usually measured, the T1 longitudinal relaxation time constant, the T2 transverse relaxation time constant and the ^{1}H-^{15}N heteronuclear NOEs (Nuclear Overhauser effect). The intrinsically disordered proteins are very dynamic macromolecules which exist as an ensemble of conformations lacking of stable structure over time. Therefore, the ^{15}N relaxation measurements are conducted routinely as part of their structural characterization. Figure 31 illustrates the R1, R2 relaxation rates and heteronuclear NOEs plots for a protein when it is well-folded (a) and when is partially disordered (b). The differences in the range of the values in all three plots and the values' fluctuation of the protein in panel (b) with higher R1 values and lower values for R2 and heteronuclear NOEs indicates its partially disordered nature.
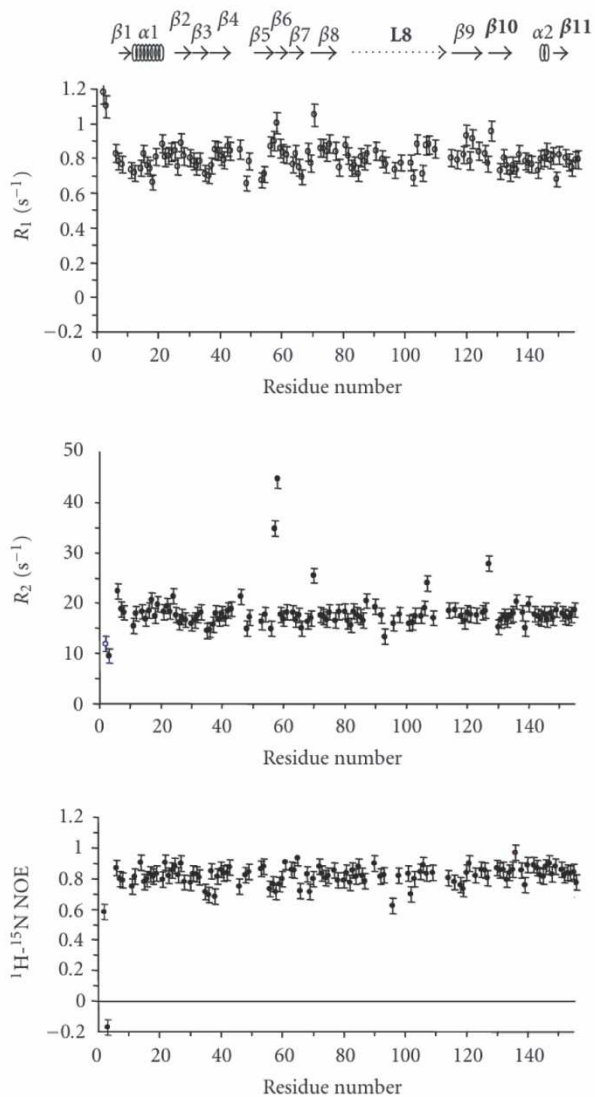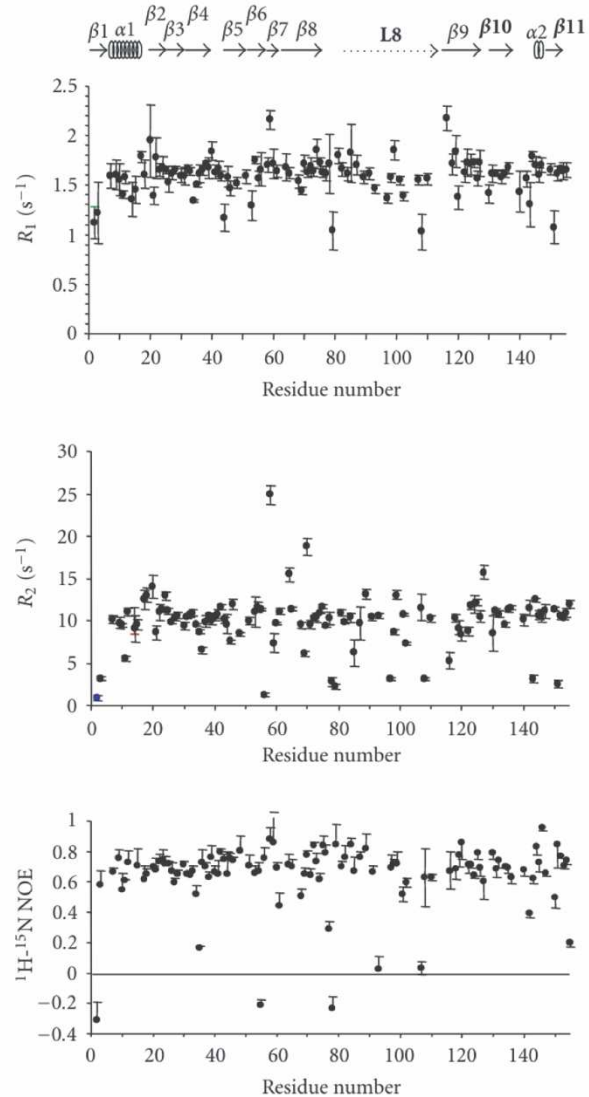
Figure 31. [15]N relaxation parameters, R1, R2 relaxation rates and heteronuclear NOEs plots, for a protein when it is well-folded (a) and when it is partially disordered (b). On the top, the secondary structural elements are shown as α1 and α2 for α-helices, β1-β11 for β-strands and L8 for a loop. Modified from Alcaraz et al.[181], available under a Creative Commons Attribution License (CC BY 3.0).

Objectives

ORF3 protein is a small regulatory multifunctional protein essential in the life cycle of HEV and poorly characterized. Previous studies have shown that ORF3 is mainly involved in the release of the infectious viral particles and interacts with other viral and host proteins inside the cell. In addition, it is reported that ORF3 protein is associated with the cellular membranes either via its oligomerization and transmembrane insertion, or via palmitoylation of its N-terminal Cysteine-rich region[81,82].

In this study, a detailed molecular characterization of the ORF3 protein in order to decipher its functional role(s) during the HEV life cycle and the mechanism for the release of infectious viral particles is achieved using NMR spectroscopy and other biophysical techniques.

The first and main aim is the structural and dynamic characterization of HEV ORF3 protein. The expression and the purification optimization as well the structural characterization of the protein using NMR spectrometry are needed due to the lack of biophysical data.

Secondly, the determination of the association of ORF3 protein to the membrane is studied. The structural study of ORF3 protein anchoring will provide a better understanding of its function and elucidate the preferrable mode out of the two membrane-anchoring ones that have been proposed in the literature[81,82]. In order to investigate the ORF3 membrane association, the Nanodisc (ND) technology is used.

Finally, the last important aim of this study is the molecular interaction of HEV ORF3 protein with different proteins of host cells and especially with human Tsg101 UEV protein using NMR spectroscopy, Isothermal Titration Calorimetry (ITC) and X-ray crystallography. The determination of binding sites in both proteins, the affinity of the interaction as well the crystal structure of the complex will provide more in-depth information about the virus-host proteins interaction which represents an interesting drug target for anti-HEV compounds.

Materials and Methods

1. HEV ORF3 protein

1.1 HEV ORF3 constructs

In this study, ORF3 protein sequence of Genotype 3 that infects animals (zoonic infection) and indirectly humans by consumption of undercooked meat from infected animals is used. Genotype 3 is found mainly on developed countries, such as Europe, United States of America (USA), Australia and Russia[182,24].

Various constructs of ORF3 protein were designed for recombinant expression of the protein in *E. coli*, mandatory for biophysical analyses as they require important quantity of purified protein and/or isotopic labelling. As shown in Figure 32, the constructs used in this study are the ORF3 WT, the ORF3 Cter which is the C-terminal region of the ORF3 protein (aa 48-113), the ORF3 C8A in which all Cysteines are mutated to Alanine residues, the ORF3 C8S in which all Cysteines are mutated to Serine residues and the ORF3 C20 and the His-ORF3 C20 in which all Cysteines are mutated to Serine except of Cysteine in position 20 of the protein sequence. All the ORF3 constructs have been designed in the laboratory and then have been synthesized by GeneCust company using pET24a(+) as the host vector. The pET24a(+) vector contains a kanamycin gene as antibiotic resistance gene.
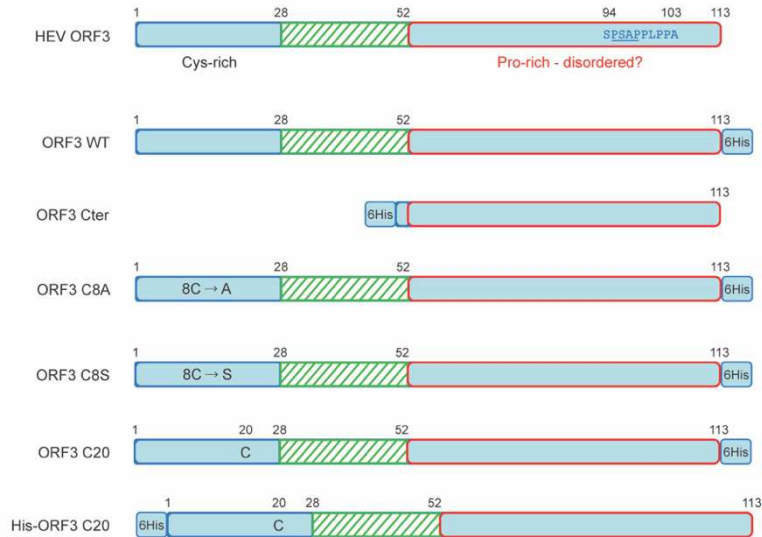
Figure 32. HEV ORF3 genotype 3 sequence (top) and designed constructs used in this study (below).

Apart from the ORF3 Cter protein, all the constructs have in C-terminal region of the protein a PreScission Protease cleavage site inserted between the protein sequence and the 6xHis-tag having as a result the expression of the ORF3–PreScission cleavage site–6xHis-tag protein (Figure 33). ORF3 Cter protein has the same cleavage site and a 6xHis-tag in the N-terminus as shown in Figure 34. His-ORF3 C20 protein does not contain any cleavage site and the 6xHis-tag is in the N-terminal region (Figure 35).

ORF3 WT

```
         10          20          30          40          50          60
MGSPCALGLF CCCSSCFCLC CPRHRPASRL AVVVGGAAAV PAVVSGVTGL ILSPSPSPIF

         70          80          90         100         110         120
IQPTPSPPIS FHNPGLELAL GSRPAPLAPL GVTSPSAPPL PPAVDLPQLG LRRGLEVLFQ
```

PreScission Protease cleavage site

GPGHHHHHH

6xHis-tag

Figure 33. Sequence of expressed ORF3 WT protein. The 6xHis-tag is colored with cyan, the PreScission Protease cleavage site with red and the purple star is the last residue of the protein sequence. ORF3 C8S, ORF3 C8A and ORF3 C20 have the same cleavage site and 6xHis-tag positions.

ORF3 Cter

6xHis-tag   PreScission Protease cleavage site

```
         10          20          30          40          50          60
MGSSHHHHHH SSGLEVLFQG PTGLILSPSP SPIFIQPTPS PPISFHNPGL ELALGSRPAP

         70          80
LAPLGVTSPS APPLPPAVDL PQLGLRR
```

Figure 34. Sequence of expressed ORF3 Cter protein. The 6xHis-tag is colored with cyan, the PreScission Protease cleavage site with red and the purple star is the first residue of the protein sequence.

His-ORF3 C20

6xHis-tag

```
         10          20          30          40          50          60
MGSHHHHHHH GSPSALGLFS SSSSSFSLCS PRHRPASRLA VVVGGAAAVP AVVSGVTGLI

         70          80          90         100         110         120
LSPSPSPIFI QPTPSPPISF HNPGLELALG SRPAPLAPLG VTSPSAPPLP PAVDLPQLGL
```

RR

Figure 35. Sequence of expressed His-ORF3 C20 protein. The 6xHis-tag is colored with cyan and the purple star is the first residue of the protein sequence.

1.2 HEV ORF3 expression and purification protocol

The plasmid harboring the ORF3 coding sequence is transformed into chemically competent *E. coli* BL21 (DE3) cells for overexpression. Pre-cultures are grown in Lysogeny Broth (LB) medium at 37°C using colonies from the agar plate with shaking at 180 rpm overnight containing 25 mg/L kanamycin. In this study, unlabeled and labeled protein samples for biochemical and biophysical experiments are produced. For the unlabeled samples, 10 mL of the overnight pre-culture are added in 1 L LB medium containing 25 mg/L kanamycin. In order to use NMR Spectroscopy, labeled protein samples with different labeling scheme are prepared. For these samples, 20 mL of the overnight pre-culture are centrifuged at 1,800 xg for 10 min at 4°C and then the cell pellet is resuspended in 1 L M9 minimal medium containing 1 g/L $^{15}NH_4Cl$ (Sigma-Aldrich) as nitrogen source, 4 g/L ^{12}C D-glucose or 3 g/L ^{13}C D-glucose (Sigma-Aldrich) as carbon source, 42 mM Na_2HPO_4, 22 mM KH_2PO_4, 8.56 mM NaCl, 1 mM $MgSO_4$, 0.1 mM $CaCl_2$, 1X MEM vitamins, 0.5 g/L ^{15}N Isogro or 0.5 g/L ^{15}N ^{13}C Isogro (Sigma-Aldrich) and 25 mg/L kanamycin for producing ^{15}N or double ^{15}N, ^{13}C labeled protein, respectively. The cells are grown at 37°C at 180 rpm until the OD_{600} reached ~1.0 and the induction starts adding 0.4 mM isopropyl-β-D-thiogalactopyranoside (IPTG). After 5 hours at 37°C, the cells are harvested at 6,000 xg for 20 min at 4°C. The cell pellet is resuspended with ~40 mL of 1x PBS buffer, transferred in a new 50 mL tube and kept frozen at -80 °C until starting the purification procedure.

Because of the lack of a purification protocol, the optimal conditions and steps for ORF3 protein purification had to be found. In the resuspended cell pellet that contains the expressed protein of interest, DNase I 22.9 mg/L (EUROMEDEX) and RNase A 13.3 mg/L (Sigma-Aldrich) is added to digest the nucleic acids. The cells are passed six times in the Avestin EmulsiFlex C3 Homogenizer at 4°C for their lysis and thus the cellular content is released. In order to separate the soluble fraction from the cell membranes and debris, the lysed cells are centrifuged at 39,000 xg for 40 min at 4°C. The first observation is that the protein is insoluble and stuck to the *E. coli* cellular debris. In order to recover the protein from the insoluble fraction, a denaturing buffer containing 6 M Urea is used. After the resuspension of the pellet, a centrifugation step at 10,000 xg for 1 hour at 4°C is done for separation of the ORF3 protein from the insoluble cell materials. The

supernatant is filtered with 0.45 μm filter before its injection in the ÄKTA pure chromatography system (Cytiva) and using a 1 mL HisTrap column (Cytiva) to further purify the protein of interest monitoring the 215 nm, 260 nm and 280 nm UV absorbance. ORF3 protein does not contain any aromatic residues in its sequence and thus, the unspecific 215nm wavelength is important for the protein detection. In order to successfully isolate the ORF3 protein from the *E. coli* extract, the column has to be equilibrated with urea-containing buffer (Buffer A1: 50 mM Tris-HCl pH 6.8, 50 mM NaCl, 10 mM Imidazole, 6 M Urea), the denaturing agent is then slowly removed by washing the column with Buffer B containing 50 mM Tris-HCl pH 6.8, 50 mM NaCl and 10 mM Imidazole using a linear gradient over 30 Column Volumes (CV). Then ORF3 is step-eluted with an imidazole-containing buffer (Buffer A2: 20 mM Tris-HCl pH 6.8, 50 mM NaCl, 300 mM Imidazole) and 0.5 mL fractions are collected in 96-well plate. Different fractions along the purification steps are then checked with SDS-PAGE (Sodium dodecyl-sulfate polyacrylamide gel electrophoresis) to determine the fractions containing the ORF3 protein. All the elution fractions containing the protein with the correct molecular weight are pooled in a new 50 mL tube and they can be frozen using liquid Nitrogen and kept at -80°C until proceeding to the next step. In order to increase the protein purity, a second step of purification is performed by Reverse Phase (RP) Chromatography. The addition of 0.1% TFA (final concentration) is needed in order to lower the pH of the sample before the RP run. Using a Zorbax C8 column and 5 mL loop for protein injection, the column is equilibrated with Buffer A containing 0.1% TFA and 5% Acetonitrile. After the injection of all the sample, the column is washed with a linear gradient toward 30% Buffer B (0.1% TFA, 80% Acetonitrile) for ~1 CV. The elution step is done with increasing concentration of Buffer B from 30% to 100% for ~5 CV. Three UV absorbance curves (215 nm, 260 nm and 280 nm) are monitoring the purification of ORF3 protein. The elution fractions are checked with SDS-PAGE, are pooled and diluted with final NMR buffer containing 50 mM NaPi pH 6.1, 50 mM NaCl, 0.1 mM EDTA, 1 mM DTT. The next step is the overnight dialysis at 4°C using 6-8 kDa cut-off membrane against 3 L NMR Buffer. The concentration of the protein is then performed using Vivaspin Turbo concentrator with 5 kDa cut-off membrane at 3,900 xg at 4°C for 20-min runs. The estimation of the ORF3 protein concentration is done with two different ways because of the absence of aromatic residues. The first one is the Bradford assay estimation using the calibration

curve specifically built for ORF3 protein as described in the results section. It is used before the concentration procedure, but also for the calculation of the concentration of the final sample. The second technique for the concentration' estimation includes a 4-20% SDS-PAGE with increasing volume of ORF3 sample. Comparing the bands with the band of the molecular marker with fixed concentration, the concentration of the final sample is determined. Combining the results of these two tools, the estimation of ORF3 concentration is as accurate as it could be, and protein aliquots are prepared to be flash freeze with liquid Nitrogen and be stored at -80°C until further use.

The biophysical characteristics of uncleaved ORF3 constructs calculated using Expasy Protparam Tool[183] are shown in Table 1.

Table 1. Biophysical characteristics of uncleaved ORF3 constructs based on Expasy Protparam Tool[183].

ORF3 constructs						
	ORF3 WT	ORF3 Cter	ORF3 C8A	ORF3 C8S	ORF3 C20	His-ORF3 C20
Amino acids (aa)	129	87	129	129	129	122
Molecular weight (Da)	13,150.46	9,020.37	12,893.98	13,021.98	13,038.04	12,321.16
Theoretical pI	8.54	7.07	11.70	11.70	11.30	11.70

2. Membrane Scaffold Protein (MSP)

2.1 MSP constructs

Membrane Scaffold Proteins (MSPs) derived from the Apolipoprotein, ApoA1, and used for Nanodisc (ND) assembly. Different engineered truncated MSP variants are available (Figure 36). The length of the MSP determines the diameter of the ND which varies from 6 to 10 nm.
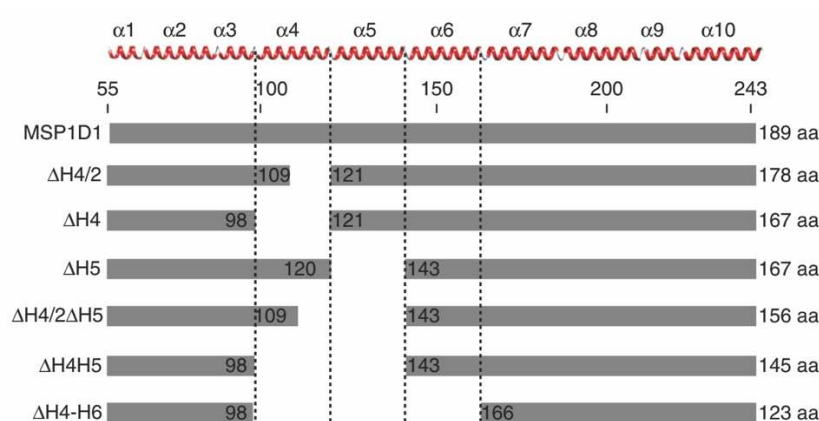
Figure 36. Engineered truncated MSP constructs. MSPD1 variants with deletion of different helix(es) are available for Nanodisc assembly. From Hagn et al.[184], with permission from Nature Protocols, Copyright 2018.

Based on the literature, MSP1D1ΔH5 protein, MSP1D1 variant with deletion of helix 5, that gives NDs of about 7-8 nm diameter is stable enough over time and is suitable for NMR Spectroscopy studies[184]. The plasmid encoding the MSPD1ΔH5 protein is available in Addgene (pET28a-MSP1D1deltaH5, plasmid #71714). In the N-terminal region of the protein, a Tobacco Etch Virus (TEV) cleavage site was inserted between the 6xHis-tag and the protein sequence having as a result the expression of the 6xHis-tag−TEV cleavage site−MSP1D1ΔH5 protein (Figure 37).

MSPD1ΔH5

```
            10          *20          30          40          50          60
MGSSHHHHHH  ENLYFQGSTF  SKLREQLGPV  TQEFWDNLEK  ETEGLRQEMS  KDLEEVKAKV

            70          80          90         100         110         120
QPYLDDFQKK  WQEEMELYRQ  KVEPLGEEMR  DRARAHVDAL  RTHLAPYSDE  LRQRLAARLE

           130         140         150         160         170         180
ALKENGGARL  AEYHAKATEH  LSTLSEKAKP  ALEDLRQGLL  PVLESFKVSF  LSALEEYTKK

LNTQ
```

Figure 37. Sequence of expressed MSPD1ΔH5 protein. The 6xHis-tag is colored with cyan, the TEV Protease cleavage site with red and the purple star is the first residue of the protein sequence.

2.2 MSPD1ΔH5 protein expression and purification

The recombinant plasmid pET28a(+) is transformed into chemically competent *E. coli* BL21 (DE3) cells for overexpression. For protein expression, pre-cultures are grown in LB medium at 37ºC using colonies from the agar plate with shaking at 180 rpm overnight containing 25 mg/L kanamycin based on the gene antibiotic resistance. Adding 10 mL of the overnight pre-culture in 1L LB medium containing 25 mg/L kanamycin, the cells are grown at 37 ºC at 180 rpm until the OD_{600} reached 0.6-0.8 and then the induction starts by adding 1 mM IPTG in the culture. After 4-5 hours at 37 ºC, the cells are harvested at 5,000 xg for 20 min at 4 ºC. The cell pellet is resuspended with ~40 mL of Buffer A containing 20 mM Tris-HCl pH 8, 500 mM NaCl, 10 mM Imidazole and a 1x protease inhibitors tablet is added (cOmplete™, EDTA-free Protease Inhibitor Cocktail tablets, Roche Diagnostics GmbH). The 50 mL tube with the resuspended cell pellet kept frozen at -80 ºC until starting the purification procedure.

The purification procedure of MSP1D1ΔH5 protein is based on Hagn *et al.* 2018[184], lasts three days and is described below in details. The first day of the purification starts thawing the cell pellet and adding 22.9 mg/L DNase I and 13.3 mg/L RNase A for nucleic acids' digestion. The lysis of the cells is performed using the Avestin EmulsiFlex C3 homogenizer at 4ºC. After the cell lysis, 1% Triton X-100 is added and then the cells are centrifuged twice at 10,000 xg for 20 min at 4ºC keeping the supernatant containing the bacterial proteins and protein of interest in new 50 mL tubes. The filtration of the supernatant with 0.45 μm filter is done prior to the 5 mL HisTrap column (Cytiva) purification using the ÄKTA pure chromatography system (Cytiva). The protein purification is monitored using the 215 nm, 260 nm and 280 nm UV absorbance. The filtered supernatant is passed through the column equilibrated with Buffer A. The column is firstly washed with 50 mL Buffer A containing 1% Triton X-100 and then with 50 mL Buffer A containing 50 mM sodium cholate. The next wash step of the column is done with Buffer A for 10 CV and then with 20% Buffer B (50 mM Tris-HCl pH 8, 500 mM NaCl, 500 mM Imidazole) until all the bacterial proteins removed from the HisTrap column. The MSP1D1ΔH5 protein is step-eluted with Buffer B and 2 mL fractions are collected in 96-well plate. Different fractions in the purification steps are then checked with SDS-PAGE to determine the fractions containing the

MSP1D1ΔH5 protein. All the elution fractions containing the protein with the correct molecular weight are pooled and 1 mM DTT is added to the sample. The next step is the N-terminal 6xHis-tag cleavage using TEV Protease in ratio 1:100 w/w (TEV protease: target protein) which specifically cleaves between the Glutamine (Q) and Glycine (G) of the amino-acid sequence ENLYFQG of the cleavage site prior to the MSP1D1ΔH5 protein sequence. The cleavage is done at 15°C overnight during dialysis against 3 L Dialysis Buffer (20 mM Tris-HCl pH 7.5, 100 mM NaCl) using 6-8 kDa cut-off membrane. The next morning the efficiency of the TEV Protease cleavage is checked by SDS-PAGE analysis with samples before and after the TEV Protease addition. The sample is passed again through the 5 mL HisTrap column using the ÄKTA pure chromatography system to separate the cleaved protein from the TEV Protease which binds to the column because of the presence of a 6xHis-tag in the N-terminus, the cleaved N-terminal 6xHis-tag part and any remaining proteins. The fractions are checked using SDS-PAGE and the flow-through and wash-fractions that contain the cleaved protein is directly dialyzed against 3 L MSP buffer containing 20 mM Tris-HCl pH 7.5, 100 mM NaCl, 0.5 mM EDTA using 6-8 kDa cut-off membrane overnight at 4°C. The last day of the purification includes the concentration of the protein. The dialyzed sample is transferred to a new 50 mL tube and based on the Beer-Lambert law (Beer's law), measuring the absorbance at 280 nm, the initial concentration is calculated. The Vivaspin Turbo concentrator with 5 kDa cut-off membrane is then used to centrifuge at 3,900 xg for 15 min runs until the final concentration to be around 500-600 μM. After the verification of the final concentration is reached, aliquots of MSP1D1ΔH5 protein in appropriate volume for one Nanodisc assembly reaction are made, flash frozen with liquid Nitrogen and stored at -80°C until further use. The biophysical characteristics of MSP1D1ΔH5 protein before and after 6xHis-tag cleavage by TEV Protease calculated using Expasy Protparam Tool[183] are shown in Table 2.

Table 2. Biophysical characteristics of MSPD1ΔH5 protein based on Expasy Protparam Tool[183].

MSP1D1ΔH5 protein		
	Uncleaved	*Cleaved*
Amino acids (aa)	184	168
Molecular weight (Da)	21,468.12	19,488.01
Theoretical pI	5.94	5.54
Extinction coefficient, ε (M^{-1} cm^{-1})	19,940	18,450

2.3 Nanodiscs assembly procedure – Attachment of ORF3 protein

The assembly of the Nanodiscs (NDs) requires the MSP protein, the lipids and the protein of interest. Apart from the different constructs of MSP protein, variety of lipids depending on their charge and their length as well as mixture of different kinds can be used for the ND assembly. The rational of ND assembly is the tendency of MSP protein to wrap lipid bilayer, characteristic of ApoA1 protein, and create stable well-defined diameter particles depending on the MSP length. The assembly protocol is based on Hagn *et al.* 2018 and two different procedures are conducted[184]. The first one includes the protein of interest, ORF3 protein, in the assembly while the second one is a two-step procedure, preparing an "empty" Nanodisc and then attach the ORF3 protein on it through modified lipids.

In this study, various lipids are used for ND assembly and have been purchased from Avanti Polar Lipids. The pairs of neutral and negative charged lipids used are the 1,2-dimyristoyl-sn-glycero-3-phosphocholine (14:0 PC – DMPC) with the 1,2-dimyristoyl-sn-glycero-3-phospho-(1'-rac-glycerol) (sodium salt) (14:0 PG) and the 1,2-dioleoyl-sn-glycero-3-phosphocholine (18:1 (Δ9-Cis) PC – DOPC) with the 1,2-dioleoyl-sn-glycero-3-phospho-L-serine (sodium salt) (18:1 PS – DOPS). The modified lipids used for ORF3 protein attachment are the 1,2-dioleoyl-sn-glycero-3-phosphoethanolamine-N-[4-(p-maleimidomethyl)cyclohexane-carboxamide] (sodium salt) (18:1 PE-MCC) or the 1,2-dioleoyl-sn-glycero-3-[(N-(5-amino-1-carboxypentyl)iminodiacetic acid)succinyl] (nickel salt) (18:1 DGS-NTA(Ni)).

The lipids in powder form are resuspended with 100 mM sodium cholate in MSP buffer for stock solutions with 50 mM final concentration and further dissolved -if needed- using the sonication bath for 10-20 min. The lipids in chloroform have to be dried under nitrogen flux heated in glass tube and continue with lyophilization overnight to get rid of any remaining chloroform. The dried film is dissolved with 100 mM sodium cholate in MSP buffer to prepare lipid stock with final concentration of 50 mM.

Regarding the first procedure which includes the ORF3 protein, the ND components are mixed in an Eppendorf tube with the following order, firstly the MSP buffer, secondly the cholate buffer

(total final concentration 20 mM), thirdly the lipid mixture (8 mM final concentration), then the MSP protein (200 µM final concentration) and finally the ORF3 protein (100 µM final concentration) in final volume of 300 µL. The DMPC and PG lipids are used in ratio 3:1 in powder form.

The second procedure starts with the formation of "empty" Nanodiscs, their two components are the MSP1D1ΔH5 protein (200 µM final concentration) and the lipids (10 mM final concentration) in MSP buffer with total volume 300 µL. The DMPC, PG and PE-MCC lipids in ratio 73:25:2 in powder form, the DOPC, DOPS and PE-MCC lipids in ratio 73:25:2 and 7:2:1 in chloroform form and DOPC and DGS-NTA(Ni) lipids in ratio 19:1, 9:1 and 4:1 in chloroform form are used.

The mixture is incubated at 23°C for 2 h with shaking at 750 rpm on Thermomixer Eppendorf device. The detergent, the sodium cholate, is replaced by lipids either using Bio-Beads SM-2 (Bio-Rad) or dialysis against 2 L MSP buffer as a result the Nanodiscs formation. Both options for detergent removal are effective with the dialysis to be more time-consuming as it lasts about two days compared to the Bio-beads SM-2 which trap the sodium cholate within few hours. Therefore, the Bio-beads SM-2 are used for detergent removal, but they need some wash steps before use. They are firstly washed with 20 mL methanol, secondly with 20 mL ethanol twice, then with 20 mL milliQ water four (4) times and finally with 20 mL MSP buffer twice. For 300 µL reaction, 0.3-0.4 g of wet Bio-beads SM-2 are added in two steps, the half of the amount is added to the mixture after the 2-hour incubation for another 1-hour shaking at 23°C at 750 rpm and then the second half afterwards for another 2 h. After the incubation with the Bio-beads, the reaction is carefully transferred in a new Eppendorf tube to remove the Bio-beads from the mixture. The Bio-beads are washed twice with 150 µL MSP buffer to recover as possible reaction. After a high-speed centrifugation at 16,100 xg for 10 min at 4°C of the mixture and remove all the precipitates and remaining beads, the Nanodiscs are purified by Superdex 200 10/300 GL (Cytiva) size exclusion column using the ÄKTA pure chromatography system (Cytiva) with 500 µL injection loop collecting 300 µL fractions in 96-well plate. The fractions of the peak correspond to the well-assembled ND are pooled and interact with ORF3 protein at 23°C overnight with

shaking at 300 rpm. The next day, the interaction is purified again with Superdex 200 10/300 GL (Cytiva) size exclusion column. The shifted peak includes the ND with attached ORF3 protein is checked with a SDS-PAGE and then the fractions are concentrated to reach the appropriate volume and concentration to further characterize the interaction using NMR Spectrometry.

The purified "empty" Nanodiscs can be stored for up to two (2) months at 4°C.

3. Tsg101 UEV protein

3.1 Tsg101 UEV constructs

The DNA encoding the human UEV domain of Tsg101 protein (residues 1-145) (Uniprot accession number Q99816) is designed with different affinity purification tags for recombinant expression and then synthesized by GeneCust using pET28a(+) or pGEX-6P-1 as host vector. Three different constructs are produced in this study, the Tsg101 UEV, the Tsg101 UEV FLAG-tag and the Tsg101 UEV GST-tag. For Tsg101 UEV, in the N-terminal region of the protein, a TEV cleavage site is inserted between the 6xHis-tag and the protein sequence having as a result the expression of the 6xHis-tag−TEV cleavage site−UEV domain of Tsg101 protein and the host vector is the pET28a(+) (Figure 38). For Tsg101 UEV FLAG-tag, between the TEV cleavage site and the protein sequence is inserted an extra affinity tag, the FLAG-tag, having as a result the expression of the 6xHis-tag−TEV cleavage site−FLAG-tag−UEV domain of Tsg101 protein and the host vector is the pET28a(+) (Figure 39). For the Tsg101 UEV GST-tag, in the N-terminal region of the protein, the GST-tag and a PreScission Protease cleavage site are prior to the protein sequence having as a result the expression of GST-tag−PreScission cleavage site−UEV domain of Tsg101 protein and the host vector is the pGEX-6P-1 (Figure 40).

Tsg101 UEV

```
      6xHis-tag      TEV Protease cleavage site
        10           20      *    30           40           50           60
MGSSHHHHHH  SSGENLYFQG  AMAVSESQLK  KMVSKYKYRD  LTVRETVNVI  TLYKDLKPVL

        70           80           90          100          110          120
DSYVFNDGSS  RELMNLTGTI  PVPYRGNTYN  IPICLWLLDT  YPYNPPICFV  KPTSSMTIKT

       130          140          150          160
GKHVDANGKI  YLPYLHEWKH  PQSDLLGLIQ  VMIVVFGDEP  PVFSRP
```

Figure 38. Sequence of expressed Tsg101 UEV protein. The 6xHis-tag is colored with cyan, the TEV Protease cleavage site with red and the purple star is the first residue of the protein sequence.

Tsg101 UEV – FLAG-tag

6xHis-tag   TEV Protease cleavage site   FLAG-tag

```
          10         20         30        * 40         50         60
MGSSHHHHHH SSGENLYFQG ASGDYKDDDD KGSGSMAVSE SQLKKMVSKY KYRDLTVRET

          70         80         90        100        110        120
VNVITLYKDL KPVLDSYVFN DGSSRELMNL TGTIPVPYRG NTYNIPICLW LLDTYPYNPP

         130        140        150        160        170        180
ICFVKPTSSM TIKTGKHVDA NGKIYLPYLH EWKHPQSDLL GLIQVMIVVF GDEPPVFSRP
```

Figure 39. Sequence of expressed Tsg101 UEV – FLAG-tag protein. The 6xHis-tag is colored with cyan, the TEV Protease cleavage site with red, the FLAG-tag with green and the purple star is the first residue of the protein sequence.

Tsg101 UEV – GST-tag

GST-tag

```
          10         20         30         40         50         60
MSPILGYWKI KGLVQPTRLL LEYLEEKYEE HLYERDEGDK WRNKKFELGL EFPNLPYYID

          70         80         90        100        110        120
GDVKLTQSMA IIRYIADKHN MLGGCPKERA EISMLEGAVL DIRYGVSRIA YSKDFETLKV

         130        140        150        160        170        180
DFLSKLPEML KMFEDRLCHK TYLNGDHVTH PDFMLYDALD VVLYMDPMCL DAFPKLVCFK
```

PreScission Protease cleavage site

```
         190        200        210        220        230      * 240
KRIEAIPQID KYLKSSKYIA WPLQGWQATF GGGDHPPKSD LEVLFQGPLG SMAVSESQLK

         250        260        270        280        290        300
KMVSKYKYRD LTVRETVNVI TLYKDLKPVL DSYVFNDGSS RELMNLTGTI PVPYRGNTYN

         310        320        330        340        350        360
IPICLWLLDT YPYNPPICFV KPTSSMTIKT GKHVDANGKI YLPYLHEWKH PQSDLLGLIQ

         370
VMIVVFGDEP PVFSRP
```

Figure 40. Sequence of expressed Tsg101 UEV – GST-tag protein. The GST-tag is colored with cyan, the PreScission Protease cleavage site with red and the purple star is the first residue of the protein sequence.

3.2 Tsg101 UEV expression and purification protocols

The recombinant plasmids are transformed into chemically competent *E. coli* BL21 (DE3) cells for overexpression. The pET28a(+) vector contains a kanamycin gene as antibiotic resistance gene while the pGEX-6P-1 vector contains an ampicillin gene. Pre-cultures are grown in LB medium at 37°C using colonies from the agar plate with shaking at 180 rpm overnight containing 25 mg/L kanamycin or 100 mg/L ampicillin depending on the vector's resistance gene. Both unlabeled (all constructs) and labeled (only for Tsg101 UEV and Tsg101 UEV FLAG-tag) protein samples are produced for biophysical and structural characterization. For the unlabeled samples, 10 mL of the overnight pre-culture are added in 1 L LB medium that contains the appropriate antibiotic. For labeled samples, 20 mL of the overnight pre-culture are centrifuged at 1,800 xg for 10 min at 4°C and then the cell pellet is resuspended in 1 L M9 minimal medium containing 1 g/L $^{15}NH_4Cl$, 4 g/L ^{12}C D-glucose or 3 g/L ^{13}C D-glucose, 42 mM Na_2HPO_4, 22 mM KH_2PO_4, 8.56 mM NaCl, 1 mM $MgSO_4$, 0.1 mM $CaCl_2$, 1X MEM vitamins, 0.5 g/L ^{15}N Isogro or 0.5 g/L ^{15}N ^{13}C Isogro and 25 mg/L kanamycin for producing ^{15}N or double ^{15}N ^{13}C labeled protein, respectively. The cells are grown at 37°C at 180 rpm until the OD_{600} reached ~1.0 and the induction starts adding 0.4 mM IPTG. After 4-5 hours at 37 °C, the cells are harvested at 6,000 xg for 20 min at 4 °C. The cell pellet is resuspended with ~40 mL of Resuspend Buffer (50 mM Sodium Phosphate pH 7.8, 500 mM NaCl, 10 mM Imidazole) for Tsg101 UEV and the Tsg101 UEV FLAG-tag constructs or ~40 mL 1X PBS buffer for Tsg101 UEV GST-tag with addition of a 1x protease inhibitors tablet (cOmplete™, EDTA-free Protease Inhibitor Cocktail tablets, Roche Diagnostics GmbH). The 50 mL tube with the resuspended cell pellet kept frozen at -80 °C until starting the purification procedure.

The purification procedures of the Tsg101 UEV constructs are similar, the protocol for the Tsg101 UEV and the Tsg101 UEV FLAG-tag proteins lasts three days while for the Tsg101 UEV GST-tag lasts two days with slight differences. For all the constructs, the steps before the affinity column chromatography are the same with first step the addition of 22.9 mg/L DNase I and 13.3 mg/L RNase A in the cell pellet and the lysis of the cells using the Avestin EmulsiFlex C3 homogenizer at 4°C. In order to separate the soluble cellular content from the cell membranes and debris, the lysed cells are centrifuged at 39,000 xg for 40 min at 4°C and then the supernatant containing the

soluble fraction -including bacterial and protein of interest- is transferred in new 50 mL tubes. After filtration of the supernatant with 0.45 µm filter, a 1 mL HisTrap column (Cytiva) for the two constructs containing 6xHis-tag in the N-terminal or a 5 mL GSTrap column (Cytiva) for Tsg101 UEV GST-tag is used with the ÄKTA pure chromatography system (Cytiva) to purify the protein of interest, with monitoring the 215 nm, 260 nm and 280 nm UV absorbance. For Tsg101 UEV and the Tsg101 UEV FLAG-tag, the HisTrap column is pre-equilibrated with Buffer A (50 mM Sodium Phosphate pH 7.6 and 400 mM NaCl) in presence of 4% of Buffer B (50 mM Sodium Phosphate pH 7.6, 300 mM NaCl and 300 mM Imidazole). Two wash steps of the column, first with 4% Buffer B for ~40 CV and then with 15% of Buffer B for ~30 CV, are done to remove all bacterial contaminants from the HisTrap column. The Tsg101 UEV protein is then eluted with increasing concentration of Buffer B for 20 CV and 0.5 mL fractions are collected in 96-well plate. For Tsg101 UEV GST-tag, the GSTrap column is equilibrated with 1X PBS buffer and the wash step is done with the same buffer for 5 CV. The protein is step-eluted with Elution Buffer containing 50 mM Tris-HCl pH 8, 10 mM reduced Glutathione, GSH (Sigma-Aldrich) and 0.8 mL fractions are collected in 96-well plate. Different fractions through the purification steps are checked using SDS-PAGE and the fractions containing the Tsg101 UEV protein are pooled in a new 50 mL tube. After this step, the two protocols have the main difference. The 6xHis-tag in the N-terminal region is further cleaved while the GST-tag is remaining for experimental purpose.

For Tsg101 UEV and Tsg101 UEV FLAG-tag constructs, in order to simultaneously eliminate the Imidazole in the buffer and cleave the 6xHis-tag, the appropriate amount of TEV 6xHis-tag protease is added to the sample that is then dialyzed overnight at 15°C with 6-8 kDa cut-off membrane against 3 L Dialysis Buffer containing 50 mM Tris-HCl pH 6.4 or pH 8 (for Tsg101 UEV and Tsg101 UEV FLAG-tag, respectively), 250 mM NaCl and 5 mM β-Mercaptoethanol. The next morning, using SDS-PAGE the efficiency and the level of the TEV cleavage is analyzed. The cleaved protein is passed again through the 1 mL HisTrap column to remove both the TEV 6xHis-tag protease and the cleaved His-tag from the sample. The following step for unlabeled and labeled Tsg101 UEV and labeled Tsg101 UEV FLAG-tag protein samples is the overnight dialysis of cleaved protein with 6-8 kDa cut-off membrane at 4°C against 3 L NMR Buffer (50 mM Sodium Phosphate pH 6.1, 50 mM NaCl, 0.1 mM EDTA). For unlabeled Tsg101 UEV protein sample used for X-ray

crystallography experiments, the dialysis is performed in 3 L Buffer containing 10 mM Tris-HCl pH 6.52, 100 mM NaCl. The unlabeled Tsg101 UEV FLAG-tag protein is further purified using HiLoad 16/600 Superdex 75 pg (Cytiva) size exclusion column in the ÄKTA pure chromatography system (Cytiva) with SEC Buffer (50 mM Tris-HCl pH 7.8, 50 mM NaCl) and 5 mL injection loop(s) collecting 1 mL fractions in 96-well plate. For Tsg101 UEV GST-tag construct, based on the SDS-PAGE, the pooled fractions containing the protein are directly injected in HiLoad 16/600 Superdex 75 pg (Cytiva) size exclusion column in the ÄKTA pure chromatography system (Cytiva) with SEC Buffer (50 mM Tris-HCl pH 7.8, 50 mM NaCl) and 5 mL injection loop(s) collecting 1 mL fractions in 96-well plate.

The last step of the purification is the concentration of the protein in the final concentration depending on the experiment to be used. Measuring the absorbance at 280 nm and based on the Beer-Lambert law (Beer's law), the initial concentration is calculated. The Vivaspin Turbo concentrator with 5 kDa cut-off membrane is then used to centrifuge at 3,900 xg for 15 min runs until the target final concentration. After the verification that the final concentration is reached, the protein is split in aliquots which then are flash frozen with liquid Nitrogen and stored at -80°C until further use.

The biophysical characteristics of Tsg101 UEV constructs calculated using Expasy Protparam Tool[183] are shown in Table 3.

Table 3. Biophysical characteristics of Tsg101 UEV constructs based on Expasy Protparam Tool[183].

	Tsg101 UEV constructs		
	Tsg101 UEV	Tsg101 UEV FLAG-tag	Tsg101 UEV GST-tag
	Cleaved	*Cleaved*	*Uncleaved*
Amino acids (aa)	147	161	376
Molecular weight (Da)	16,742.51	18,169.87	43,438.55
Theoretical pI	8.81	7.01	6.52
Extinction coefficient, ε (M^{-1} cm^{-1})	25,900	27,390	68,760

4. NMR Spectroscopy

4.1 NMR Spectrometers

The Integrative Structural Biology group can mainly use two (2) NMR spectrometers, a 900 MHz and a 600 MHz, for biomolecular characterization of the proteins located on Campus CNRS of Haute Borne and on Pasteur Institute of Lille, respectively. Both spectrometers are equipped with a 5 mm cryogenic triple resonance probe for higher sensitivity. Specifically, the Bruker Avance Neo 900 MHz spectrometer is equipped with a 5 mm CPTCI (^{1}H, ^{15}N, ^{13}C) cryoprobe and the Bruker 600 MHz AvanceIII HD spectrometer is equipped with a 5 mm CPQCI (^{1}H, ^{15}N, ^{13}C, ^{19}F) cryoprobe. In addition, the 900 MHz and 600 MHz spectrometers are equipped with SampleJet and SampleCase, respectively, sample changers that give the opportunity to record many samples automatically. Both spectrometers are used for ORF3 and Tsg101 UEV protein characterization.

In this study, the ORF3 C20 protein is further characterized in collaboration with Dr. Anja Böckmann and her team, Molecular Microbiology and Structural Biochemistry – Protein Solid State NMR group in Lyon using solid-state NMR Spectrometry. The experiments are conducted in the 800 MHz spectrometer using the 3.2 mm HCN probe.

4.2 Sample preparation

For liquid-state NMR experiments, ^{15}N or double ^{15}N, ^{13}C labeled protein samples are prepared in NMR buffer, 50 mM Sodium Phosphate pH 6.1, 50 mM NaCl, 0.1 mM EDTA. For recording any NMR spectrum, 12 mM 3-(Trimethylsilyl)propanoic acid (TMSP) is added as the internal standard for proton chemical shift referencing and 5% (v/v) D$_2$O for lock purposes. There are various NMR tubes used such as 5 mm Shigemi tube, 5 mm tube and 3 mm tube, each one has different volume limits. The lower volume can be used in a 5 mm Shigemi tube is 250 μL while the volume for 5 mm tube is 500 μL and for 3 mm tube 200 μL.

For solid-state NMR experiments, 25mg of ^{15}N, ^{13}C ORF3 C20 labeled protein is precipitated with 100 mM cholate buffer in MSP buffer and is filled into a 3.2 mm rotor. DSS is added for the calibration of the spectra.

4.3 NMR Experiments

All the spectra are acquired at 293K except of the ones for ORF3 WT protein recorded at 298K and processed with Bruker TopSpin software package 4.0.7 and 3.6.2. The analysis of the spectra is done using the NMRFAM-Sparky software 3.19 (T. D. Goddard and D. G. Kneller, SPARKY 3, University of California, San Francisco)[185].

For NMR backbone protein assignments of ORF3 C20 and Tsg101 UEV proteins, the 2D ^{1}H, ^{15}N HSQC and 3D ^{1}H, ^{15}N, ^{13}C HNCACB, HN(CO)CACB, HNCO, HN(CA)CO, HNHA, HN(CA)NNH are the main experiments needed. Because of the disordered nature of ORF3 C20, the 3D ^{1}H, ^{15}N, ^{13}C (H)NCANNH, H(N)CANNH, HCBCACON, (H)NCOCANNH, H(N)COCANNH and HACONH are recorded to help the assignment procedure providing more information for each peak. For both proteins, the proline assignments are obtained analyzing the carbon detection 3D ^{1}H, ^{15}N, ^{13}C (H)CACON and HCAN and the 2D ^{15}N, ^{13}C NCO and ^{13}C,^{13}C CACO experiments. All 3D NMR spectra are recorded using non-uniform sampling and the methyl signal of TMSP is used as proton chemical shift reference at 0 ppm.

The interaction between the ORF3 C20 and the Tsg101 UEV proteins is studied also using NMR Spectroscopy. When the Tsg101 UEV protein is ^{15}N labeled and addition of increasing concentration of unlabeled ORF3 C20 protein is made, only 1D and 2D ^{1}H, ^{15}N HSQC spectra are recorded for each titration point. When the interaction is studied from the ORF3 C20 side, a double ^{15}N, ^{13}C labeled protein is used and the addition of increasing concentration of unlabeled Tsg101 UEV protein is monitoring using both 2D ^{1}H, ^{15}N HSQC and 2D ^{15}N, ^{13}C NCO spectra in order to also detect the proline affected peaks. For all experiments, each point of interaction is prepared separately in 5 mm tubes and the SampleJet in 900 MHz spectrometer is used to record the series of the spectra.

4.4 Backbone and Proline assignments

4.4.1 ORF3 C20 protein

NMR experiments for backbone and proline assignments of ORF3 C20 protein are recorded at 293K on Bruker 600 MHz AvanceIII HD spectrometer equipped with a 5 mm CPQCI (^{1}H, ^{15}N, ^{13}C, ^{19}F) cryoprobe. The ^{15}N, ^{13}C ORF3 C20 protein sample is at 100 μM in NMR buffer, 20 mM Sodium Phosphate pH 6.1, 50 mM NaCl, 0.1 mM EDTA, 2 mM THP (Tris(hydroxypropyl)phosphine), 5% (v/v) D$_2$O placed in a Shigemi tube. All the experiments for backbone and proline assignments are recorded with the same protein sample.

The 2D ^{1}H, ^{15}N HSQC spectrum is acquired with 2048 and 256 complex points in the direct and indirect dimensions, respectively, and 16 scans. The spectral width (SW) in indirect ^{15}N dimension is 23 ppm and the position of the carrier (O1P) is at 118 ppm. The 3D ^{1}H, ^{15}N, ^{13}C HNCACB and HN(CO)CACB spectra are acquired with 1024, 96 and 256 complex points in the direct, indirect (^{15}N) and indirect (^{13}C) dimensions, respectively, and 32 scans. The SW in indirect ^{15}N dimension is 23 ppm and the position of the carrier is at 118 ppm while the SW in indirect ^{13}C dimension in HNCACB is 70 ppm with the carrier at 39 ppm and the SW in indirect ^{13}C dimension in HN(CO)CACB is 58 ppm with the carrier at 41 ppm. The 3D HNCO spectrum is recorded with 1022, 96 and 128 complex points in the direct, indirect (^{15}N) and indirect (^{13}C) dimensions, respectively, and 16 scans and the 3D HNCACO spectrum with 1176, 96 and 128 complex points and 64 scans. In both HNCO and HNCACO spectra, the SW in indirect ^{15}N dimension is 23 ppm with the carrier at 118 ppm and in indirect ^{13}C dimension is 10 ppm with the carrier at 172.5 ppm. The 3D (H)NCANNH and H(N)CANNH spectra are acquired with 2048, 128 and 128 complex points in the direct, indirect (^{15}N) and indirect (^{15}N and ^{1}H, respectively) dimensions, respectively, and 32 scans. The SW in indirect ^{15}N dimension for both spectra is 23 ppm and the position of the carrier is at 118 ppm while the SW in indirect ^{15}N dimension in (H)NCANNH is 23 ppm with the carrier at 118 ppm and the SW in indirect ^{1}H dimension in H(N)CANNH is 12 ppm with the carrier at 4.697 ppm. The 3D HCBCACON spectrum is recorded with 2048, 120 and 120 complex points in the direct, indirect (^{15}N) and indirect (^{13}C) dimensions, respectively, and 32 scans. The SW in indirect ^{15}N dimension in is 35 ppm with the carrier at 123.5 ppm and in indirect ^{13}C dimension in is 60

ppm with the carrier at 40 ppm. The 3D (H)NCOCANNH spectrum is acquired with 2048, 140 and 140 complex points in the direct, indirect (^{15}N) and indirect (^{15}N) dimensions, respectively, and 64 scans. The SW in both indirect ^{15}N dimension in is 23 ppm with the carrier at 118 ppm. The 3D H(N)COCANNH spectrum is recorded with 2048, 140 and 160 complex points in the direct, indirect (^{15}N) and indirect (^{1}H) dimensions, respectively, and 64 scans. The SW in indirect ^{15}N dimension in is 23 ppm with the carrier at 118 ppm and in indirect ^{1}H dimension in is 10 ppm with the carrier at 4.698 ppm. The 3D HNHA spectrum is acquired with 2048, 256 and 128 complex points in the direct, indirect (^{1}H) and indirect (^{15}N) dimensions, respectively, and 32 scans. The SW in indirect ^{1}H dimension in is 12 ppm with the carrier at 4.697 ppm and in indirect ^{15}N dimension in is 23 ppm with the carrier at 118 ppm. The 3D HACAN spectrum is recorded with 2048, 128 and 128 complex points in the direct, indirect (^{13}C) and indirect (^{15}N) dimensions, respectively, and 32 scans. The SW in indirect ^{13}C dimension in is 40 ppm with the carrier at 52 ppm and in indirect ^{15}N dimension in is 35 ppm with the carrier at 123.5 ppm. The 3D HACONH spectrum is recorded with 2048, 64 and 512 complex points in the direct, indirect (^{15}N) and indirect (^{1}H) dimensions, respectively, and 32 scans. The SW in indirect ^{15}N dimension in is 23 ppm with the carrier at 118 ppm and in indirect ^{1}H dimension in is 3 ppm with the carrier at 4.698 ppm. Finally, the 2D carbon detection ^{15}N, ^{13}C NCO spectrum is acquired with 1024 and 520 complex points in the direct (^{13}C) and indirect (^{15}N) dimensions, respectively, and 160 scans. The SW in direct ^{13}C dimension in is 40 ppm with the carrier at 173.5 ppm and in indirect ^{15}N dimension in is 47 ppm with the carrier at 123 ppm.

Combined the information of these all experiments, the backbone and proline assignments of ORF3 C20 protein are obtained.

4.4.2   ORF3 Cter protein

In addition, the backbone and proline assignments of ORF3 Cter protein are obtained recording the same data set on Bruker 900 MHz Avance Neo spectrometer equipped with a 5 mm CPTCI (^{1}H, ^{15}N, ^{13}C) cryoprobe at 293K using a 300 μM ^{15}N, ^{13}C ORF3 Cter labeled protein sample in 50mM Sodium Phosphate pH 6.8, 50 mM NaCl, 0.5 mM EDTA, 5% (v/v) D$_2$O placed in a Shigemi tube.

The 2D ^1H, ^{15}N HSQC spectrum is acquired with 3072 and 128 complex points in the direct and indirect dimensions, respectively, and 32 scans. The spectral width (SW) in indirect ^{15}N dimension is 22 ppm and the position of the carrier (O1P) is at 118 ppm. The 3D ^1H, ^{15}N, ^{13}C HNCACB and HN(CO)CACB spectra are acquired with 1540, 62 and 128 complex points in the direct, indirect (^{15}N) and indirect (^{13}C) dimensions, respectively, and 128 scans. The SW in indirect ^{15}N dimension is 22 ppm and the position of the carrier is at 118 ppm while the SW in indirect ^{13}C dimension is 56 ppm with the carrier at 44 ppm. The 3D HNCO spectrum is recorded with 1536, 62 and 96 complex points in the direct, indirect (^{15}N) and indirect (^{13}C) dimensions, respectively, and 64 scans and the 3D HNCACO spectrum with 1536, 62 and 128 complex points and 192 scans. In both HNCO and HNCACO spectra, the SW in indirect ^{15}N dimension is 22 ppm with the carrier at 118 ppm and in indirect ^{13}C dimension is 12 ppm with the carrier at 174 ppm. The 3D HN(CA)NNH spectrum is recorded with 2048, 128 and 128 complex points in the direct, indirect (^{15}N) and indirect (^{15}N) dimensions, respectively, and 64 scans. The SW in both indirect ^{15}N dimension in is 22 ppm with the carrier at 118 ppm. The 3D HACAN spectrum is recorded with 2048, 98 and 128 complex points in the direct, indirect (^{13}C) and indirect (^{15}N) dimensions, respectively, and 32 scans. The SW in indirect ^{13}C dimension in is 30 ppm with the carrier at 53.2 ppm and in indirect ^{15}N dimension in is 36 ppm with the carrier at 123 ppm. The 3D HNHA spectrum is acquired with 2048, 98 and 256 complex points in the direct, indirect (^1H) and indirect (^{15}N) dimensions, respectively, and 32 scans. The SW in indirect ^1H dimension in is 6 ppm with the carrier at 4.697 ppm and in indirect ^{15}N dimension in is 22 ppm with the carrier at 118 ppm. Finally, the 2D carbon detection ^{15}N, ^{13}C NCO spectrum is acquired with 1024 and 256 complex points in the direct (^{13}C) and indirect (^{15}N) dimensions, respectively, and 32 scans. The SW in direct ^{13}C dimension in is 40 ppm with the carrier at 173.5 ppm and in indirect ^{15}N dimension in is 47 ppm with the carrier at 123 ppm.

Combined the information of these all experiments, the backbone and proline assignments of ORF3 Cter protein are obtained.

4.4.3 ORF3 WT protein

Furthermore, the backbone assignments of ORF3 WT protein are obtained using the backbone assignments of ORF3 C20 protein and recording the 2D HSQC and 3D HNCO and HNCACB spectra on Bruker 900 MHz Avance Neo spectrometer equipped with a 5 mm CPTCI (^{1}H, ^{15}N, ^{13}C) cryoprobe at 298K using a 75 µM ^{15}N, ^{13}C ORF3 WT protein sample in 50 mM Sodium Phosphate pH 6.1, 50 mM NaCl, 5 mM DTT, 0.1 mM EDTA, 5% (v/v) D_2O in a closed 5 mm Shigemi tube.

The 2D ^{1}H, ^{15}N HSQC spectrum is acquired with 3072 and 256 complex points in the direct and indirect dimensions, respectively, and 16 scans. The spectral width (SW) in indirect ^{15}N dimension is 22 ppm and the position of the carrier (O1P) is at 118 ppm. The 3D ^{1}H, ^{15}N, ^{13}C HNCACB spectrum is acquired with 1542, 100 and 180 complex points in the direct, indirect (^{15}N) and indirect (^{13}C) dimensions, respectively, and 256 scans. The SW in indirect ^{15}N dimension is 22 ppm and the position of the carrier is at 118 ppm while the SW in indirect ^{13}C dimension is 65 ppm with the carrier at 39 ppm. The 3D HNCO spectrum is recorded with 1536, 88 and 128 complex points in the direct, indirect (^{15}N) and indirect (^{13}C) dimensions, respectively, and 128 scans. The SW in indirect ^{15}N dimension is 22 ppm with the carrier at 118 ppm and in indirect ^{13}C dimension is 15 ppm with the carrier at 173.5 ppm.

Comparing the 3D data set of ORF3 WT protein with the corresponding ones of ORF3 C20 protein, the backbone assignments of ORF3 WT are reliably transferred on the 2D HSQC spectrum.

4.4.4 Tsg101 UEV protein

NMR experiments for backbone and proline assignments of Tsg101 UEV protein are recorded at 298K on Bruker 600 MHz AvanceIII HD spectrometer equipped with a 5 mm CPQCI (^{1}H, ^{15}N, ^{13}C, ^{19}F) cryoprobe. The ^{15}N, ^{13}C Tsg101 UEV protein sample is at 330 µM in NMR buffer, 50 mM Sodium Phosphate pH 6.1, 50 mM NaCl, 0.1 mM EDTA, 5% (v/v) D_2O placed in a Shigemi tube. All the experiments for backbone and proline assignments are recorded with the same protein sample.

The 2D ^{1}H, ^{15}N HSQC spectrum is acquired with 2048 and 256 complex points in the direct and indirect dimensions, respectively, and 16 scans. The spectral width (SW) in indirect ^{15}N dimension is 28 ppm and the position of the carrier (O1P) is at 117.5 ppm. The 3D ^{1}H, ^{15}N, ^{13}C HNCACB and HN(CO)CACB spectra are acquired with 2048, 82 and 120 complex points in the direct, indirect (^{15}N) and indirect (^{13}C) dimensions, respectively, and 48 scans. In both spectra, the SW in indirect ^{15}N dimension is 28 ppm with the carrier at 117.5 ppm and in indirect ^{13}C dimension is 60 ppm with the carrier at 42 ppm. The 3D HNCO spectrum is recorded with 1432, 82 and 128 complex points in the direct, indirect (^{15}N) and indirect (^{13}C) dimensions, respectively, and 16 scans. The SW in indirect ^{15}N dimension is 30 ppm with the carrier at 118 ppm and in indirect ^{13}C dimension is 11 ppm with the carrier at 173 ppm. The 3D HNCACO spectrum is acquired with 1426, 82 and 128 complex points in the direct, indirect (^{15}N) and indirect (^{13}C) dimensions, respectively, and 64 scans. The SW in indirect ^{15}N dimension is 28 ppm with the carrier at 117.5 ppm and in indirect ^{13}C dimension is 14 ppm with the carrier at 173.5 ppm. The 3D HN(CA)NNH spectrum is recorded with 2048, 128 and 128 complex points in the direct, indirect (^{15}N) and indirect (^{15}N) dimensions, respectively, and 48 scans. The SW in both indirect ^{15}N dimension in is 28 ppm with the carrier at 117.5 ppm. The 3D HNHA spectrum is acquired with 2048, 256 and 128 complex points in the direct, indirect (^{1}H) and indirect (^{15}N) dimensions, respectively, and 32 scans. The SW in indirect ^{1}H dimension in is 12 ppm with the carrier at 4.7 ppm and in indirect ^{15}N dimension in is 28 ppm with the carrier at 117.5 ppm. The 3D HACAN spectrum is recorded with 2048, 128 and 128 complex points in the direct, indirect (^{13}C) and indirect (^{15}N) dimensions, respectively, and 32 scans. The SW in indirect ^{13}C dimension in is 40 ppm with the carrier at 52 ppm and in indirect ^{15}N dimension in is 28 ppm with the carrier at 117.5 ppm. The 2D carbon detection ^{15}N, ^{13}C NCO spectrum is acquired with 1024 and 160 complex points in the direct (^{13}C) and indirect (^{15}N) dimensions, respectively, and 160 scans. The SW in direct ^{13}C dimension in is 40 ppm with the carrier at 173.5 ppm and in indirect ^{15}N dimension in is 47 ppm with the carrier at 123 ppm. The 2D carbon detection ^{13}C, ^{13}C CACO spectrum is acquired with 724 and 256 complex points in the direct (^{13}C) and indirect (^{13}C) dimensions, respectively, and 160 scans. The SW in direct ^{13}C dimension in is 20 ppm with the carrier at 173.5 ppm and in indirect ^{13}C dimension in is 50 ppm with the carrier at 173.5 ppm. Finally, the 3D carbon detection ^{13}C ,^{15}N, ^{13}C HCACON spectrum is

acquired with 1024, 128 and 64 complex points in the direct (^{13}C), indirect (^{15}N) and indirect (^{13}C) dimensions, respectively, and 32 scans. The SW in direct ^{13}C dimension in is 40 ppm with the carrier at 173.5 ppm, in indirect ^{15}N dimension in is 43 ppm with the carrier at 123 ppm and in indirect ^{13}C dimension in is 40 ppm with the carrier at 173.5 ppm.

The backbone and proline assignments of Tsg101 UEV protein are obtained and deposited in the Biological Magnetic Resonance Data Bank (BMRB) under accession code 50765.

Results

1. Molecular characterization of HEV ORF3 protein in solution

1.1 HEV ORF3 sequence analysis

In this study, as mentioned in Materials and Methods, the protein sequence of HEV ORF3 protein is derived from Genotype 3 (HEV3) found mainly in developed countries and transmitted indirectly to humans by consuming of uncooked or undercooked meat from infected animals[182,24].

An alignment of ORF3 amino acid sequence of the first four HEV genotypes (HEV1-HEV4) isolated in human using the multiple sequence alignment tool Clustal Omega[126,127] shows an identity more than 73% along these genotypes (Figure 41).

```
HEV4    MEMPPCALGLFCFCSSCFCLCCPRHRPVSRLAVAAGKRG--AAVVSGVTGLILSPSPSPI    58
HEV3    -MGSPCALGLFCCCSSCFCLCCPRHRPASRLAVVVGGAAAVPAVVSGVTGLILSPSPSPI    59
HEV1    MGSRPWALGLFCCCSSCFCLCCSRHRPVSRLAAVVGGAAAVPAVVSGVTGLILSPSQSPI    60
HEV2    MGSPPCALGLFCCCSSCFCLCCPRHRPVSRLAAVVGGAAAVPAVVSGVTGLILSPSQSPI    60
        * ****** ********* ****.****...*   .   ************** ***


HEV4    FIQPTPSHLTFQPPPGLELALGSQSVHSAPLGVTSPSAPPLPPVVDLPQLGLRR         112
HEV3    FIQPTPSPPISFHNPGLELALGSRPAPLAPLGVTSPSAPPLPPAVDLPQLGLRR         113
HEV1    FIQPTPSPRMSPLRPGLDLVFANPSDHSAPLGATRPSAPPLPHVVDLPQLGPRR         114
HEV2    FIQPTPLPQTLPLRPGLDLAFANQPGHLAPLGEIRPSAPPLPPVADLPQPGLRR         114
        ******      ***:*.:..    ****   ******* ..**** * **
```

Figure 41. Alignment of ORF3 amino acid sequence of the first four HEV genotypes (HEV1-HEV4) that infect humans using the multiple sequence alignment tool Clustal Omega[126,127]. HEV1: GenBank accession # O90299, HEV2: GenBank accession # Q03499, HEV3: GenBank accession # ADV71353 and HEV4: GenBank accession # Q91VZ7. The star (*) sign indicates the conserved residues, the colon (:) residues with conservation between groups of strongly similar properties and the period (.) conservation between groups of weaky similar properties.

The sequence of HEV3 ORF3 Wild Type (WT) protein is shown in Figure 42.
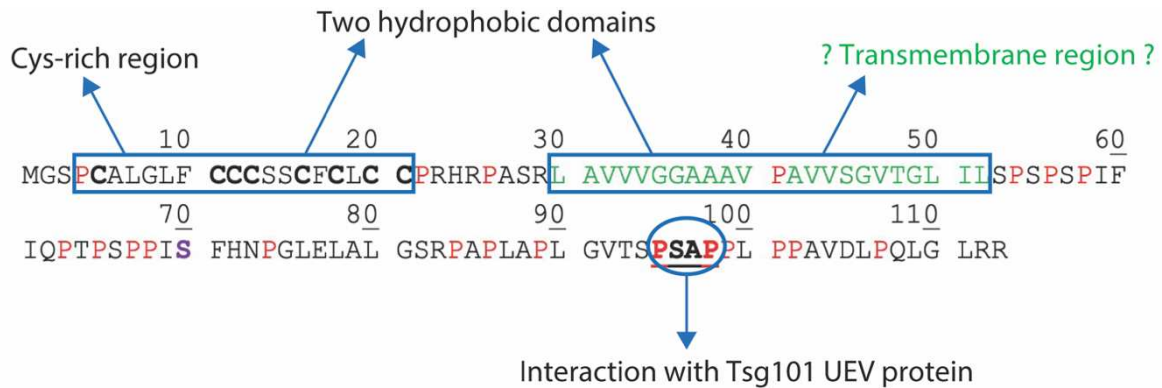
Figure 42. HEV ORF3 WT Genotype 3 sequence with blue boxes the hydrophobic domains, with green the transmembrane region, with red the Proline residues and with the blue circle the PSAP motif involved in the interaction with Tsg101 UEV protein.

Analyzing the protein sequence, two hydrophobic domains in blue boxes are identified. The first domain is a Cysteine-rich region that contains eight (8) Cysteine residues (in bold in Figure 42). Based on Gouttenoire *et al.* 2018 studies, this region is involved in the anchoring of the protein to the membrane by a post-translational modification, the palmitoylation of one or multiple Cysteine residues[82]. The second hydrophobic domain (marked also in green in Figure 42) is predicted to be transmembrane and because of this region, Ding *et al.* 2017 proposed to a transmembrane insertion of ORF3 protein and its oligomerization that forms an ion channel and correlated the ORF3 function with a viroporin function[81].

Using the online TMHMM v2.0 server[186], the prediction of the transmembrane region of HEV ORF3 WT protein is obtained (Figure 43). The probability of the region between residues Leucine 30 and Leucine 52 (Leu50-Leu52) to be transmembrane helix is high with an overall score of 0.8.

HEV ORF3 WT	TMHMM2.0	inside	1	29
HEV ORF3 WT	TMHMM2.0	TMhelix	30	52
HEV ORF3 WT	TMHMM2.0	outside	53	113
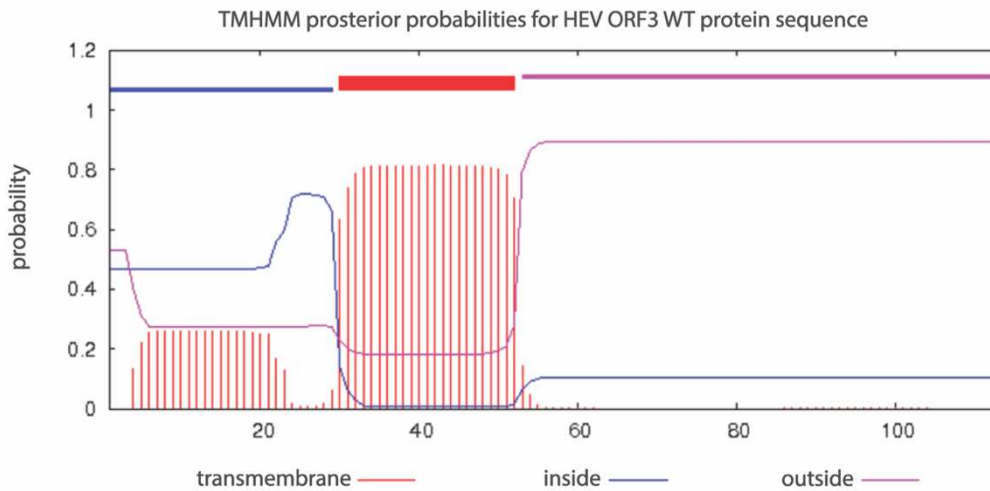
Figure 43. Transmembrane prediction of HEV ORF3 WT protein using the online TMHMM v2.0 server[186].

In Figure 42, the C-terminal region contains many Proline residues, colored in red and constitute the 18.6% of the total sequence (21 Proline residues out of total 113 residues). Analyzing the protein sequence using the online GeneSilico MetaDisorder server[187], the C-terminus is predicted to be disordered – a protein that does not have a stable 3D conformation over the time – as well the first residues in N-terminus (Figure 44).
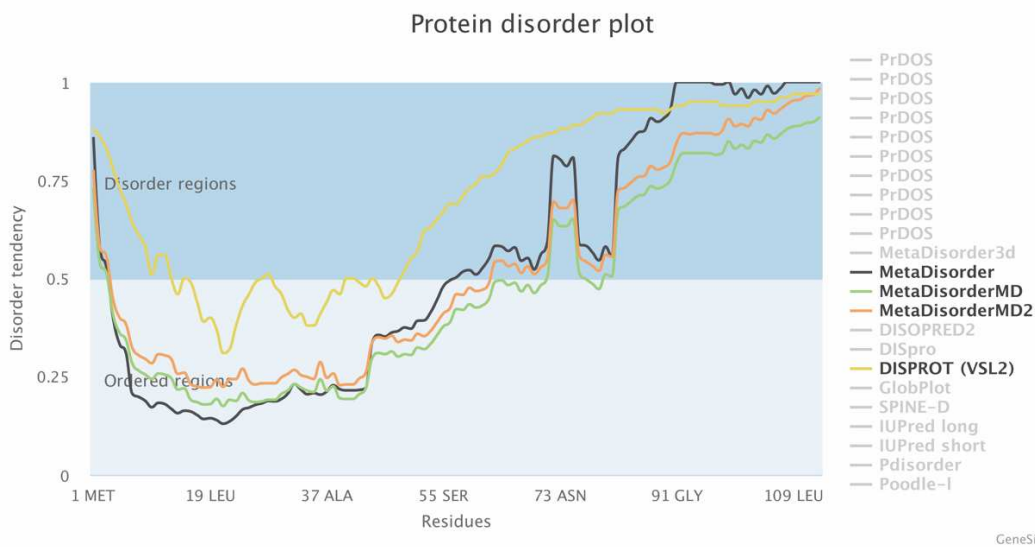
Figure 44. Prediction of the disordered regions of HEV ORF3 WT protein using the online GeneSilico MetaDisorder server[187].

109

In addition, the C-terminal region contains a PSAP motif pointed in a blue circle in Figure 42 which is close to the PTAP motif found in other viruses, such as HIV[146], Ebola[51] etc., and through this motif they interact with the Ubiquitin E2 Variant (UEV) domain of human tumor susceptibility gene 101 (Tsg101) protein, essential protein for the viral secretion.

As mentioned in Material and Methods section, the constructs of HEV ORF3 protein used are the ORF3 WT, the ORF3 Cter (C-terminal region, aa 48-113), the ORF3 C8A (all Cysteines are mutated to Alanine residues), the ORF3 C8S (all Cysteines are mutated to Serine residues), the His-ORF3 C20 and the ORF3 C20 proteins (all Cysteines are mutated to Serine residues except of Cysteine in position 20) with the latter to be the most characterized protein in this study.

1.2 Structure prediction for HEV ORF3 protein

The sequence analysis of the HEV ORF3 protein as mentioned above shows that the C-terminus is predicted to be disordered. A structure prediction is performed in the easy-to-use fast online ColabFold platform, which combines the multiple sequence alignment MMseqs2 (Many-against-Many searching) software with AlphaFold2, for further structural analysis[188–192]. Providing the ORF3 WT protein sequence (113 aa) in the online platform, the prediction was performed in less than 15 min.

Figure 45 illustrates the results of HEV ORF3 disordered protein. Specifically, the MMseqs2 module first provides a multiple sequence alignment as a graph shown in Figure 45a and it will be used by AlphaFold2. The number of found sequences and the sequence identity of the query are represented with a preferable profile of cyan-bluish color at the top and a flatter smooth black line. Regarding the ORF3 protein, the multiple sequence alignment contains low number of sequences on the prediction and the black line fluctuates along the sequence. Next, the AlphaFold2 predicts five (5) models of the protein structure (Figure 45b). For each model, an average predicted Local Distance Difference Test (LDDT) that predicts the quality of the model at each amino acid of the protein is provided in the graph (c) in Figure 45. The predicted LDDT score of the models of ORF3 protein ranges from 40 to 60. An overall AlphaFold prediction with a pLDDT score lower or equal to 50 highlights the fact that there is a high probability that ORF3 protein is disordered. Finally, the Predicted Aligned Error (PAE) plots for each predicted model is provided and gives the distance error for every pair of residues in 0-35 Å range values (Figure 45d). For ORF3 protein models, the PAE plots are colored in red along all the sequence, more than 25 Å distance error for every pair of residues, indicating that the relative positions and/or orientations on all pairs in the structure are uncertain which is expected because of the disorder nature and non-stable 3D conformation over the time.
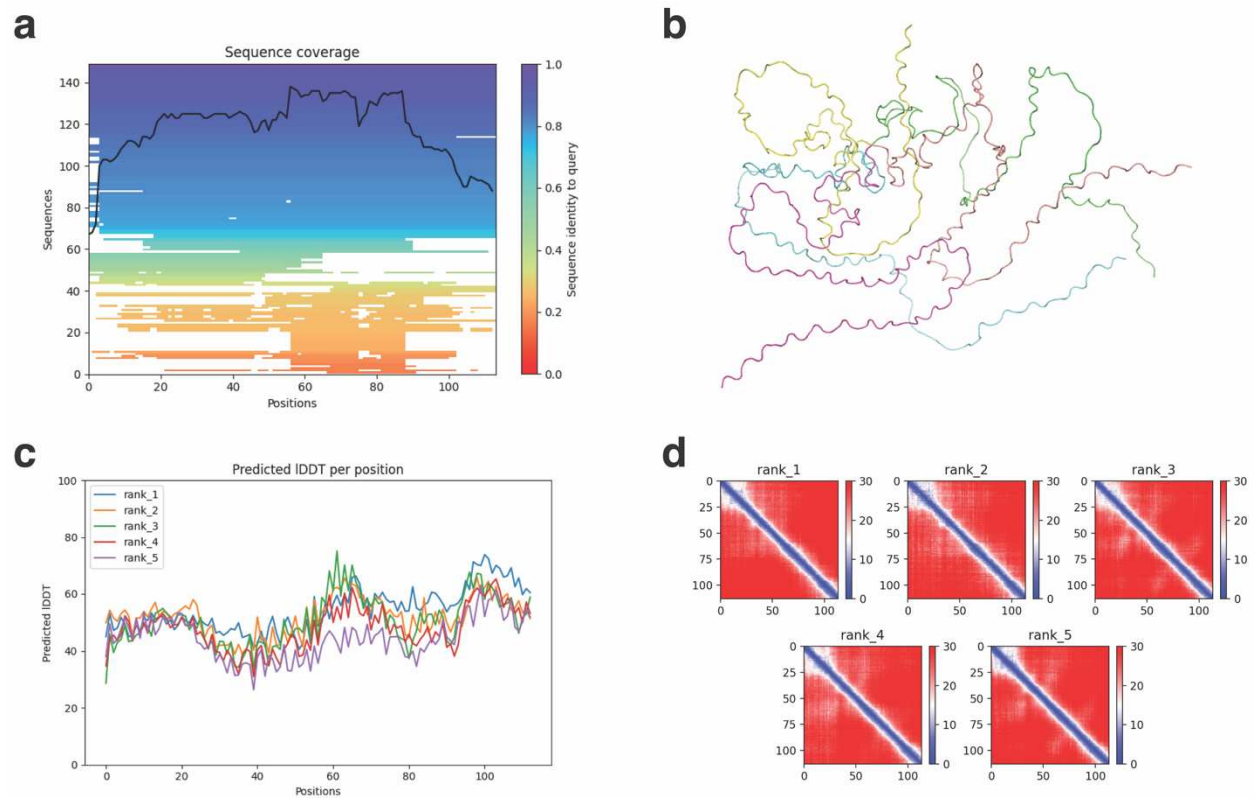
Figure 45. Results of structure prediction by ColabFold[188–192] for ORF3 protein. (a) Multiple Sequence Alignment graphs generated by MMseqs2 module. (b) Predicted (5) models of protein structures by AlphaFold2. (c) Plots of predicted Local Distance Difference Test (LDDT) scores by AlphaFold2. (d) Predicted Aligned Error (PAE) plots by AlphFold2.

1.3 HEV ORF3 protein purification and concentration estimation

Before starting my PhD, preliminary experiments were conducted using the ORF3 Cter protein. This construct was chosen in order to be able to express and purify the protein without the possibility of failure due to the hydrophobic domains in N-terminal region, the cysteine-rich and the transmembrane regions.

In order to work with the full-length protein and also avoid purification issues, the mutation of all Cysteines to Alanines, the ORF3 C8A construct was then used. The ORF3 C8A protein had good expression, but it was soluble only in presence of detergent (2% octyl-β-D-glucopyranoside, β-OG) throughout the purification steps. Specifically, this protein has to be eluted using an imidazole-containing buffer in presence of the detergent and then to dialyze against NMR buffer including 2% β-OG. The main issue is that the amount of the detergent in the final NMR sample was not same among the different protein samples prepared based on 1D NMR spectra recorded. In addition, the quality and the peaks of the 2D ^{1}H, ^{15}N HSQC spectra of ORF3 C8A labeled samples differed concluding that the amount of detergent is important. Because it was impossible to control the exact amount of detergent added to the sample during the dialysis step, the purification of this construct was not performed again.

In order to continue working with the full-length and trying to solve the solubility problem, the Cysteines in N-terminus were mutated to Serine residues, the ORF3 C8S construct. The expression level of this construct was very low. Looking back to the vectors design process, the *E. coli* codon optimization of the DNA sequence of ORF3 C8S differed from the DNA codon sequence of the ORF3 C8A which had good expression level. Therefore, the DNA sequence of ORF3 C8A was used as template and the codon of the amino acids that have to be changed in order to construct the ORF3 C8S is manually modified. In particular, the codon of the eight Alanine residues in the N-terminal, mutations of the WT protein sequence, are replaced by the codon for Serine residues. Using this strategy, the yield of the protein production was recovered in high levels and it was soluble without the presence of the detergent, but this protein was not further used.

The "new" ORF3 C8S DNA sequence is used for the design of the ORF3 C20 construct, the most characterized protein in this study. It contains only one Cysteine residue in position 20 of the protein sequence used to study the membrane anchoring by its palmitoylation. Finally, the ORF3 WT protein is also studied in order to confirm that the ORF3 C20 is a valid protein model.

In Material and Methods section, the expression and purification protocol are described in details and the steps of the final optimum purification procedure for ORF3 protein are shown in Figure 46. Many steps had to be optimized in order to obtain stable ORF3 protein samples. All the ORF3 samples, unlabeled and labeled, have followed the same described procedure. The following data sets are derived from ORF3 C20 protein purifications unless otherwise stated.
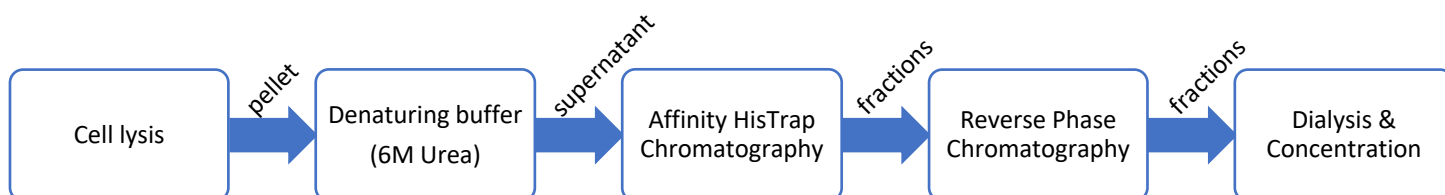
Figure 46. Diagram of the purification steps of ORF3 protein.

The sequence of ORF3 C20 protein is shown in Figure 47. Between the ORF3 sequence and the 6xHis-tag in the C-terminal region, there is a PreScission Protease cleavage site, Leu-Glu-Val-Leu-Phe-Gln-Gly-Pro (LEVLFQGP). All the constructs, apart from the His-ORF3 C20 protein, contain this cleavage site in their sequence. The PreScission Protease specifically cleaves between Glutamine (Q) and Glycine (G) residues of the cleavage site at 4°C in a Cleavage Buffer containing 50 mM Tris-HCl pH 8, 150 mM NaCl, 1 mM EDTA, 1 mM DTT.

Figure 47. Sequence of ORF3 C20 protein. The 6xHis-tag is colored with cyan, the PreScission Protease cleavage site with red and the purple star is the last residue of the ORF3 sequence.

The cleavage of the 6xHis-tag for obtaining the protein as close possible in the nature sequence was performed after the affinity chromatography during overnight dialysis at 4°C against 3 L Cleavage Buffer using 6-8 kDa cut-off membrane. The efficiency of the cleavage was checked using the before and after PreScission Protease addition samples in a 4-20% SDS-PAGE (Figure 48). During the overnight 6xHis-tag cleavage, an initial partial precipitation of the protein was observed. After the Reverse Phase and during the protein concentration step, the cleaved protein was precipitated and was very unstable even at 4°C. Therefore, the 6xHis-tag cleavage is not performed due to the instability of the cleaved protein and it is not included in the purification protocol. The uncleaved ORF3 C20 protein is used in all the following experiments unless otherwise stated.
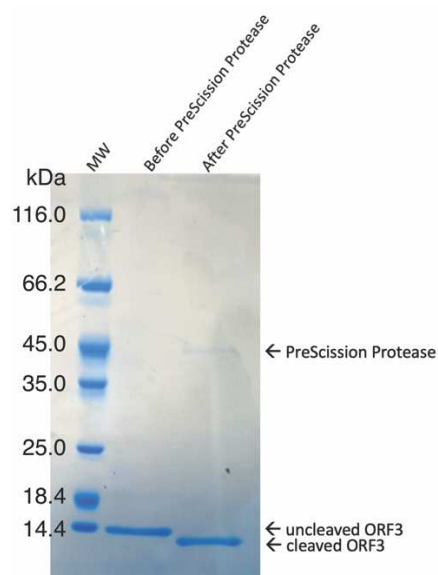
Figure 48. 4-20% SDS-PAGE of PreScission Protease cleavage of ORF3 C20 protein with Coomassie blue staining.

As mentioned in the protocol, the first obstacle that had to be resolved was the insolubility of the protein in the bacterial extract. The resuspension with the denaturing buffer helps to separate ORF3 protein from the *E. coli* cellular debris, but also adds some extra wash steps during the HisTrap affinity purification which can be successful only if the denaturing agent was slowly removed from the column. The chromatogram of the ORF3 C20 HisTrap purification with the three UV absorbance curves at 280 nm, at 260 nm and at 215 nm and the 4-20% SDS-PAGE with different fractions are shown in Figure 49. Since ORF3 protein does not contain any aromatic

residues in its sequence, it is not possible to detect the protein with classical 280 nm wavelength and therefore the unspecific 215 nm wavelength is used to monitor the chromatography.
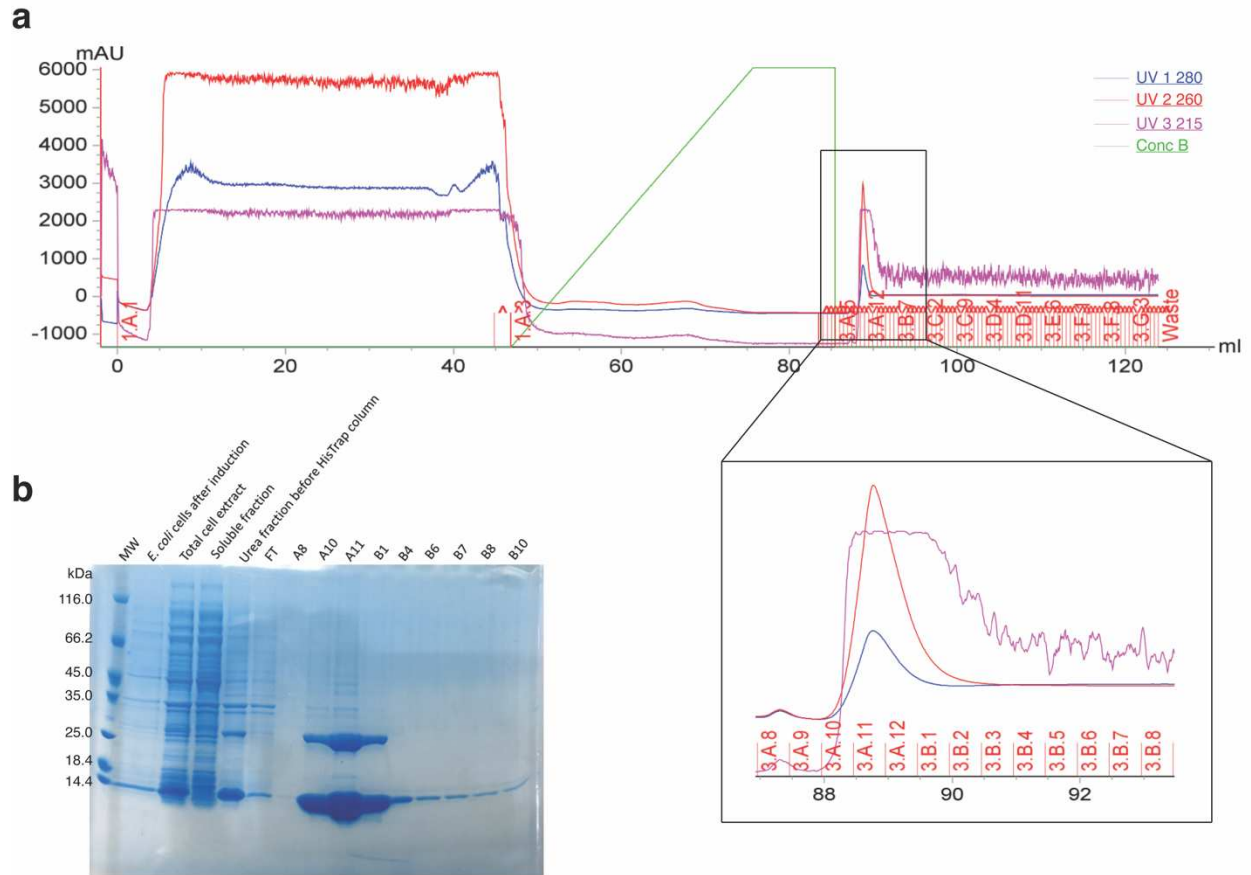
Figure 49. (a) Chromatogram of affinity HisTrap purification of ORF3 C20 protein. In the black box, the elution peak is shown zoomed. Blue: 280 nm, red: 260 nm, magenta: 215 nm and green: concentration of Buffer B. (b) 4-20% SDS-PAGE of fractions based on the chromatogram (a) with Coomassie blue staining.

In order to increase the protein purity, a second step of purification is performed by Reverse Phase (RP) Chromatography. Using a 5 mL loop, all the ORF3 C20 sample is injected on a Zorbax C8 column. Then the protein is eluted with a gradient of acetonitrile. The ORF3 protein is monitored uniquely with the 215 nm absorbance curve. In Figure 50, the chromatogram of the RP purification step as well the 4-20% SDS-PAGE substantiate the purity of the protein sample after the second step of the purification.
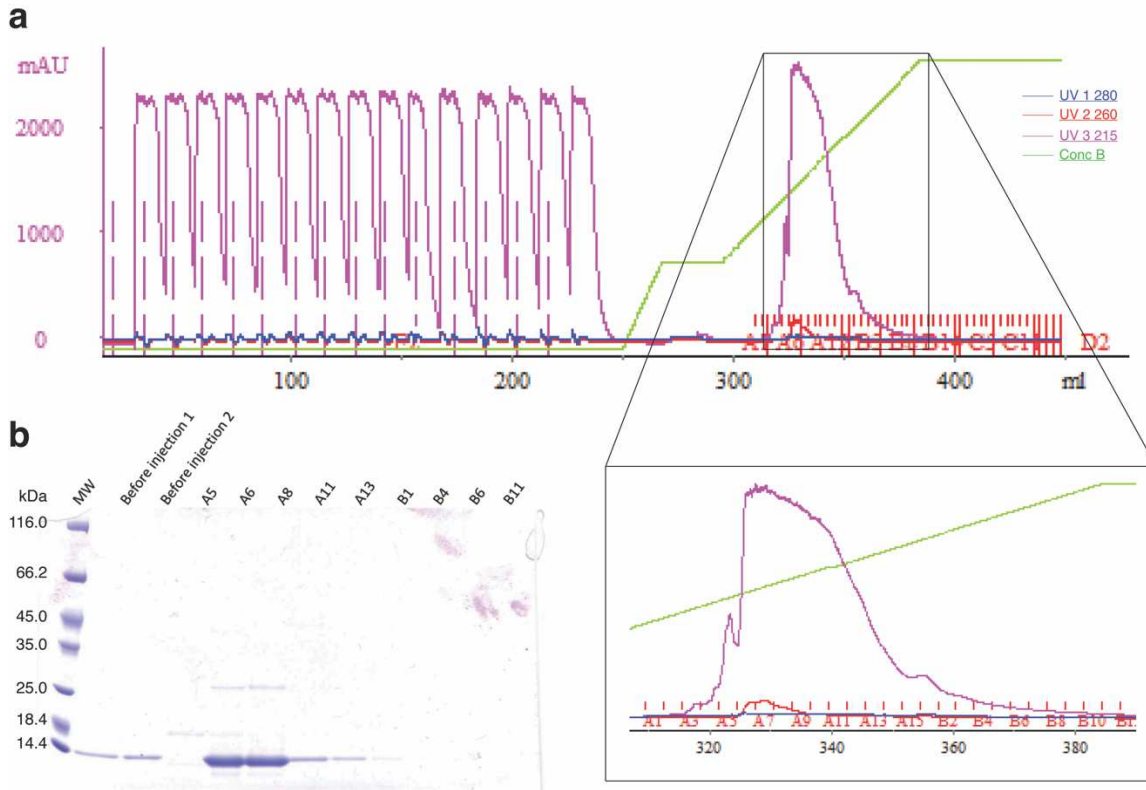
Figure 50. (a) Chromatogram of Reverse Phase purification of ORF3 C20 protein. In the black box, the elution peak is shown zoomed. Blue: 280 nm, red: 260 nm, magenta: 215 nm and green: concentration of Buffer B (0.1% TFA, 80% Acetonitrile). (b) 4-20% SDS-PAGE of fractions based on the chromatogram (a) with Coomassie blue staining.

In order to estimate the protein concentration in the pooled fractions containing the ORF3 C20 protein after the RP chromatography step and because of absence of aromatic residues and thus, unable to use the 280 nm UV absorbance and Beer's Law, a 4-20% SDS-PAGE with increasing volume of the protein is done and the sample bands are compared with the ones of the molecular marker with fixed concentration (Broad Range Protein Molecular Weight Markers, Promega). Specifically, protein samples of 1, 2, 3, 4, 5, 6 and 10 µL were used as well 10 µL of the Promega molecular weight marker with nine identifiable protein bands at 10, 15, 25, 35, 50, 75, 100, 150 and 225 kDa molecular weight. All the proteins of the marker are at 0.1 µg/µL final concentration except of the 50 kDa protein which is in threefold concentration, at 0.3 µg/µL and therefore this band is more intense that the others (Figure 51).
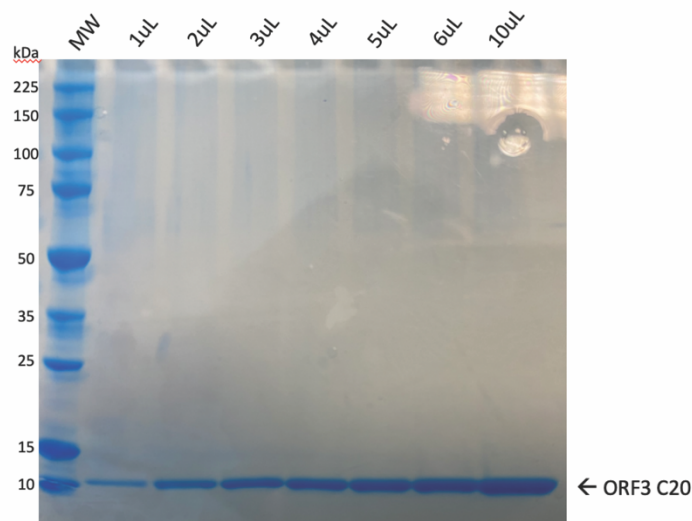
Figure 51. 4-20% SDS-PAGE of ORF3 C20 protein for the concentration' estimation after the Reverse Phase chromatography with Coomassie blue staining. For the Molecular Weight marker (MW), the Broad Range Protein Molecular Weight Markers (Promega) is used.

Moreover, the initial protocol included a lyophilization step after the RP purification taking advantage of having the protein in Acetonitrile buffer which easily evaporates under vacuum conditions. The protein is splitted in 15 mL tubes with specific amount based on the concentration' estimation, flash freeze with liquid Nitrogen and lyophilized. This step was important because the protein was stored in powder form at -20°C for months and thus it could be resuspended in any buffer of interest depending on the experiment in which it would be used. This step was followed in the first purifications until the partial protein resuspension caused problems and losing protein for the following experiments. Although different buffers and different final concentrations of the protein were tested, this main problem could not be resolved and therefore this step was replaced by the dilution and dialysis with NMR Buffer overnight at 4°C directly after the RP chromatography to get rid of the RP buffer components.

In order to be more precise and create a tool for the estimation of the final protein concentration for all future purifications, we performed amino acid analysis determining the exact concentration of the protein sample that was used in the 4-20% SDS-PAGE analysis with increasing volume of the purified ORF3 sample and in the creation of a calibration curve using Bradford reagent. Exceptionally, these experiments were performed for both cleaved and uncleaved ORF3 C20 protein samples derived from the same expression culture. After the affinity

purification step, the protein was splitted in two samples, the first one directly proceeded to the RP chromatography (uncleaved sample) and the second one incubated with PreScission Protease for the 6xHis-tag removal and then passed through the RP column (cleaved sample). For both samples, after the RP purification step, tubes with an estimation of 1 mg of ORF3 were lyophilized. The resuspension of both samples was done with the appropriate amount of 50 mM Sodium Phosphate pH 6.1, 50 mM NaCl buffer to prepare approximative 2 mg/mL stock protein samples. During the resuspension, the proteins could not be completely dissolved with a remaining small pellet for the uncleaved sample and a big pellet for the cleaved sample even after two sonication runs in a heated sonication bath. Using the supernatant for both samples and assuming that the concentration was at 2 mg/mL, three experiments were conducted and described below.

Firstly, the 4-20% SDS-PAGE with increasing volume of both proteins was performed in order to determine if the estimation of 2 mg/mL was correct. Secondly, assuming that both proteins are at 2 mg/mL, the calibration curves with increasing amount of the protein in Bradford reagent were created. Finally, the amino acid analysis for both proteins was conducted by an external laboratory, Chemistry of Biomolecules Unit, CNRS UMR 3523, Department of Structural Biology and Chemistry, Institute Pasteur in Paris, and provides the exact concentration of the protein sample. This technique is based on the detection and quantification of free amino acids of the studied protein after the sample hydrolysis[193].

In Figure 52, the 4-20% SDS-PAGE for uncleaved (a) and cleaved (b) ORF3 C20 protein samples for concentration' estimation are shown. Preparing the protein samples for both proteins as described above and comparing their bands with the ones of the Promega protein marker, the concentration of uncleaved protein is estimated to be around 2 mg/mL as expected while the concentration of cleaved protein is dramatically lower than 2 mg/mL, close to 0.3 mg/mL.
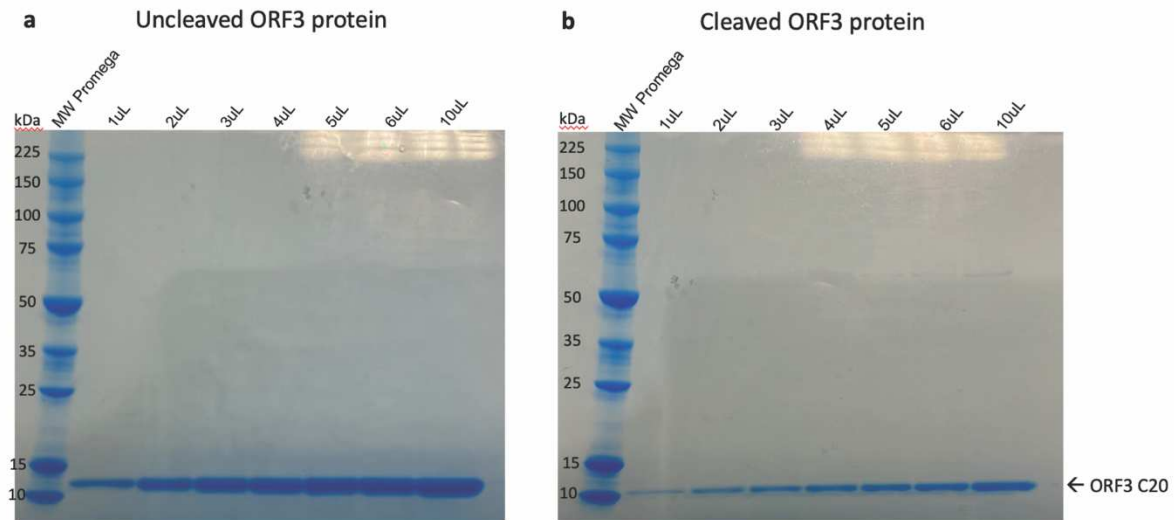
Figure 52. 4-20% SDS-PAGE of (a) uncleaved and (b) cleaved ORF3 C20 protein for the concentration' estimation with Coomassie blue staining. For the Molecular Weight marker (MW), the Broad Range Protein Molecular Weight Markers (Promega) is used.

Meanwhile, the calibration curves for both uncleaved and cleaved proteins using the Bradford reagent were created. Protein samples of 0, 1, 2, 5, 10, 15, 20 and 25 µg were prepared and incubated with 1X Bradford reagent for 5 min at room temperature on a dark place. The samples were transferred in cuvettes and the absorbance at 595 nm was measured in duplicates. Based on the UV measurements, the calibration curve for the uncleaved ORF3 C20 protein was calculated to be $y=0.0076x+0.0278$ ($R^2=0.9488$) and for the cleaved ORF3 C20 protein $y=0.003x+0.0428$ ($R^2=0.8669$). The measurements of the cleaved protein for each point have higher standard deviation, the R^2 of the calibration curve is low and the line is not well-fitted in the measurement points. The reason of the low value could be the wrong assumption of the protein concentration and therefore the measurements are in the lower limit of detection of the absorbance.

The results of the amino acid analysis received couple days later reveal the exact concentration of the protein samples. In Figure 53, the results of uncleaved ORF3 C20 protein show that the concentration is 1.17 mg/mL instead of 2 mg/mL while in Figure 54, the concentration of cleaved ORF3 C20 protein is 0.12 mg/mL.

ANALYSE DES ACIDES AMINES
Unité de Chimie des Biomolécules
Institut Pasteur, 28 rue du Dr Roux, 75724 Paris Cedex 15

6025

Demandeur : **Dr. Xavier HANOULLE, Ph.D.**

Unité : Integrative Structural Biology, CNRS ERL 9002 - INSERM U1167- Institut Pasteur de Lille
xavier.hanoulle@univ-lille.fr
Référence échantillon : ORF3-C20 **1**
Quantité :   25 µL (d'une solution à   2 mg/mL) lyophilisation   nmoles

Analyse demandée :   O   Hydrolyse HCl 6N

                            O   Hydrolyse HCl 6N phénol(Tyr), 20h →   500 µL

                            O   Oxydation performique (Cys, Met)*

                            O   Détermination concentration en protéine (+NorLeu) → 10 nmoles

                                                                        soit   20   µL

Date :11/02/2021

MW : 13038

- Opérateur : Christelle Ganneau
              Françoise Baleux

- Date d'hydrolyse : 11/02/2021

- Date d'analyse : 16/02/2021

- Analyseur : *Hitachi L-8880*

- Colonne : *Hitachi 2620MSC- 4,6x80*

- Volume reprise échantillon : 350 µl

6025   Volume injecté : 50 µl

   Volume injecté :

2,236 nmoles peptide/protéine pour 25 µL
⇨ 89,446 nmoles peptide/protéine par mL
soit   1,17 mg/mL

Théo	AA			
	CysSO3H			
	CMC			
1+1	Asx	D+N	4,3	
	MetSO2			
3	Thr	T	5,2	
20	Ser	S	18,6	
2+3	Glx	E+Q	9,1	
13	Gly	G	14,1	
12	Ala	A	12,8	
1	Cys *	C	nd	
10	Val	V	9,9	
1	Met *	M	nd	
4	Ile	I	4,5	
17	Leu	L	17,5	
	NorLeu	REEL	1,389	
		THEO	1,429	
	Tyr	Y		
5	Phe	F	**5**	
	ßAla/GlucoNH2			
	Lys	K	1,2	
8	His	H	7,0	
	NH3			
6	Arg	R	4,6	
22	Pro	P	21,7	
	Trp	W		
	GalNHAc			

Figure 53. Results of Amino Acid Analysis of uncleaved ORF3 C20 protein conducted by an external laboratory, Chemistry of Biomolecules Unit, CNRS UMR 3523, Department of Structural Biology and Chemistry, Institute Pasteur in Paris.

ANALYSE DES ACIDES AMINES

Unité de Chimie des Biomolécules
Institut Pasteur, 28 rue du Dr Roux, 75724 Paris Cedex 15

6026

Demandeur : **Dr. Xavier HANOULLE, Ph.D.**
Unité : Integrative Structural Biology, CNRS ERL 9002 - INSERM U1167- Institut Pasteur de Lille
xavier.hanoulle@univ-lille.fr
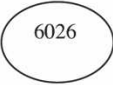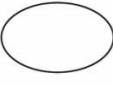
Référence échantillon : ORF3-C20 **2**
Quantité :   140 µL (d'une solution à   0,3 mg/mL) Lyophilisation   nmoles

Analyse demandée :   O   Hydrolyse HCl 6N

O   Hydrolyse HCl 6N phénol(Tyr), 20h →   500 µL

O   Oxydation performique (Cys, Met)*

O   Détermination concentration en protéine (+NorLeu) → 10 nmoles

soit   20 µL

Date :11/02/2021

MW : 12004

- Opérateur : Christelle Ganneau
              Françoise Baleux

- Date d'hydrolyse : 11/02/2021

- Date d'analyse : 16/02/2021

- Analyseur : *Hitachi L-8880*

- Colonne : *Hitachi 2620MSC- 4,6x80*

- Volume reprise échantillon : 350 µl

6026   Volume injecté : 50 µl

          Volume injecté :

1,351 nmoles peptide/protéine pour 140 µL
⇨ 9,652 nmoles peptide/protéine par mL
soit   0,12 mg/mL

Théo	AA			
	CysSO3H			
	CMC			
1+1	Asx	D+N	3,4	
	MetSO2			
3	Thr	T	3,2	
20	Ser	S	16,4	
2+3	Glx	E+Q	6,2	
11	Gly	G	12,5	
12	Ala	A	13,4	
1	Cys *	C	nd	
10	Val	V	9,7	
1	Met *	M	nd	
4	Ile	I	3,8	
17	Leu	L	19,3	
	NorLeu	REEL	1,449	
		THEO	1,429	
	Tyr	Y		
5	Phe	F	**5**	
	ßAla/GlucoNH2			
	Lys	K	1,2	
2	His	H	1,4	
	NH3			
6	Arg	R	6,9	
21	Pro	P	21,8	
	Trp	W		
	GalNHAc			

Figure 54. Results of Amino Acid Analysis of cleaved ORF3 C20 protein conducted by an external laboratory, Chemistry of Biomolecules Unit, CNRS UMR 3523, Department of Structural Biology and Chemistry, Institute Pasteur in Paris.

Using the real exact concentration of the protein samples, the calibration curves were calculated again having as a result the calibration curve for uncleaved ORF3 C20 protein to be $y=0.0163x+0.0278$ ($R^2=0.9488$) and for cleaved ORF3 C20 protein to be $y=0.0504x+0.0428$ ($R^2=0.8669$) (Figure 55).
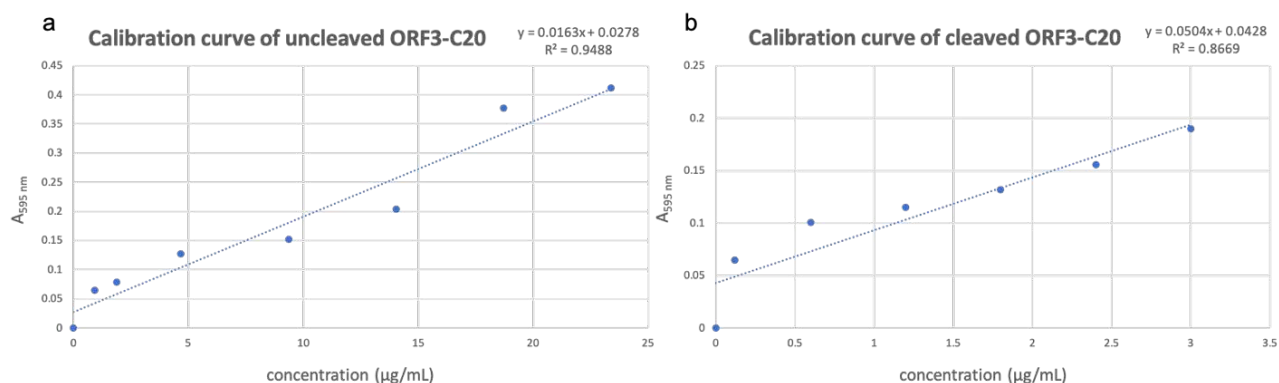
Figure 55. Calibration curves for (a) uncleaved and (b) cleaved ORF3 C20 protein using Bradford reagent.

Because the 6xHis-tag cleavage is not performed in the future experiments, only the Bradford curve for uncleaved ORF3 protein combined with the results of 4-20% SDS-PAGE with increasing volume of ORF3 sample in each purification is used for the determination of the protein concentration.

Another observation during the concentration step of the uncleaved ORF3 protein sample was that when the concentration exceeds the value of 200 µM, the protein starts to precipitate concluding that there is a concentration limit at ~200 µM (~2.6 mg/mL). In few cases, it was also observed an inexplicable phenomenon. The addition of 3 mM THP on the protein has as a result the fogginess of the sample that was disappeared after its incubation for couple of days at 4°C without any loss of protein based on the absorbance at 595 nm with Bradford reagent after a high-speed centrifuge run (16,100 xg) for 10 min at 4°C (Figure 56).
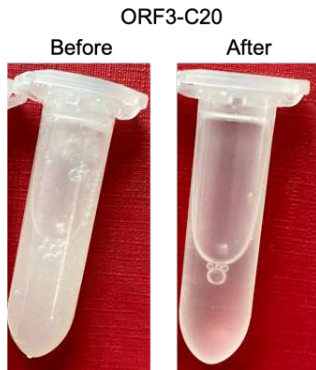
Figure 56. ORF3 C20 protein at 190 µM after concentration step before and after incubation for couple of days at 4ºC.

The only difference in the purification procedure of the ORF3 WT protein is the amount of DTT used in purification buffers. Because it includes eight Cysteine residues in the N-terminal region in total, 5 mM DTT final concentration instead of 1 mM DTT is used in order to reduce all the Cysteines during its purification. For the purification of ORF3 Cter and the His-ORF3 C20 proteins, the protocol was followed without any changes in any step.

To conclude, we managed to express HEV ORF3 protein in *E. coli* and successfully purify unlabeled and labeled protein samples obtaining a yield of ~22-25 mg, ~17-19 mg and ~13 mg per L of culture for unlabeled, ^{15}N labeled and ^{15}N, ^{13}C double labeled samples, respectively.

1.4 NMR analysis of HEV ORF3 protein constructs

In order to performed a structural and functional characterization of HEV ORF3 we mainly used solution-state NMR Spectrometry. The preparation of the labeled protein sample, the acquisition of the NMR experiments, the analysis of the NMR data and finally the structural determination are the main steps for NMR structural characterization. The first step, the preparation of the sample was described in the previous chapter. Here, the NMR analysis of the ORF3 protein constructs is described in details.

<u>ORF3 Cter protein</u>

The NMR analysis of the shortest construct, the ORF3 Cter (residues T48 to R113), is performed using a 300 μM ^{15}N, ^{13}C ORF3 Cter labeled uncleaved protein sample at 293K on 900 MHz Spectrometer before starting my PhD from other lab members. For backbone and proline assignments, 2D and 3D classical and carbon detection NMR spectra have to be recorded and especially a complete NMR dataset containing the 2D ^{1}H, ^{15}N HSQC, 3D ^{1}H, ^{15}N, ^{13}C HNCO, HNCACO, HNCACB, HN(CO)CACB, HN(CA)NNH, HNHA, HACAN and 2D carbon detection ^{15}N, ^{13}C NCO spectra was collected. The 2D ^{1}H, ^{15}N HSQC spectrum is the first and main 2D experiment recorded and is considered the "fingerprint" of the protein because it is unique for each protein. It provides the correlation between the nitrogen (^{15}N) and the amide proton (^{1}H) of the protein. Therefore, each resonance or peak of the 2D ^{1}H, ^{15}N HSQC spectrum corresponds to a residue of the ORF3 Cter sequence except of Prolines. In addition, between each 3D experiment, a HSQC spectrum is acquired to check the stability and the quality of the sample.

For backbone protein assignments, the standard 2D ^{1}H, ^{15}N HSQC spectrum is first recorded. All the nitrogen (^{15}N) and proton (^{1}H) correlations are shown in the spectrum, which are mainly backbone amide groups, but also the side chain of Asparagine (Asn), Glutamine (Gln) and Tryptophan (Trp) residues are visible. The 3D HNCO spectrum is the most sensitive triple-resonance experiment and provides the backbone carbonyl (^{13}CO), amide (^{15}N) and amide proton (^{1}H) correlations. In each NH HNCO strip, only the carbonyl group of the preceding residue (CO$_{i-1}$) is detected. This spectrum is used in conjunction with 3D HNCACO in which two carbonyl

groups are visible for each NH strip in different intensities, one of the same residue (CO_i) and one for the preceding one (CO_{i-1}). The most intense peak belongs to the former (CO_i) and can be distinguished comparing the peaks on a NH strip of the two spectra. The next experiment, the 3D HNCACB is very important spectrum for the protein assignment while in each NH strip two sets of Cα and Cβ peaks are usually visible, the Cα_i and Cβ_i of the same residue that are also more intense and the Cα_{i-1} and Cβ_{i-1} of the preceding one. The Cα_i and Cβ_i resonances of the same residue are always detectable in each NH strip of the spectrum and the distinction from the resonances of the preceding residue is achieved by using the 3D HN(CO)CACB spectrum which provides only the latter. The 3D HN(CA)NNH spectrum provides the correlation of the backbone amide group of the same residue (N_i) with the amide group of the preceding (N_{i-1}) and the following (N_{i+1}) residues. In the 3D HNHA spectrum, two peaks are found in each NH strip corresponded to the Hα_i of the same residue and the Hα_{i-1} of the preceding residue. The 3D HACAN spectrum provides the correlation between the amide group of the same residue (N_i) and the carbon Cα_{i-1} and proton Hα_{i-1} of the preceding residue and is secondarily used to obtain more information.

Based on the sequence analysis of the ORF3 protein, Proline residues constitutes the 18.6% of the full protein sequence and for ORF3 Cter construct the 25.8% of the sequence. Therefore, the Proline assignment is significant for the study of these proteins. Consequently, the 2D carbon detection ^{15}N, ^{13}C NCO spectrum is recorded and provides the correlation between the amide (N_i) of the same residue and the backbone carbonyl (CO_{i-1}) of the preceding residue. In this experiment, the Proline residues are visible in ^{15}N region ~132-142 ppm. Combined the 2D NCO spectrum with either the 3D HNCO or HNCACO experiment, the Proline residues could be assigned.

Based on the information provided on all these experiments, the backbone and proline assignments are achieved for ORF3 Cter protein.

In Figure 57, the assigned 2D ^{1}H, ^{15}N HSQC spectrum of ORF3 Cter (T48-R113) protein is shown using the numbering of the ORF3 protein sequence. In addition, the assigned pairs of NH_2 side chains (sc) of all Asn and Gln residues are depicted in the 2D spectrum.
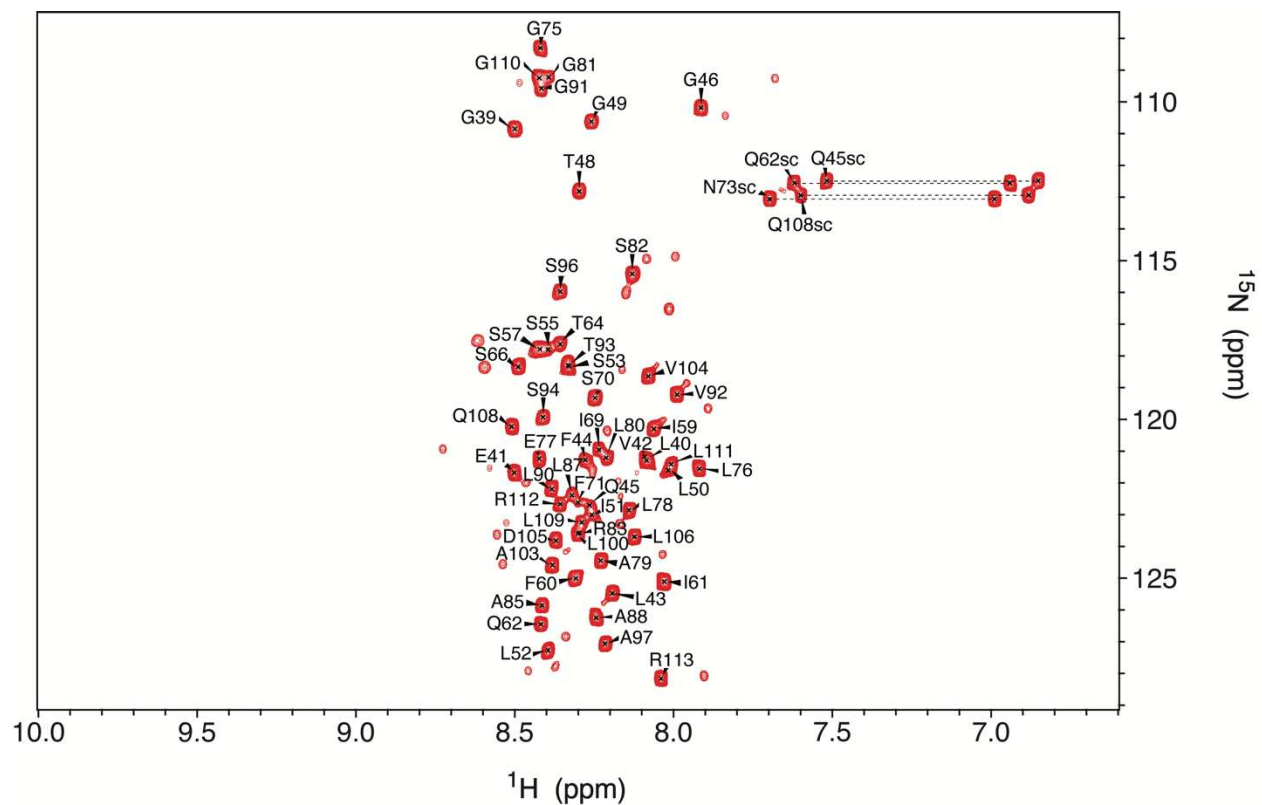
Figure 57. 2D ¹H, ¹⁵N HSQC assigned spectrum of ¹⁵N, ¹³C ORF3 Cter protein. The NH₂ side chains (sc) of all Asn and Gln residues are assigned and their pairs are labeled with dashed line.

The spectrum has a narrow dispersion in its proton dimension that is characteristic for the disordered proteins. This is the first experimental evidence that the C-terminal region of the protein is disordered.

All Prolines are assigned as shown in ¹⁵N region 134-140 ppm of the 2D carbon detection ¹⁵N, ¹³C NCO spectrum in Figure 58.
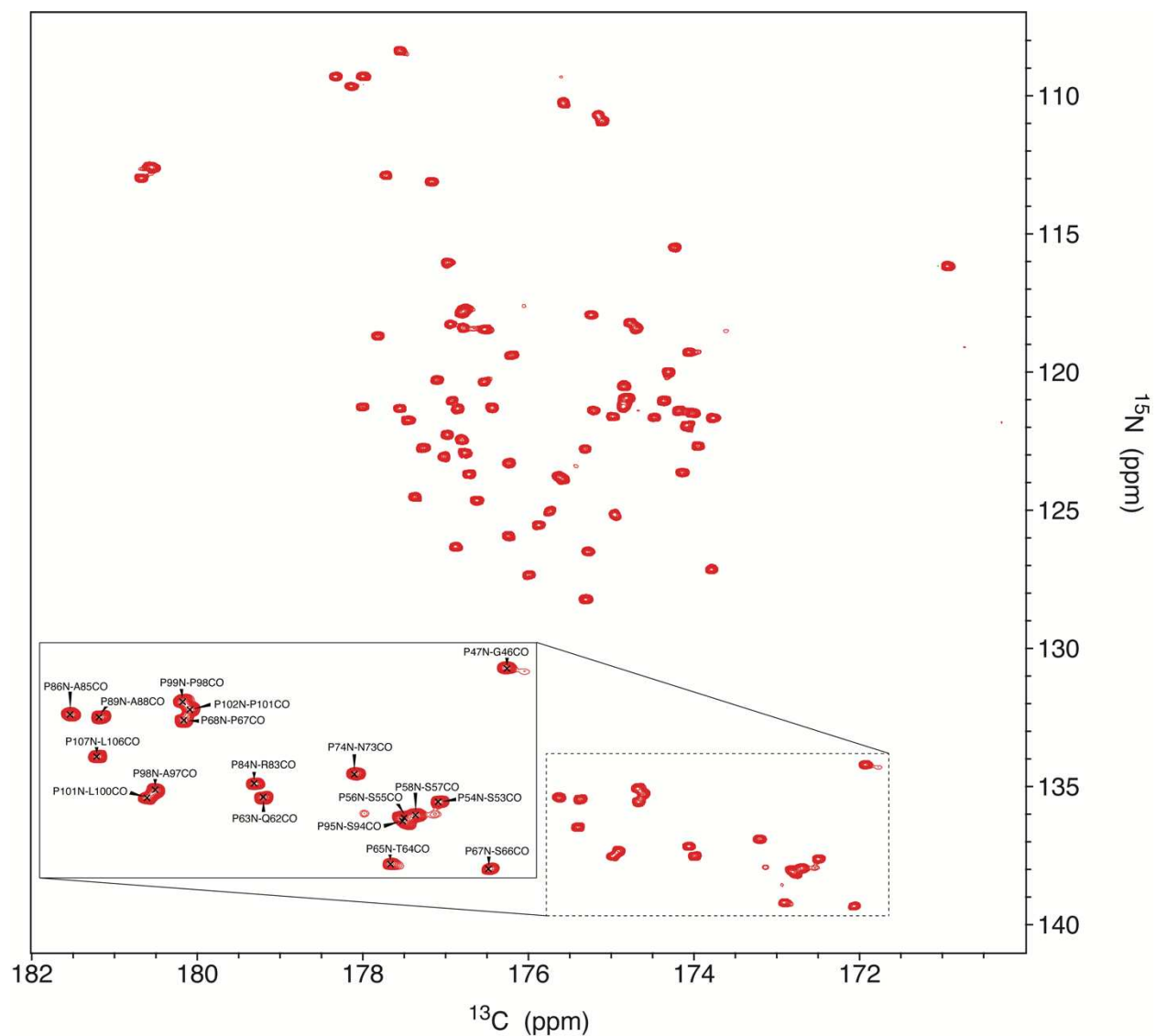
Figure 58. 2D carbon detection ^{15}N, ^{13}C NCO spectrum of ^{15}N, ^{13}C ORF3 Cter protein. All 18 Prolines residues are assigned shown in the zoomed box.

The ORF3 Cter protein sequence contains 87 residues in total of which the 18 are Proline residues (25.8% of the sequence). In the 2D ^{1}H, ^{15}N HSQC spectrum, 55 residues are assigned out of 69 remaining residues. The ones that could not be detected are the first 12 residues in the N-terminal region as well the Histidine in position 72 (His72) and the Asparagine in position 73 (Asn73).

Based on all the spectra, 56 out of 69 $^{1}H^N$ resonances (81%), 71 out of 87 $^{1}H\alpha$ resonances (82%), 75 out of 87 ^{15}N resonances (86%), 72 out of 87 $^{13}C\alpha$ resonances (83%), 64 out of 79 $^{13}C\beta$ resonances (81%), 76 out of 87 ^{13}CO resonances (87%) and 2 out of 8 $^{1}H\alpha_2$ resonances (25%) for

Gly residues are assigned in total. Regarding to the side chain assignment of Asn and Gln residues, the $^{13}C\gamma$ of all Asn and Gln residues, the $^{13}C\delta$, $^{15}N\epsilon$ and $^{1}H\epsilon$ of all Gln and the $^{15}N\delta$ and $^{1}H\delta$ of the Asn resonances are achieved.

Figure 59 indicates the assigned (in black) and unassigned (in red) residues of ORF3 Cter protein.

ORF3 Cter

```
      36          46    *    56          66          76          86
MGSSHHHHHH SSGLEVLFQG PTGLILSPSP SPIFIQPTPS PPISFHNPGL ELALGSRPAP

      96          106
LAPLGVTSPS APPLPPAVDL PQLGLRR
```

Figure 59. ORF3 Cter protein sequence with assigned (in black) and unassigned (in red) residues. The purple star shows the first residue of the actual protein sequence.

ORF3 C8A protein

A 100 μM ^{15}N ^{13}C ORF3 C8A labeled protein in the presence of 2% β-OG detergent was prepared in order to record 2D and 3D NMR experiments for protein backbone and proline assignment purposes before starting my PhD from other lab members. The sample was place on a Shigemi tube and 2D and 3D dataset were recorded at 293K on 900 MHz Spectrometer.

Although the resolution and the sensitivity of the first 2D ^{1}H, ^{15}N HSQC spectrum were satisfying and proceeding to the 3D dataset, the sample was precipitated over the time and thus, only the 3D HNCO and HNCACB spectra were recorded with low signal-to-noise ratio. Therefore, a new 110 μM ^{15}N ^{13}C ORF3 C8A labeled sample was prepared to record again the NMR experiments. However, the 2D ^{1}H, ^{15}N HSQC spectrum of this sample had lower quality with broad line widths and lower peak intensities. The overlay of the 2D ^{1}H, ^{15}N HSQC of the two ^{15}N ^{13}C ORF3 C8A protein samples is shown in Figure 60.
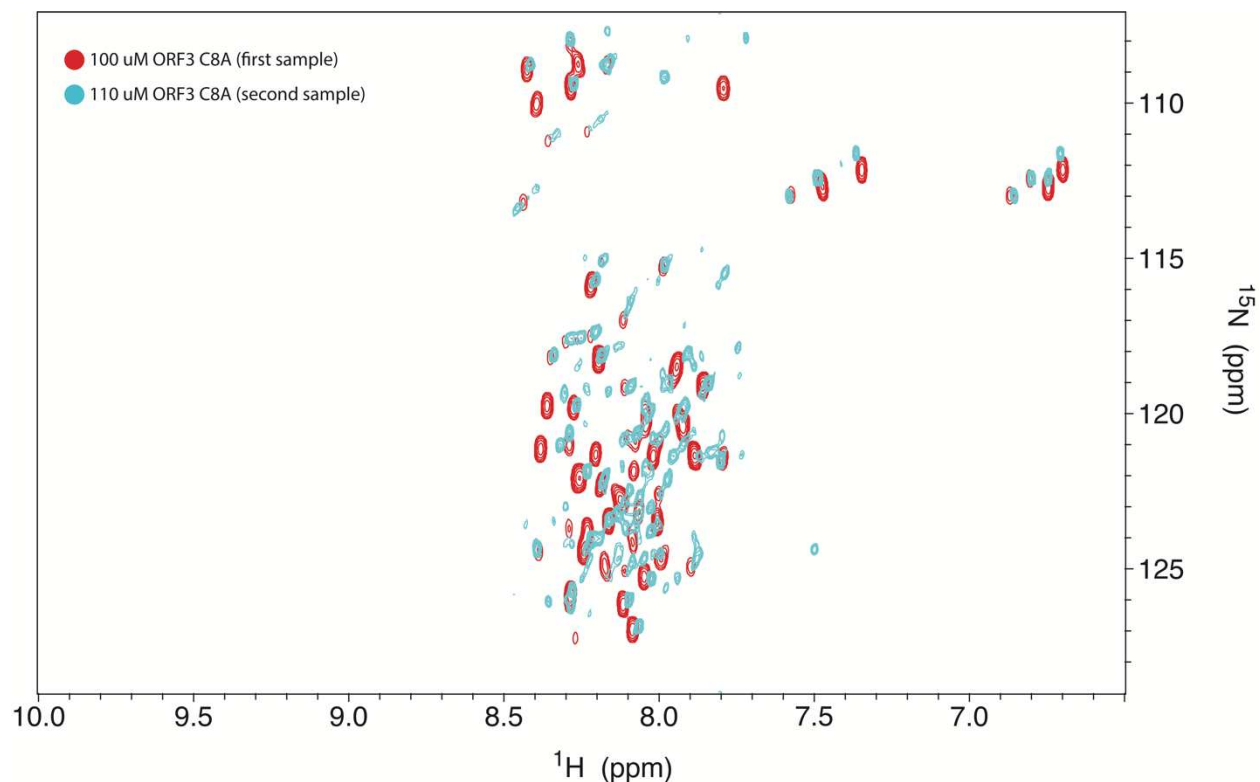
Figure 60. Overlay of 2D ^1H, ^{15}N HSQC spectra of 100 µM ^{15}N, ^{13}C ORF3 C8A protein (first sample) in red and 110 µM ^{15}N, ^{13}C ORF3 C8A protein (second sample) in cyan.

This difference of the spectrum quality was potentially related to the difference of the concentration of the detergent on the sample and consequently in sample homogeneity. Figure 61 depicts the overlay of the 1D spectra for the two samples the zoomed boxes corresponded to the β-OG detergent peaks. The three group peaks at 0.85-1.65 ppm region, specially the first at 0.85-0.88 ppm, second one at 1.25-1.37 ppm and the third one at 1.61-1.65 ppm region, and the peak at 5.8 ppm are characteristic for octyl-β-D-glucopyranoside detergent[194]. Based on the overlay of the 1D spectra, the first ORF3 C8A labeled sample in red is in threefold concentration of the detergent while in the second protein sample in cyan the peak at 5.8 ppm barely exists.
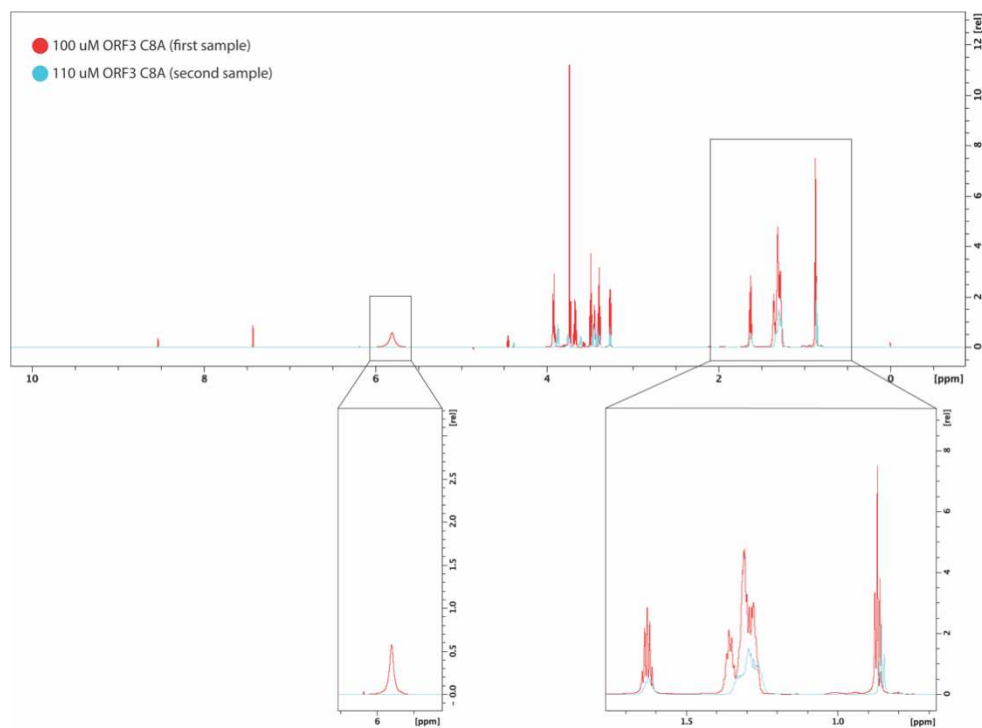
Figure 61. Overlay of 1D spectra of 100 μM ^{15}N, ^{13}C ORF3 C8A protein (first sample) in red and 110 μM ^{15}N, ^{13}C ORF3 C8A protein (second sample) in cyan.

The second labeled sampled had lower concentration of the detergent, the protein was less stable and the 2D HSQC spectrum had lower quality. Based on these results, higher concentration of the detergent in the sample, better quality NMR spectra are recorded. Because of the difficulty to control with precision the amount of the detergent in the protein sample and the preliminary results obtained showed that the sample homogeneity and stability are utmost importance for qualitative NMR data sets, the ORF3 C8A protein construct was not further studied. However, the 2D ^{1}H, ^{15}N HSQC spectrum of ORF3 C8A protein has a narrow dispersion in its proton dimension that is characteristic for the disordered proteins, but it had to be further investigated.

ORF3 C20 protein

In this study, the ORF3 C20 protein is the most biophysically and structurally characterized. This construct was selected for further analysis because it is close to the ORF3 WT protein sequence containing one Cysteine (in position 20) needed for further membrane anchoring studies and the remaining 7 Cysteines are mutated to Serine residues.

A ^{15}N, ^{13}C ORF3 C20 double-labeled uncleaved protein sample at 100 µM is prepared for acquiring the NMR experiments needed for backbone and proline assignments at 293K on 600 MHz Spectrometer. A complete NMR dataset containing the 2D ^{1}H, ^{15}N HSQC, 3D ^{1}H, ^{15}N, ^{13}C HNCO, HNCACO, HNCACB, HN(CO)CACB, HNHA, HACAN, HCBCACON, HACONH, (H)NCANNH, (H)NCOCANNH, H(N)CANNH, H(N)COCANNH and 2D carbon detection ^{15}N, ^{13}C NCO spectra is recorded. Apart from the spectra acquired for ORF3 Cter assignments, the 3D (H)NCANNH, (H)NCOCANNH, H(N)CANNH, H(N)COCANNH, HACONH and HCBCACON spectra helped to assign more resonances for ORF3 C20 protein. The 3D HACONH spectrum provides the correlation of the backbone amide group (N_i) and the amide proton (H_i) of the same residue with the proton ($H\alpha_{i-1}$) of the preceding one. The 3D HCBCACON spectrum provides the correlation of the backbone amide group of the same residue (N_i) with the carbon $C\alpha_{i-1}$ and $C\beta_{i-1}$ and the proton ($H\alpha_{i-1}$) of the preceding residue. The 3D (H)NCANNH spectrum provides the correlation of the backbone amide group of the same residue (N_i) with the amide group of the preceding (N_{i-1}) and the following (N_{i+1}) residues. This spectrum is used in conjunction with the 3D (H)NCOCANNH spectrum in which only the correlation with the amide group of the following (N_{i+1}) residue is visible. The 3D H(N)CANNH spectrum provides the correlation of the backbone amide group of the same residue (N_i) with the amide proton of the preceding (H_{i-1}) and the following (H_{i+1}) residues. This spectrum is used in conjunction with the 3D H(N)COCANNH spectrum in which only the correlation with the amide proton of the following (H_{i+1}) residue is visible.

Combining the information of all these experiments, the backbone and proline assignments are achieved for ORF3 C20 protein. In Figure 62, the assigned 2D ^{1}H,^{15}N HSQC spectrum of uncleaved ORF3 C20 protein is depicted and the assigned pairs of NH$_2$ side chains (sc) of all Asn and Gln residues are represented.
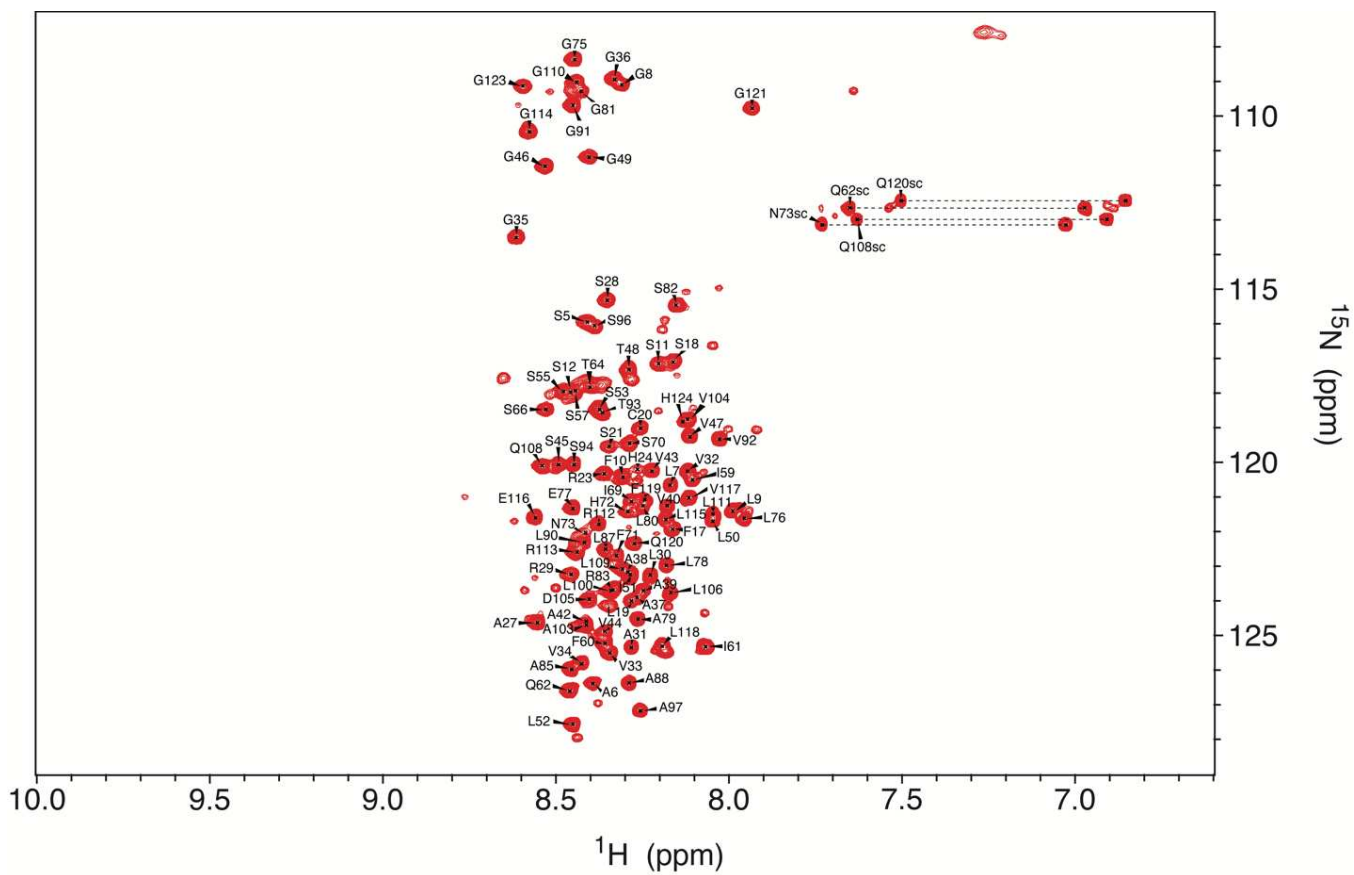
Figure 62. 2D ¹H, ¹⁵N HSQC assigned spectrum of ¹⁵N, ¹³C ORF3 C20 protein. The NH₂ side chains (sc) of all Asn and Gln residues are assigned and their pairs are labeled with dashed line.

The 2D HSQC spectrum of ORF3 C20 protein has also a narrow dispersion in its proton dimension and thus, it suggests that the full-length ORF3 C20 protein is mainly disordered without excluding the possibility of the presence of helical segments.

All 22 Prolines of ORF3 C20 protein are assigned as shown in ¹⁵N region 133-140 ppm of the 2D carbon detection ¹⁵N, ¹³C NCO spectrum in Figure 63.
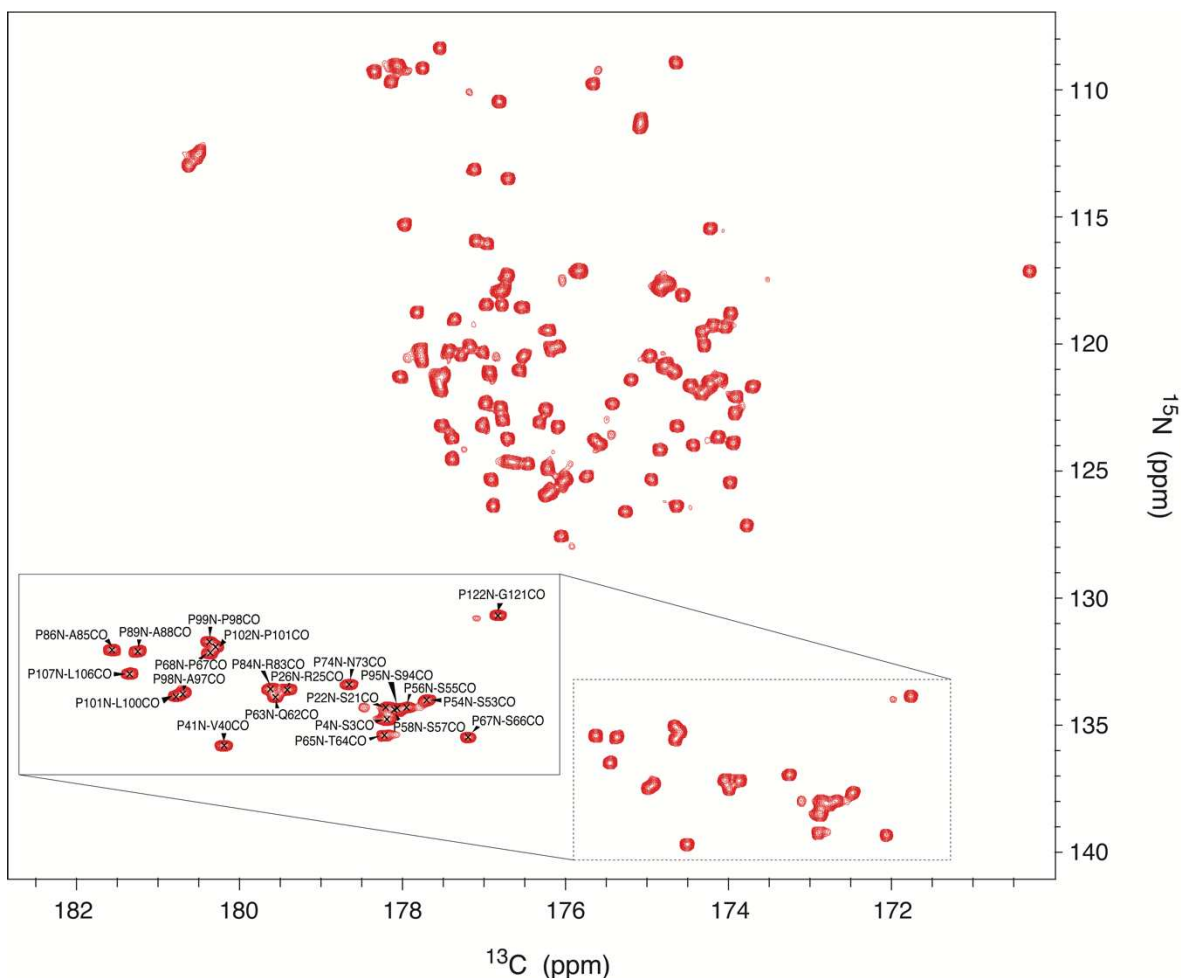
Figure 63. 2D carbon detection ^{15}N, ^{13}C NCO spectrum of ^{15}N, ^{13}C ORF3 C20 protein. All 22 Prolines residues are assigned shown in the zoomed box.

The ORF3 C20 protein sequence contains 129 residues in total of which the 22 are Proline residues. In the 2D ^1H, ^{15}N HSQC spectrum, 94 residues are assigned out of 107 remaining residues. Apart from the first residue of the sequence (Met1) that it is not detectable in the HSQC spectrum because of the fast exchange of the NH^{3+} protons of the N-terminal with water molecules, the Gly2, Ser3 and the Ser13-Ser16 of the Serine stretch in the N-terminal region, the Arg25 and the 5 Histidine residues of the 6xHis-tag (His125-His129) in the C-terminus could not be assigned.

Based on all the recorded assigned spectra, 99 out of 107 $^1H^N$ resonances (93%), 114 out of 129 $^1H\alpha$ resonances (88%), 116 out of 129 ^{15}N resonances (90%), 112 out of 129 $^{13}C\alpha$ resonances (87%), 98 out of 116 $^{13}C\beta$ resonances (84%), 117 out of 129 ^{13}CO resonances (91%) are assigned

in total. Regarding to the side chain assignment of Asn and Gln residues, the ^{13}Cγ of all Asn and Gln residues, the ^{13}Cδ, ^{15}Nε and ^{1}Hε of all Gln and the ^{15}Nδ and ^{1}Hδ of the Asn resonances are achieved. In addition, concerning the Prolines residues and their side chain assignments, 11 out of 22 ^{13}Cδ and ^{1}Hδ resonances (50%) are achieved using the 3D HACAN spectrum.

Figure 64 indicates the assigned (in black) and unassigned (in red) residues of ORF3 C20 protein.

ORF3 C20

```
         10         20         30         40         50         60
MGSPSALGLF  SSSSSSFSLC  SPRHRPASRL  AVVVGGAAAV  PAVVSGVTGL  ILSPSPSPIF

         70         80         90        100        110      *  120
IQPTPSPPIS  FHNPGLELAL  GSRPAPLAPL  GVTSPSAPPL  PPAVDLPQLG  LRRGLEVLFQ


GPGHHHHHH
```

Figure 64. ORF3 C20 protein sequence with assigned (in black) and unassigned (in red) residues. The purple star shows the last residue of the actual protein sequence.

Figure 65 illustrates the overlay of the 2D ^{1}H, ^{15}N HSQC spectra of ^{15}N, ^{13}C ORF3 Cter (in red) and ^{15}N, ^{13}C ORF3 C20 (in cyan) proteins. Based on this overlay, we could conclude that the disordered nature of ORF3 Cter is independent from the hydrophobic N-terminal region of the protein.
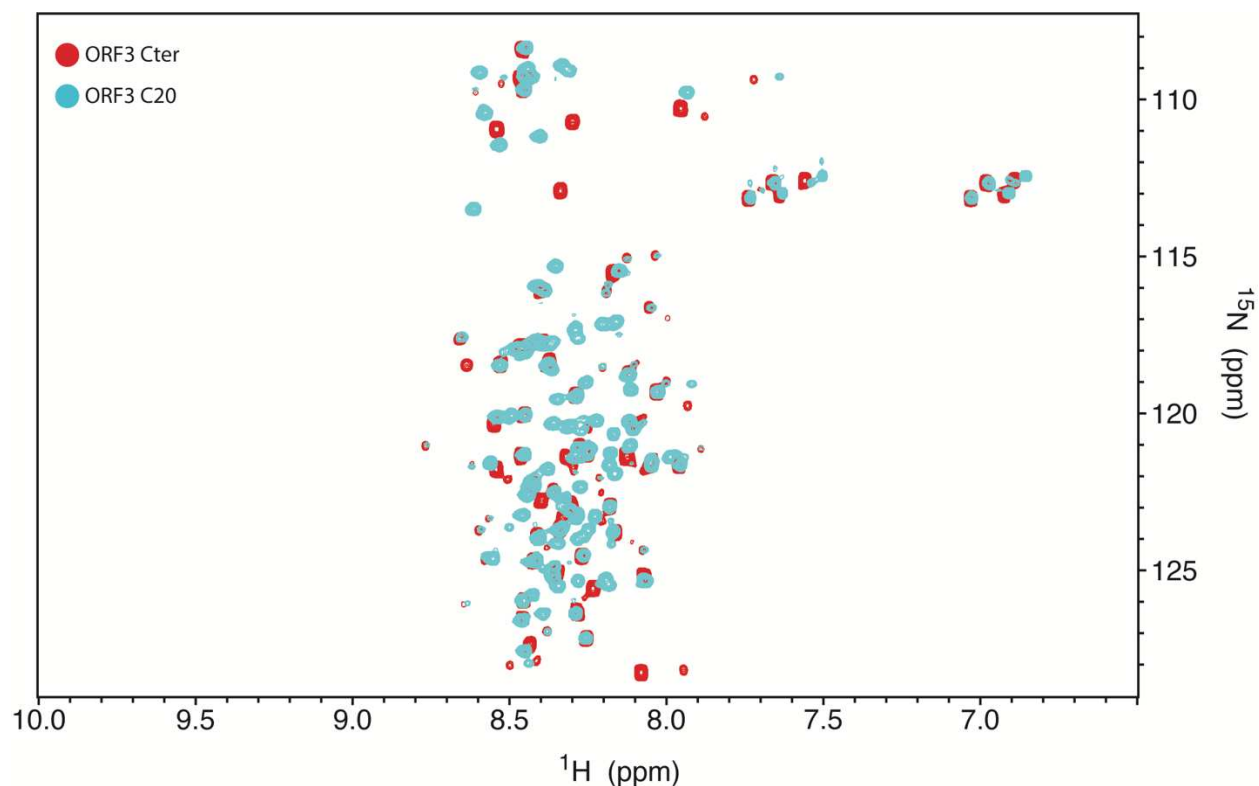
Figure 65. Overlay of 2D ^1H, ^{15}N HSQC spectra of ^{15}N, ^{13}C ORF3 Cter protein in red and ^{15}N, ^{13}C ORF3 C20 protein in cyan.

The experimental NMR chemical shifts contain information on the secondary structure elements of the protein. Disordered proteins do not adopt stable 3D structure over time, they rather exist as a mixture of dynamic conformers that can contain residual secondary structures, regions that adopt helical or extended structure part of the time. In order to gain more information about the residual secondary structure of ORF3 C20 protein, the Secondary Structure Propensities (SSP) analysis is used to estimate the fraction of alpha-helices and β-strand or extended region along the protein sequence[195]. This method combines the experimental NMR chemical shifts from different nuclei into a single secondary structure propensity (SSP) score at a given residue[195]. In the SSP analysis, positive and negative scores, ranged from 1 to -1, indicate helical and extended regions propensities, respectively. Although for well-folded proteins all the available chemical shifts (^{15}N, $^1H^N$, $^{13}C\alpha$, $^{13}C\beta$, ^{13}CO and $^1H\alpha$) are used, for disordered proteins, the experimental $^{13}C\alpha$, $^{13}C\beta$ and $^1H\alpha$ chemical shifts and in some cases also the ^{13}CO chemical shifts, are only used as input data for SSP calculation. Specifically, using the experimental ^1H and ^{13}C chemical shifts ($^{13}C\alpha$, $^{13}C\beta$, ^{13}CO and $^1H\alpha$) of ORF3 C20 protein in the SSP program, the SSP score values for the assigned residues through the sequence are calculated and the plot is shown in Figure 66.
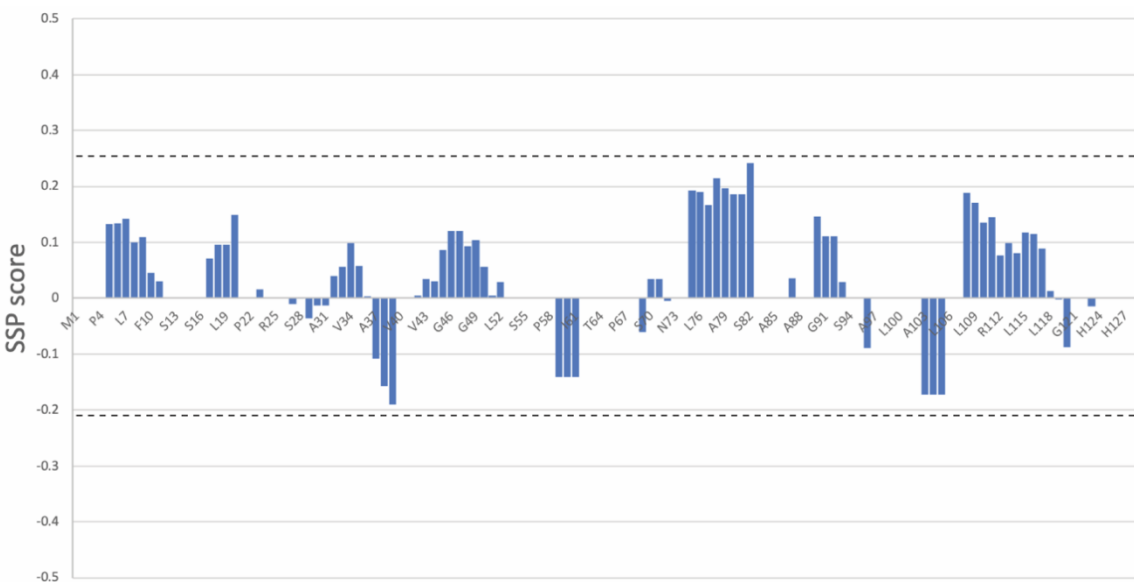
Figure 66. Secondary Structure Propensities (SSP) values for ORF3 C20 protein calculated with the SSP program using the experimental $^{13}C\alpha$, $^{13}C\beta$, ^{13}CO and $^{1}H\alpha$ chemical shifts[195]. The range of the SSP score is indicated with dashed lines.

Generally, a SSP score at a given residue of 1 or -1 indicates fully formed α-helix or β-strand, respectively, while if the score is equal to 0.5 that means that the 50% of the conformers in the disordered state ensemble are helical at that position. The first observation on the plot with the SSP score values of ORF3 C20 protein is the overall low SSP values for all residues are lower than 0.2/-0.2 which is further experimental evidence of the disordered nature of the ORF3 protein. Secondly, despite the low values, it seems that few regions have the propensity to adopt helical conformation, such as the regions Gly75-Ser82 and Gln108-Leu118.

After having analyzed the secondary structural propensities from the chemical shifts, ^{15}N relaxation data was recorded at 293K on 600 MHz Spectrometer using a 170 μM ^{15}N ORF3 C20 protein sample in NMR buffer (50 mM Sodium Phosphate pH 6.1, 50 mM NaCl, 0.1 mM EDTA, 3 mM THP, 5% D_2O) placed in Shigemi tube. For disordered proteins, ^{15}N relaxation measurements are conducted routinely as part of their structural characterization measuring usually three parameters, the T1 longitudinal relaxation time constant, the T2 transverse relaxation time constant and the ^{1}H-^{15}N heteronuclear NOEs. The ^{15}N heteronuclear relaxation rates (R1, R2) and heteronuclear NOE ({^{1}H}-^{15}N NOE) of backbone amides of the assigned and not overlapped residues of ORF3 C20 protein are shown in Figure 67.

137
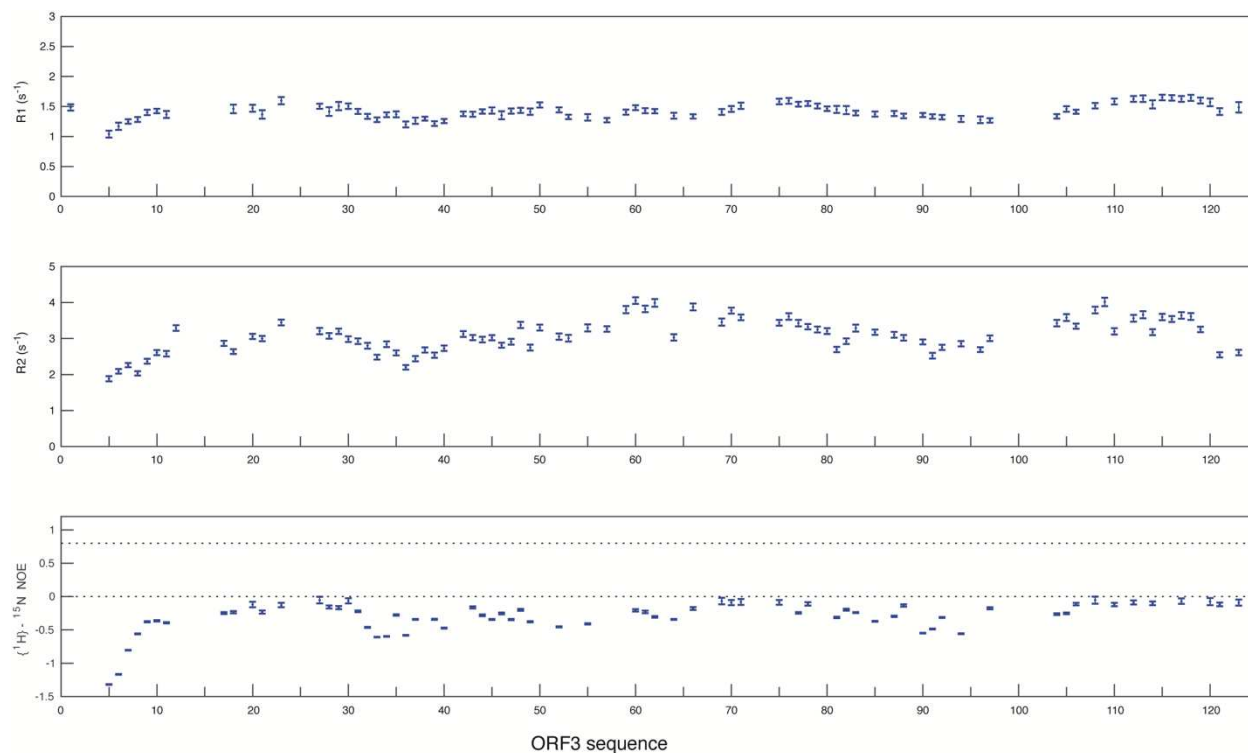
Figure 67. Backbone ^{15}N relaxation data of ORF3 C20 protein recorded on 600 MHz Spectrometer. Top plot: R1 relaxation rates; middle plot: R2 relaxation rates; bottom plot: heteronuclear NOE ({^{1}H}-^{15}N NOE).

Looking the relaxation data along the ORF3 C20 sequence, only small variations can be found. The values of R1 relaxation rate range from 1 to 1.5 s^{-1} for the assigned residues while the variation in the R2 relaxation rate is a little bit higher with the values range from 2 to 4 s^{-1} for the assigned residues with the highest values found in the middle and at the termini of the sequence. For folded proteins, the heteronuclear NOE values are close to 1 (higher dashed line in the plot). The overall negative heteronuclear NOE values and the R1 and R2 small variation' values of all the ORF3 C20 residues along the sequence reflect the disordered behavior of the protein.

During the optimization of the purification protocol, the need of the Reverse Phase (RP) Chromatography step and the differences on NMR spectra were tested. Two 100 µM ^{15}N ORF3 C20 protein samples were prepared, one following the optimized procedure, the ORF3 C20 RP sample, and the second one without the RP step dializing the protein against NMR buffer directly after the affinity HisTrap Chromatography step, the ORF3 C20 HisTrap sample. For both samples, 1D and 2D ^{1}H, ^{15}N HSQC spectra were recorded and compared in order to determine if the RP step interferes with the potential "structure" of the ORF3 protein. In the overlay of the 1D spectra

(Figure 68a), the imidazole peaks at ~7.28 and ~8.18 ppm are intense despite the overnight dialysis, although in the overlay of the 2D ^{1}H, ^{15}N HSQC spectra of the samples (Figure 68b), there are not differences, only few minor peaks in ORF3 C20 RP sample detected.
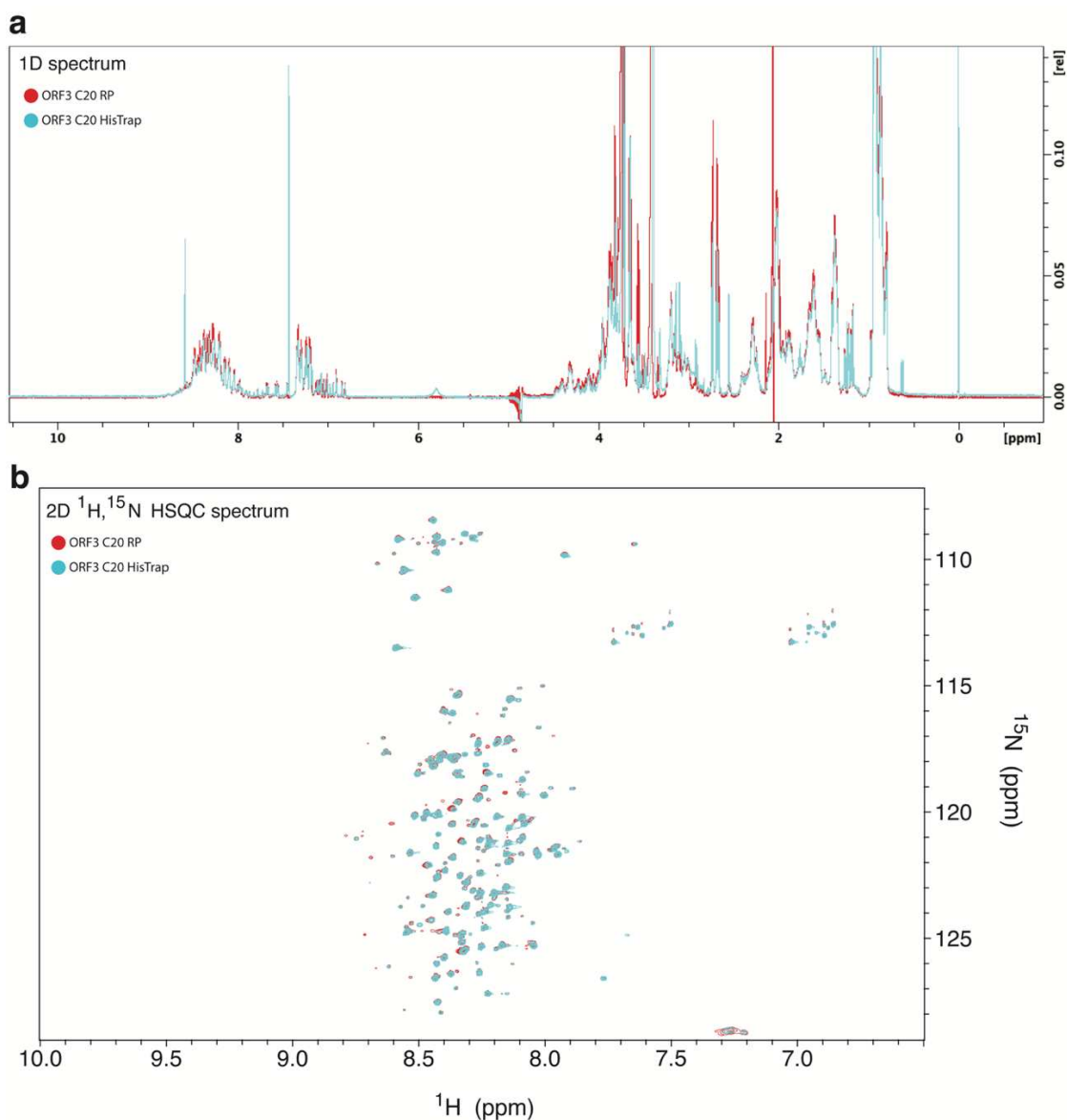
Figure 68. Overlay of (a) 1D spectra and (b) 2D ^{1}H, ^{15}N HSQC spectra of ^{15}N ORF3 C20 RP protein in red and ^{15}N ORF3 C20 HisTrap protein in cyan.

Because the two HSQC spectra are almost identical and due to the presence of the imidazole peaks in the ORF3 C20 HisTrap protein spectra, the Reverse Phase Chromatography step does not interfere with the potential ORF3 "structure" and was kept in the protocol for further protein purification.

ORF3 WT protein

In this study, the most characterized protein is the ORF3 C20 mutant. However, the ORF3 WT protein including all 8 Cysteine residues in the N-terminal region is also studied using the NMR Spectrometry in order to confirm that the ORF3 C20 is a valid protein model.

Firstly, a ^{15}N ORF3 WT labeled protein sample at 171 µM in NMR buffer (50 mM Sodium Phosphate pH 6.1, 50 mM NaCl, 0.1 mM EDTA, 5 mM DTT, 5% D$_2$O) is purified and prepared for recording a 2D ^{1}H, ^{15}N HSQC spectrum at 293K on 900 MHz Spectrometer placed in 5 mm tube. Overlaying the 2D HSQC spectrum of ORF3 WT protein with the one of the assigned ORF3 C20 mutant, the obtained assignments of the latter could be transferred in the WT construct. Figure 69 indicates this overlay with ORF3 C20 spectrum in red and ORF3 WT spectrum in cyan.
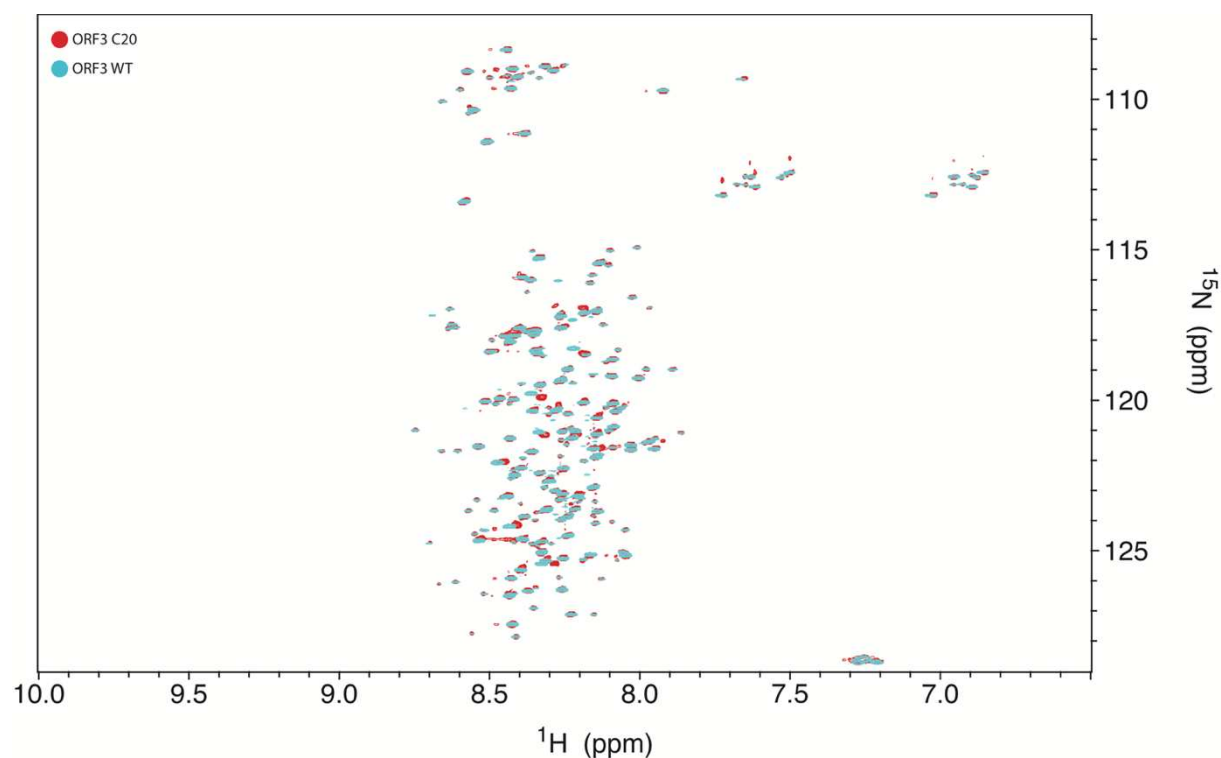
Figure 69. Overlay of 2D ^{1}H, ^{15}N HSQC spectra of ^{15}N ORF3 C20 protein in red and ^{15}N ORF3 WT protein in cyan.

The first observation on this spectrum is the narrow dispersion in the proton dimension as the one of ORF3 C20 construct concluding that ORF3 WT protein is mainly disordered. Few peaks on these two spectra are different, this was expected because of the Cysteine mutations, and few shifted peaks are noticed. In order to reliably transfer the backbone assignments in the ORF3 WT

protein, a double labeled sample is prepared and two 3D experiments are recorded, a 3D HNCO and a 3D HNCACB spectra. The 75 µM ^{15}N, ^{13}C ORF3 WT sample is placed in a Shigemi tube and the data are recorded at 298K on 900 MHz Spectrometer. The recording temperature is higher than in the other experiments in order to increase the signal to noise ratio on the 3D spectra. The 3D HNCO spectrum is selected because of its sensitivity while the 3D HNCACB because of the information obtained. The carbonyl group of the preceding residue (CO$_{i-1}$) detected in the HNCO spectrum in conjunction with the two sets of Cα and Cβ peaks, from the same (Cα$_i$ and Cβ$_i$) and from the preceding residue (Cα$_{i-1}$ and Cβ$_{i-1}$) detected in the HNCACB spectrum are compared with the corresponding assigned ones of the ORF3 C20 protein. This information is sufficient to transfer 76 peaks on the 2D ^{1}H, ^{15}N HSQC spectrum of ORF3 WT protein with the N-terminal region that could not be assigned (Figure 70).
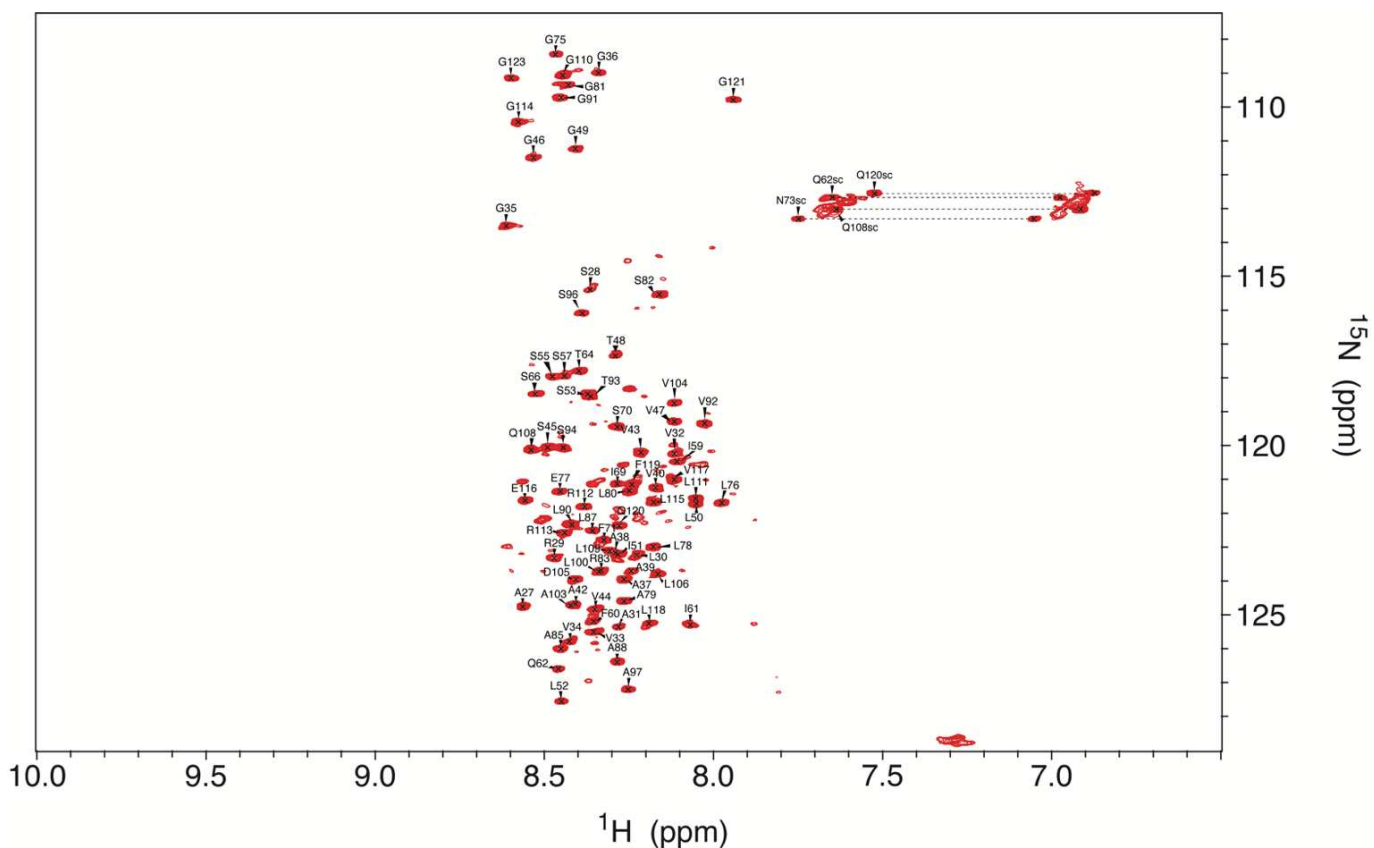
Figure 70. 2D ^{1}H, ^{15}N HSQC assigned spectrum of ^{15}N, ^{13}C ORF3 WT protein. The NH$_2$ side chains (sc) of all Asn and Gln residues are assigned and their pairs are labeled with dashed line.

The presence of many minor peaks on 2D HSQC spectrum of ORF3 WT protein because of the cis-trans isomerization of Proline residues and the lower stability of this construct are the reasons for continuing work with ORF3 C20 protein.

Figure 71 indicates the assigned (in black) and unassigned (in red) residues of ORF3 WT protein.

ORF3 WT

```
        10          20          30          40          50          60
MGSPCALGLF  CCCSSCFCLC  CPRHRPASRL  AVVVGGAAAV  PAVVSGVTGL  ILSPSPSPIF

        70          80          90         100         110    *    120
IQPTPSPPIS  FHNPGLELAL  GSRPAPLAPL  GVTSPSAPPL  PPAVDLPQLG  LRRGLEVLFQ


GPGHHHHHH
```
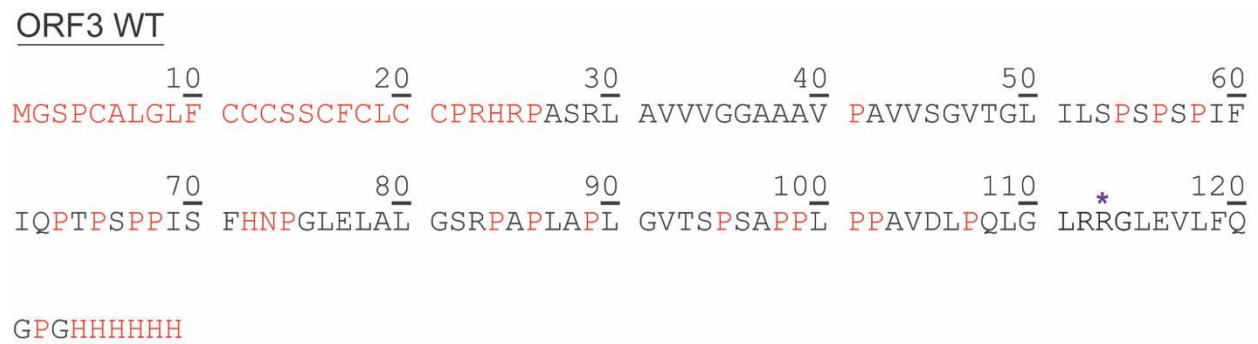
Figure 71. ORF3 WT protein sequence with assigned (in black) and unassigned (in red) residues. The purple star shows the last residue of the actual protein sequence.

142

2. Membrane anchoring of HEV ORF3 protein

In order to determine the association of HEV ORF3 protein with the membrane, the Nanodisc (ND) technology is mainly used which mimics a bilayer membrane and thus, is closed to the native environment of membrane proteins[184]. This technology enables the possibility of the investigation of the two different approaches in order to study both the ORF3 transmembrane oligomerization model[81] and the model of ORF3 protein association via post-translational modification (palmitoylation)[82] that have been proposed in the literature. As mentioned in Material and Methods, the assembly of the NDs requires the MSP protein, particularly the MSP1D1ΔH5 protein, the lipids, such as DMPC, PG, DOPC, DOPS, PE-MCC and DGS-NTA(Ni) mixed in different ratio, and the protein of interest, the ORF3 C20 protein. In order to study the first hypothesis corresponding to the transmembrane oligomerization of ORF3, it requires the simultaneous mixture of the MSP protein, the lipids and the ORF3 protein. If ORF3 possess a transmembrane segment, then the protein should be incorporated into the NDs during the assembly procedure. In order to study the second hypothesis corresponding to the membrane association of ORF3 via its cysteine(s) palmitoylation, we used a two-step procedure, preparing an "empty" Nanodisc and then attached the ORF3 protein on it through different means (*in vitro* palmitoylation of ORF3 C20, use of modified lipids containing either a cysteine-reactive function or a NTA group). In Figure 72, the transmembrane insertion (a) and the association via post-translational modification (b), the two hypotheses of ORF3 protein anchoring to the membrane using the Nanodisc technology, are shown.
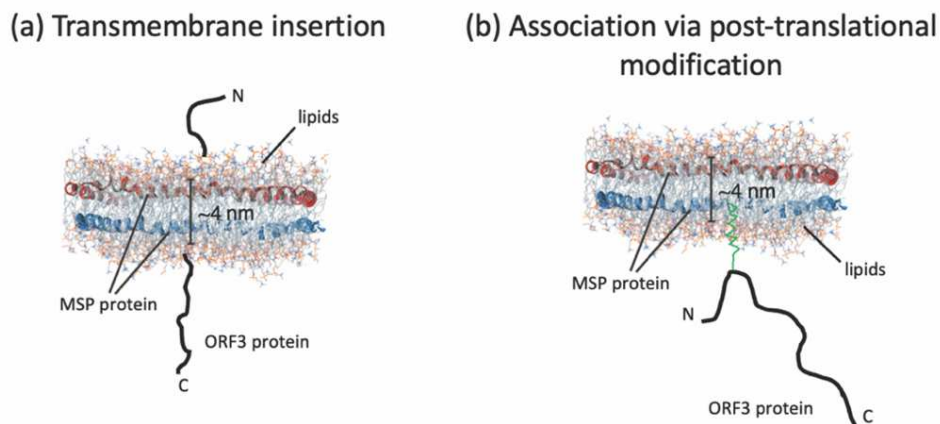
Figure 72. The two approaches of ORF3 protein anchoring to the membrane using the ND technology, (a) the transmembrane insertion and (b) the association via post-translational modification.

2.1 MSP1D1ΔH5 protein purification

The purification protocol for MSP1D1ΔH5 protein, MSP1D1 variant with deletion of helix 5 which gives NDs of about 7-8 nm diameter suitable for NMR studies[184], is described in details in the Material and Methods section. It lasts three days including affinity HisTrap chromatography steps, 6xHis-tag cleavage using TEV Protease, overnight dialysis step and protein concentration to a final concentration of 500-600 µM for further ND assembly.

After the cell lysis and the separation from the insoluble *E. coli* cellular debris, the supernatant is purified by affinity HisTrap step including wash steps with Buffer A (20 mM Tris-HCl pH 8, 500 mM NaCl, 10 mM Imidazole) containing different detergents, firstly with 1% Triton X-100 and then with 50 mM sodium cholate, which replace any cell bound lipids from the MSP1D1ΔH5 protein before the elution from the column. The chromatogram of the MSP1D1ΔH5 HisTrap purification with the three (3) UV absorbance curves at 280 nm, at 260 nm and at 215 nm (a) and the 4-20% SDS-PAGE with different fractions (b) are shown in Figure 73.
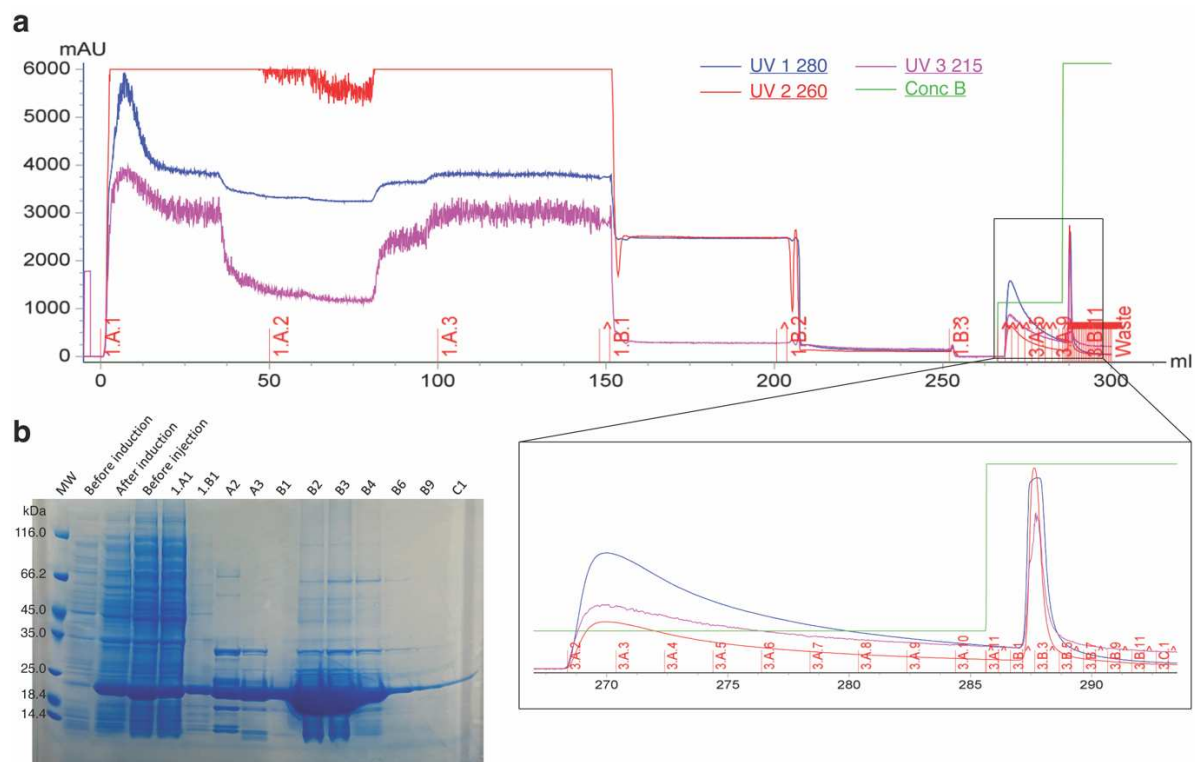
Figure 73. (a) Chromatogram of affinity HisTrap purification of MSP1D1ΔH5 protein. In the black box, the elution peak is shown zoomed. Blue: 280 nm, red: 260 nm, magenta: 215 nm and green: concentration of Buffer B. (b) 4-20% SDS-PAGE of fractions based on the chromatogram (a) with Coomassie blue staining.

The efficiency of the 6xHis-tag overnight cleavage using TEV Protease on the pooled protein fraction is checked by SDS-PAGE before the second affinity purification step (Figure 74a). The chromatogram of the second HisTrap purification step monitoring using the 280 nm, the 260 nm and the 215 nm UV absorbance (b) and the 4-20% SDS-PAGE with the fractions (c) are shown in Figure 74.
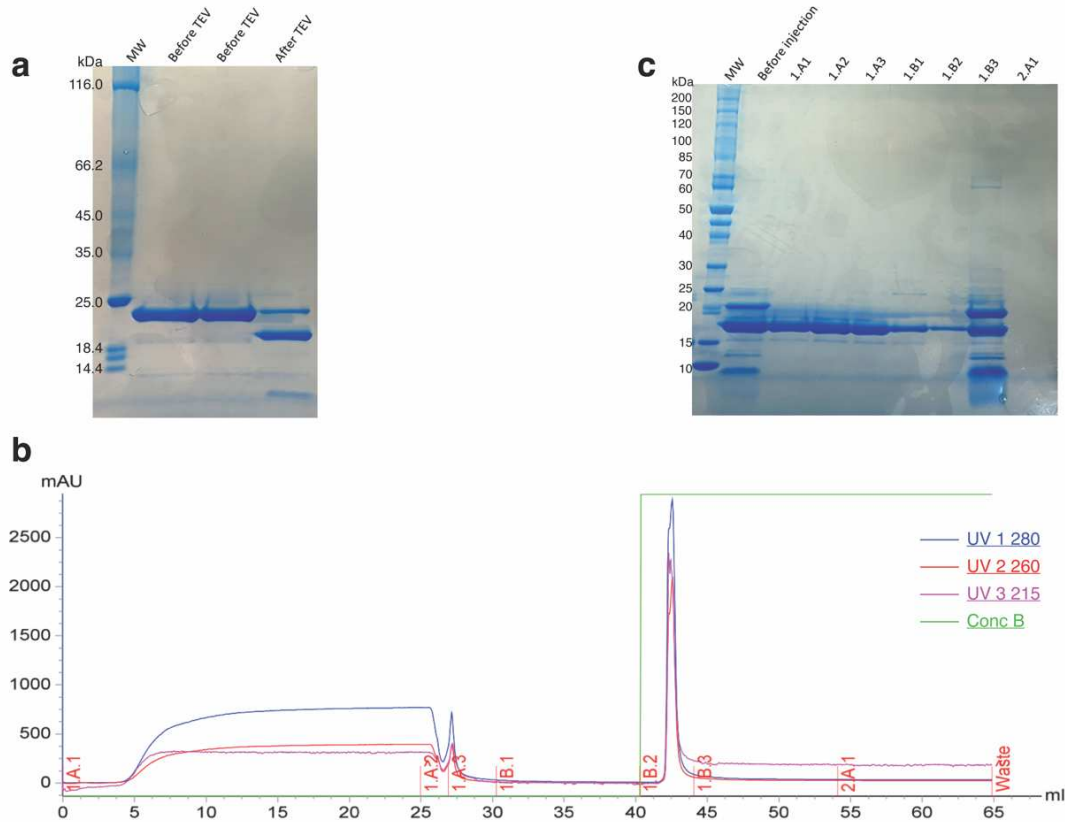
Figure 74. (a) 4-20% SDS-PAGE of pooled MSP1D1ΔH5 protein fractions before and after TEV cleavage with Coomassie blue staining. (b) Chromatogram of second HisTrap purification step of MSP1D1ΔH5 protein after TEV cleavage. Blue: 280 nm, red: 260 nm, magenta: 215 nm and green: concentration of Buffer B. (b) 4-20% SDS-PAGE of fractions based on the chromatogram (b) with Coomassie blue staining.

The last steps as described above corresponds to the overnight dialysis against the MSP buffer (20 mM Tris-HCl pH 7.5, 100 mM NaCl, 0.5 mM EDTA), the concentration of the protein and the preparation of the aliquots used for ND assembly. The yield of the expressed protein is ~30-35 mg per L of culture. Note that in this protein purification shown in the figures above, the yield was higher than usual and therefore the affinity purification steps and the TEV cleavage step were performed twice in order to purify the maximal amount of the protein sample.

2.2 Nanodiscs assembly procedure – Attachment of ORF3 protein

Having prepared the MSP1D1ΔH5 protein, the lipids stocks and the ORF3 C20 protein, the two ND assembly procedures, corresponding to the two hypotheses, can be followed as described in Material and Methods section.

2.2.1 Transmembrane insertion

In order to study the possibility that ORF3 is a transmembrane protein, the ND assembly in the presence of ORF3 protein is performed by mixing all the components in a specific order as described. The detergent removal, to induce the NDs assembly, could be done either using Bio-Beads SM-2 (Bio-Rad) or dialysis against MSP buffer with the former being preferred due to the procedure time[184].

Before starting the assembly, the interaction of ORF3 protein with the Bio-Beads SM-2 has to be tested. After adding the Bio-Beads into the protein in two steps (performed as in the ND assembly protocol) and incubation at 23°C for 2 h at 750 rpm each step, the ORF3 C20 protein, which contains a 6xHis-tag in the C-terminal region, is mixed with Ni-NTA beads and is eluted with imidazole-containing buffer to its recovery. Figure 75 illustrates the procedure followed to check the interaction as well the SDS-PAGE with samples of ORF3 protein before and after and the Bio-beads after the interaction. In the ORF3 eluted sample after purification with the Ni-NTA beads (sample 2, third lane in the SDS-PAGE), no band could be detected while a small amount of the protein appears in the Bio-beads sample. This shows that ORF3 protein interacts with Bio-beads. Moreover, Bradford dye reagent is added to the Bio-beads which turned blue, signifying the presence of ORF3 protein on the beads (Figure 75, right).
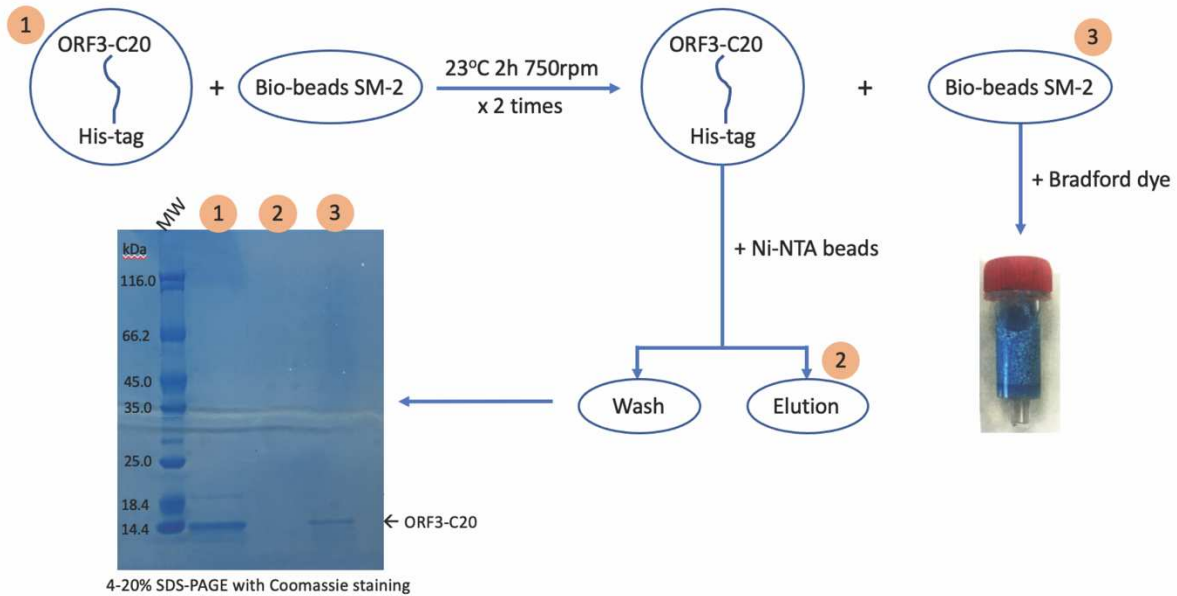
Figure 75. Diagram of the procedure followed to check if ORF3 C20 protein interacts with Bio-beads SM-2. (left) 4-20% SDS-PAGE with ORF3 C20 protein and Bio-beads samples as labeled in the diagram with Coomassie blue staining. (right) Bio-beads used for the interaction test with Bradford dye reagent.

Because of the interaction of ORF3 protein with Bio-beads, the dialysis procedure has to be used for the detergent removal in the ND assembly. Mixing all the components with the following order, firstly the MSP buffer, secondly the sodium cholate buffer, thirdly the lipid mixture (DMPC:PG in ratio 3:1), then the MSP1D1ΔH5 protein and finally the ORF3 C20 protein, a direct precipitation of the ORF3 protein is observed. After a high-speed centrifugation (at 16,100 xg) for 10 min at 4°C, SDS-PAGE analysis showed that all the amount of ORF3 protein added is precipitated. In order to determine which of the component(s) is (are) responsible for the precipitation, ORF3 protein is mixed with all the components separately. A sedimentation of the protein is observed in all samples prepared. The pellet of each mixture contains the entire amount of ORF3 protein added based on the SDS-PAGE shown in Figure 76. The component that tends the ORF3 C20 protein to precipitate is probably the sodium cholate buffer because it is present in each sample.
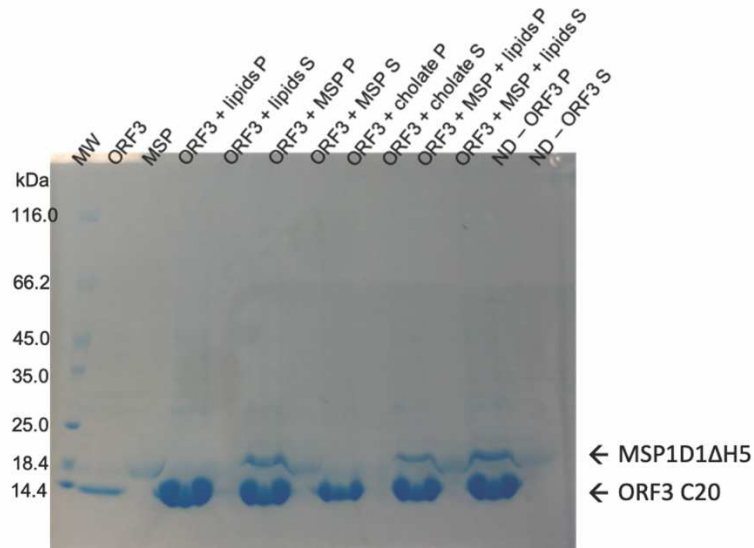
Figure 76. 4-20% SDS-PAGE for ORF3 C20 protein and every component of Nanodisc assembly with Coomassie staining. P: pellet, S: supernatant.

Therefore, the transmembrane insertion approach cannot be studied using the Nanodisc technology.

2.2.2 Association via post-translational modification, palmitoylation

The second approach is the ORF3 protein membrane anchoring via palmitoylation of its Cysteine-rich region[82]. The procedure involves two steps, first the formation of "empty" Nanodiscs (NDs) containing only the MSP1D1ΔH5 protein and the lipids and then the attachment of ORF3 C20 protein. A prerequisite for the second step is the existence of modified lipids on the "empty" NDs or the modification of ORF3 C20 using *in vitro* palmitoylation assay.

The *in vitro* palmitoylation of the ORF3 C20 protein is firstly performed using the protocol described by Yousefi-Salakdeh, Johansson and Strömberg[196] by other lab members before starting my PhD. The lyophilized ORF3 C20 protein was resuspended with 100% TFA to final concentration of 600 µM. The one half was used as a control sample and the other half was mixed with excess of palmitoyl chloride (6.6 mM final concentration) and incubated for 10 min at room temperature. The reaction was terminated by addition of 80% ethanol, the two samples were purified by Reverse Phase (RP) Chromatography using the Zorbax C8 column and finally checked by MALDI-TOF (Matrix Assisted Laser Desorption Ionization-Time Of Flight mass spectrometry) analysis. In the MALDI-TOF mass spectra, the palmitoylated ORF3 C20 protein could not be detected. A second experiment with 100-time more palmitoyl chloride and overnight incubation at room temperature has the same negative results.

Due to the unsuccessful *in vitro* palmitoylation of ORF3 protein, the modified lipids containing either a cysteine-reactive function (PE-MCC) or a NTA group (DGS-NTA(Ni)) are used, included in the lipid mixture in the formation of "empty" NDs. The PE-MCC lipids contain a maleimide group, mimic the palmitoyl and react with the only Cysteine of ORF3 protein. On the other hand, the DGS-NTA(Ni) lipids is used as a last option for ORF3 attachment via its 6xHis-tag. However, the interaction of ORF3 protein with "empty" Nanodiscs is studied before introducing these modified lipids due to the precipitation of the protein during the transmembrane insertion.

2.2.2.1 Interaction of ORF3 protein with "empty" NDs with DMPC and DMPC/PG lipids

Due to the interaction and precipitation of ORF3 protein during the transmembrane insertion process, it was important to test whether ORF3 protein also interacts, maybe in an unspecific manner, with "empty" NDs using NMR Spectroscopy before proceeding to the introduction of the modified lipids. So, we started with the preparation of "empty" DMPC and DMPC/PG lipids.

Followed the protocol described in the Material and Methods section, "empty" Nanodiscs with DMPC lipids and mixture of DMPC/PG lipids in ratio 3:1 (stocks in powder form) are prepared using Bio-beads SM-2 for detergent removal and then purified by Superdex 200 5/150 GL (Cytiva) Size Exclusion Chromatography (SEC) column in order to check the quality of the assembly and remove any unbound components. The chromatogram of SEC purification for the "empty" NDs with DMPC lipids monitoring using the 280 nm absorbance curve is shown in Figure 77. The single peak at ~1.68 mL corresponds to the assembled "empty" NDs.
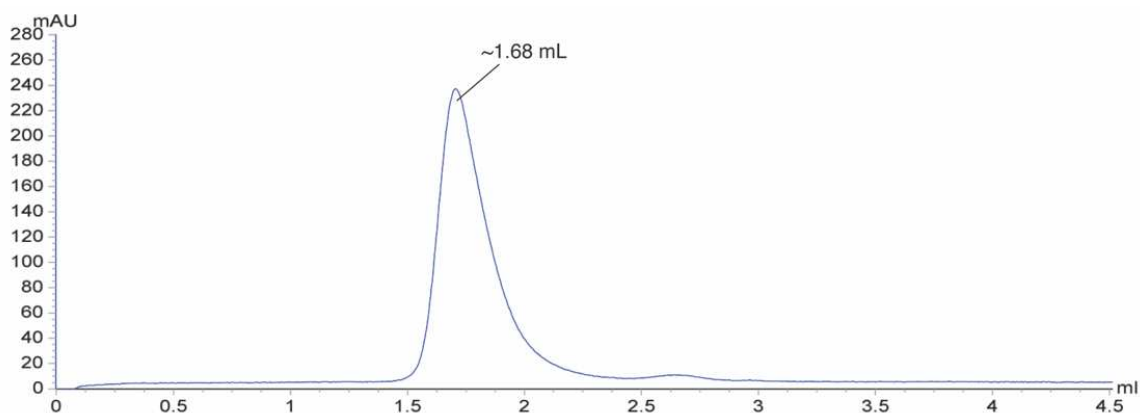
Figure 77. Chromatogram of SEC purification using Superdex 200 5/150 GL column for the "empty" NDs with DMPC lipids prepared using Bio-beads SM-2. The 280 nm absorbance curve is used to monitor the purification.

In addition, the size of the "empty" NDs with mixture of DMPC/PG lipids is measured by Dynamic Light Scattering (DLS) technique resulting in a diameter of about 8 nm as expected. The chromatogram of SEC step (a) and the DLS size distribution by mass plot (b) for the "empty" NDs with mixture of DMPC/PG lipids are shown in Figure 78.
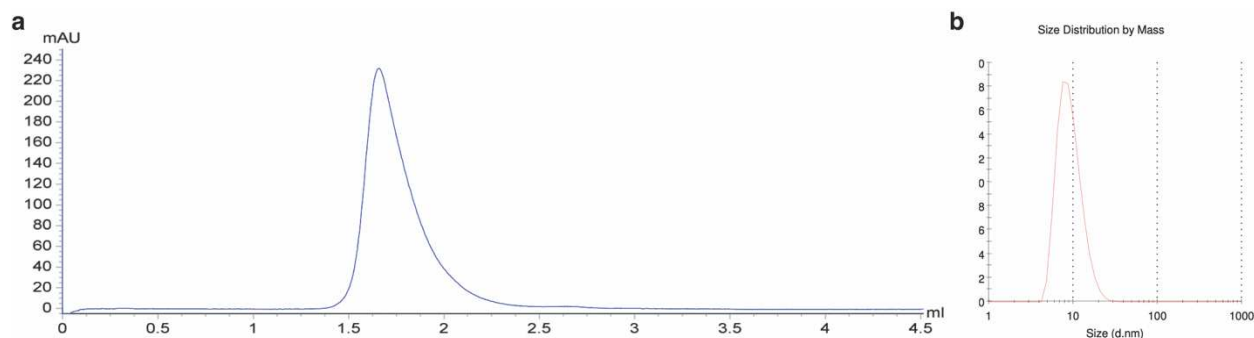
Figure 78. (a) Chromatogram of SEC purification using Superdex 200 5/150 GL column for the "empty" NDs with mixture DMPC/PG lipids prepared using Bio-beads SM-2. The 280 nm absorbance curve is used to monitor the purification. (b) DLS size distribution by mass plot for "empty" NDs with mixture of DMPC/PG lipids prepared using Bio-beads SM-2.

Moreover, "empty" NDs with mixture DMPC/PG lipids using the dialysis procedure for detergent removal are prepared. The SEC profile of the assembled NDs is the same with the corresponding one using the Bio-beads as shown in Figure 79. As mentioned previously, this option is more time-consuming as it lasts about two days compared to the Bio-beads (few hours) and therefore the latter is used for the ND assembly.
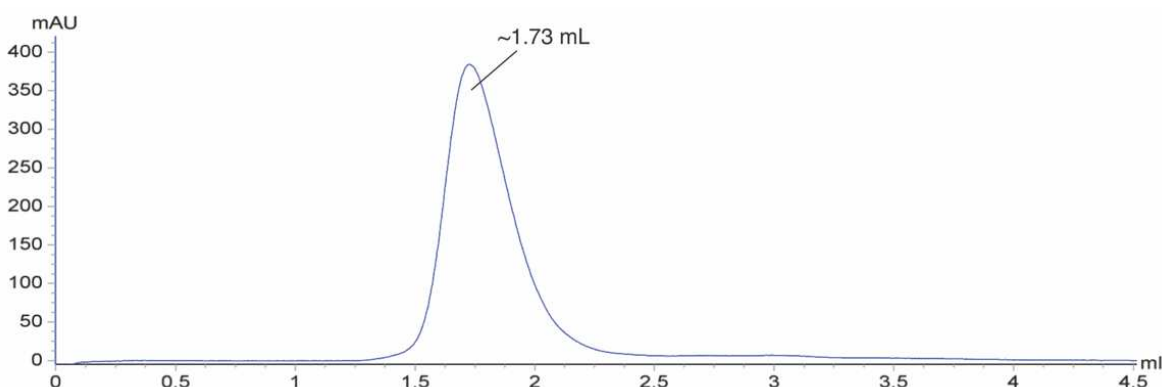
Figure 79. Chromatogram of SEC purification using Superdex 200 5/150 GL column for the "empty" NDs with mixture DMPC/PG lipids prepared using dialysis against MSP buffer. The 280 nm absorbance curve is used to monitor the purification.

Next to test the potential interaction between ORF3 C20 and empty NDs we used NMR spectroscopy. "Empty" NDs first with mixture of DMPC/PG lipids (a) and then with DMPC lipids (b) are prepared and mixed with 100 μM ^{15}N uncleaved ORF3 C20 labeled sample. 2D ^{1}H, ^{15}N HSQC spectra of ORF3 C20 protein in presence and absence of "empty" NDs (for both (a) and (b) NDs) are recorded at 293K on 900 MHz Spectrometer. The overlay of the HSQC spectra demonstrates the shift of some peaks of ORF3 protein indicating interaction with "empty" NDs (Figure 80).
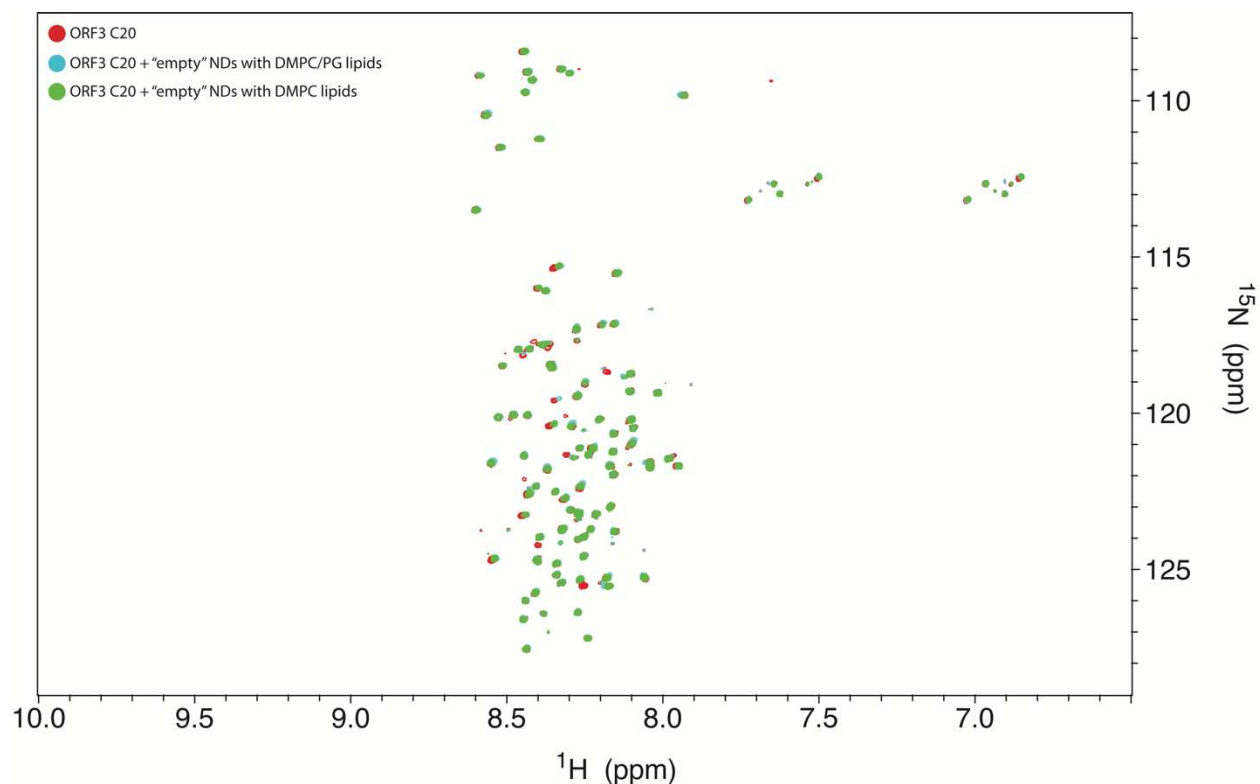
Figure 80. Overlay of 2D ^1H, ^{15}N HSQC spectra of ^{15}N ORF3 C20 protein alone in red, in presence of "empty" NDs with DMPC/PG lipids in cyan and in presence of "empty" NDs with DMPC lipids in green.

Analyzing the chemical shift perturbations (CSP) of the assigned residues of ORF3 C20 protein induced by "empty" NDs with mixture of DMPC/PG lipids, the interaction is detected in the N-terminal and C-terminal region of the ORF3 C20 protein (Figure 81). These regions contain residues with positive charges, such as Arg23 and Arg29 in the N-terminus and Arg113 in the C-terminus where the 6xHis-tag is also located. The assembled NDs include neutral DMPC lipids, which contain one positive and one negative charge, and negative charged PG lipids, which contain only a negatively charged hydrophilic head group with a single negative charge. As a result, the "empty" NDs have a total net negative charge which could be the reason of the observed interaction.
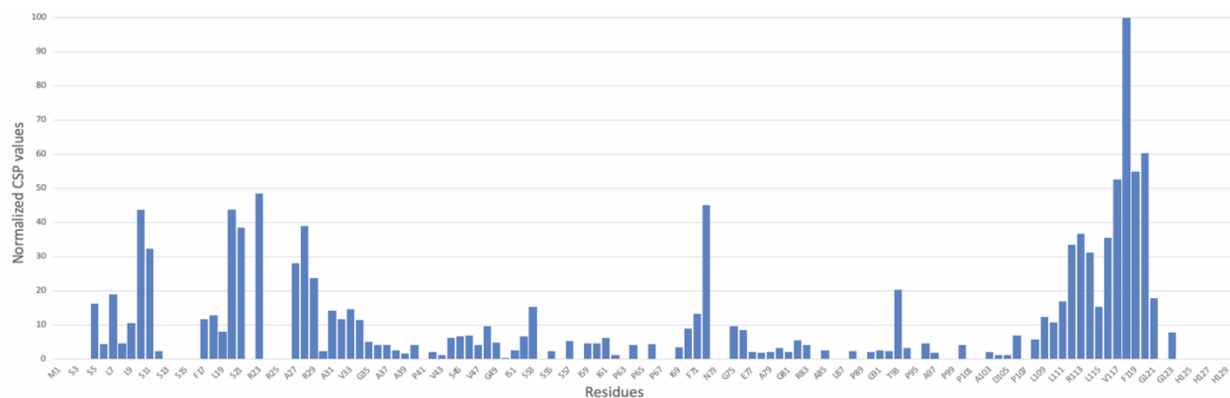
Figure 81. Normalized CSP values for the assigned residues of ^{15}N uncleaved ORF3 C20 protein induced by "empty" NDs with mixture of DMPC/PG lipids.

The same experiments and data analysis are conducted for the ORF3 C20 protein induced by "empty" NDs with only neutral DMPC lipids in order to elucidate if the charge of lipids is responsible for the shifted peaks. In Figure 82, the normalized CSP values for the assigned residues of ORF3 protein show similar results with same previously detected regions of interaction concluding that the charges could be the reason of the observed interaction in both experiments.
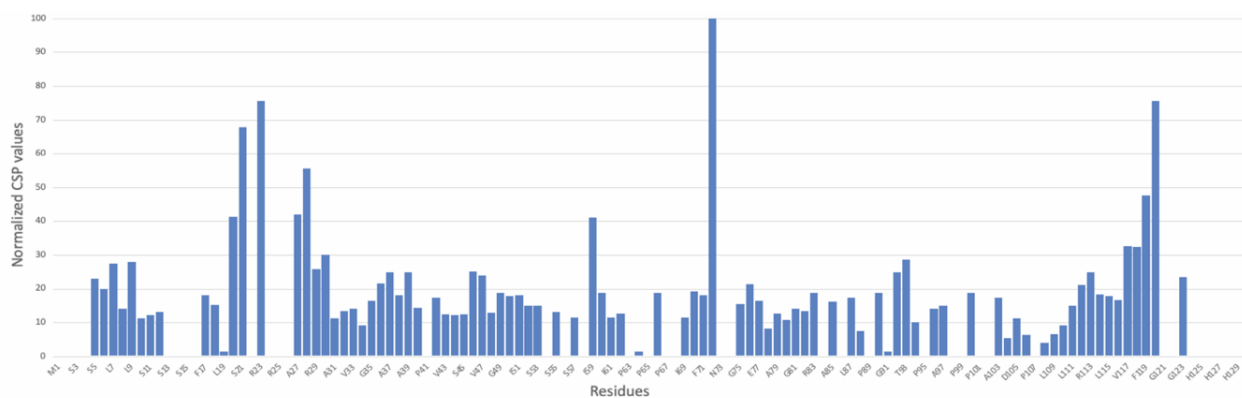
Figure 82. Normalized CSP values for the assigned residues of ^{15}N uncleaved ORF3 C20 protein induced by "empty" NDs with DMPC lipids.

The C-terminal positive-charged 6xHis-tag of ORF3 C20 protein could be responsible for the observed interaction with the "empty" NDs in this region. The comparison of the results of the same experiments with cleaved ORF3 protein would clarify whether this hypothesis is valid, but due to the instability of the protein, these experiments could not be conducted.

Even if we monitored an interaction between "empty" NDs and ORF3 protein, this does not induce any precipitation of ORF3, probably because the hydrophobic parts of both lipids and MSP1D1ΔH5 protein are hidden in the assembly. Thus, we decided to pursue our idea to attach ORF3 to the NDs using modified lipids.

2.2.2.2 "Empty" Nanodiscs with DMPC/PG/PE-MCC lipids – Attachment of ORF3 protein

The next experiment in the ND assembly is the introduction of a modified lipid, PE-MCC, which contains a maleimide group that will react with the thiol group of the Cysteine of the ORF3 C20 protein (Figure 83).
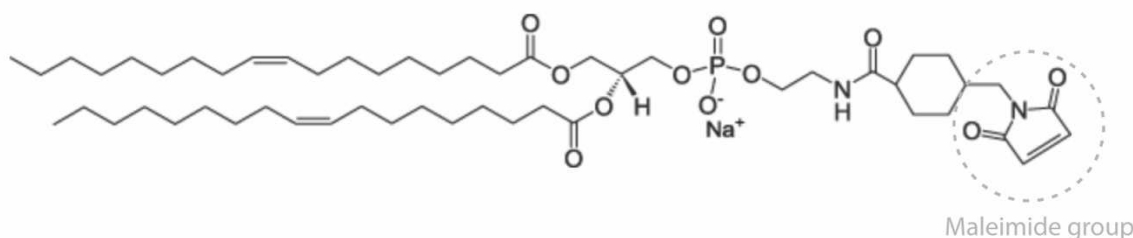
Figure 83. Structure of PE-MCC modified lipid with gray dashed circle indicating the maleimide group. Available on Avanti Polar Lipids website[197].

The goal is the attachment of one protein molecule per Nanodisc surface, so only one modified lipid in each ND side. Using the MSP1D1ΔH5 protein, the inner diameter of the Nanodisc is 6 nm and the total lipid area is 28 nm^2 which can fit ~52 DMPC molecules as described by Hagn *et al.*[184] (Figure 84). Using DMPC, PG and the modified PE-MCC lipids, the ratio calculated is 38:13:1 (molecules per ND surface).
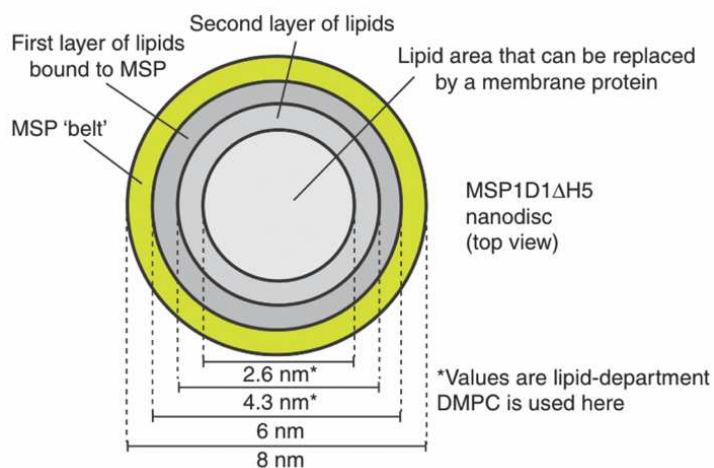
Figure 84. Top view of an assembled Nanodisc using MSP1D1ΔH5 protein with total lipid area 28 nm^2 which can fit ~52 DMPC lipid molecules. From Hagn et al.[184], with permission from Nature Protocols, Copyright 2018.

The PE-MCC modified lipid stock was purchased in powder form. Attempting to prepare a 50 mM stock solution in sodium cholate buffer results in partial dissolution of the lipids even after two 15 min sonication runs in a heated sonication bath. Thus, two ND assembly reactions are

performed in parallel, one with the DMPC/PG stock solution and the supernatant of PE-MCC after a high-speed centrifugation run (16,100 xg) (ND-1) and a second one with DMPC/PG and the partially dissolved PE-MCC solution (ND-2) (Figure 85a, chromatograms in magenta). After the SEC purification of the "empty" NDs and assuming they contain the modified lipids, the attachment of ORF3 protein is performed using ^{15}N uncleaved ORF3 C20 protein in ratio 1:1 overnight at RT (~18°C) on the bench. The reaction samples are further purified by SEC column while 100 μL fractions are collected to be checked with 4-20% SDS-PAGE analysis. In Figure 85, the overlay of the 280 nm chromatograms for both assembled NDs (ND-1 and ND-2) before and after the ORF3 incubation (a) and the 4-20% SDS-PAGE with the fractions collected in the SEC step for the reaction samples (b) are shown.
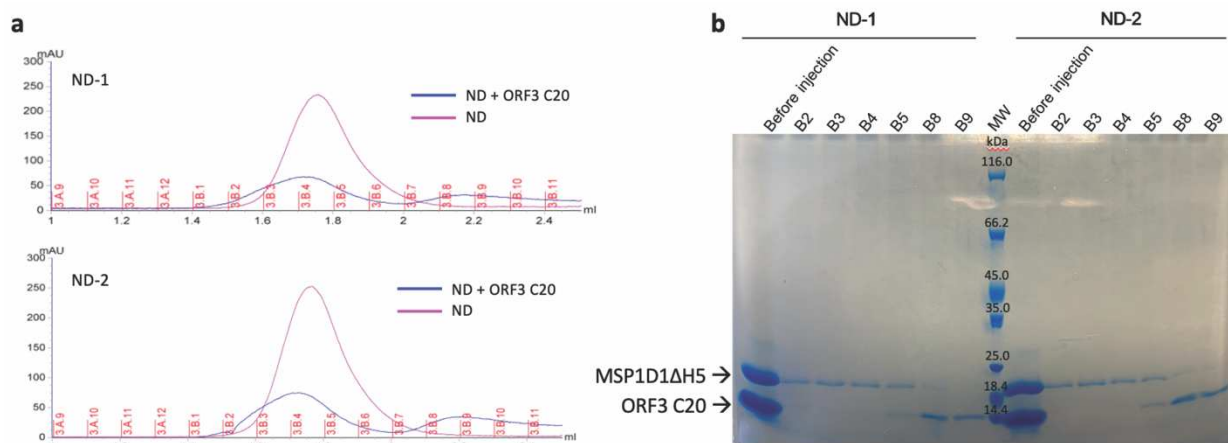
Figure 85. (a) Overlay of 280 nm chromatograms of "empty" Nanodiscs (magenta) and reactions of NDs with ^{15}N ORF3 C20 protein (blue) using two different stocks of PE-MCC lipids, supernatant (ND-1) and mixed solution (ND-2). (b) 4-20% SDS-PAGE with fractions collected in the SEC step for the reaction samples for both ND-1 (left) and ND-2 (right) with Coomassie staining.

In the chromatograms of the reaction samples (chromatograms in blue), no peak shift is detected and in the SDS-PAGE analysis, there is no fraction containing both MSP1D1ΔH5 and ORF3 C20 proteins. Therefore, the protein is not attached to the NDs, but also the presence of modified lipids is unclear. In addition, NDs with mixture of DMPC and PE-MCC lipids in three different ratio, 98% DMPC/2% PE-MCC, 90% DMPC/10% PE-MCC and 80% DMPC/20% PE-MCC to determine the appropriate amount of the modified lipids, are prepared and then incubated with ^{15}N ORF3 C20 protein at a ratio 1:1 at 23°C for 16 h at 300 rpm using the Eppendorf ThermoMixer machine. The corresponding overlay of the chromatograms before and after the incubation with ORF3 protein as well the SDS-PAGE with the SEC fractions had the same negative results.

Because of the partial dissolution of the PE-MCC lipids in powder, a stock of PE-MCC lipids dissolved in chloroform is purchased and the lipid preparation protocol is changed to ensure the introduction of the modified lipids into the NDs assembly. This protocol, as described in the Material and Methods section, involves the evaporation of the chloroform making a dried film of mixture of lipids and then proceeds to the NDs assembly. For this purpose, the DOPC and DOPS lipids dissolved in chloroform are also purchased.

2.2.2.3 "Empty" Nanodiscs with DOPC and DOPC/DOPS lipids

Following the protocol for the preparation of the lipid mixture dissolved in chloroform as described in Material and Methods section, "empty" Nanodiscs with DOPC lipids are firstly prepared. The chromatogram of SEC purification using Superdex 200 5/150 GL for the "empty" NDs with DOPC lipids monitoring using the 280 nm absorbance curve is shown in Figure 86.
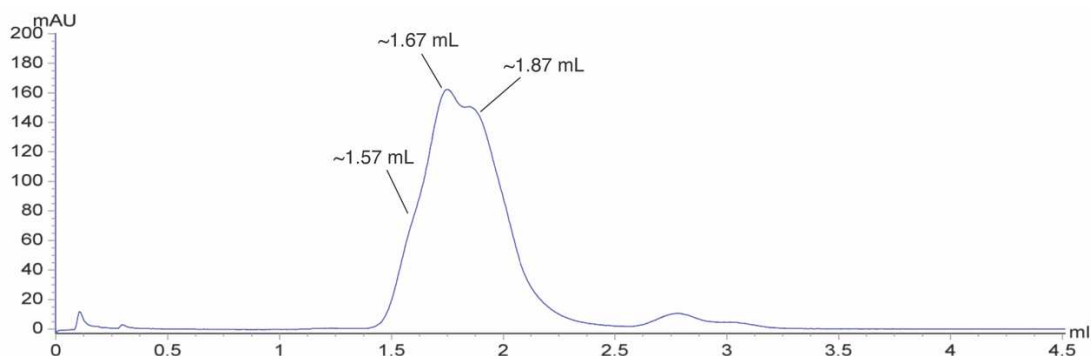
Figure 86. Chromatogram of SEC purification using Superdex 200 5/150 GL column for the "empty" NDs with DOPC lipids. Three eluted peaks at ~1.57 mL for MSP protein aggregation, at ~1.67 mL for assembled NDs and at ~1.87 mL for MSP protein. The 280 nm absorbance curve is used to monitor the purification.

In the SEC chromatogram, three peaks are eluted, one at ~1.57 mL corresponds to aggregates of MSPD1ΔH5 protein, a second one at ~1.67 mL corresponds to the assembled "empty" NDs and one at ~1.87 mL corresponds to MSPD1ΔH5 protein based on their molecular weight. The SEC profile is the same for assembled "empty" NDs with DOPC lipids in powder form. The existence of the two additional peaks in the NDs assembly needs to be further studied. In order to find the correct ratio of MSP1D1ΔH5 protein to lipids and achieve a single peak, different conditions are tested. The ratio of MSP1D1ΔH5 protein to lipids used for NDs assembly is 1 to 50 as previously mentioned.

Firstly, the amount of the MSP1D1ΔH5 protein used is reduced by keeping the concentration of the lipids constant, namely two different ratio 3:200 and 1:100 of MSP protein to lipids resulting in MSP protein aggregation at higher levels (Figure 87).
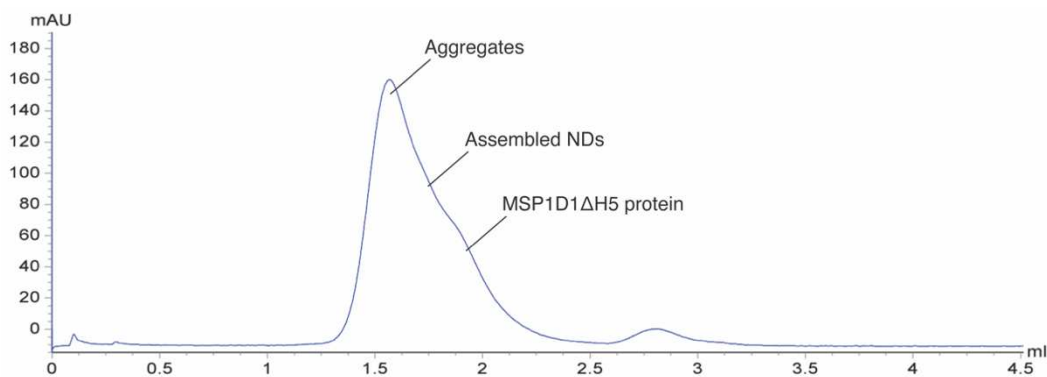
Figure 87. Chromatogram of SEC purification using Superdex 200 5/150 GL column for the "empty" NDs at ratio of MSP protein to DOPC of 3:200. The 280 nm absorbance curve is used to monitor the purification. Same SEC profile obtained for the "empty" NDs at ratio 1:100.

In the next experiments, the concentration of the protein is the same and the amount of the lipids is reduced and the ratios tested are 1:40 and 1:25. The level of ND assembly is lower while the MSP protein in free state is at higher levels in these cases (Figure 88).
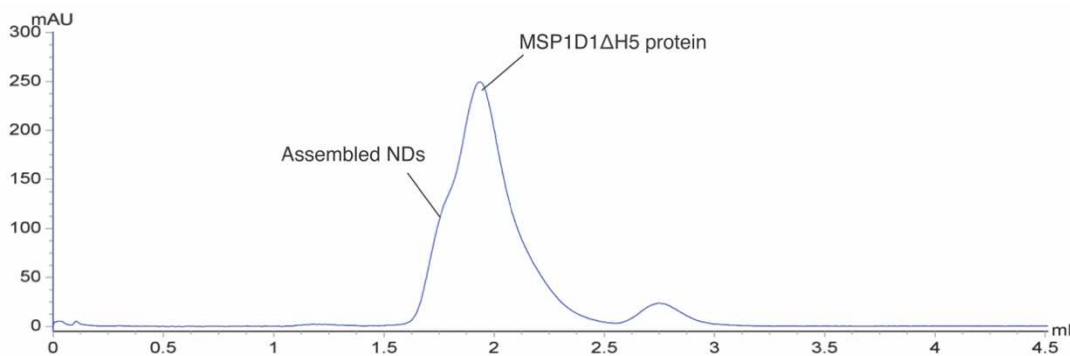
Figure 88. Chromatogram of SEC purification using Superdex 200 5/150 GL column for the "empty" NDs at ratio of MSP protein to DOPC of 1:25. The 280 nm absorbance curve is used to monitor the purification. Same SEC profile obtained for the "empty" NDs at ratio 1:40.

The same results with "free" MSP protein are observed when the concentration of the protein is increased and the amount of lipids are constant at ratios used of 3:100 and 4:100 (Figure 89).

159
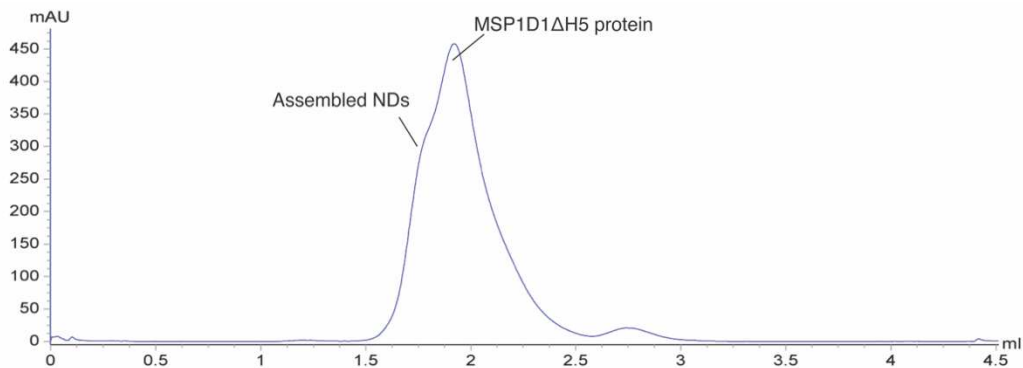
Figure 89. Chromatogram of SEC purification using Superdex 200 5/150 GL column for the "empty" NDs at ratio of MSP protein to DOPC of 4:100. The 280 nm absorbance curve is used to monitor the purification. Same SEC profile obtained for the "empty" NDs at ratio 3:100.

Moreover, "empty" NDs with mixture of DOPC/DOPS lipids in ratio 3:1 are prepared with the same three eluted peaks as shown in Figure 90.
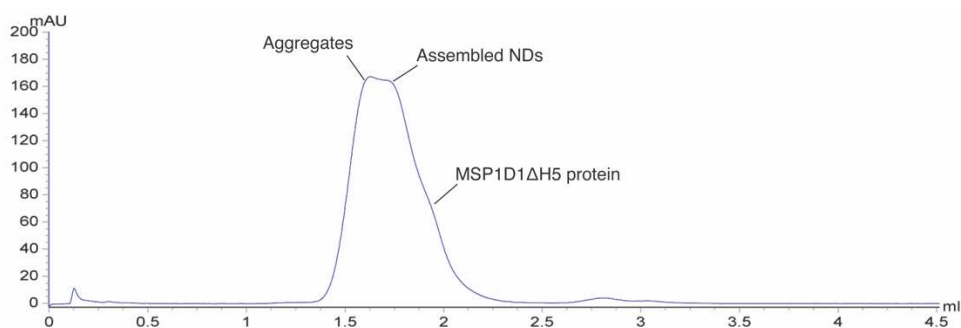
Figure 90. Chromatogram of SEC purification using Superdex 200 5/150 GL column for the "empty" NDs with mixture DOPC/DOPS lipids. Three eluted peaks at ~1.62 mL for MSP protein aggregation, at ~1.74 mL for assembled NDs and at ~1.94 mL for MSP protein. The 280 nm absorbance curve is used to monitor the purification.

Other experiments in which the amount of MSP1D1ΔH5 protein is reduced and the DOPC/DOPS lipid concentration is constant in the final mixture, namely two ratio of MSP protein to lipids 3:200 and 1:100 are used, have the same results as the assembled NDs with only DOPC lipids at these ratios (Figure 87).

Based on these results, the NDs assembly using DOPC lipids or mixture of DOPC/DOPS lipids cannot be achieved without a level of aggregation or free MSP protein in the mixture. Therefore, the initial MSP protein to lipids 1:50 ratio is used for the further NDs preparation and introduction of the modified PE-MCC lipids for ORF3 C20 protein attachment.

2.2.2.4 "Empty" Nanodiscs with DOPC/DOPS/PE-MCC lipids – Attachment of ORF3 protein

As described above, for NDs assembly with modified PE-MCC lipids in powder form, 38 molecules of DMPC, 13 molecules of PG and 1 molecule of PE-MCC are mixed for the attachment of one protein molecule per Nanodisc surface. Using the same lipid ratio, "empty" NDs are assembled with 73% DOPC lipids, 25% DOPS lipids and 2% modified PE-MCC lipids. The chromatogram of SEC purification using Superdex 200 5/150 GL monitoring using the 280 nm absorbance curve has the same three eluted peaks as shown in Figure 91.
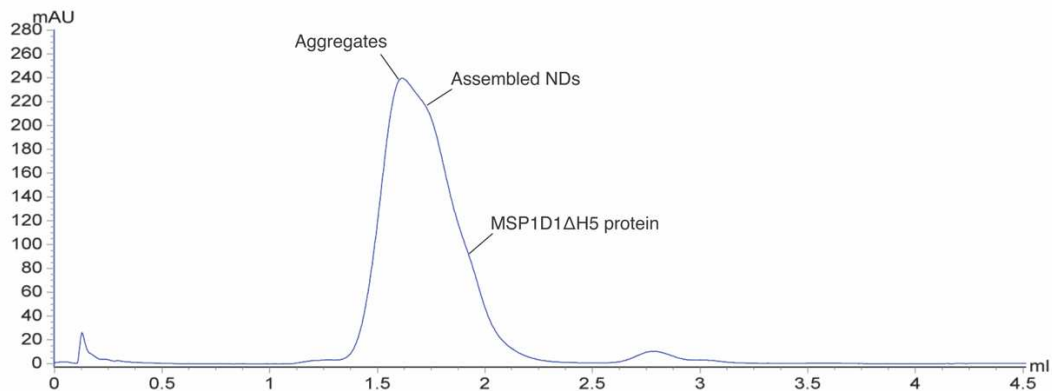
Figure 91. Chromatogram of SEC purification using Superdex 200 5/150 GL column for the "empty" NDs with mixture of 73% DOPC/25% DOPS/2% PE-MCC lipids. Three eluted peaks at ~1.61 mL for MSP protein aggregation, at ~1.75 mL for assembled NDs and at ~1.94 mL for MSP protein. The 280 nm absorbance curve is used to monitor the purification.

In order to separate these peaks, the assembled NDs are purified by Superdex 200 10/300 GL column collecting 300 μL fractions. Figure 92 illustrates the SEC chromatogram in 8-15 mL region where the eluted peaks are detected (a) as well the 4-20% SDS-PAGE with the collected fractions (b).
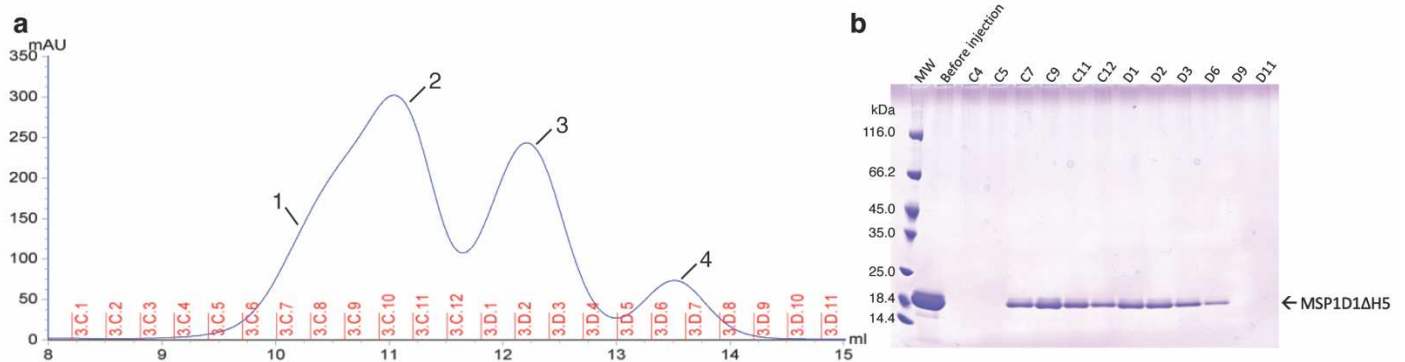
Figure 92. (a) Chromatogram of SEC purification using Superdex 200 10/300 GL column for the "empty" NDs with mixture of 73% DOPC/25% DOPS/2% PE-MCC lipids collecting 300 μL fractions. The 280 nm absorbance curve is used to monitor the purification. (b) 4-20% SDS-PAGE with fractions collected in the SEC purification (a) with Coomassie staining.

Based on the SEC profile, there are four (4) peaks with the first two (2) to correspond to the MSP protein aggregation (fractions C6-C8 and C9-C11), the third one to the assembled NDs (fractions D1-D4) and the fourth one to "free" MSP1D1ΔH5 protein (fractions D5-D8). The 4 pooled fractions' samples are concentrated up to 50 μL and incubated with 100 μL of 217.7 μM ORF3 C20 protein at 23°C overnight at 300 rpm using the Eppendorf ThermoMixer machine to check the reaction with ORF3 protein. The next day, each reaction is further purified by Superdex 200 5/150 GL column collecting 100 μL fractions which are then checked by SDS-PAGE. The four SEC chromatograms (a) and the 4-20% SDS-PAGE (b) are shown in Figure 93.
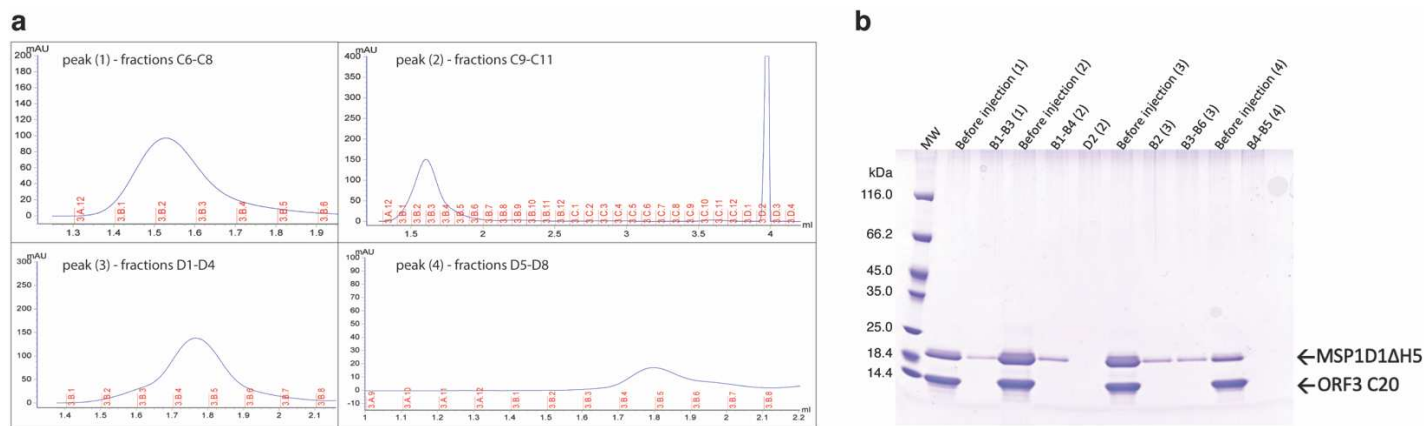
Figure 93. (a) Chromatograms of SEC purification using Superdex 200 5/150 GL column for the 4 reactions with ORF3 C20 protein collecting 100 μL fractions. The 280 nm absorbance curve is used to monitor the purification. (b) 4-20% SDS-PAGE with fractions collected in the SEC purification (a) with Coomassie staining.

The eluted peaks of the first two reactions are at ~1.52 mL (reaction 1) and ~1.60 mL (reaction 2) correspond to the MSP protein aggregation peak while in the SDS-PAGE analysis only MSPD1ΔH5 protein band is detected. In the reaction (3), a peak at ~1.77 mL is eluted corresponding to the assembled "empty" NDs peak with only MSPD1ΔH5 protein band detected. The eluted peak in the reaction (4) at ~1.80 mL is hardly detected on SDS-PAGE (slight band) because of the low concentration and corresponds to the "free" MSPD1ΔH5 protein. Based on these results, the ORF3 C20 protein is not attached to the NDs.

The presence of modified PE-MCC lipids is verified by SAMSA Fluorescein (A-685) reagent (5-((2-(and-3)-S-(acetylmercapto)succinoyl)amino)fluorescein) (Molecular Probes). SAMSA fluorescein reagent reacts with the maleimide groups forming thiol-containing fluorescent protein conjugates with excitation at 495 nm and emission at 520 nm. Its molecular weight is 521 Da and

the extinction coefficient is ~80,000 cm^{-1}M^{-1} at 495 nm. The removal of the acetyl protecting group with 0.1 M NaOH activates the SAMSA fluorescein which is then neutralized with 6 M HCl and buffered with 0.5 M Sodium Phosphate pH 7 before use[198]. The conjugates will be then confirmed with MALDI-TOF (Matrix Assisted Laser Desorption Ionization-Time Of Flight mass spectrometry) analysis and the detection of the molecular weight difference.

The fractions of the 4 reactions of NDs with ORF3 protein are incubated with the activated SAMSA fluorescein at 10-fold molar excess at 23°C for 30 min at 300 rpm in the dark using the Eppendorf ThermoMixer machine. Then, the unbound SAMSA fluorescein reagent is removed by buffer exchange with MSP buffer using a 0.5 mL Zeba Spin Desalting column (7K MWCO) (Thermo Fisher Scientific). The verification of the existence of the modified lipids is performed using MALDI-TOF analysis. The samples are prepared using Reversed-Phase ZipTip$_{C4}$ pipette tips and then measured using the α-Cyano-4-hydroxycinnamic acid (HCCA) (Sigma-Aldrich) matrix in ratio 1:1. In the MALDI-TOF mass spectra, no peak is detected and therefore either the modified lipids are not included in the ND assembly or the procedure was not conducted correctly.

Due to these results, the amount of modified PE-MCC lipids on the NDs assembly procedure is increased to 10% while the amount of DOPC and DOPS lipids decreases to 70% and 20%, respectively. The assembled "empty" NDs are purified by Superdex 200 10/300 GL column and the fractions C6-C11 (peak 1), C12-D4 (peak 2) and D5-D7 (peak 3) are pooled (Figure 94).
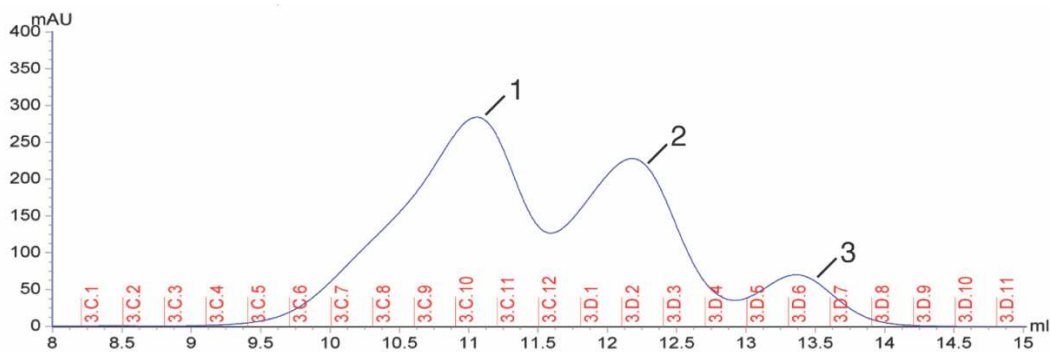
Figure 94. (a) Chromatogram of SEC purification using Superdex 200 10/300 GL column for the "empty" NDs with mixture of 70% DOPC/20% DOPS/10% PE-MCC lipids collecting 300 µL fractions. The 280 nm absorbance curve is used to monitor the purification.

The two thirds of pooled fractions corresponding to peak 1 of the SEC chromatogram are then incubated with 200 µL of 217.7 µM ORF3 C20 protein at 23°C overnight at 300 rpm. Both samples

before and after the addition of ORF3 protein are purified by Superdex 200 5/150 GL column and the chromatograms are overlayed as shown in Figure 95a. A 4-20% SDS-PAGE with the fractions collected in the both SEC runs are shown in Figure 95b.
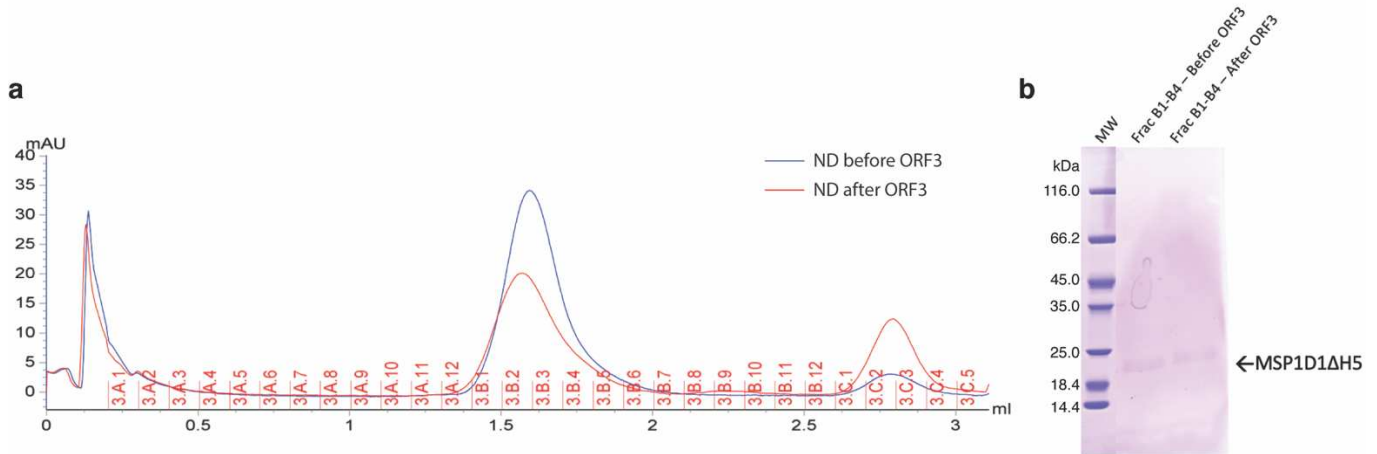
Figure 95. (a) Overlay of SEC chromatograms before (in blue) and after (in red) incubation with ORF3 C20 protein of assembled NDs containing 70% DOPC/20% DOPS/10% PE-MCC lipids using Superdex 200 5/150 GL column. The 280 nm absorbance curve is used to monitor the purifications. (b) 4-20% SDS-PAGE with fractions collected in the SEC purifications (a) with Coomassie staining.

The exact overlap of the SEC peaks and also the absence of the ORF3 band on the SDS-PAGE indicate that the ORF3 C20 protein is not conjugated through the thiol group of Cys20 on the maleimide group of the PE-MCC modified lipids of the "empty" NDs.

Therefore, the optimal conditions for the ORF3 attachment in the NDs assembly containing modified PE-MCC lipids have to be further investigated.

2.2.2.5 "Empty" Nanodiscs with DOPC/DSG-NTA(Ni) lipids – Attachment of ORF3 protein

Due to the negative results using the PE-MCC modified lipids and attachment using ta cysteine-reactive function, the strategy for the protein attachment on the NDs assembly had to be changed. Liu *et al*. 2019 used the modified DGS-NTA(Ni) phospholipids in a mixture with DOPC lipids[199]. DGS-NTA(Ni) lipids contain nickel ion and thus recombinant proteins containing 6xHis-tag can be bound (Figure 96).
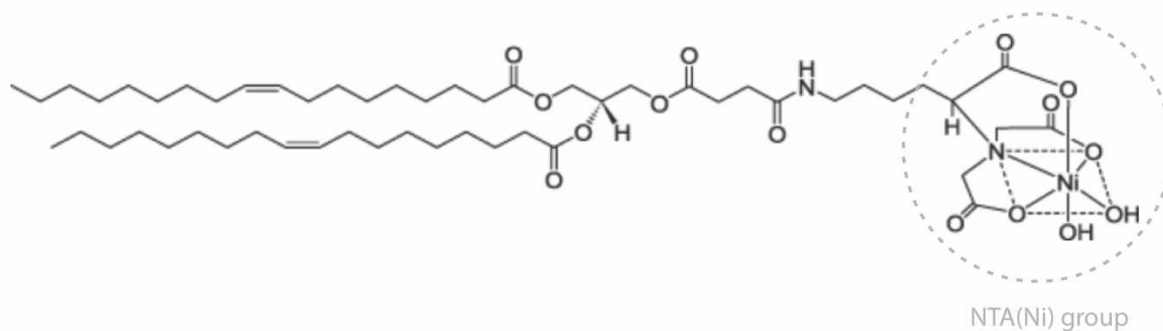
Figure 96. Structure of DGS-NTA(Ni) modified lipid with gray dashed circle indicating the NTA(Ni) group. Available on Avanti Polar Lipids website[200].

The ORF3 C20 construct used in the NDs assembly contains a C-terminal 6xHis-tag. The two hydrophobic regions of HEV ORF3 protein, the Cysteine-rich and the predicted transmembrane region, involved in the membrane anchoring, are located in the N-terminus. In order to bind the ORF3 protein and bring the N-terminal region closer to the ND assembly, the His-ORF3 C20 construct, which contains the 6xHis-tag in N-terminal region without any cleavage site, was designed.

In order to increase the possibility of the protein attachment and based on Liu *et al*. 2019 studies[199], the percentage of the mixture of DOPC and modified DGS-NTA(Ni) lipids in NDs assembly initially tested is 95% and 5% (ratio 19:1), respectively. The 300 µL assembled NDs sample is directly incubated with 260 µL of 216.1 µM His-ORF3 C20 protein at 23°C at 300 rpm overnight using the Eppendorf ThermoMixer machine. Analytical SEC purifications before and after the addition of protein using the Superdex 200 5/150 GL column as well a 4-20% SDS-PAGE with the fractions collected in the second SEC run are performed and shown in Figure 97.
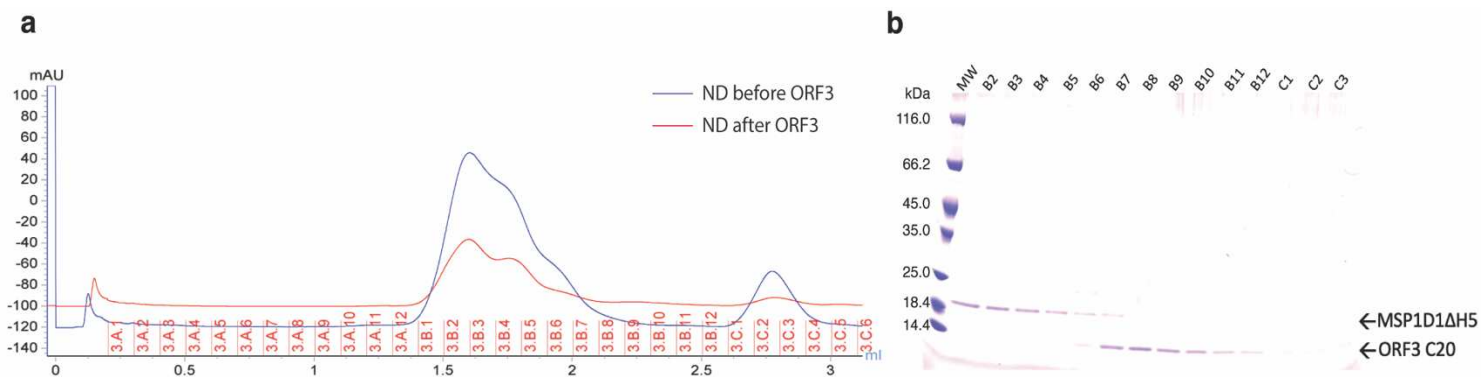
Figure 97. (a) Overlay of SEC chromatograms before (in blue) and after (in red) incubation with His-ORF3 C20 protein of assembled NDs containing 95% DOPC/5% DGS-NTA(Ni) lipids using Superdex 200 5/150 GL column. The 280 nm absorbance curve is used to monitor the purifications. (b) 4-20% SDS-PAGE with fractions collected in the second SEC run (in red) with Coomassie staining.

In the SEC profile of these samples, the MSP protein aggregation, assembled NDs and "free" MSP1D1ΔH5 protein peaks are firstly observed and eluted in the same size as the previous ones. In addition, based on the overlay of the SEC chromatograms and the SDS-PAGE, the protein does not bind to the "empty" NDs and therefore in the next experiment the amount of DOPC to DGS-NTA(Ni) lipids is adjusted to 90% to 10% (ratio 9:1). The assembled NDs is purified by Superdex 200 10/300 GL column and the fractions of C6-C11 (peak 1), C12-D4 (peak 2) and D5-D7 (peak 3) are pooled. The two thirds of the first eluted fraction (C6-C11) are then incubated with 200 µL of 216.1 µM His-ORF3 C20 protein at 23ºC overnight at 300 rpm and further purified by Superdex 200 10/300 GL column collecting 300 µL fractions. In Figure 98, the overlay of the SEC chromatogram of the "empty" assembled NDs with the chromatogram of the reaction with His-ORF3 C20 protein (a) and the 4-20% SDS-PAGE with the fractions of the reaction' SEC run (b) are shown.

166
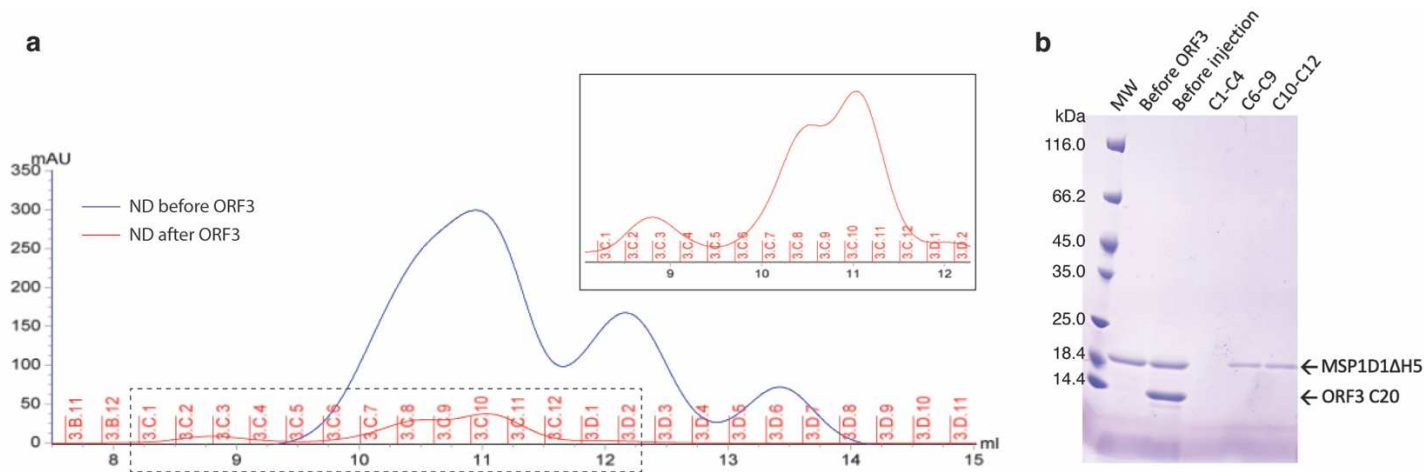
Figure 98. (a) Overlay of SEC chromatograms before (in blue) and after (in red – zoomed panel) incubation with His-ORF3 C20 protein of assembled NDs containing 90% DOPC/10% DGS-NTA(Ni) lipids using Superdex 200 10/300 GL column. The 280 nm absorbance curve is used to monitor the purifications. (b) 4-20% SDS-PAGE with fractions collected in the second SEC run (in red) with Coomassie staining.

In the SEC chromatogram of the reaction with His-ORF3 C20 protein, a small shifted peak is detected at ~8.80 mL. Although the ORF3 band is not detected in the SDS-PAGE because of the low concentration of the sample, this is the first clue of the protein attachment on the NDs.

Due to these results, the NDs assembly are then prepared using 80% DOPC and 20% DGS-NTA(Ni) lipids (ratio 4:1). The purification of assembled NDs by Superdex 200 10/300 GL column and the reaction of peak 1 (fractions C6-C11) with ^{15}N His-ORF3 C20 labeled protein (two thirds of the peak 1 sample incubates with 200 µL of 216.1 µM His-ORF3 C20 protein at 23°C at 300 rpm overnight) are performed as described above with DOPC/DGS-NTA(Ni) lipids ratio 9:1. Figure 99 illustrates the overlay of the SEC chromatogram of the "empty" assembled NDs with the chromatogram of the reaction with His-ORF3 C20 protein (500 µL injection run) (a) and the 4-20% SDS-PAGE with the fractions of the reaction' SEC run (b).
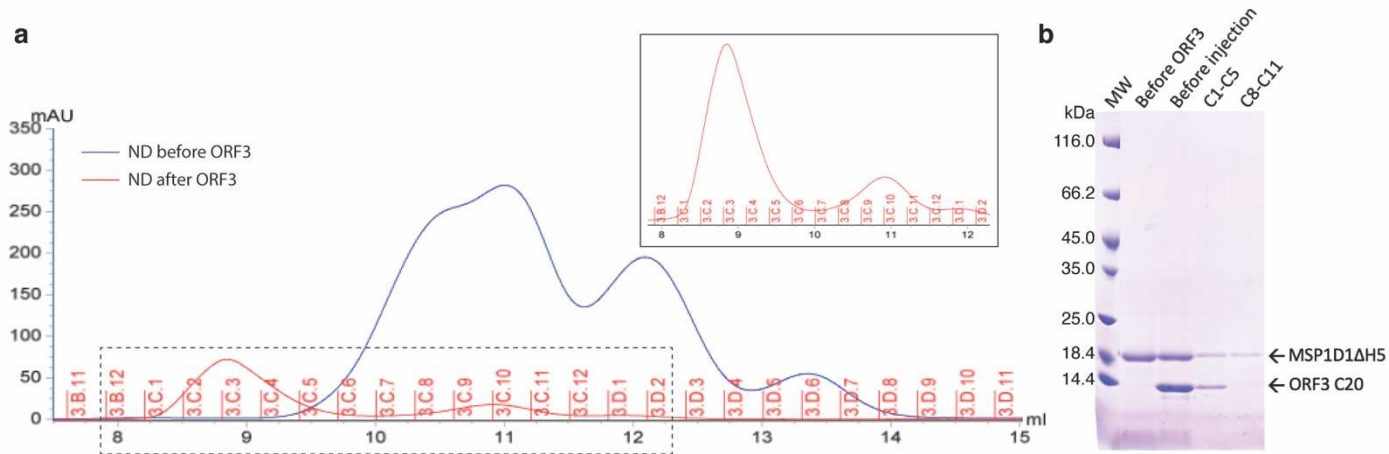
Figure 99. (a) Overlay of SEC chromatograms before (in blue) and after (in red – zoomed panel) incubation with His-ORF3 C20 protein of assembled NDs containing 80% DOPC/20% DGS-NTA(Ni) lipids using Superdex 200 10/300 GL column. The 280 nm absorbance curve is used to monitor the purifications. (b) 4-20% SDS-PAGE with fractions collected in the second SEC run (in red) with Coomassie staining.

The second band in the fourth lane of the 4-20% SDS-PAGE (Figure 99b) confirms that the shifted peak eluted at ~8.88 mL (fractions C1-C5) contains the assembled NDs with His-ORF3 C20 protein attached. The overlay of the analytical runs (100 µL injection run using Superdex 200 5/150 GL column) before and after the His-ORF3 C20 protein incubation clearly shows the peak shift and is further confirmed by SDS-PAGE (Figure 100).
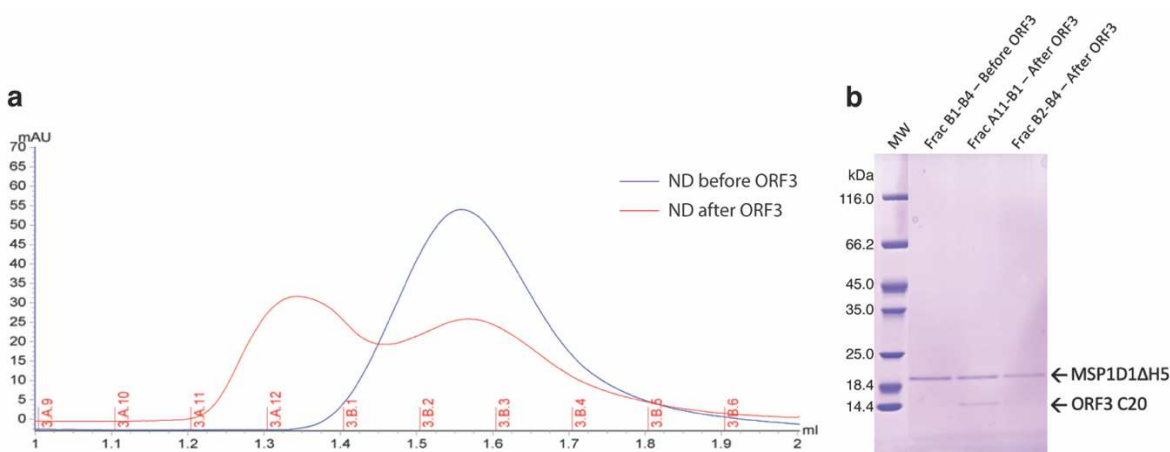
Figure 100. (a) Overlay of SEC chromatograms before (in blue) and after (in red) incubation with His-ORF3 C20 protein of assembled NDs containing 80% DOPC/20% DGS-NTA(Ni) lipids using Superdex 200 5/150 GL column (analytical runs). The 280 nm absorbance curve is used to monitor the purifications. (b) 4-20% SDS-PAGE with fractions collected in the SEC runs with Coomassie staining.

After incubating the remaining peak 1 sample of the initial SEC step with ^{15}N His-ORF3 C20 protein and performing SEC purifications for the entire sample, the fractions C1-C5 of all runs are concentrated up to 280 µL and placed in a Shigemi tube to record a 2D ^{1}H, ^{15}N HSQC spectrum at 293K on 900 MHz Spectrometer. The same spectrum is acquired for ^{15}N His-ORF3 C20 protein

168

alone because its protein sequence slightly differs from ORF3 C20 protein in which the backbone assignments are already obtained. This experiment will provide us more information about the affected residues of ORF3 C20 protein when it is attached in the NDs and therefore about the

The overlay of the two spectra is shown in Figure 101, the control spectrum in red and the one with assembled NDs with ^{15}N His-ORF3 C20 protein attached in cyan.
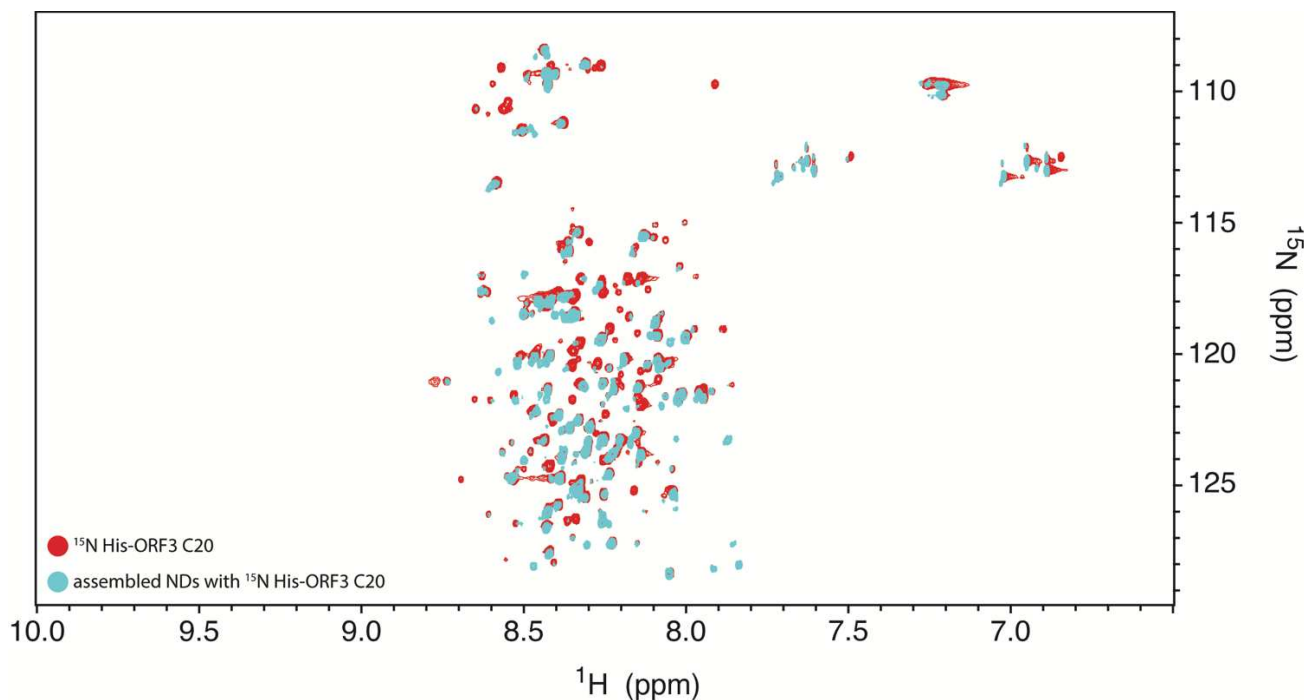
Figure 101. Overlay of ^{1}H, ^{15}N HSQC spectra of ^{15}N His-ORF3 C20 protein in red and assembled NDs with ^{15}N His-ORF3 C20 protein attached in cyan.

Looking closer to the Glycine area in the 2D HSQC spectrum of the control sample (in red), more Glycine peaks than expected are observed. Figure 102 shows the sequence alignment of ORF3 C20 and His-ORF3 C20 proteins using the multiple sequence alignment tool Clustal Omega[126,127]. The total number of glycine residues is 11 and 13 in His-ORF3 C20 and ORF3 C20, respectively. When compared to ORF3 C20, in His-ORF3 C20 there is an extra Gly in the N-terminal region before the 6xHis-tag (Gly2 in His-ORF3 C20 sequence) while 3 Gly are missing at the end (Gly114, Gly121 and Gly123 in ORF3 C20 sequence). The Glycine residues differ in the two constructs are marked with a red box in Figure 102.

```
His-ORF3-C20    MGSHHHHHHHGSPSALGLFSSSSSSSFSLCSPRHRPASRLAVVVGGAAAVPAVVSGVTGLI    60
ORF3-C20        ---------MGSPSALGLFSSSSSSSFSLCSPRHRPASRLAVVVGGAAAVPAVVSGVTGLI    51
                         *********************************************************

His-ORF3-C20    LSPSPSPIFIQPTPSPPISFHNPGLELALGSRPAPLAPLGVTSPSAPPLPPAVDLPQLGL    120
ORF3-C20        LSPSPSPIFIQPTPSPPISFHNPGLELALGSRPAPLAPLGVTSPSAPPLPPAVDLPQLGL    111
                ***********************************************************

His-ORF3-C20    RR----------------        122
ORF3-C20        RRGLEVLFQGPGHHHHHH        129
                **
```

Figure 102. Sequence alignment of His-ORF3 C20 and ORF3 C20 proteins using the multiple sequence alignment tool Clustal Omega[126,127].

The 2D ^{1}H, ^{15}N HSQC spectrum of ^{15}N His-ORF3 C20 protein (in red) is then compared to the one for ^{15}N, ^{13}C ORF3 C20 protein recorded for the assignment procedure (in cyan) as shown in Figure 103. In the zoom panel, the assigned Gly residues of ORF3 C20 protein are marked. The Gly114, Gly121 and Gly123 peaks are present and exactly overlap in the 2D HSQC spectrum of ^{15}N His-ORF3 C20 protein, and there is also an additional peak possibly corresponding to the Gly2 of the His-ORF3 sequence.
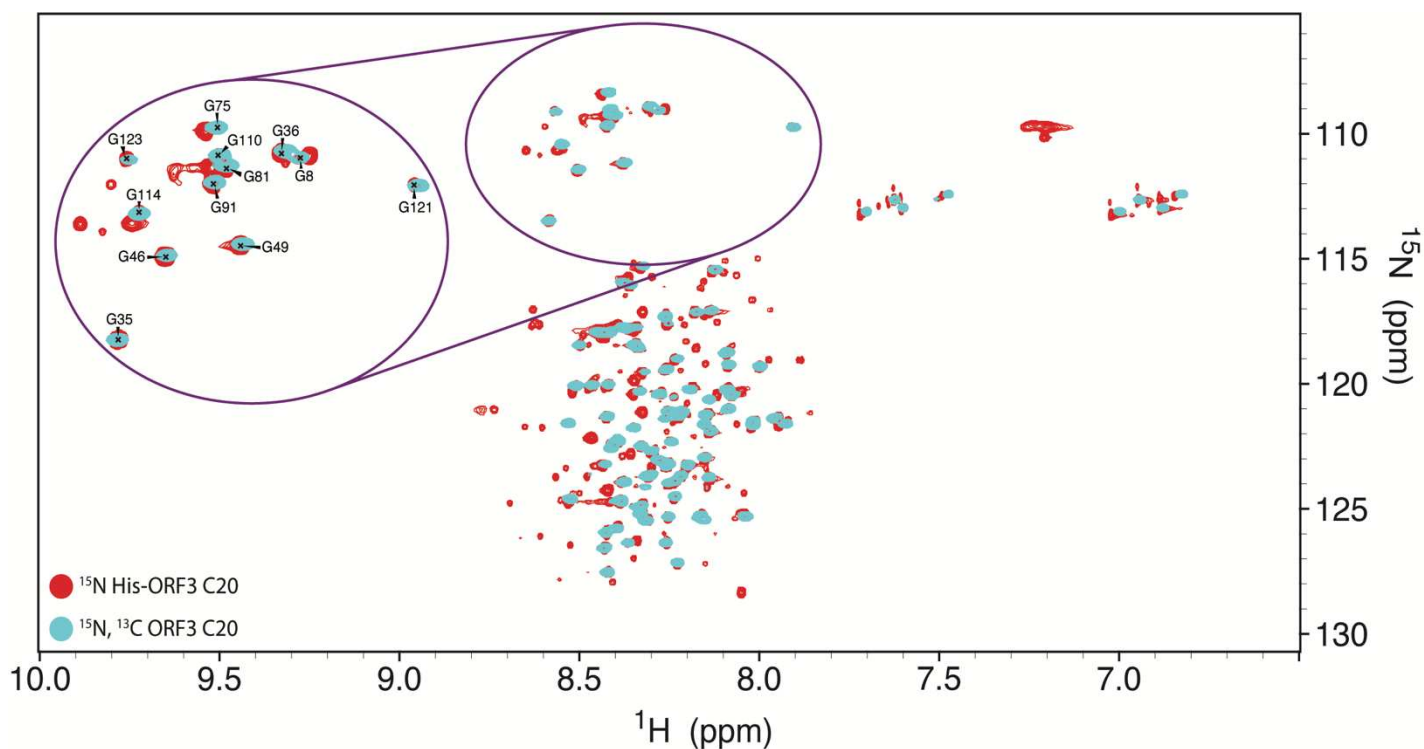
Figure 103. Overlay of ^{1}H, ^{15}N HSQC spectra of ^{15}N His-ORF3 C20 protein in red and ^{15}N, ^{13}C ORF3 C20 protein in cyan.

The existence of the Gly peaks of the C-terminal region of ORF3 C20 protein may indicates, for an unknown reason, the presence of both ORF3 protein constructs in the NMR sample. In order to further investigate this assumption, MALDI-TOF measurements are performed. We analyzed the NMR ^{15}N His-ORF3 C20 sample, a sample from the same ^{15}N His-ORF3 C20 protein stock, a ^{15}N ORF3 C20 sample and a mixture of ^{15}N His-ORF3 C20 and ^{15}N ORF3 C20 in ratio 1:1. The samples are prepared using Reversed-Phase ZipTip$_{C4}$ pipette tips and then crystalized using the α-Cyano-4-hydroxycinnamic acid (HCCA) (Sigma-Aldrich) matrix in ratio 1:1 (v/v). The overlay of the MALDI-TOF mass spectra of all four samples and the main peaks detected during the measurements are shown in Figure 104.
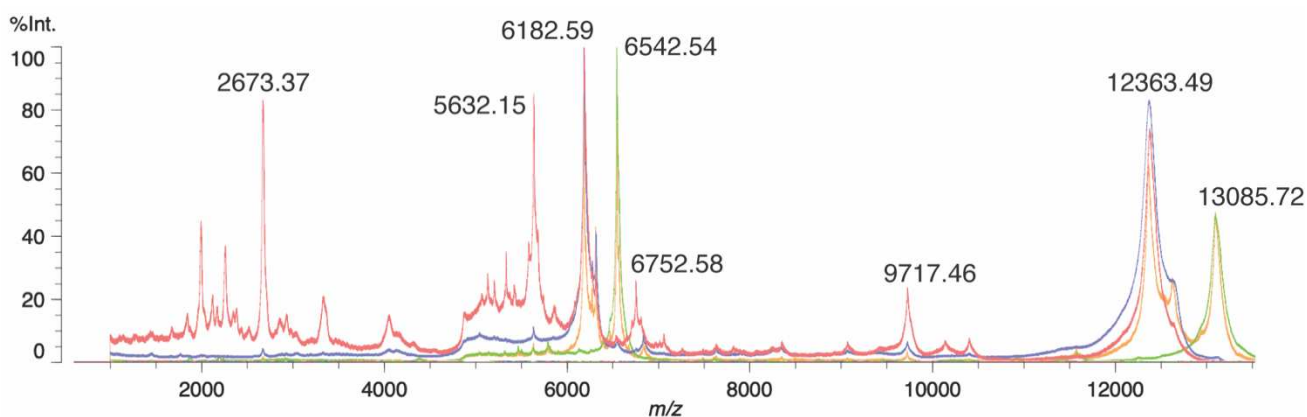
Expected Molecular Weight of labeled proteins

^{15}N His-ORF3 C20 : 12348.64 Da

^{15}N ORF3 C20 : 13071.46 Da

— NMR ^{15}N His-ORF3 C20 sample
— ^{15}N His-ORF3 C20
— ^{15}N ORF3 C20
— ^{15}N His-ORF3 C20 + ^{15}N ORF3 C20

Figure 104. Overlay of mass spectra of the NMR ^{15}N His-ORF3 C20 sample in red, a sample from the same ^{15}N His-ORF3 C20 protein stock in magenta, a ^{15}N ORF3 C20 sample in green and a mixture of ^{15}N His-ORF3 C20 and ^{15}N ORF3 C20 in ratio 1:1 sample in orange using MALDI-TOF analysis. The main peaks detected in the measurements are marked on the overlay.

In the mass spectrum of the NMR ^{15}N His-ORF3 C20 sample (in red), one peak at 12363.49 Da is detected which correspond to the molecular weight of ^{15}N His-ORF3 C20 protein (including ^{15}N labeling and removal of the first Met) and thus the NMR tube contains only this construct. There are other four main peaks at 2673.37 Da (named as peak 1), 5632.15 Da (named as peak 2), 6752.58 Da (named as peak 3) and 9717.46 Da (named as peak 4) which the combination of two and especially the peaks 1 with 4 and peaks 2 with 3, correspond to the molecular weight of ^{15}N His-ORF3 C20 protein. This observation indicates that the protein is probably partially cleaved

into two regions. Of note, the MS analysis was performed on the NMR tube several weeks after it has been prepared for NMR experiments. Using the Expasy PeptideCutter tool[183] providing the His-ORF3 C20 protein sequence, the potential cleavage sites along the sequence caused by various proteases are predicted. Using the information provided from this tool and after several trials using the Protein Parameter Calculator[201] tool which provides the molecular weight of samples with isotopic labeling schemes, for each pair of peaks two potential cleavage regions of ^{15}N His-ORF3 C20 protein are predicted (Table 4).

Table 4. Potential cleavage regions of ^{15}N His-ORF3 C20 protein predicted by using Protein Parameter Calculator[201].

Sequence	Molecular Weight (Da)	MALDI-TOF mass peak	Cleavage position (aa)
^{2}GSHHHHHHHGSPSALGLFSSSSSSF^{26}	2660.42	1	
^{27}SLCSPRHRPASRLAVVVGGAAAVPAVVSGVTGLILSPSPSPIFIQPTPSPPISFHNPGLELALGSRPAPLAPLGVTSPSAPPLPPAVDLPQLGLRR^{122}	9706.24	4	26
^{2}GSHHHHHHHGSPSALGLFSSSSSSFSLCSPRHRPASRLAVVVGGAAAVPAVVSGVTGLILSPSPSPIFIQPTPSPPISFHNPGLELALGSRPAPL^{96}	9713.84	4	96
^{97}APLGVTSPSAPPLPPAVDLPQLGLRR^{122}	2652.81	1	
^{2}GSHHHHHHHGSPSALGLFSSSSSSFSLCSPRHRPASRLAVVVGGAAAVPAVVSGVT^{57}	5669.56	2	57
^{58}GLILSPSPSPIFIQPTPSPPISFHNPGLELALGSRPAPLAPLGVTSPSAPPLPPAVDLPQLGLRR^{122}	6697.09	3	
^{2}GSHHHHHHHGSPSALGLFSSSSSSFSLCSPRHRPASRLAVVVGGAAAVPAVVSGVTGLILSPSPSPI^{68}	6742.74	3	68
^{69}FIQPTPSPPISFHNPGLELALGSRPAPLAPLGVTSPSAPPLPPAVDLPQLGLRR^{122}	5623.92	2	

The protein cleavage could be occurred because of the protein storage at 4°C for about two weeks before performing the MALDI-TOF analysis. Therefore, the recording of a 2D ^{1}H, ^{15}N HSQC spectrum, a MALDI-TOF measurement as well SDS-PAGE analysis for a new ^{15}N His-ORF3 C20 protein sample have to be conducted again in the same day. Using a new protein stock to perform these experiments, the HSQC spectrum is identical with the one recorded previously. Also, the mass spectrum obtained by MALDI-TOF measurement contains two peaks at 12363.49 Da [M+H]$^+$ and 6180.78 Da [M+2H]$^{2+}$ as expected. In the SDS-PAGE, a slight difference between the His-ORF3 C20 and ORF3 C20 proteins can be detected and also the two ^{15}N His-ORF3 C20 samples from the two NMR samples prepared for the control experiment are in the same size.

In addition, in the mass spectra of the ^{15}N His-ORF3 C20 sample (in magenta) and ^{15}N ORF3 C20 sample (in green), two peaks at 12363.49 Da [M+H]$^+$ and 6182.59 Da [M+2H]$^{2+}$ and two peaks at 13085.72 Da [M+H]$^+$ and 6542.54 Da [M+2H]$^{2+}$, respectively, are detected as expected. The mass spectrum of the mixture of ^{15}N His-ORF3 C20 and ^{15}N ORF3 C20 in ratio 1:1 sample (in orange) contain these four peaks as expected.

Based on all the above information and regarding the 2D HSQC spectrum of assembled NDs with ^{15}N His-ORF3 C20 protein attached in Figure 101, a reliable conclusion could not be drawn.

2.3 Transmembrane insertion study using liposomes

In order to reconstitute the ORF3 protein in a lipid environment and thus, study transmembrane insertion and simultaneously its oligomerization state, a preparation of liposomes using *E. coli* polar lipid extract (Avanti Polar Lipids) is performed. It is the total lipid extract derived from *E. coli*B (ATCC 11303) during the growth phase of Kornberg Minimal medium culture incubated at 37ºC, mimics the *E. coli* inner membrane and contains phosphatidylethanolamine (PE), phosphatidylglycerol (PG) and cardiolipin lipids. A 10 mg/mL stock in sodium cholate buffer was prepared after vortex and sonication steps in a heated sonication bath for 20 min total time. Three different lipid concentration (1, 2 and 5 mg/mL) samples were mixed with 100 µM ORF3 C20 protein, which immediately precipitated. The foggy samples were incubated at 23ºC at 300 rpm overnight and the next day both supernatant and pellet samples for the three mixtures and also a ORF3 C20 protein sample (as control sample) were loaded in a native PAGE. In native PAGE analysis, proteins are separated depending their size, net charge and eventually their structure while it is performed at 4ºC with constant 90 V for 3 hours. No band could be detected after the Coomassie blue staining without reasonable explanation. The *E. coli* polar lipid extract was insoluble in other buffers containing detergents such as Triton X-100, so this approach could not be further studied.

2.4 Solid-state NMR analysis of ORF3 C20 protein membrane anchoring

Although the protein is pure and soluble in NMR buffer (50 mM Sodium Phosphate pH 6.1, 50 mM NaCl, 0.1 mM EDTA, 1 mM DTT) up to ~200 μM concentration, ORF3 C20 protein possesses a strong tendency to interact with hydrophobic (or amphiphilic) molecules, such as membrane lipids and detergents, having as a result the precipitation of the protein as described above. In collaboration with Dr. Anja Böckmann and two members of her team, Dr. Marie-Laure Fogeron (research assistant) and Dr. Lauriane Lecoq (researcher), Molecular Microbiology and Structural Biochemistry – Protein Solid State NMR group in Lyon, solid-state NMR experiments of ORF3 C20 protein are performed to gain more information regarding its membrane determinant(s). The objective was to reconstruct and then sediment the ORF3 C20 protein in lipid mixture (phosphatidylcholine-cholesterol) and record the 2D ^{13}C-^{13}C DARR, 2D HC-INEPT, 2D NCA, 2D NCO, 3D NCACX, 3D NCOCX and 3D CANCO spectra on a ^{15}N, ^{13}C labeled sample which would be used for the assignments of ORF3 resonances in interaction with lipids. Compared with the obtained liquid-state NMR data and protein assignments in absence of lipids, the residues potentially involved in the interaction with the membrane during HEV infection could be determined. Two labeled ORF3 C20 samples, 10 mg (213 μM) of ^{15}N ORF3 C20 and 25 mg (203 μM) of ^{15}N, ^{13}C ORF3 C20 protein, are prepared and sent in Lyon. Marie-Laure Fogeron conducted the tests for the protein preparation using ^{15}N ORF3 C20 sample while Lauriane Lecoq was responsible for carrying out the NMR experiments using ^{15}N, ^{13}C ORF3 C20 sample.

The first experiment conducted was the addition of PC/Chol at LPR0.5 or 1 (Lipid-to-Protein Ratio) lipids in presence and absence of mβCD to ORF3 C20 sample to achieve the protein sedimentation. In Figure 105, the protocol scheme of sample preparation (a) and the 4-20% SDS-PAGE (b) for all the tested conditions' samples are shown resulting in non-precipitation of ORF3 protein in any case.

a

^{15}N ORF3 C20 protein ──────────────────────→ SDS-PAGE
sent to Lyon @ 4°C   (6 μg, input)
frozen in liquid nitrogen upon arrival, and stored @ -80°C
8x 100 μg in 36 μL

PC/Chol in Triton X-100 @ 5 mg/mL

incubate 2h @ 4°C

conditions w/o mβCD   **conditions with mβCD**

30 min @ 20.000*g*, 4°C   add mβCD stepwise (7 steps, 1h each) :

                                 5 μL @ 10% for LPR0.5, and 10 μL @ 10% for LPR1

Pellet   **SN**   incubate 36h @ 4°C

5 min @ 20.000*g*, 4°C   30 min @ 20.000*g*, 4°C

Pellet   **SN** (pooled with first SN)

                                Pellet   **SN**

SDS-PAGE   5 min @ 20.000*g*, 4°C

                                Pellet   **SN** (pooled with first SN)

                                        SDS-PAGE

b

#1: no lipids, no mBCD
#2: PC/Chol LPR0.5, no mBCD
#3: PC/Chol LPR1, no mBCD

#1: no lipids, no mBCD, 36h @ 4°C
#4: PC/Chol LPR0.5 + mBCD
#5: PC/Chol LPR1 + mBCD
#6: no lipids, mBCD (a)
#7: no lipids, mBCD (b)

Figure 105. (a) Protocol scheme for sample preparation with addition of PC/Chol at LPR0.5 or 1 lipids in presence (with) and absence (w/o) of mβCD. (b) 4-20% SDS-PAGE for all tested conditions' samples indicated on the scheme (a) with Coomassie blue staining. P: pellet, SN: supernatant. Figure prepared by Dr. Marie-Laure Fogeron and Dr. Anja Böckmann.

176

Because the protein stayed soluble with the addition of the PC/Chol lipids and knowing that sodium cholate buffer induces the ORF3 precipitation, the next experiment was the addition of sodium cholate buffer (100 mM sodium cholate in MSP buffer containing 20 mM Tris-HCl pH 7.7, 100 mM NaCl, 0.5 mM EDTA) in presence and absence of PC/Chol at LPR0.8 or 1.4 lipids. In Figure 106, the protocol scheme of sample preparation (a) and the 4-20% SDS-PAGE (b) for all the tested conditions' samples are shown. Although the ORF3 protein precipitates in presence of sodium cholate buffer without the PC/Chol lipids as expected, interestingly the sample was soluble in the presence of both cholate buffer and lipids.
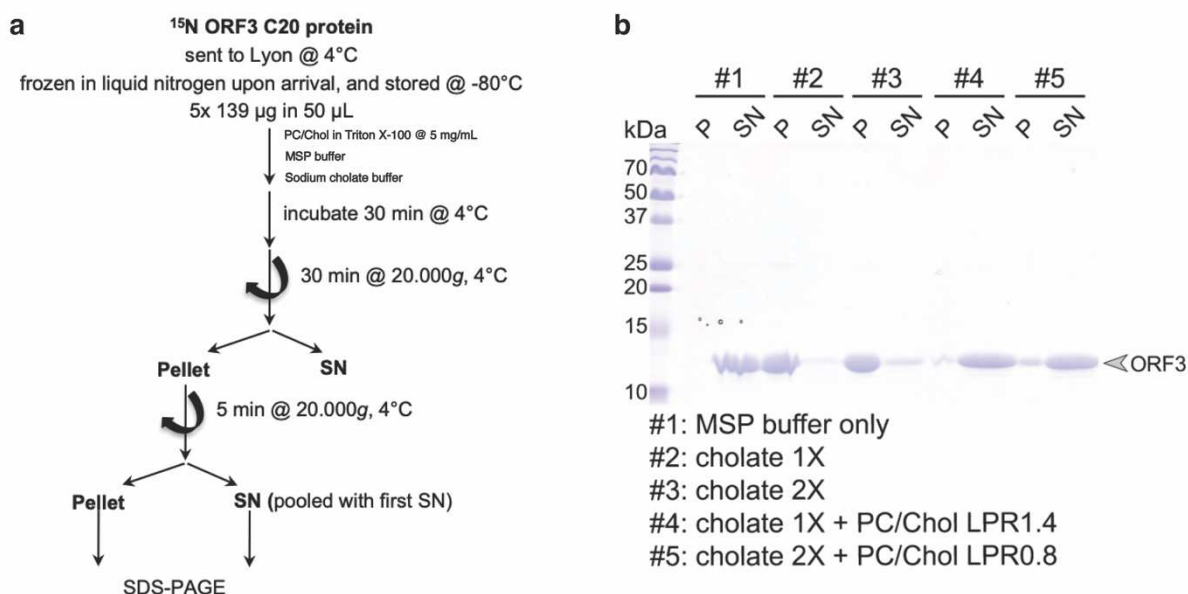
Figure 106. (a) Protocol scheme for sample preparation with addition of sodium cholate buffer in presence and absence of PC/Chol at LPR0.8 or 1.4 lipids. (b) 4-20% SDS-PAGE for all tested conditions' samples indicated on the scheme (a) with Coomassie blue staining. P: pellet, SN: supernatant. Figure prepared by Dr. Marie-Laure Fogeron and Dr. Anja Böckmann.

The last test performed before the final preparation of the double-labeled sample for recording the NMR data was the ultracentrifugation of the samples instead of high-speed centrifugation steps at 20,000 xg. The protein alone, the protein in presence of PC/Chol lipids, the protein in presence of PC/Chol lipids and mβCD and the protein in presence of PC/Chol lipids and sodium cholate buffer were prepared and ultracentrifuged at 200,000 xg for 4 h at 4°C. In Figure 107, the protocol scheme of sample preparation (a) and the 4-20% SDS-PAGE (b) for all the tested conditions' samples are shown.
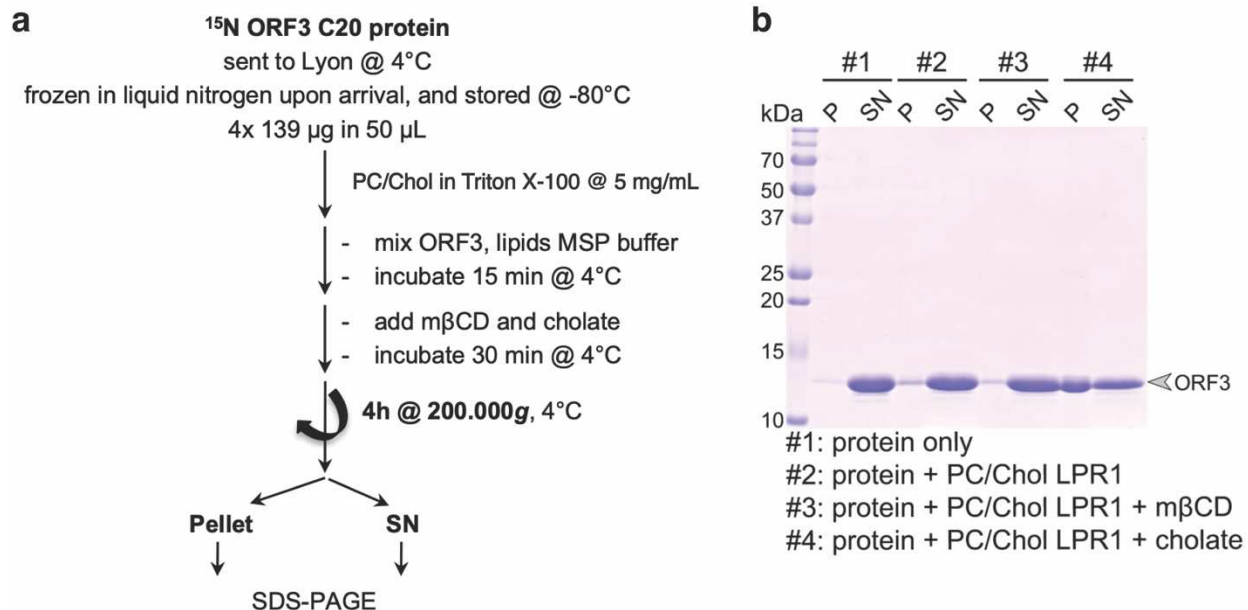
Figure 107. (a) Protocol scheme for sample preparation using an ultracentrifugation step at 200,000 xg for 4h at 4ºC. (b) 4-20% SDS-PAGE for all tested conditions' samples indicated on the scheme (a) with Coomassie blue staining. P: pellet, SN: supernatant. Figure prepared by Dr. Marie-Laure Fogeron and Dr. Anja Böckmann.

The ORF3 C20 protein alone and in presence of PC/Chol LPR1 lipids in with and without mβCD was soluble after the ultracentrifugation step. The only observation for these samples was that the addition of mβCD has a result the formation of a transparent pellet corresponding to Triton X-100-depleted PC/Chol lipids. However, adding the sodium cholate buffer to the protein-lipid mixture and followed the ultracentrifugation step, ORF3 C20 could be only partially sedimented.

Based on the results of all these tests and specifically the unsuccessful sedimentation of the protein without sodium cholate buffer, the ^{15}N, ^{13}C ORF3 C20 labeled protein was mixed only with sodium cholate buffer in order to effectively precipitate all the sample, then a 3.2-mm NMR rotor was filled and finally 2D and 3D solid-state NMR experiments were collected in Lyon by Dr. Lauriane Lecoq. In Figure 108, the procedure for sample preparation (a) and the 4-20% SDS-PAGE (b) for the supernatant samples of each step are shown. The precipitation of the protein with the addition of sodium cholate buffer and the filling of the rotor were sufficient allowed to proceed to the NMR measurements.
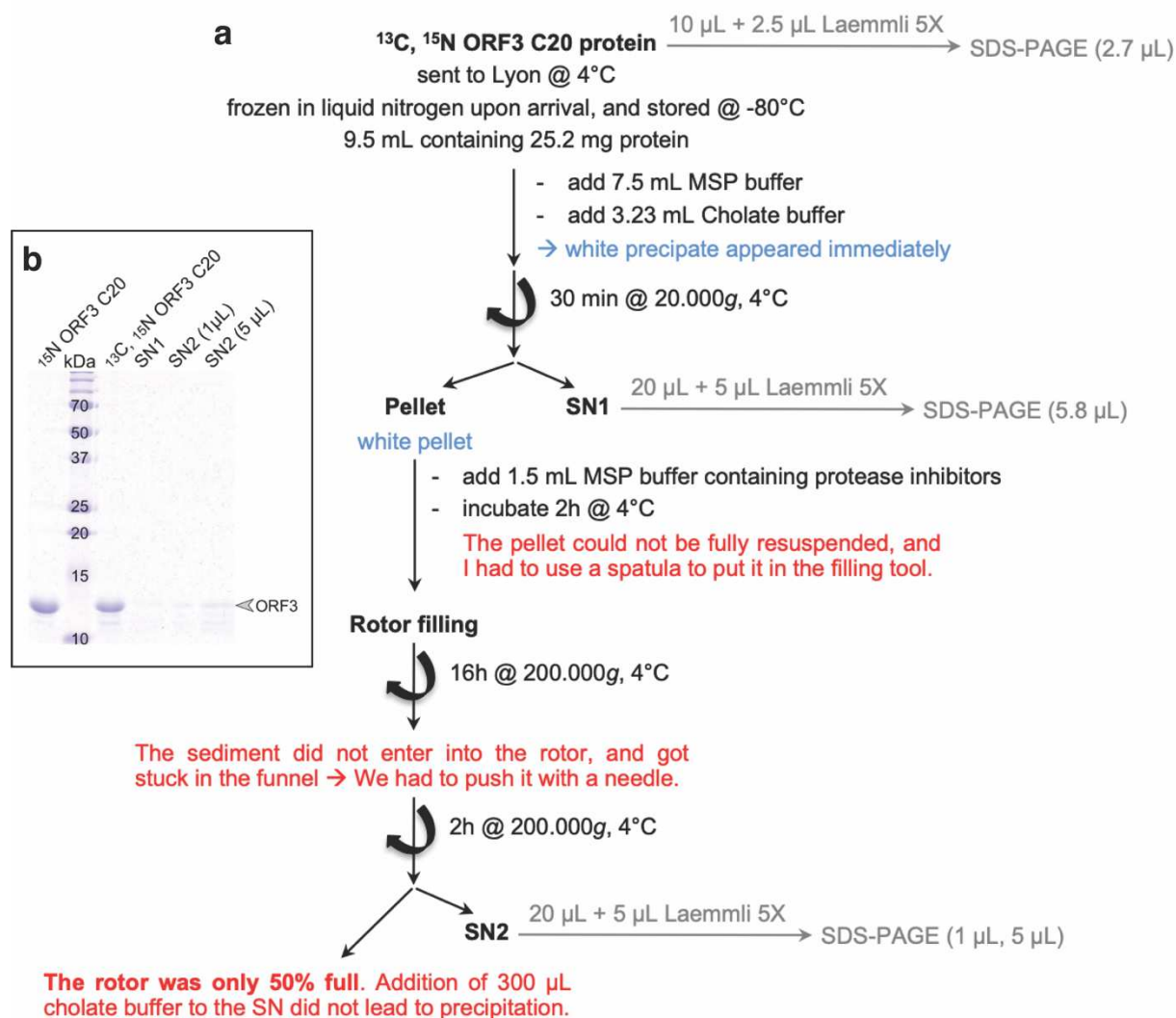
Figure 108. (a) Preparation procedure of the ^{15}N, ^{13}C ORF3 C20 sample for recording 2D and 3D NMR dataset using a 3.2-mm rotor. (b) 4-20% SDS-PAGE for the supernatant samples of each step indicated on the scheme (a) with Coomassie blue staining. Figure prepared by Dr. Marie-Laure Fogeron and Dr. Anja Böckmann.

Although the rotor was only 50% full, it was adequate to record a 2D ^{13}C-^{13}C DARR and a 2D NCA spectra on the 800 MHz Spectrometer with the temperature to be around 293K in the rotor. The NMR signal in both experiments was good as shown in panel (a) in Figure 109, but the profile of ORF3 C20 protein is typical of an aggregated protein with overlapped not dispersed peaks and therefore, the inter-residue contacts for assignment could not be obtained. In panel (b) in Figure 109, the corresponding 2D spectra of a well-folded protein with identical experimental times are represented. Because of the aggregation of the protein, the 3D NMR data could not be recorded.
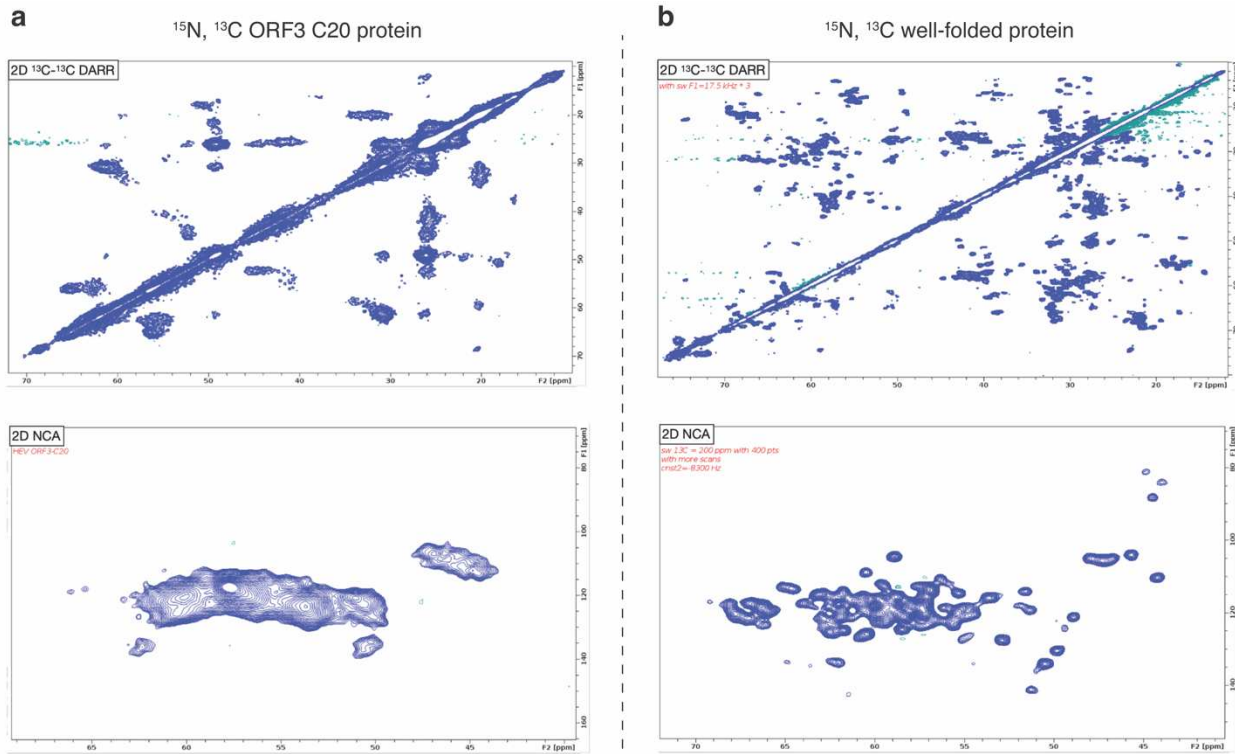
Figure 109. 2D ^{13}C-^{13}C DARR (top) and 2D NCA (bottom) NMR spectra recorded with identical experimental times at 293K at 800 MHz Spectrometer for ^{15}N, ^{13}C ORF3 C20 protein (a) and well-folded protein (b) filled in a 3.2-mm rotor. Figure prepared by Dr. Lauriane Lecoq and Dr. Anja Böckmann.

3. Interaction of HEV ORF3 protein with human Tsg101 UEV domain

The last aim of this study is the interaction of HEV ORF3 protein with other protein partners and especially with Ubiquitin E2 Variant (UEV) domain of human Tumor susceptibility gene 101 (Tsg101) protein, mention as Tsg101 UEV protein, an important ESCRT-I complex component and essential for viral secretion. As mentioned in the HEV ORF3 sequence analysis, a PSAP motif is located in the C-terminal region of HEV ORF3 protein and is close to the PTAP motif found in other viruses, such as HIV[146], Ebola[51] etc., and through this motif they interact with the Tsg101 UEV protein. The characterization of the interaction of the two proteins is performed by various biophysical techniques as solution-state NMR Spectroscopy, Isothermal Titration Calorimetry (ITC), X-ray Crystallography, Thermal Shift Assay (TSA) and Fluorescence Polarization (FP) assays and the obtained results are presented in the following chapters.

3.1 Human Tsg101 UEV protein purification

The purification protocol for the preparation of labeled and unlabeled Tsg101 UEV samples is described in details in Material and Methods. The structural characterization and the interaction with HEV ORF3 protein is performed on Tsg101 UEV construct (147 amino acids, ~16.7 kDa protein). The two other constructs, the Tsg101 UEV FLAG-tag and the Tsg101 UEV GST-tag, are prepared for screening purposes and further described in the Homogeneous Time Resolved Fluorescence (HTRF) chapter. The purification steps for the three constructs differ depending on the presence of TEV Protease cleavage site and the purpose of the sample preparation. The main Tsg101 UEV purification procedure lasts three days including affinity HisTrap chromatography steps, 6xHis-tag cleavage using TEV Protease, overnight dialysis step and protein concentration.

After the cell lysis and the separation from the insoluble *E. coli* cellular debris, the supernatant is purified by affinity HisTrap column and the corresponding chromatogram monitoring using the 280 nm, the 260 nm and the 215 nm UV absorbance (a) and the 4-20% SDS-PAGE with different fractions (b) are shown in Figure 110.
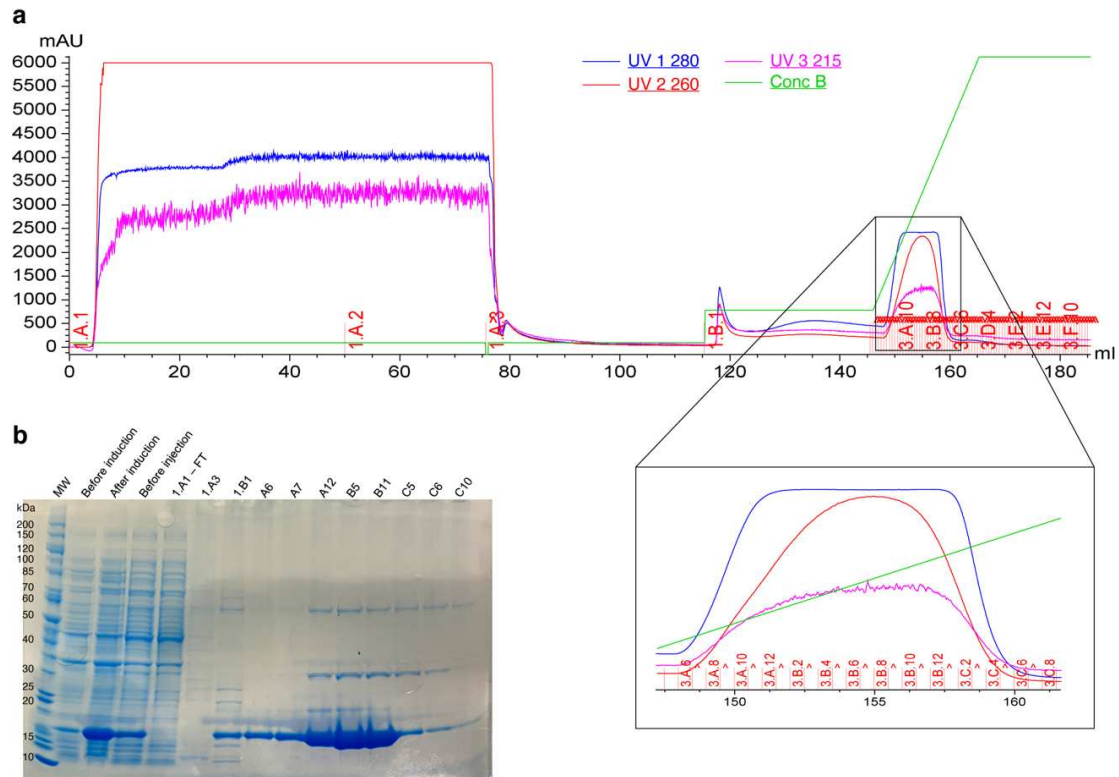
Figure 110. (a) Chromatogram of affinity HisTrap purification of Tsg101UEV protein. In the black box, the elution peak is shown zoomed. Blue: 280 nm, red: 260 nm, magenta: 215 nm and green: concentration of Buffer B. (b) 4-20% SDS-PAGE of fractions based on the chromatogram (a) with Coomassie blue staining.

Then the 6xHis-tag of the purified protein is cleaved using TEV Protease before the second affinity purification step (Figure 111a). The chromatogram of the second HisTrap purification step with the three UV absorbance curves at 280 nm, at 260 nm and at 215 nm (b) and the 4-20% SDS-PAGE with the fractions (c) are shown in Figure 111.
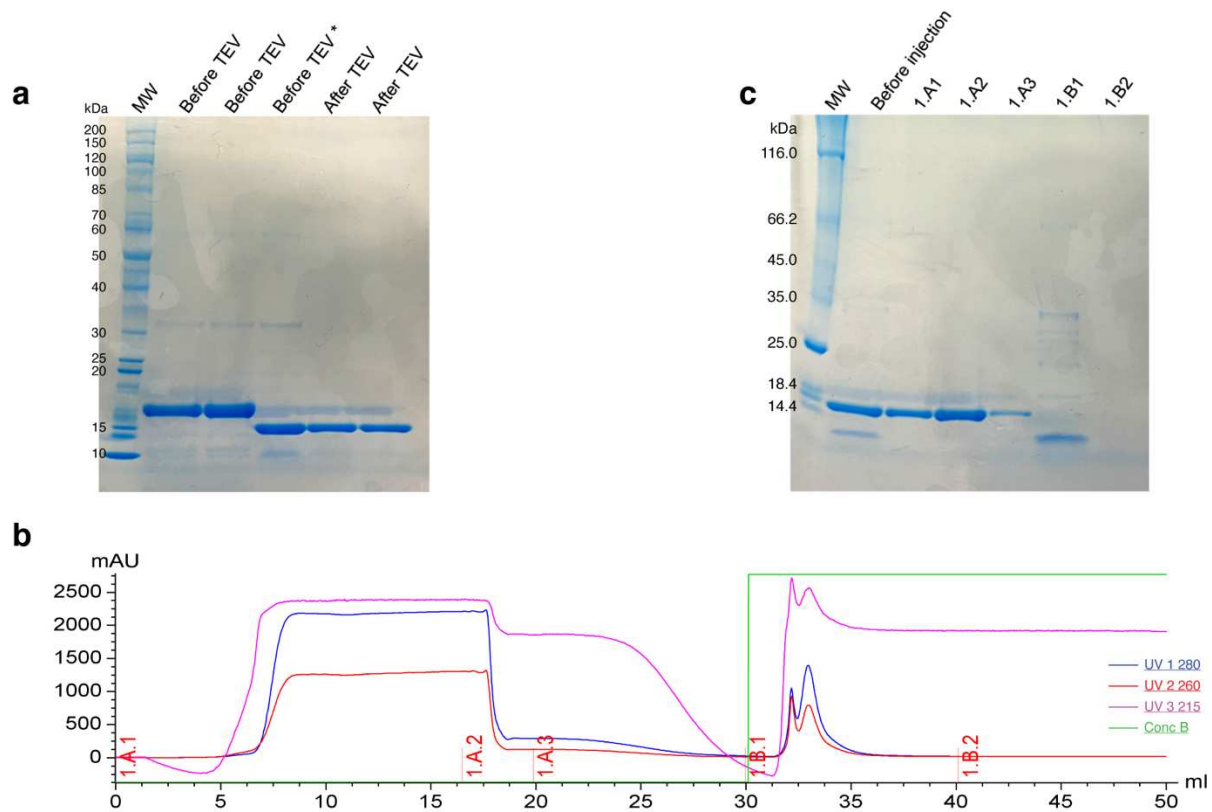
Figure 111. (a) 4-20% SDS-PAGE of pooled Tsg101 UEV protein fractions before and after TEV cleavage with Coomassie blue staining. (b) Chromatogram of second HisTrap purification step of Tsg101 UEV protein after TEV cleavage. Blue: 280 nm, red: 260 nm, magenta: 215 nm and green: concentration of Buffer B. (b) 4-20% SDS-PAGE of fractions based on the chromatogram (b) with Coomassie blue staining. (*) In (a) 4-20% SDS-PAGE, the Before TEV* sample is a sample from a tube that is incubated at room temperature with the protein and TEV protease left on the tube.

The last steps include the overnight dialysis against the NMR buffer (50 mM Sodium Phosphate pH 6.1, 50 mM NaCl, 0.1 mM EDTA), the concentration of the protein and the preparation of the aliquots stored at -80°C until further use.

3.2 NMR characterization of human Tsg101 UEV domain

The first technique used for the characterization of the interaction of HEV ORF3 protein with human Tsg101 UEV domain is NMR Spectrometry. In order to study this interaction process, we need both purified partners and at least one should be with an isotopic labeling. In addition, if the NMR assignments are known, then we can identify, at a per residue level, the binding site on the protein of interest. Regarding ORF3 we indeed purified the protein with various isotopic enrichments and we performed the assignment, which include the proline residues. Due to the fact that the backbone and proline assignments of free human Tsg101 UEV domain are not available in the Biological Magnetic Resonance Data Bank (BMRB) or in any publicly available database, before proceeding with the interaction experiments, we performed the NMR assignment of Tsg101 UEV NMR spectra.

A ^{15}N, ^{13}C Tsg101 UEV labeled protein sample at 330 μM is prepared for recording the NMR experiments needed for backbone and proline assignments at 298K on 600 MHz Spectrometer. A complete NMR dataset containing the 2D ^{1}H, ^{15}N HSQC, 3D ^{1}H, ^{15}N, ^{13}C HNCO, HNCACO, HNCACB, HN(CO)CACB, HNHA, HACAN, HN(CA)NNH, the 2D carbon detection ^{15}N, ^{13}C NCO and ^{13}C, ^{13}C CACO and the 3D carbon detection ^{13}C ,^{15}N, ^{13}C HCACON spectra is recorded. The 2D carbon detection ^{13}C, ^{13}C CACO spectrum provides the correlation between the carbonyl group (CO$_i$) and the carbon Cα (Cα$_i$) of the same residue while the 3D carbon detection ^{13}C ,^{15}N, ^{13}C HCACON spectrum provides the correlation of the carbonyl group (CO$_i$) and the carbon Cα (Cα$_i$) of the same residue with the amide group of the following (N$_{i+1}$) residue.

After TEV cleavage, there are two extra remaining residues in N-terminal region that are not considered in the numbering for the protein assignment. Therefore, the first residue (residue 1) is the Methionine (Met) of human Tsg101 UEV domain and this residue is detectable in the ^{1}H, ^{15}N HSQC spectrum because of the presence of the two extra residues from the N-terminal tag.

Combining the information of all these experiments, the backbone (including proline residues) assignments are achieved for Tsg101 UEV protein in its apo-state. Figure 112 illustrates the assigned 2D ^{1}H, ^{15}N HSQC spectrum of Tsg101 UEV protein.

Figure 112. 2D ^{1}H, ^{15}N HSQC assigned spectrum of ^{15}N, ^{13}C Tsg101 protein. The NH$_2$ side chains (sc) of Trp, Asn and Gln residues are also assigned and the pairs of Asn and Gln side chains are labeled with dashed line.

The 2D ^{1}H, ^{15}N HSQC spectrum of Tsg101 UEV protein is a classical HSQC spectrum of a well-structured protein with dispersed peaks from ~6.5 ppm to ~11 ppm in the proton dimension. All the protein residues, but Asn45 are assigned in the HSQC. Also, the NH$_2$ resonances from side chains of all Trp and Gln are assigned while the 6 out of 7 NH$_2$ resonances from side chains of Asn are assigned (the side chain of Asn54 is not detected).

In addition, all 13 Prolines of Tsg101 UEV protein are assigned as shown in ^{15}N region 132-143 ppm of the 2D carbon detection ^{15}N, ^{13}C NCO spectrum in Figure 113.

Figure 113. 2D carbon detection ^{15}N, ^{13}C NCO spectrum of ^{15}N, ^{13}C Tsg101 UEV protein. All 13 Prolines residues are assigned shown in the zoomed box.

Based on all the recorded spectra, 131 out of 132 $^{1}H^{N}$ resonances (99%), 138 out of 145 $^{1}H\alpha$ resonances (95%), 145 out of 145 ^{15}N resonances (100%), 145 out of 145 $^{13}C\alpha$ resonances (100%), 133 out of 138 $^{13}C\beta$ resonances (96%), 144 out of 145 ^{13}CO resonances (99%) and 5 out of 7 $^{1}H\alpha_2$ resonances (71%) for Gly residues are assigned in total. Regarding to the side chain assignment of Trp, Asn and Gln residues, the $^{15}N\epsilon$ and $^{1}H\epsilon$ resonances (100%) of all Trp residues, the $^{13}C\gamma$, $^{13}C\delta$, $^{15}N\epsilon$ and $^{1}H\epsilon$ resonances (100%) of all Gln residues and 6 out of 7 $^{13}C\gamma$,$^{15}N\delta$ and $^{1}H\delta$ resonances (86%) of Asn residues are achieved. In addition, concerning the Prolines residues and their side chain assignments, 8 out of 13 $^{13}C\delta$ and $^{1}H\delta$ resonances (62%) are achieved using the 3D HACAN spectrum. Looking closer to the $^{13}C\beta$ resonance of Prolines, the Proline residues Pro81 and Pro120 are in *cis* conformation with $^{13}C\beta$ to ~34 ppm. These two Prolines are also referred as *cis* peptides in the crystal structure of the Tsg101 UEV domain under PDB ID: 2f0r[144].

The backbone and proline assignments of free Tsg101 UEV protein are obtained and deposited in the Biological Magnetic Resonance Data Bank (BMRB) under accession code 50765.

Using the experimental NMR chemical shifts, the secondary structure of the protein is analyzed using online available servers, the CSI 3.0 server[202] and the TALOS+ server[203] and the Secondary Structure Propensities (SSP) program[195]. Theses analyses are compared with the secondary structure analysis of the crystal structure of the Tsg101 UEV domain (Figure 114 – top) (PDB ID: 2f0r)[144]. The CSI 3.0 web server[202] uses the backbone chemical shifts in order to predict the secondary elements of the protein. In Figure 114 (upper), it is shown that the free Tsg101 UEV domain composed by three α-helices, five edge β-strands and three more extended interior β-strands. The flexibility and the rigidity of the protein regions are predicted using TALOS+ server[203] and the values of the Random Coil Index S^2 (RCI-S^2) for all residues are shown with blue dots in Figure 114 (middle – dots). Finally, the Secondary Structure Propensities (SSP) analysis[195] was used to estimate the fraction of α-helices and extended region along the protein sequence. The grey bars in Figure 114 (lower – bars) show the SSP score values for the assigned residues through the sequence calculated based on the experimental ^{15}N, ^{1}H^N, ^{13}Cα, ^{13}Cβ, ^{13}CO and ^{1}Hα chemical shifts of the protein. Finally, all three secondary structure analysis tools are in agreement and well correlated with the crystal structure of Tsg101 UEV domain (PDB ID: 2f0r) as shown in Figure 114.
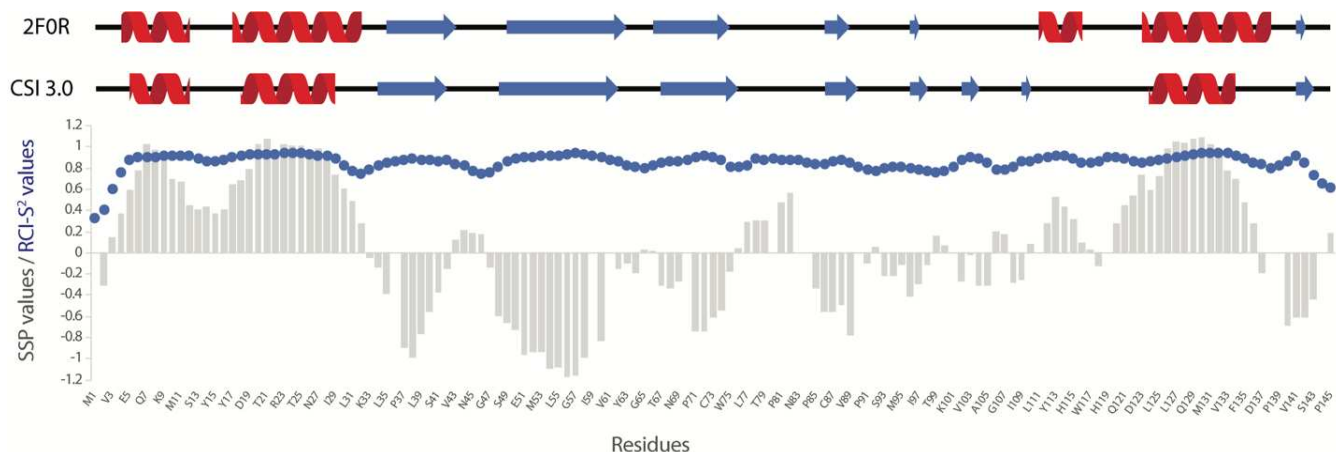
Figure 114. Secondary structure analysis of Tsg101 UEV domain. (top) Secondary structure in crystal structure of Tsg101 UEV domain (PDB ID: 2f0r)[144]. (upper) Prediction of secondary structure based on the backbone chemical shifts (^{1}H^N, ^{15}N, ^{13}Cα, ^{1}Hα, ^{13}Cβ and ^{13}CO) of the protein using the CSI 3.0 web server[202]. Red cartoon represents the α-helix and blue arrow the β-strand. (middle – blue dots) Predicted order parameter (Random Coil Index S^2, RCI-S^2) for all the residues based on the backbone chemical shifts (^{1}H^N, ^{15}N, ^{13}Cα, ^{1}Hα, ^{13}Cβ and ^{13}CO) using the TALOS+ server[203]. (lower – bars) Secondary Structure Propensities (SSP) score values for the assigned residues based on ^{1}H^N, ^{15}N, ^{13}Cα, ^{1}Hα, ^{13}Cβ and ^{13}CO chemical shifts of human Tsg101 UEV domain[195].

3.3 NMR study of the interaction between ORF3 protein and Tsg101 UEV domain

The backbone and proline assignments of ORF3 and Tsg101 UEV proteins are further used to monitor their interaction using solution-state NMR Spectroscopy. This technique requires the labeling of one protein used at a constant concentration and the addition of the second unlabeled one at increasing concentration while NMR experiments are recorded. In this study, the analysis of the interaction was performed for both proteins.

3.3.1 NMR titration of ^{15}N, ^{13}C ORF3 protein with unlabeled Tsg101 UEV domain

Due to the fact that ORF3 protein contains the PSAP motif in the C-terminal region, the first interaction experiment is conducted using ORF3 Cter construct and Tsg101 UEV protein. Exceptionally and using a specific NMR ^{1}H, ^{15}N IDIS HSQC experiment, 123 μM ^{15}N, ^{13}C ORF3 Cter double-labeled protein is mixed with 135 μM ^{15}N Tsg101 UEV labeled protein in NMR buffer (50 mM Sodium Phosphate pH 6.1, 50 mM NaCl, 0.5 mM EDTA) placed in a Shigemi tube and NMR data are recorded at 298K on 600 MHz Spectrometer. Using this pulse sequence, two ^{1}H, ^{15}N HSQC spectra are recorded at the same time and the separation of the spectra is based on the difference correlation between the ^{1}H-^{15}N(^{12}CO) and ^{1}H-^{15}N(^{13}CO) of the labeled and double-labeled proteins, respectively[204]. These ^{1}H, ^{15}N HSQC spectra are then compared to the corresponding ones of the protein in its free state and used to monitor the spectral perturbations (Chemical Shift Perturbations (CSP) and/or line broadening) induced upon binding for both proteins.

In addition, ^{15}N, ^{13}C ORF3 Cter protein is used to record the 2D carbon detection ^{15}N, ^{13}C NCO spectra before and after the addition of Tsg101 UEV domain and thus also collect the information on affected Proline residues. Figure 115 depicts the overlay of (a) the ^{1}H, ^{15}N HSQC spectra and (b) the ^{15}N, ^{13}C NCO spectra of ^{15}N, ^{13}C ORF3 Cter protein in the presence and absence of Tsg101 UEV protein.
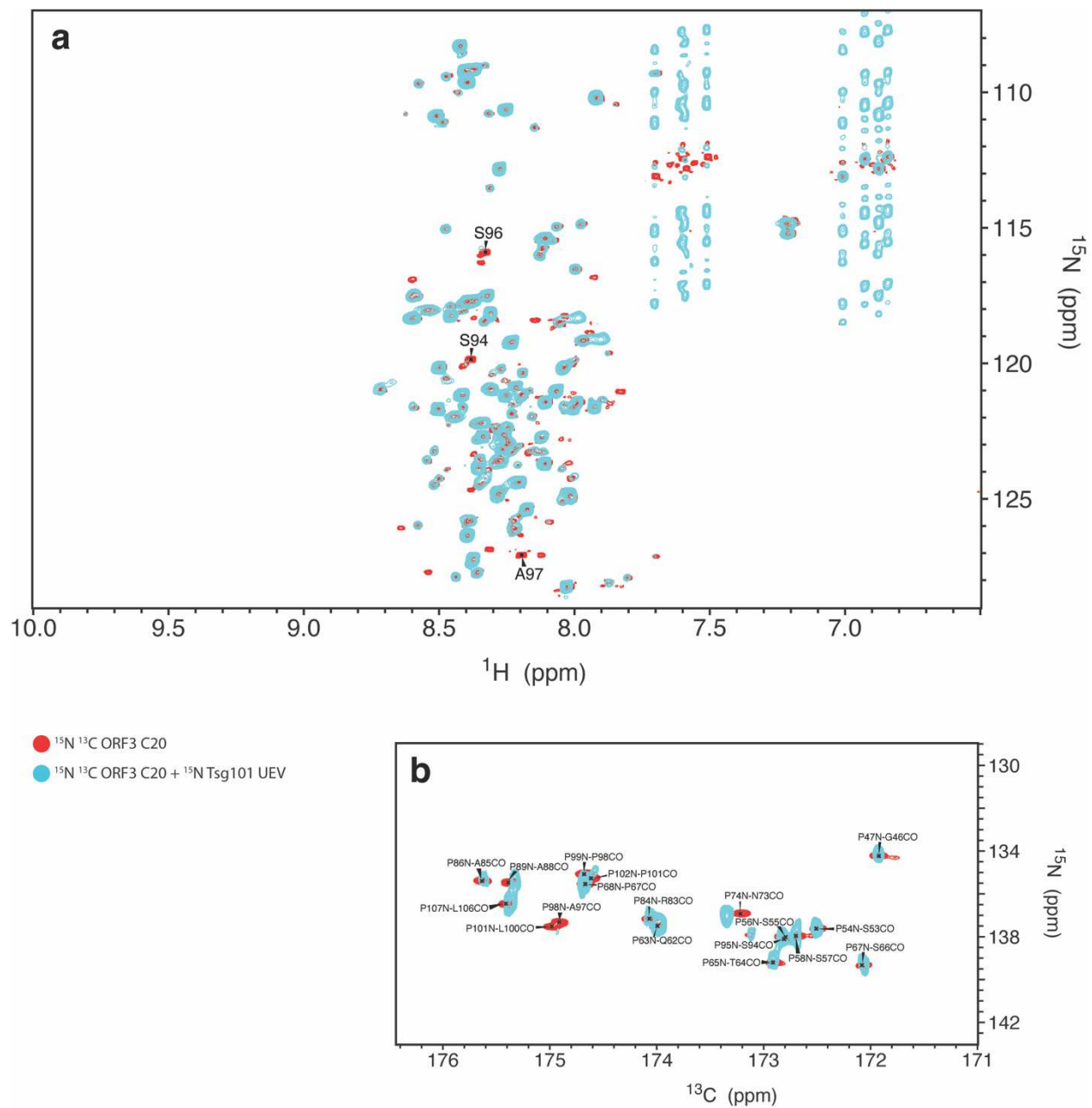
Figure 115. Overlay of (a) ^1H, ^{15}N HSQC spectra and (b) ^{15}N, ^{13}C NCO spectra of ^{15}N, ^{13}C ORF3 Cter protein in the presence (in cyan) and the absence (in red) of Tsg101 UEV domain.

The analysis of the CSP values for all assigned residues of ORF3 Cter protein induced by Tsg101 UEV domain is shown in Figure 116.
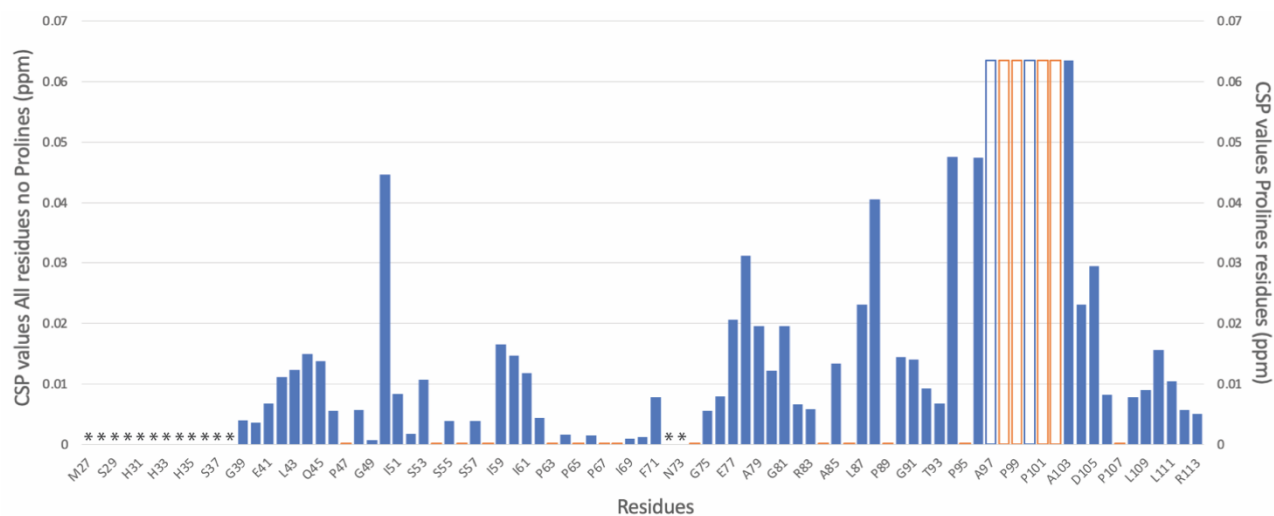
Figure 116. Chemical Shift Perturbation (CSP) values of all assigned residues of ORF3 Cter protein induced by Tsg101 UEV domain. The blue bars represent the CSP values of all residues except of Prolines while the orange bars represent the CSP values of Proline residues. The "empty" bars correspond to the residues that are disappeared with the addition of Tsg101 UEV protein. The unassigned residues of ORF3 Cter protein are marked with a star sign.

Based on the NMR data and the CSP analysis, the first conclusion drawn is that ORF3 Cter protein directly interacts with Tsg101 UEV domain. The binding site on ORF3 Cter side can be determined, it corresponds to residues Leu87-Asp105. This binding site encompasses the [95]PSAP[98] motif.

In order to further characterize the interaction with the full-length ORF3 C20 protein, a five-point NMR titration experiment of ^{15}N, ^{13}C ORF3 C20 double-labeled sample at 104 µM with addition of 20, 40, 100, 170 and 320 µM of unlabeled Tsg101 UEV domain in NMR buffer (50 mM Sodium Phosphate pH 6.1, 50 mM NaCl, 0.1 mM EDTA) is performed. For each point, a sample with the appropriate amount of each protein is placed in a 5 mm tube and both 2D ^{1}H, ^{15}N HSQC and 2D ^{15}N, ^{13}C NCO spectra are recorded at 293K on 900 MHz Spectrometer in order to detect all the affected residues including the affected proline peaks.

Figure 117 illustrates the overlay of the 2D ^{1}H, ^{15}N HSQC spectra for all titration points with the control spectrum (free 104 µM ^{15}N, ^{13}C ORF3 C20 protein) in red and the last titration point (104 µM ^{15}N, ^{13}C ORF3 C20 protein with 320 µM unlabeled Tsg101 UEV domain) in blue. In the zoom panels (magenta cycles), the affected peaks Ala88, Val 92, Ala97 and Val104 are shown.
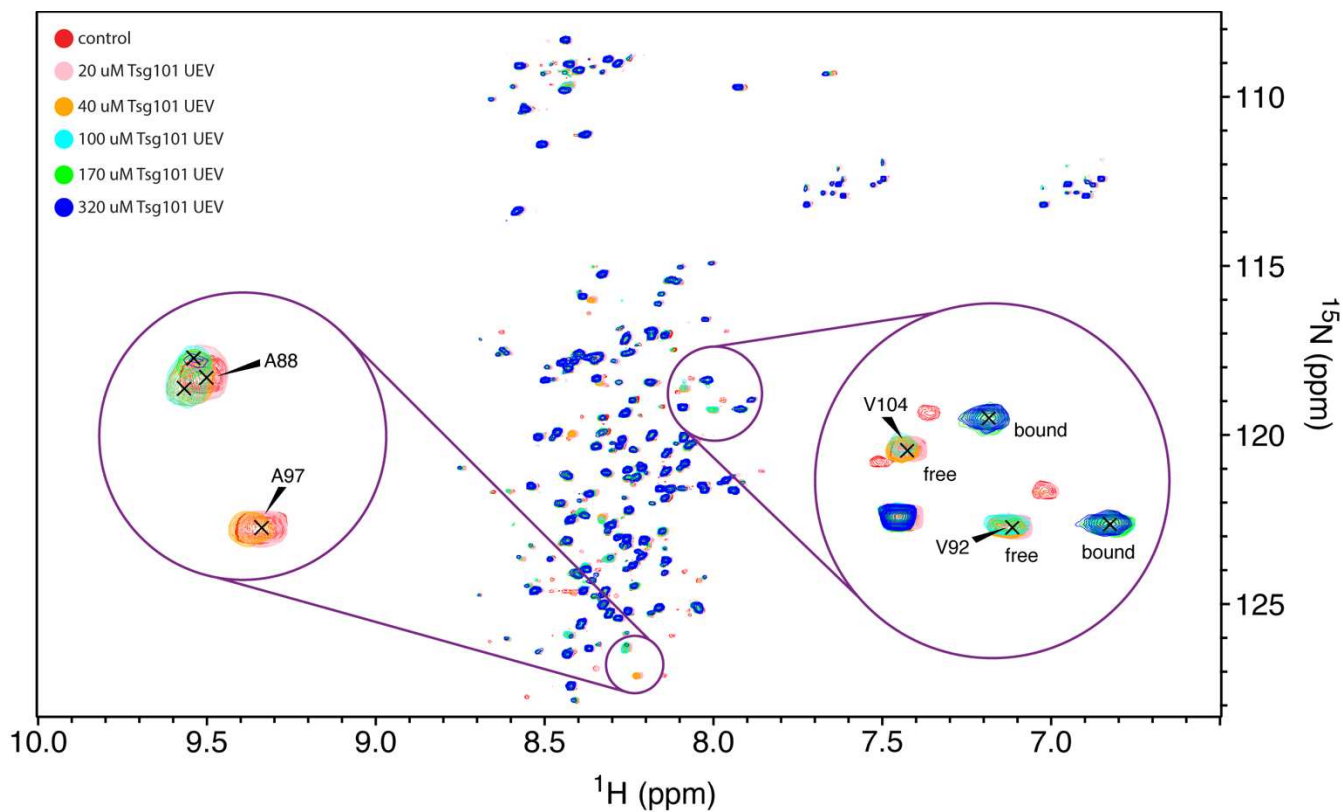
Figure 117. Overlay of 2D ^1H, ^{15}N HSQC spectra of ^{15}N, ^{13}C ORF3 C20 protein with addition of unlabeled Tsg101 UEV protein. Red: control spectrum without addition of Tsg101 UEV domain, pink: 20 uM Tsg101 UEV, orange: 40 uM Tsg101 UEV, cyan: 100 uM Tsg101 UEV, green: 170 uM Tsg101 UEV and blue: 320 uM Tsg101 UEV.

The overlay of the 2D HSQC spectra of all titration points depicts some peaks that broaden and then disappear, such as Ala97, few peaks which split, such as Ala88, few peaks that have very low intensity in the final titration point, such as Arg83 and Ala85, and peaks with slow exchange between free and bound states, such as Val92 and Val104.

In Figure 118, the overlay of the 2D ^{15}N, ^{13}C NCO spectra for all titration points with the control spectrum in red and the last titration point in blue depicts the affected Prolines peaks in the zoomed panel.
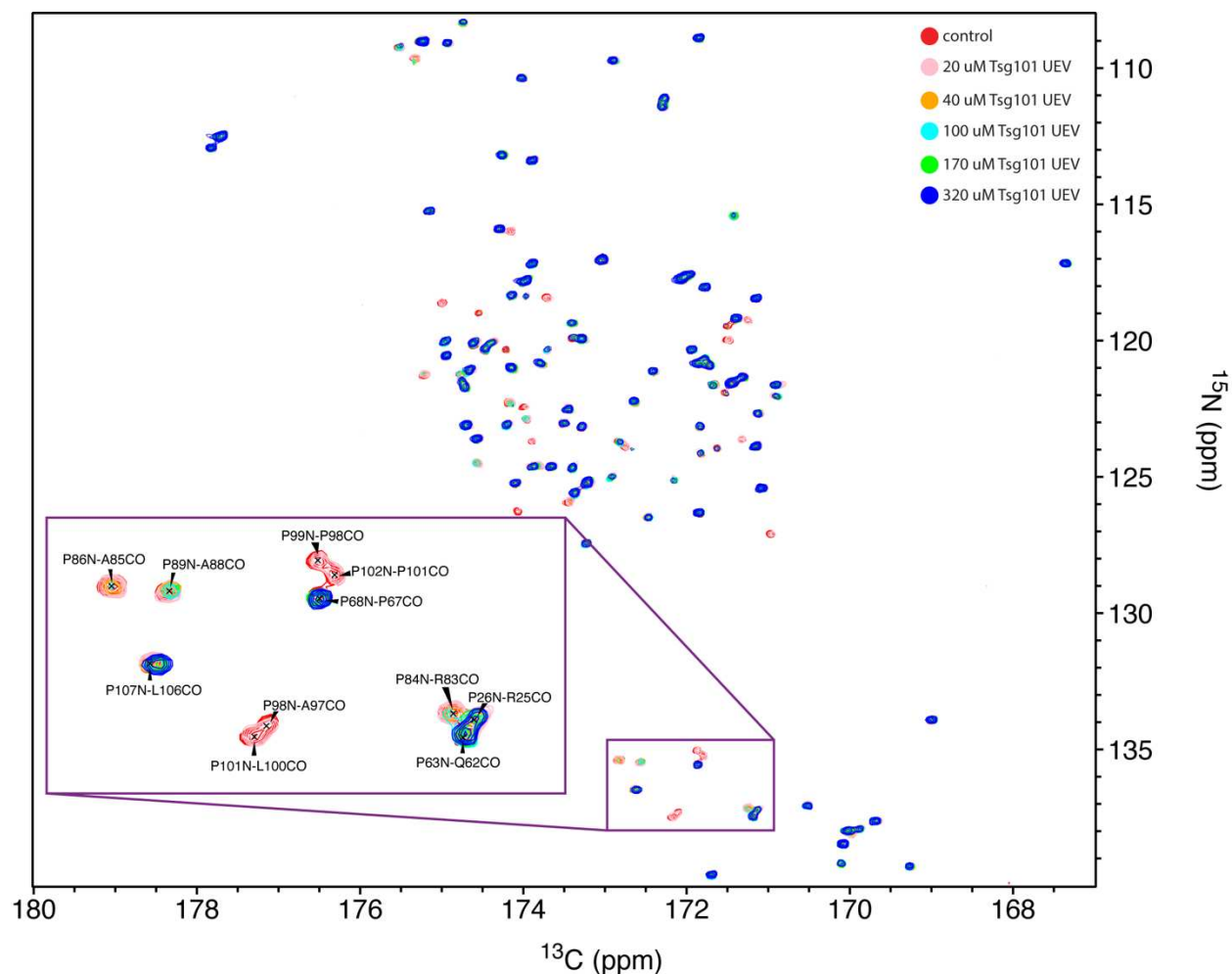
Figure 118. Overlay of 2D ^{15}N, ^{13}C NCO spectra of ^{15}N, ^{13}C ORF3 C20 protein with addition of unlabeled Tsg101 UEV protein. Red: control spectrum without addition of Tsg101 UEV domain, pink: 20 uM Tsg101 UEV, orange: 40 uM Tsg101 UEV, cyan: 100 uM Tsg101 UEV, green: 170 uM Tsg101 UEV and blue: 320 uM Tsg101 UEV.

Apart from the affected residues in HSQC spectra, the affected Proline residues which are also located in the C-terminal region of ORF3 C20 protein, can be easily detected in the overlay of the 2D NCO spectra.

The Chemical Shift Perturbation (CSP) values for all assigned residues of ORF3 C20 protein induced by Tsg101 UEV domain from both 2D HSQC and NCO spectra are analyzed and shown in Figure 119.

MGSPSALGLFSSSSSSFSLCSPRHRPASRLAVVVGGAAAVPAVVSGVTGLILSPSPSPIFIQPTPSPPISFHNPGLELALGSRPAPLAPLGVTSPSAPPLPPAVDLPQLGLRR
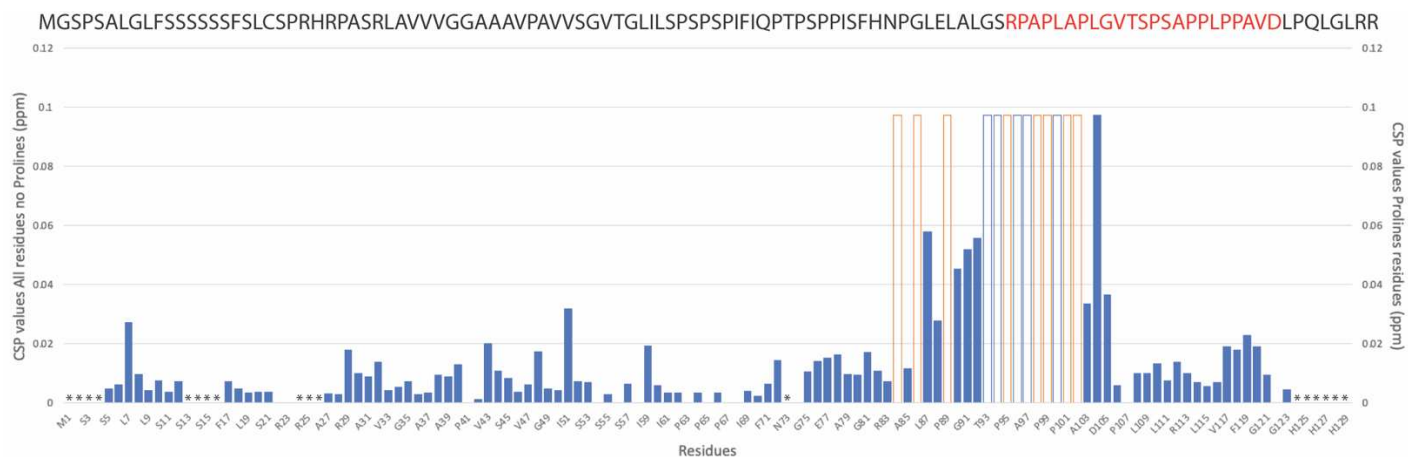
Figure 119. Chemical Shift Perturbation (CSP) values of all assigned residues of ORF3 C20 protein induced by Tsg101 UEV domain. The blue bars represent the CSP values of all residues except of Prolines while the orange bars represent the CSP values of Proline residues. The "empty" bars correspond to the residues that are disappeared with the addition of Tsg101 UEV protein. The unassigned residues of ORF3 C20 protein are marked with a star sign. On the top, the ORF3 C20 protein sequence is shown with red the affected residues, Arg83-Asp105.

The analysis of all NMR data of all titration points demonstrates the direct interaction of ORF3 C20 protein with Tsg101 UEV domain with the affected peaks in the ORF3 C20 sequence being 23 out of 129 of the total sequence and located in the C-terminal region, particularly the residues Arg83 to Asp105. This binding region, which corresponds to the region for which the NMR resonances undergone the higher perturbations, includes the ^{95}PSAP^{98} motif of ORF3 C20 protein. Due to the quick broadening and disappearance of the affected peaks in the 2D spectra, a titration curve cannot be built and therefore, the affinity of the interaction cannot be calculated using these NMR data. For further characterization of the complex of ORF3 with Tsg101 UEV protein, pepORF3, a 10-residue ORF3-derived peptide with sequence ^{93}TSPSAPPLPP^{102}, was designed and then chemically synthesized by GeneCust.

3.3.2   NMR titration of ^{15}N Tsg101 UEV domain with unlabeled ORF3 protein

As mentioned above, the first interaction experiment is performed in a mixture of ^{15}N, ^{13}C ORF3 Cter double-labeled protein and ^{15}N Tsg101 UEV labeled protein in ratio 1:1 and a specific NMR ^{1}H, ^{15}N IDIS HSQC experiment, in which two HSQC spectra are acquired in a single experiment, is recorded[204].

The 2D ^{1}H, ^{15}N IDIS HSQC spectrum of Tsg101 UEV domain (in cyan) is compared to the HSQC without the ORF3 Cter protein (in red) as shown in Figure 120. In the HSQC spectrum of Tsg101 UEV domain in presence of ORF3 Cter protein, peaks from ORF3 Cter protein can be also observed which could be due to the improper filter efficiency.
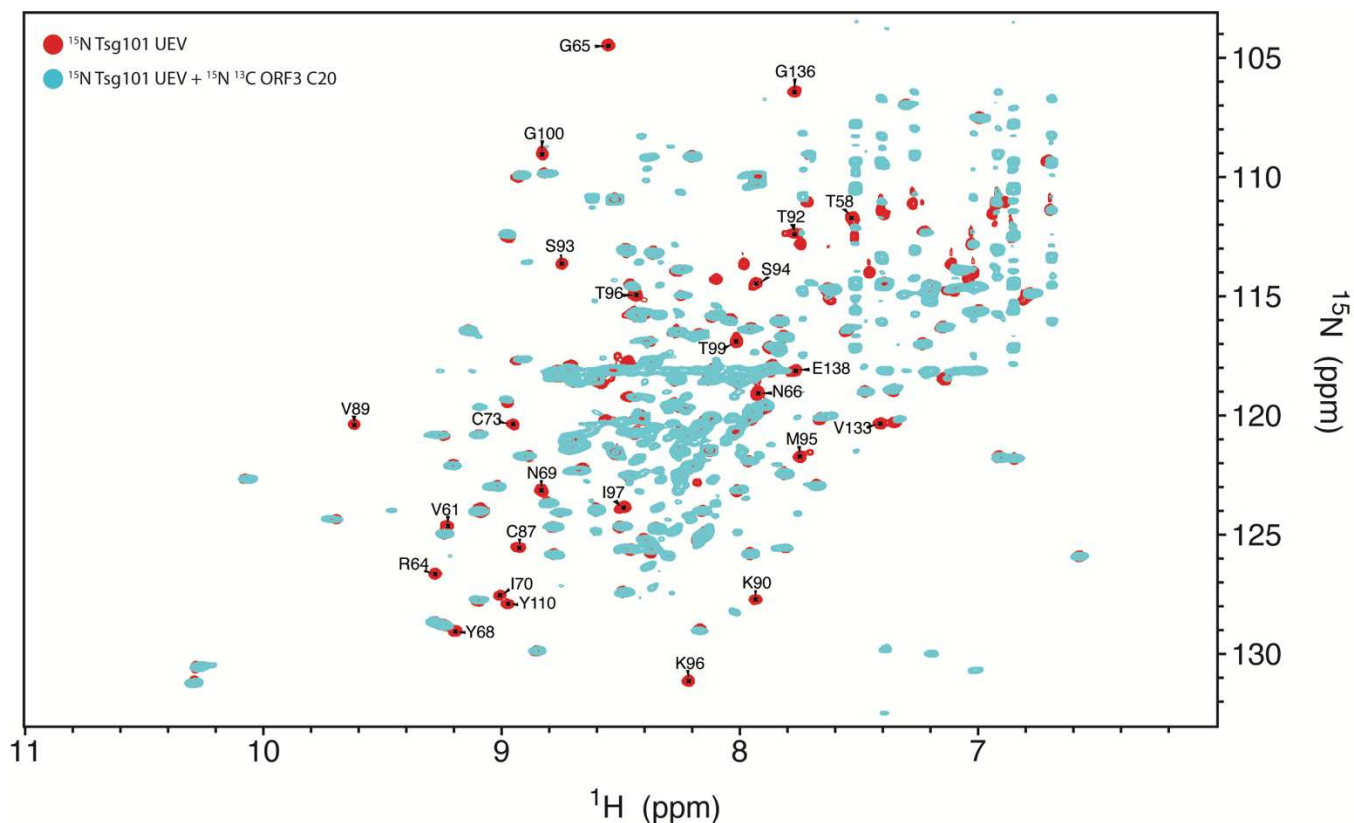
Figure 120. Overlay of ^{1}H, ^{15}N HSQC spectra of ^{15}N Tsg101 UEV domain in the presence (in cyan) and the absence (in red) of ORF3 Cter protein.

The analysis of the Chemical Shift Perturbations (CSP) values for all assigned non-Proline residues of Tsg101 UEV domain induced by ORF3 Cter protein as well the affected residues colored in the crystal structure of Tsg101 UEV domain (PDB ID: 2f0r)[144] are shown in Figure 121.
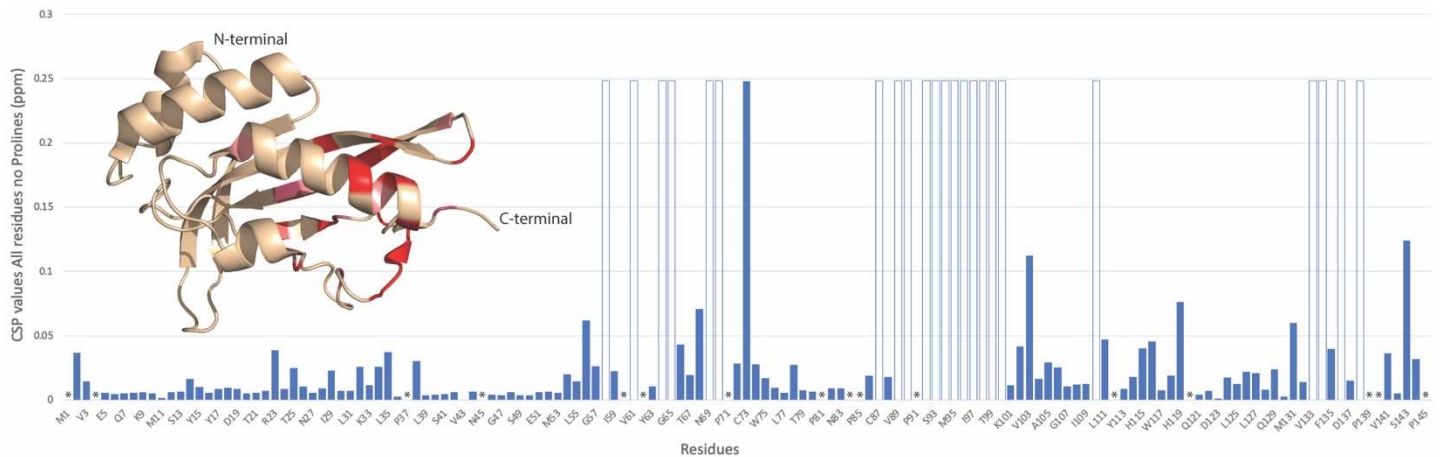
Figure 121. Chemical Shift Perturbation (CSP) values of all residues except of Prolines of Tsg101 UEV domain induced by ORF3 Cter protein. The "empty" bars correspond to the residues that are disappeared with the addition of ORF3 Cter protein. The Proline, the unassigned Asn45 and the ambiguous residues of Tsg101 UEV protein are marked with a star sign. The affected residues are colored on the Tsg101 UEV domain (PDB ID: 2f0r) with red for the residues are disappeared, raspberry color the ones with high CSP value and with pink color the ones that have very low intensity.

This first interaction experiment depicts the direct interaction of ORF3 Cter protein with Tsg101 UEV domain and the binding side of the latter protein can be also determined.

In order to study the interaction in the Tsg101 UEV side with the full-length ORF3 C20 protein, a five-point NMR titration experiment of ^{15}N Tsg101 UEV sample at 70 μM with the addition of 20, 40, 70, 120 and 160 μM of unlabeled ORF3 C20 protein is performed. For each point, a sample with the appropriate amount of each protein is placed in a 5 mm tube and a 2D ^{1}H, ^{15}N HSQC spectrum is recorded at 293K using the SampleJet on 900 MHz Spectrometer overnight.

Analyzing the 2D ^{1}H, ^{15}N HSQC spectra of the titration points, except from the control experiment, in all the spectra, the ORF3 C20 protein is detectable. Figure 122 shows the overlay of the HSQC spectrum of free 70 μM ^{15}N Tsg101 UEV protein in red, the one recorded for 70 μM ^{15}N Tsg101 UEV domain with 40 μM ORF3 C20 protein sample in cyan and the HSQC of 100 μM ^{15}N ^{13}C ORF3 C20 protein in magenta.
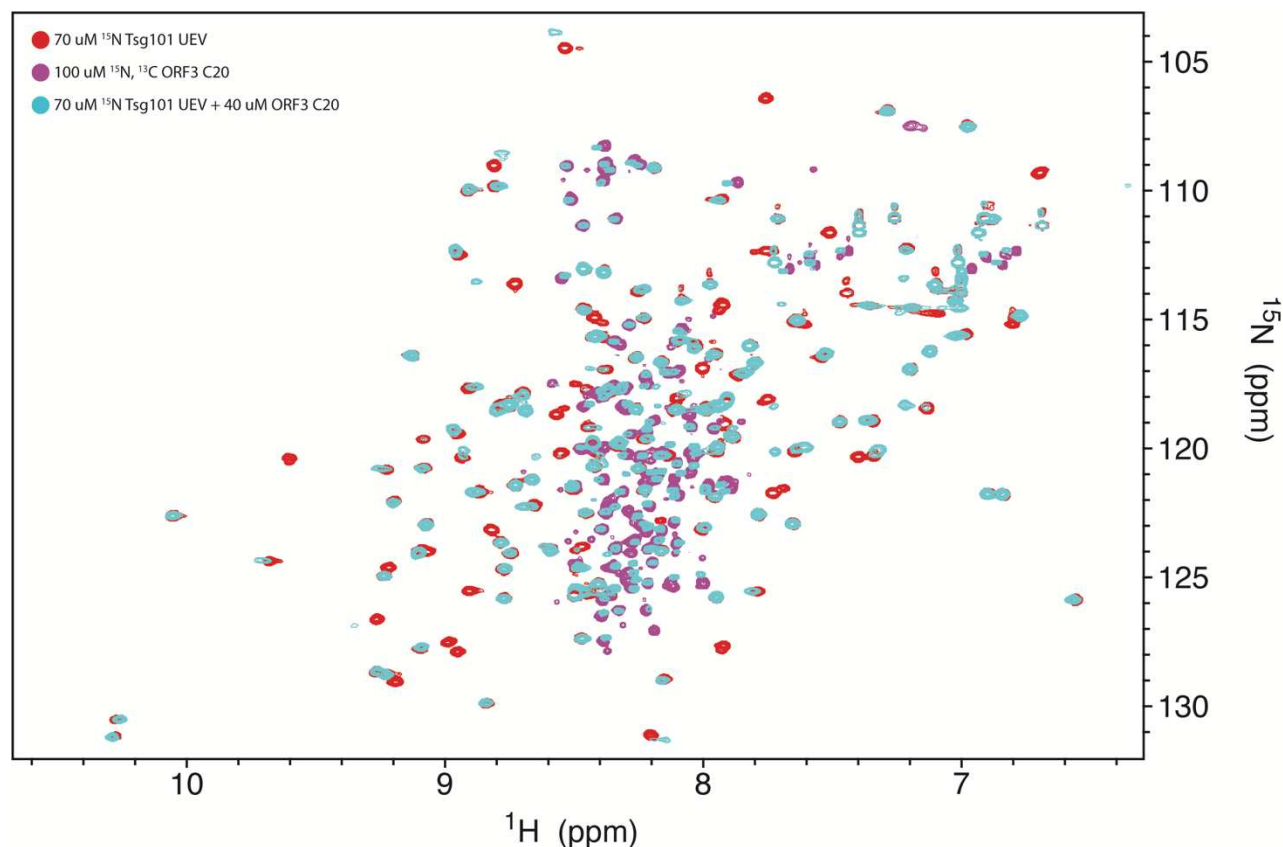
Figure 122. Overlay of ^{1}H, ^{15}N HSQC spectra of free 70 µM ^{15}N Tsg101 UEV domain in red, the 70 µM ^{15}N Tsg101 UEV domain with 40 µM ORF3 C20 protein titration point in cyan and 100 µM ^{15}N, ^{13}C ORF3 C20 recorded for backbone assignment procedure in magenta.

In the ^{1}H, ^{15}N HSQC spectrum of the titration point in cyan, both proteins can be observed with the ORF3 C20 corresponding peaks to be located in the middle of the spectrum due to its disordered nature. The reason of the presence of ORF3 C20 peaks in the HSQC was thought to be the erroneous use of ^{15}N labeled protein stock instead of the unlabeled one. Because of this unexpected result, the NMR titration experiment is repeated with freshly prepared protein samples and the same concentration points are used.

In the second NMR titration experiments, when the first points are recorded, the same results are observed, the presence of ORF3 C20 peaks in the ^{1}H, ^{15}N HSQC spectra (Figure 123).
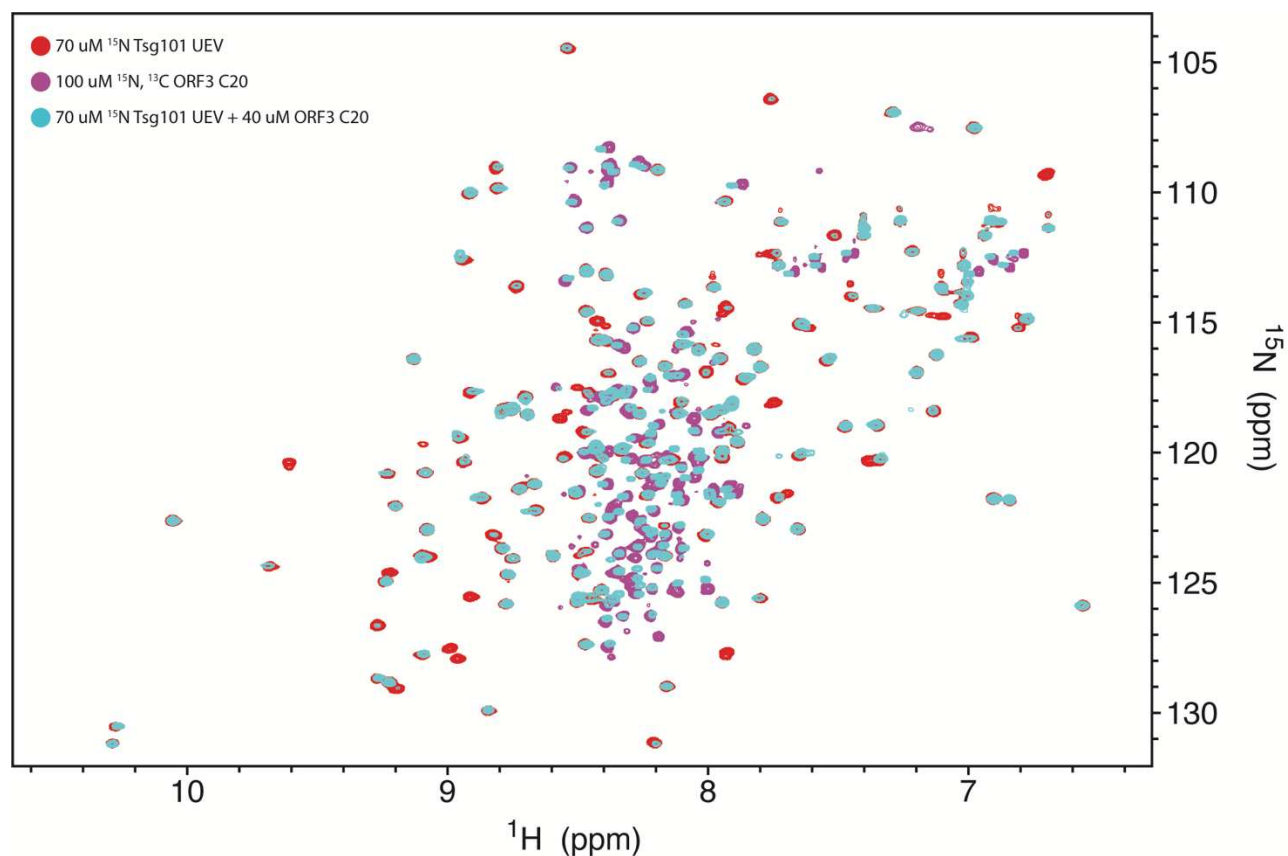
Figure 123. Overlay of ^{1}H, ^{15}N HSQC spectra of the second NMR titration experiment of free 70 μM ^{15}N Tsg101 UEV domain in red, the 70 μM ^{15}N Tsg101 UEV domain with 40 μM ORF3 C20 protein titration point in cyan and 100 μM ^{15}N, ^{13}C ORF3 C20 recorded for backbone assignment procedure in magenta.

Trying to understand the reason of the detection of ORF3 C20 protein, a 40 μM ORF3 C20 sample is placed in 5 mm tube using the same protein stock and a quick 2D ^{1}H, ^{15}N HSQC spectrum is acquired with 3072 and 64 complex points in the direct and indirect dimensions, respectively, and 4 scans with a total recording time of 5 min 9 sec. Figure 124 illustrates the ^{1}H, ^{15}N HSQC spectrum obtained for the 40 μM unlabeled ORF3 C20 protein sample.
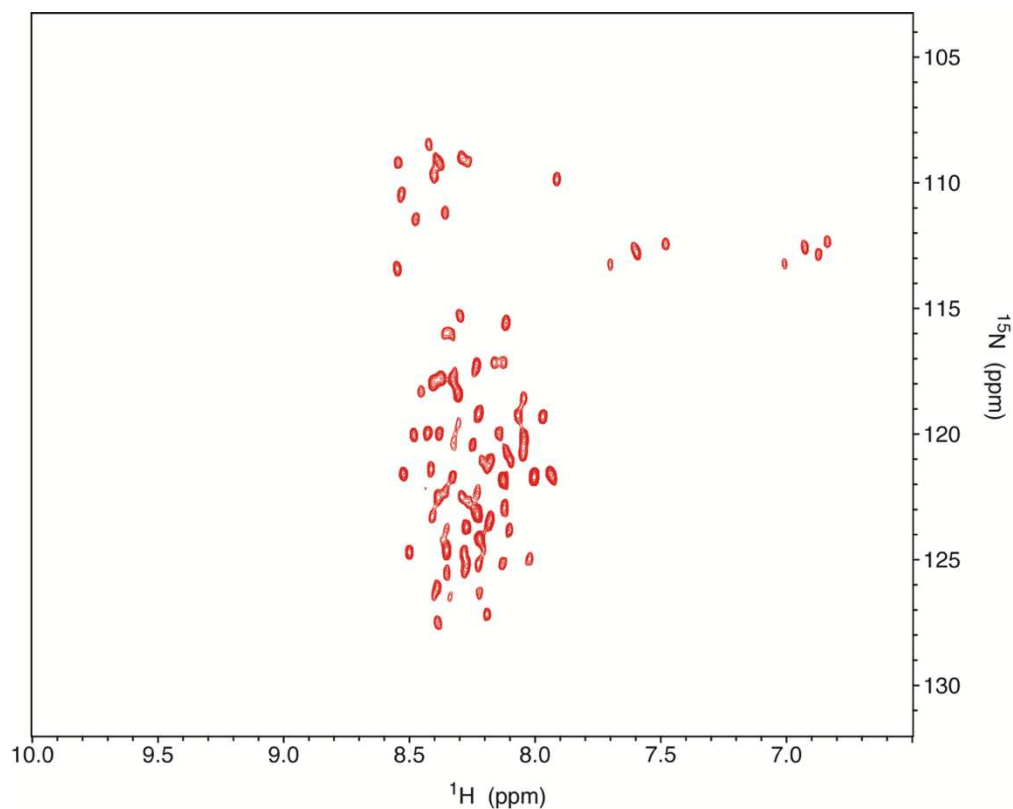
Figure 124. 2D ^1H, ^{15}N HSQC spectrum of 40 μM unlabeled ORF3 C20 protein recorded with 3072 and 64 complex points in the direct and indirect dimensions, respectively, and 4 scans with a total recording time of 5 min 9 sec.

The detection of the ORF3 C20 peaks in a 2D ^1H, ^{15}N HSQC spectrum was unexpected and unexplained because of the unlabeled origin of the sample. For this reason, 1D spectra with and without ^{15}N decoupling pulse in acquisition are recorded for the ORF3 C20 sample. The unlabeled samples have identical 1D spectra regardless the ^{15}N decoupling option as shown in Figure 125 for the 40 μM unlabeled ORF3 C20 protein sample.
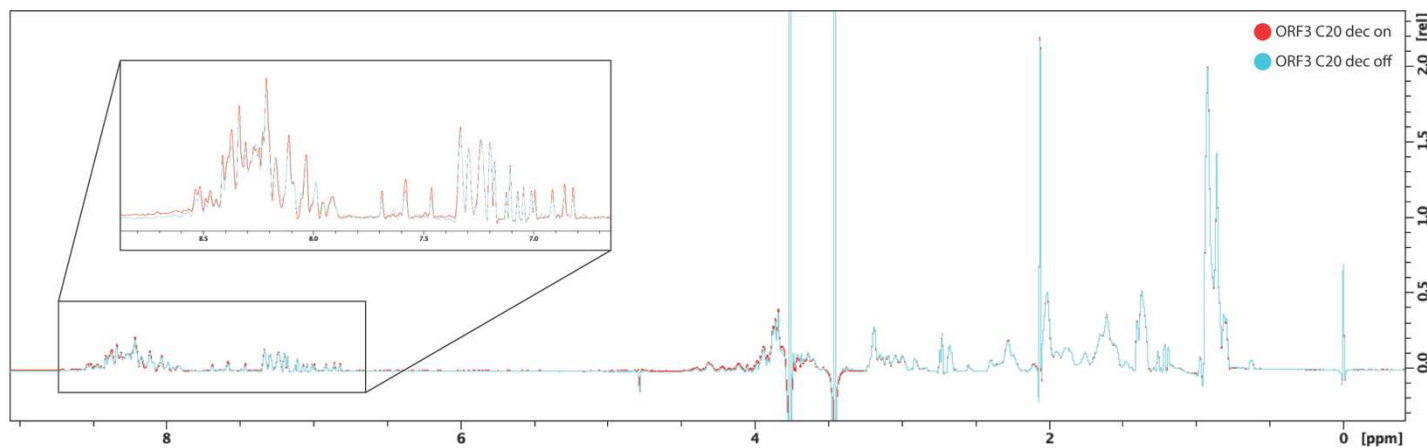
Figure 125. Overlay of 1D spectra of 40 μM ORF3 C20 protein with ^{15}N decoupling in red and without ^{15}N decoupling in cyan.

In the other hand, the 1D spectra of a ^{15}N labeled protein differ in the range 6 to 11 ppm where mainly the backbone proton amides (^{1}H^{N}) are observed. Turning off the ^{15}N decoupling has as a result the split of the peaks in this range because of the J-coupling effect[205]. Figure 126 shows the overlay of the 1D spectra for the 70uM ^{15}N Tsg101 UEV protein sample using ^{15}N decoupling in red and without in cyan.
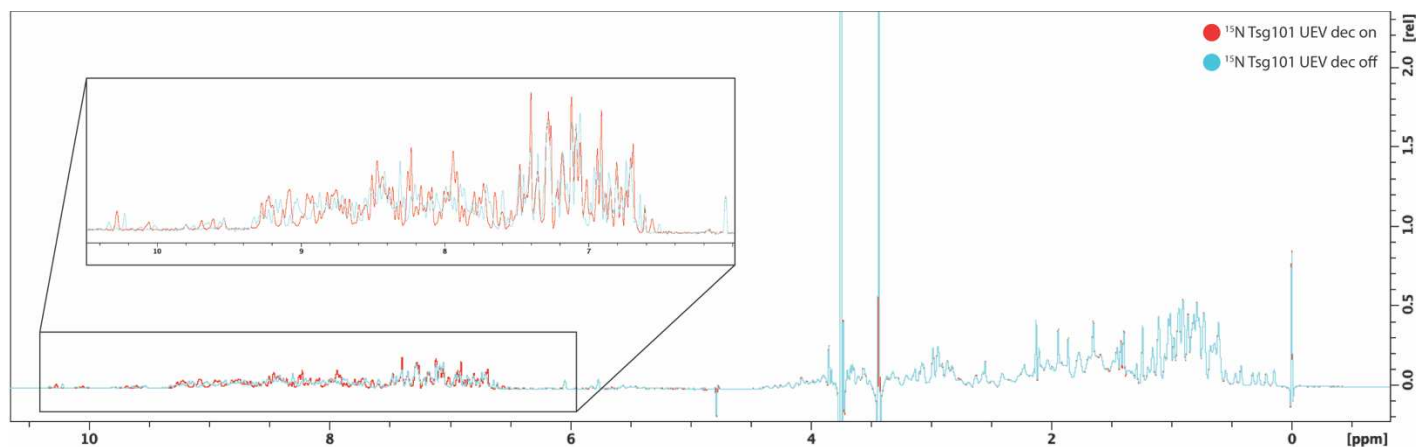
Figure 126. Overlay of 1D spectra of 70 µM ^{15}N Tsg101 UEV protein with ^{15}N decoupling in red and without ^{15}N decoupling in cyan.

To exclude the possibility of the incorrect concentration' estimation of the unlabeled ORF3 C20 protein stock, a 4-20% SDS-PAGE with increasing volume of ORF3 C20 protein sample used for recording the NMR spectra (40 µM unlabeled NMR sample) (left) and lysozyme (right) was done and the bands of the two proteins are compared to each other and with the ones of the molecular marker with fixed concentration Broad Range Protein Molecular Weight Markers, Promega) (Figure 127).
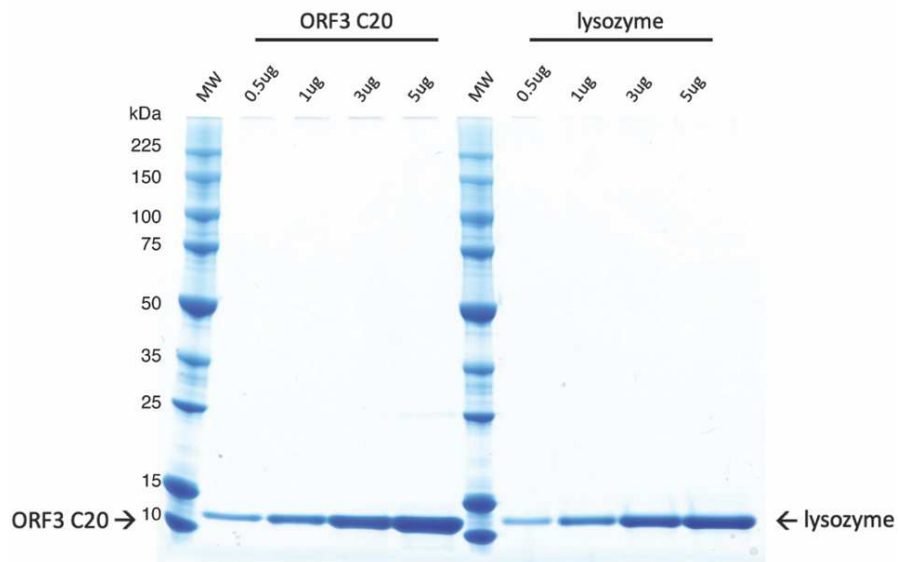
Figure 127. 4-20% SDS-PAGE with increasing volume of ORF3 C20 (left) and lysozyme (right) proteins for concentration' estimation with Coomassie blue staining. For the Molecular Weight marker (MW), the Broad Range Protein Molecular Weight Markers (Promega) is used.

Based on the SDS-PAGE, the concentration of ORF3 C20 sample was in the correct range. Therefore, the detection of ORF3 C20 protein for second time using different protein stocks is still inexplicable.

In order to successfully perform the NMR titration experiment, we decided to prepare a ^{15}N, ^{13}C Tsg101 UEV double-labeled sample and new unlabeled ORF3 C20 protein, record the specific NMR ^{1}H, ^{15}N IDIS HSQC experiment for each point in order to filter out the detection of the ORF3 C20 protein. Before setting-up the NMR IDIS HSQC experiment and in order to test if the ORF3 C20 protein is detectable again, we record a classical 2D ^{1}H, ^{15}N HSQC spectrum for the 70 µM ^{15}N, ^{13}C Tsg101 UEV with 70 µM unlabeled ORF3 C20 sample. Due to the fact that the latter protein is not detectable, for the NMR titration experiment the classical HSQC spectrum is obtained for all the titration points.

Figure 128 illustrates the overlay of the 2D ^{1}H, ^{15}N HSQC spectra for all titration points with the control spectrum (free 70 µM ^{15}N, ^{13}C Tsg101 UEV domain) in red and the last titration point (70 µM ^{15}N, ^{13}C Tsg101 UEV domain with 160 µM unlabeled ORF3 C20 protein) in blue. In the zoom panels (magenta cycles), some of the affected peaks Gly65 and Gly100 (left), Lys90, Lys98 and Lys118 (right) are shown.
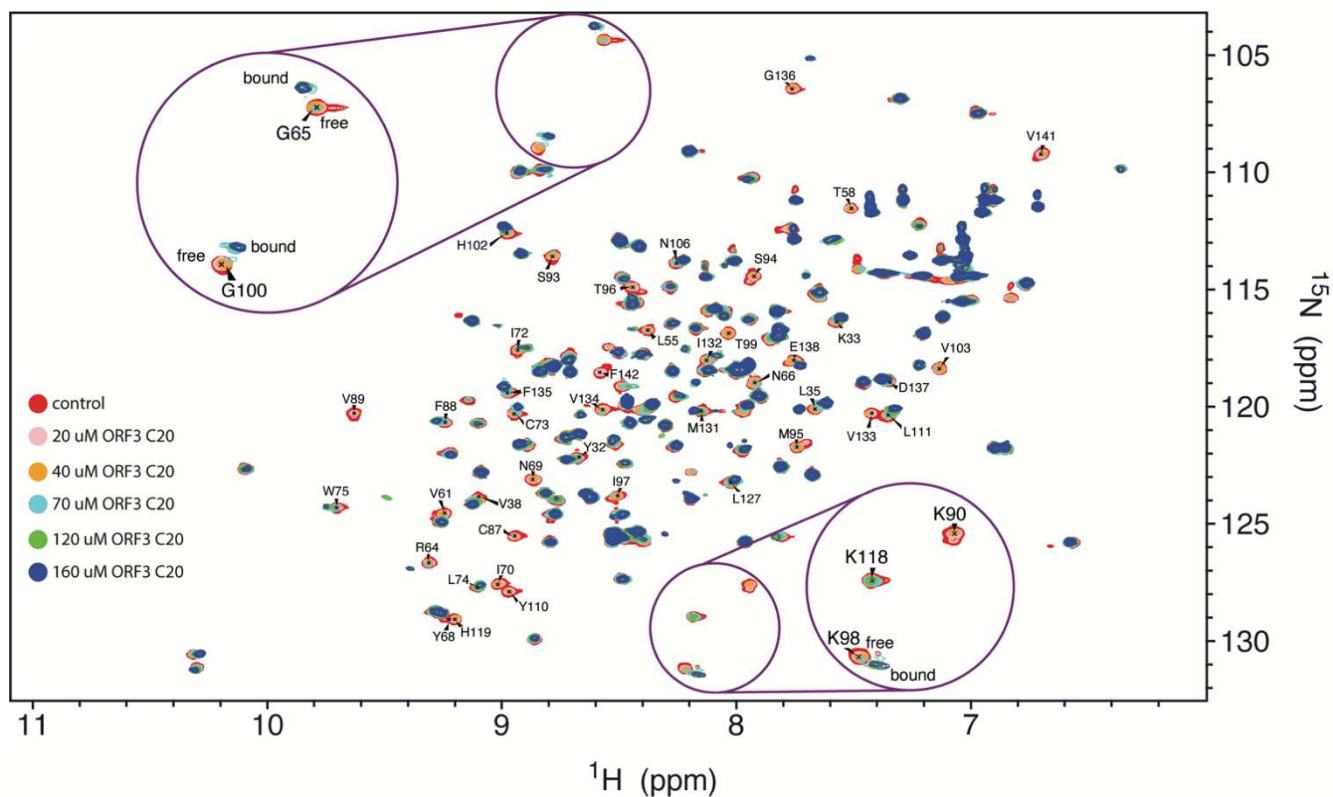
Figure 128. Overlay of 2D ^1H, ^{15}N HSQC spectra of ^{15}N, ^{13}C Tsg101 UEV domain with addition of unlabeled ORF3 C20 protein. Red: control spectrum without addition of ORF3 C20 protein, pink: 20 uM ORF3 C20, orange: 40 uM ORF3 C20, cyan: 70 uM ORF3 C20, green: 120 uM ORF3 C20 and blue: 160 uM ORF3 C20 protein.

The overlay of the 2D HSQC spectra of all titration points depicts some peaks that broaden and then disappear, such as Lys90 and Lys118, few shifted peaks, such as Phe88 and Leu111, few peaks that have very low intensity in the final titration point, such as Tyr113 and Leu114, and peaks with slow exchange between free and bound states, such as Gly65, G100 and Lys98.

The Chemical Shift Perturbation (CSP) values for all assigned non-Proline residues of Tsg101 UEV domain induced by ORF3 C20 protein based on the HSQC spectra are analyzed and shown in Figure 129 with affected residues are colored accordingly on the Tsg101 UEV domain (PDB ID: 7nlc).

201
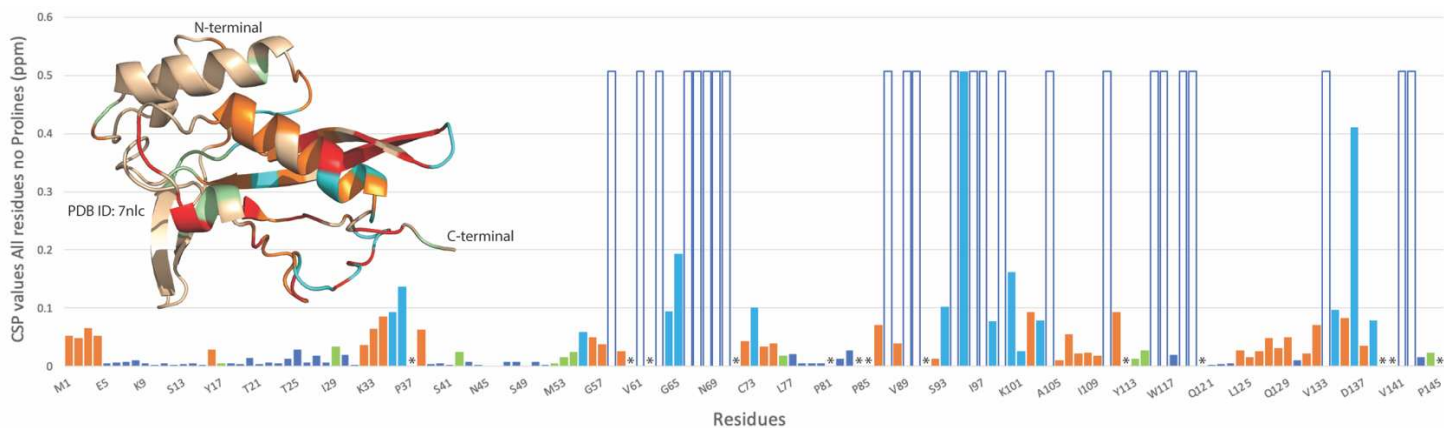
Figure 129. Chemical Shift Perturbation (CSP) values of all assigned non-Proline residues of Tsg101 UEV domain induced by ORF3 C20 protein. The "empty" bars correspond to the residues that are disappeared with the addition of ORF3 C20 protein, the orange bars to the shifted peaks, the cyan bars to the slow exchange peaks and the green ones to the peaks with low intensity in the last titration point. The Proline residues of Tsg101 UEV protein are marked with a star sign. The affected residues are colored on the Tsg101 UEV domain (PDB ID: 7nlc) with red for the disappeared residues, with orange the shifted ones, with cyan the slow exchange residues and with green color the ones that have low intensity.

The analysis of the NMR data of all titration points with full-length ORF3 C20 protein demonstrates the direct interaction and the binding site on the Tsg101 UEV domain is located in the C-terminal as shown in Figure 129 and the color-coded residues. Due to the quick broadening and disappearance of the affected peaks in the 2D spectra, a titration curve cannot be built and therefore, the affinity of the interaction cannot be calculated using these NMR data.

202

3.4 Isothermal Titration Calorimetry (ITC) experiments

Isothermal Titration Calorimetry (ITC) is a biophysical technique used for the determination of the thermodynamics parameters of an interaction process between two molecules. The titration of one molecule placed in the syringe to the other molecule placed in the sample cell results in the emission or absorption of heat and thus the raw heat pulses are translated to binding isotherm which provides useful information about the interaction as its affinity, stoichiometry and thermodynamics[206]. Because of the extreme sensitivity of the technique, the temperature during the measurement has to be stable and the samples have to be in exactly the same buffer. The schematic representation of an ITC experiment is shown in Figure 130.
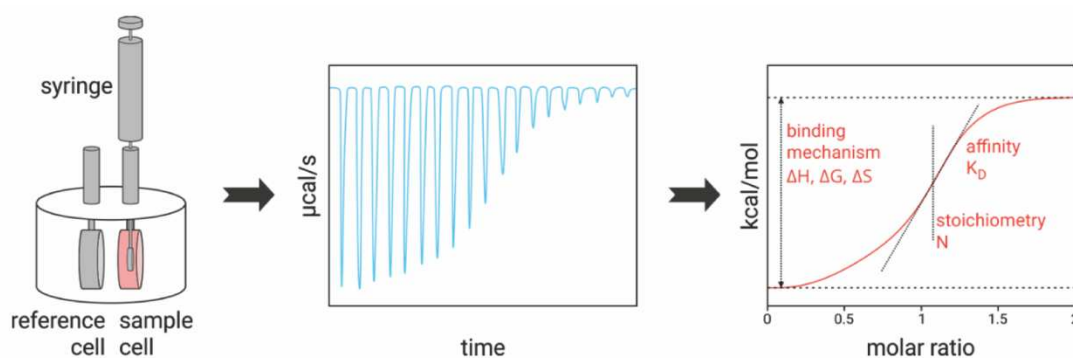
Figure 130. Schematic representation of an Isothermal Titration Calorimetry (ITC) experiment. (left) The isothermal titration calorimeter with the syringe (one molecule is placed either small molecule or a protein) and the sample cell (other molecule is placed that is protein) are shown, (middle) the raw titration thermogram, the heat per unit of time that is released after each injection, is translated to (right) the binding isotherm which provides the affinity, the stoichiometry and the thermodynamics of the interaction. (Figure derived from 2bind.com website[207]).

In this study, the NMR titration data could not be used for the determination of the affinity of the interaction between ORF3 and Tsg101 UEV due to the quick broadening of the affected peaks. Therefore, two ITC experiments are performed for further characterization of this interaction. Before the ITC experiment, both proteins are overnight dialyzed at 4°C using Spectra-Por® Float-A-Lyzer® G2 MWCO 3.5-5 kDa (1 mL) device in the same buffer containing 50 mM Sodium Phosphate pH 6.1, 50 mM NaCl. In the first experiment, the Tsg101 UEV domain is placed into the syringe at 1.2 mM final concentration and is titrated into the full-length ORF3 protein placed in the sample cell at 100 µM concentration (Figure 131a). In the second experiment, the pepORF3 peptide, designed on the basis of the NMR data, is resuspended with dialysis buffer to a final

concentration of 1.5 mM, placed in the syringe and then titrated into the sample cell that contains Tsg101 UEV protein at 100 µM concentration (Figure 131b).
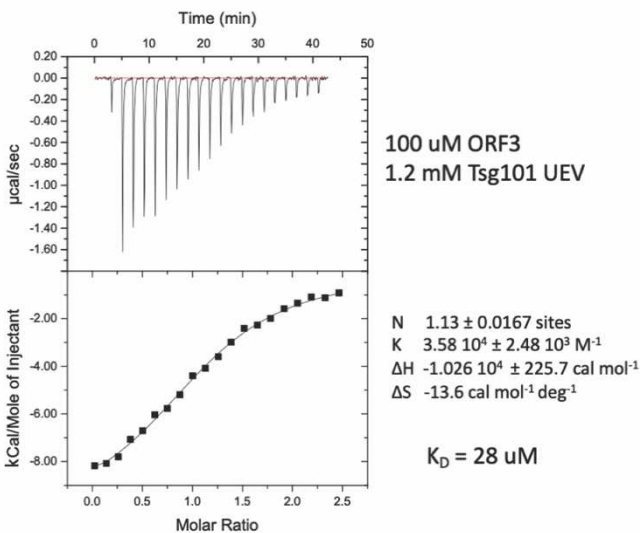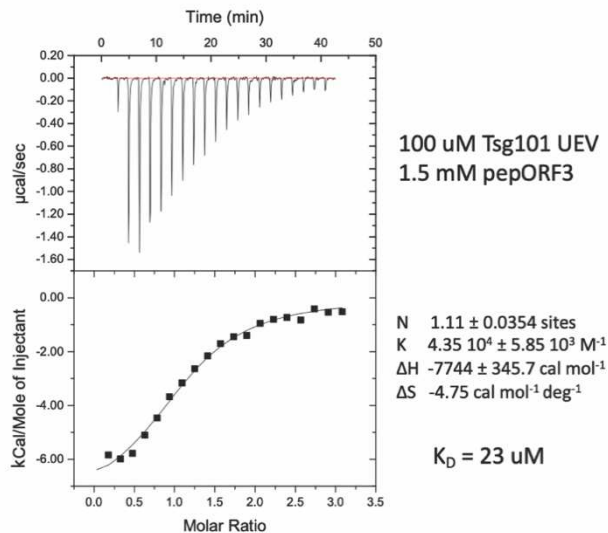
Figure 131. Isothermal Titration Calorimetry (ITC) experiments for characterization of the human Tsg101 UEV domain and ORF3 protein interaction. (a) Titration of Tsg101 UEV domain (syringe) to the full-length ORF3 protein (sample cell) with calculated K_D value of 28 µM and a stoichiometry of 1. (b) Titration of pepORF3 peptide (syringe) to the Tsg101 UEV domain (sample cell) with calculated K_D value of 23 µM and a stoichiometry of 1.

The K_D value is calculated to be 28 µM and 23 µM for the first and the second experiment, respectively. For both ITC experiments, the stoichiometry is equal to 1 meaning that one molecule of ORF3 protein interacts with one molecule of Tsg101 UEV domain. The thermodynamic ITC data of both experiments reveal hydrogen bonding with unfavorable hydrophobic interaction and conformational changes as indicated by the negative (favorable) binding enthalpy (ΔH) and positive (unfavorable) entropy factor (-TΔS). In Figure 132, a graphic representation of thermodynamic ITC data in three possible binding experiments is shown with our experimental data represented by (A) binding experiment[208].
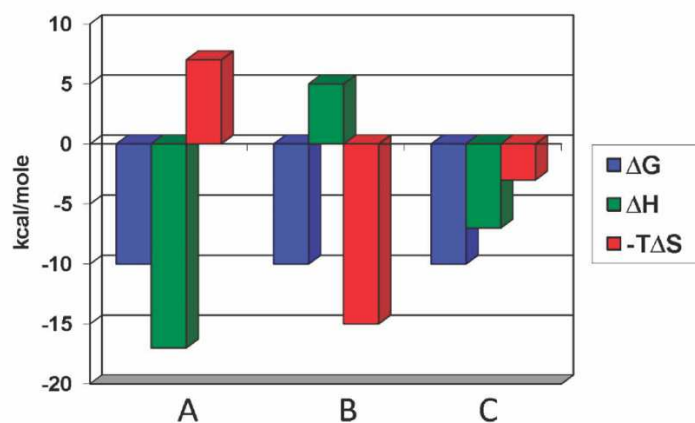
Figure 132. Graphic representation of thermodynamic ITC data, ΔG values are shown in blue bars, ΔH in green and -TΔS in red bars in (A) binding with favorable enthalpy, (B) binding with favorable entropy and (C) binding with favorable enthalpy and entropy. Favorable parameters are represented by negative values while unfavorable by positive values. From Frasca et al.[208], available under a Creative Commons Attribution License (CC BY 4.0).

The K_D values of both experiments are about the same, which is a clue that the 10-residue pepORF3 contains all the required elements to establish its proper interaction with Tsg101 UEV. Therefore, pepORF3 can be used for further study of the interaction.

3.5 Crystal structure of Tsg101 UEV domain with pepORF3 peptide

In order to get atomic details of the interaction between ORF3 and the Tsg101 UEV domain, the 10-residue ORF3-derived peptide, pepORF3 ([93]TSPSAPPLPP[102]), is used in co-crystallization experiments to obtain crystals of its complex with the human Tsg101 UEV domain.

The co-crystallization of the human Tsg101 UEV domain and the pepORF3 peptide is performed in two different set of concentrations, 20 mg/mL of Tsg101 UEV domain with 2 mM of pepORF3 and 10 mg/mL of Tsg101 UEV domain with 1 mM of pepORF3. Then, using the CyBio liquid handling system robot and a sitting-drop setting, five 96-conditions crystallization kits (JCSG+ Suite, Protein Complex Suite, pH Clear Suite, Cryos Suite and $AmSO_4$ Suite), which correspond to 480 different conditions in total, are screened at 21°C for both set of concentrations. After about 10 days, crystals in various conditions are detected, but they are not unique and look like interleaved crystals meaning the co-crystallization has to be further optimized. In the pH Clear Suite plate, in the condition of 2.4 M Ammonium sulfate pH 4 with 0.1 M Citric acid and in the drop of 10 mg/mL Tsg101 UEV with 1 mM pepORF3, big crystals are found. This crystallization condition was chosen to start the optimization using the hanging-drop setting. To optimize the crystallization condition, a screening of the Ammonium sulfate concentration from 2 to 3 M and pH 3.5, 4, 4.5 and 5 is performed. Unfortunately, no crystals were obtained after few days. However, as at the end of the process a cryoprotectant has to be included in the crystallization condition (to be able to freeze the crystal), a new optimization screen including a cryo-protectant reagent, 10 to 15% glycerol, is performed. After few days at 15°C, clear, unique, sharped, well-defined crystals in many wells for both set of concentrations of the complex were obtained (Figure 133). The range of the crystallization conditions in which the crystals were obtained is 2-3 M Ammonium sulfate pH 3.5, 0.1 M Citric acid with 10-15% glycerol. Using different size of loops and the optical microscope, several crystals from different drops of these crystallization conditions are selected and freeze with liquid Nitrogen for storage until the collection of the diffraction data. Several diffraction datasets with high resolution are collected using the Proxima 2a beamline of the SOLEIL synchrotron facility in Saint-Aubin, Paris.
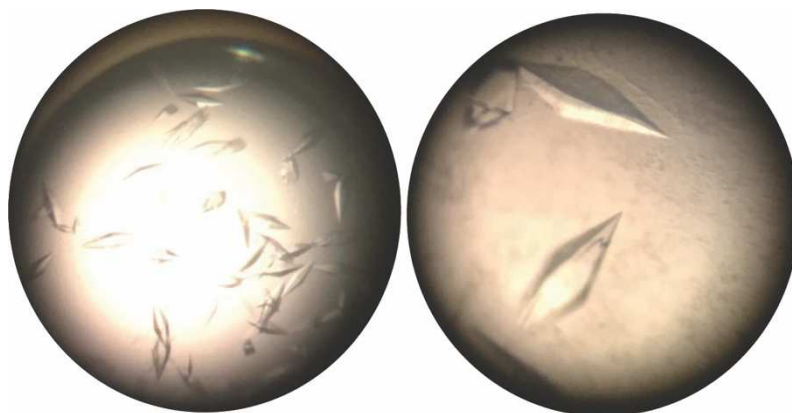
Figure 133. Crystals of human Tsg101 UEV domain grown in the presence of the pepORF3 peptide.

The next step is to solve the crystal structure and check if the pepORF3 peptide is present in the asymmetric unit of the crystal. Using the dataset with the highest resolution, corresponding to the crystal grew in 2 M Ammonium sulfate pH 3.5, 0.1M Citric acid with 13.34% glycerol and 10 mg/mL Tsg101 UEV with 1 mM pepORF3, the X-ray structure of the complex is solved using the molecular replacement method and PDB ID: 3obq as atomic model using the CCP4i2 interface[209] of the CCP4 program suite[210]. After several atomic building and refinements steps, using Coot and Refmac softwares, respectively, the final resolution of the structure is 1.4 Å. The solved structure indeed corresponds to the complex of human Tsg101 UEV protein with pepORF3 peptide (in red) as shown in Figure 134 and is deposited on Protein Data Bank (PDB) database under the accession code 7nlc (Table 5). In agreement with the ITC measurements, one molecule of Tsg101 UEV domain interacts with one molecule of pepORF3 peptide.
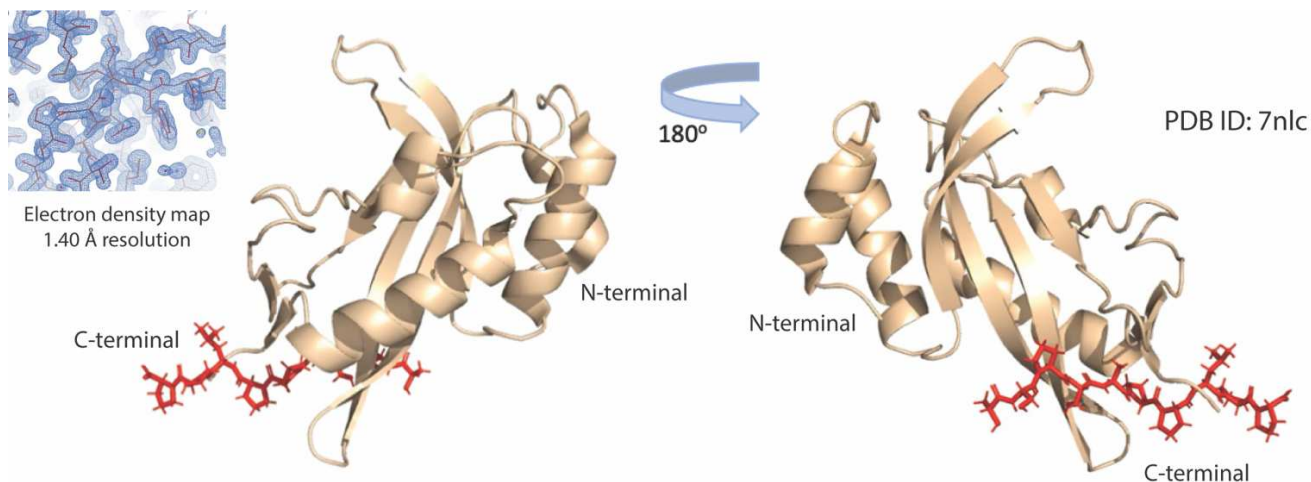
Figure 134. Crystal structure of human Tsg101 UEV domain in presence of HEV pepORF3 peptide in red (PDB ID: 7nlc).

Space group	P 41 21 2
Unit-cell parameters	
a (Å)	42.6
b (Å)	42.6
c (Å)	188.2
Resolution (Å)	47.04 - 1.40
Total observations	35531
Unique reflections	1777
% Data completeness (in resolution range)	99.9 (47.04 - 1.40)
I/σ(I)	1.34 (at 1.40 Å)
R (%)	22.9
R_{free} (%)	24.4
R.m.s.d deviations	
Bonds (Å)	0.0128
Angles (°)	1.791
F_o,F_c correlation	0.96
Total number of atoms	2905
Average B, all atoms (Å²)	27.0

Metric	Percentile Ranks	Value
Rfree		0.244
Clashscore		5
Ramachandran outliers		0
Sidechain outliers		0.7%
RSRZ outliers		10.4%

Worse — *Better*

■ Percentile relative to all X-ray structures
▯ Percentile relative to X-ray structures of similar resolution
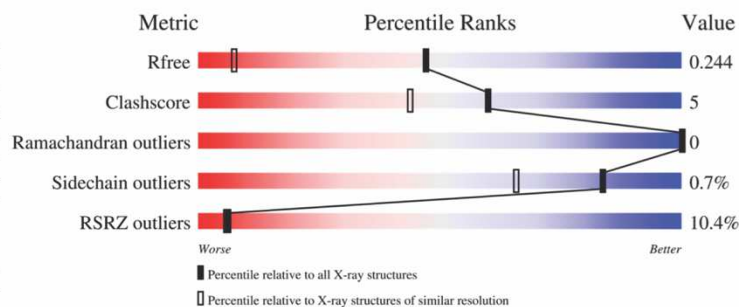
The crystal structure demonstrates more details in atomic level not only for the Tsg101 UEV domain, but also for the interaction with pepORF3 peptide. For example, the hydrogen bonds between the peptide and the residues in the binding site are detected, an information which is missing from NMR analysis. In addition, solving the crystal structure of the complex, it is shown that indeed the binding site of the pepORF3 peptide in the Tsg101 UEV domain is the same with the binding site of "late peptide" of HIV-1 Gag protein on the available structures (PDB ID: 1m4p and 1m4q)[146] and (PDB ID: 3obx, 3obq and 3obu)[147] as well with the binding site of Ebola PTAP late domain peptide (PDB ID: 4eje). Comparing this crystal structure with each of the known structures, aligning the backbone atoms of Tsg101 UEV domain, the RMSD is in the range 0.357 to 0.523, which means that the structures do not have big differences (Figure 135). Moreover, the orientation of pepORF3 peptide in the binding site is the same with the orientation of the "late peptides" and also there are hydrogen bonds between two specific Tsg101 UEV domain residues, Asn71 and Ser143, with residues of the peptides. Especially, in all structures, Ser143 binds a Proline residue in the C-terminal region of the peptides, in position 6, via a hydrogen bond, as shown in Figure 135 with yellow dashed lines.

PDB ID: 3obq
HRS PSAP peptide: PTPSAPVPL

PDB ID: 3obu
HIV-1 PTAP peptide: PEPTAPPEE

PDB ID: 7nlc
pepORF3 peptide: TSPSAPPLPP
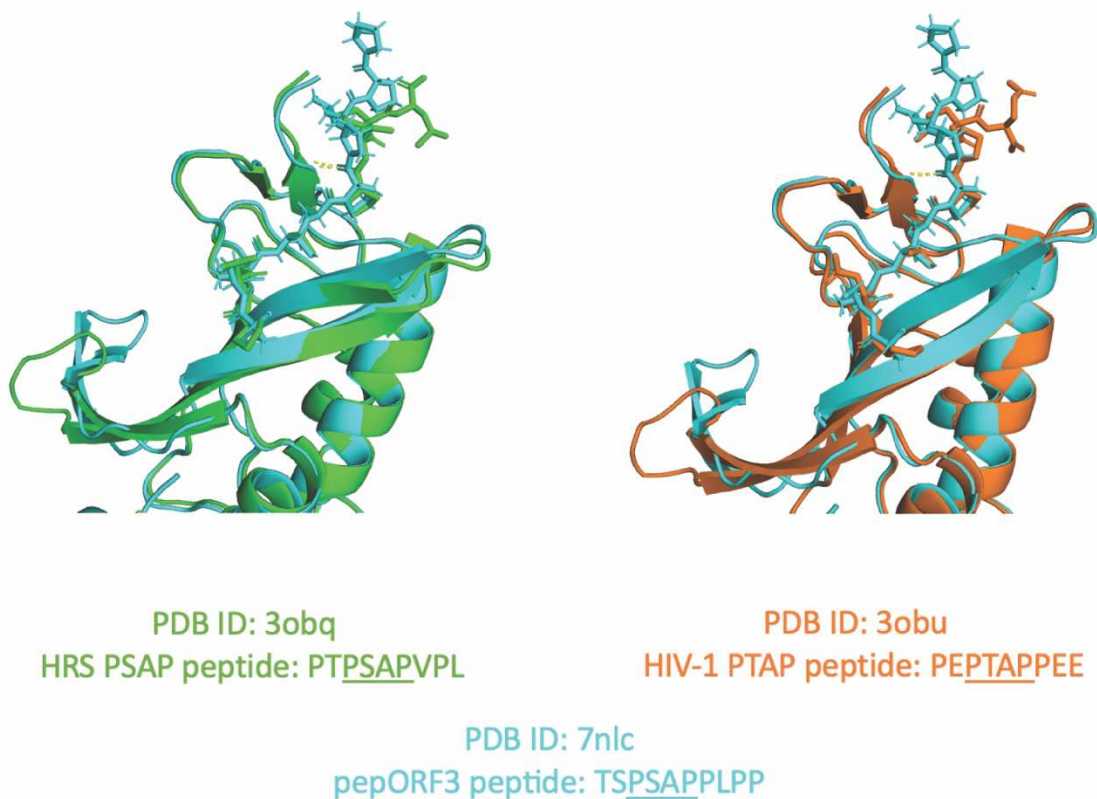
Figure 135. Comparison of crystal structures of human Tsg101 UEV in complex with "late peptides". Alignment of the backbone atoms of the Tsg101 UEV domain with HEV pepORF3 peptide in cyan (PDB ID: 7nlc) with the ones with HRS PSAP peptide in green (PBD ID: 3obq)[147] and with the ones with HIV-1 PTAP peptide in orange (PDB ID: 3obu)[147]. The yellow dashed lines correspond to the hydrogen bond of Ser143 of human Tsg101 UEV protein with Pro residue in position 6.

The binding site in which the pepORF3 peptide is located is a groove at the protein surface (Figure 136a). To break this kind of interaction it is much more difficult than in a usual protein-protein interaction involving a pocket binding site. Our high-resolution crystallographic structure of the Tsg101 UEV-pepORF3 complex could be used for the sake of new HEV inhibitors development. Indeed, specific compound(s) that can interfere with the Tsg101 UEV domain-ORF3 interaction and ultimately prevent the release of new virions from the infected cell need to be identified.

Looking closer to the solved crystal structure and Figure 136, there are two binding sites on the Tsg101 UEV domain that are very close, but also, non-overlapping. The groove in which the pepORF3 peptide is located (Figure 136a) and the ubiquitin binding site which is a binding pocket (PDB ID: 1s1q)[145] (Figure 136b). Also, a solution-state NMR structural ensemble of Tsg101 UEV domain in complex with tenatoprazole compound, an inhibitor drug candidate, shows that the latter binds in ubiquitin binding site though a disulfide bond with Cys73 of the protein (PBD ID:

5vkg). This drug proved to interfere with the early HIV-1 assembly, independently of its interaction with the HIV-1 Gag PTAP late domain[154]. In addition, Leis *et al.* published a paper last year showing that tenatoprazole and the more potent related prazole drug, ilaprazole, effectively block the release of HIV-1 and Herpes Simplex Virus (HSV) 1/ 2 infectious particles from infected cells in culture[155]. Regarding HEV infection, this class of molecules need to be further investigated.
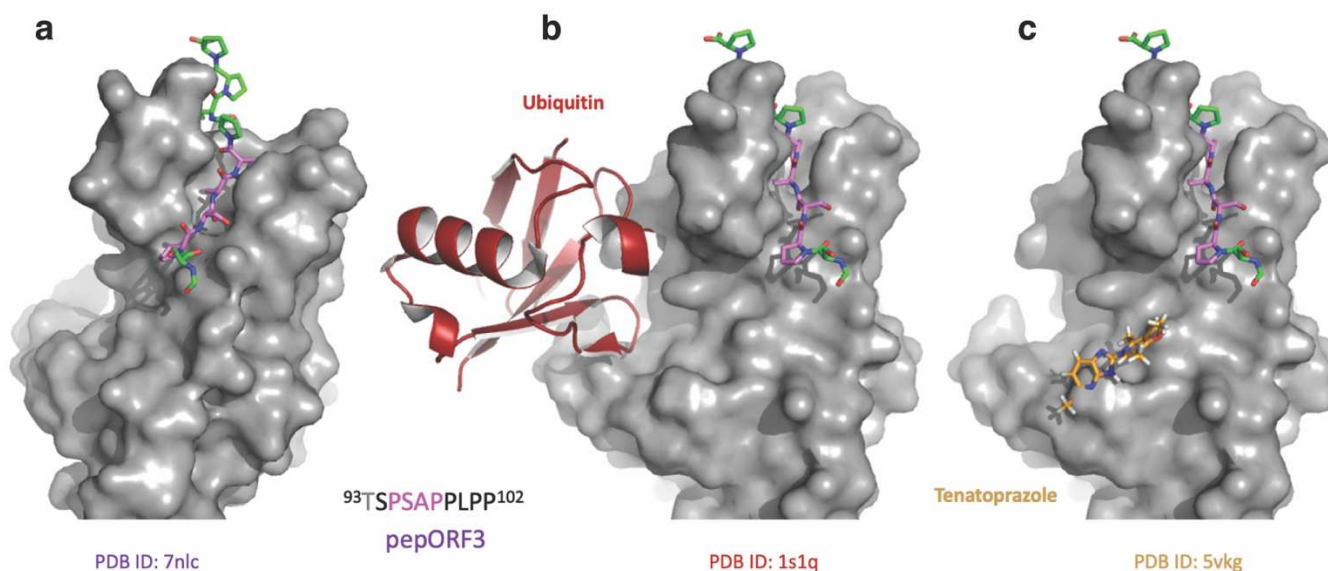
Figure 136. Binding sites of human Tsg101 UEV domain. (a) Crystal structure of Tsg101 UEV domain in presence of pepORF3 peptide in purple (PDB ID: 7nlc), (b) Crystal structure of Tsg101 UEV domain in complex with ubiquitin in red (PDB ID: 1s1q)[145], (c) Solution-state NMR structural ensemble of Tsg101 UEV domain in complex with tenatoprazole in yellow (PBD ID: 5vkg)[154].

3.6 Assay development for drug screening

Our high-resolution crystallographic structure of the Tsg101 UEV-pepORF3 complex (PDB: 7nlc) could be used for the sake of new HEV inhibitors development. Indeed, specific compound(s) that can interfere with the Tsg101 UEV domain-ORF3 interaction and ultimately prevent the release of new virions from the infected cell need to be identified.

As mentioned above, the peptide is located in a groove at the Tsg101 UEV surface which is very close, but non-overlapped to the ubiquitin binding pocket. To break this kind of the interaction and therefore block the release of new virions from the cell, is much more difficult than a usual protein-protein interaction in a pocket binding site. In order to find specific compound(s) which would only interfere with the Tsg101 UEV domain-ORF3 interaction, a high-throughput screening has to be performed. Due to the existence of two binding sites on the Tsg101 UEV domain, the screening procedure has to rely on a methodology that ascertain the targeting of the ORF3 binding groove and not the ubiquitin binding pocket. Solution-state NMR spectroscopy, using the NMR assignment of Tsg101 UEV, it will allow the direct identification of the binding site of the compounds, but it is not suited for a high-throughput screening. The other option is to use a competition assay with the displacement of the pepORF3 peptide from the Tsg101 UEV.

For this goal, we first developed a fluorescence polarization (FP) assay, in which we can monitor the displacement of a fluorescent-ORF3-derived peptide. Due to the potential autofluorescence of some compounds in the high-throughput screening, an experimental assay relying on Homogeneous Time Resolved Fluorescence (HTRF) technology was then considered.

3.6.1 Fluorescence Polarization of Tsg101 UEV domain with FITC-pepORF3 peptide – Titration and Competition Assays

In order to develop a fluorescence polarization (FP)-based competition assay, we take advantage of our crystallographic structure of the complex and we chose the minimal ORF3-derived peptide that can bind Tsg101 UEV, corresponding to the ^{94}SPSAPPLP^{102} ORF3 sequence. This peptide has been synthesized (GeneCust) with a fluorescent probe (FITC) at its N-terminus giving the Fl-pepORF3 peptide (FITC-SPSAPPLP). The FP technique is based on the detection of polarized light of a fluorescent small molecule which is inversely proportional to its tumbling rate[211]. When the fluorescent molecule is bound by a larger molecule, as a protein, its molecular tumbling rate decrease and thus the polarization of the light emitted is increased. First, we checked if this experimental setting can be used to monitor the interaction between Tsg101 UEV and the small Fl-pepORF3 peptide. Using either a 25 μM or 2.5 μM Fl-pepORF3 concentration, increasing concentrations of Tsg101 UEV protein in 1X PBS buffer (2-fold dilutions with initial concentration at 500 μM, 14 points) were added in a final reaction volume of 30 μL using a 384-well plate. The fluorescence is measured using a PHERAstar microplate-reader (BMG labtech) equipped with a FITC excitation and emission filter (FITC excitation max at 490 nm and FITC emission max 525 nm). Figure 137 illustrates the data obtained and shows that indeed we are able to monitor the interaction between Tsg101 UEV and the Fl-pepORF3 peptide using FP assay. Then, a 1:1 molecular interaction equation is used to fit the experimental points and to calculate the dissociation constant (K_D) of the interaction. Both assays with 25 μM and 2.5 μM Fl-pepORF3 peptide give similar K_D values 27.4 μM and 28.9 μM, respectively. These values are in the same range that the ones obtained using ITC experiments, 28 μM and 23 μM for full-length ORF3 and pepORF3, respectively. This means that the smaller peptide sequence is still sufficient to interact with Tsg101 UEV and that the fluorescent probe does not interfere with the interaction.
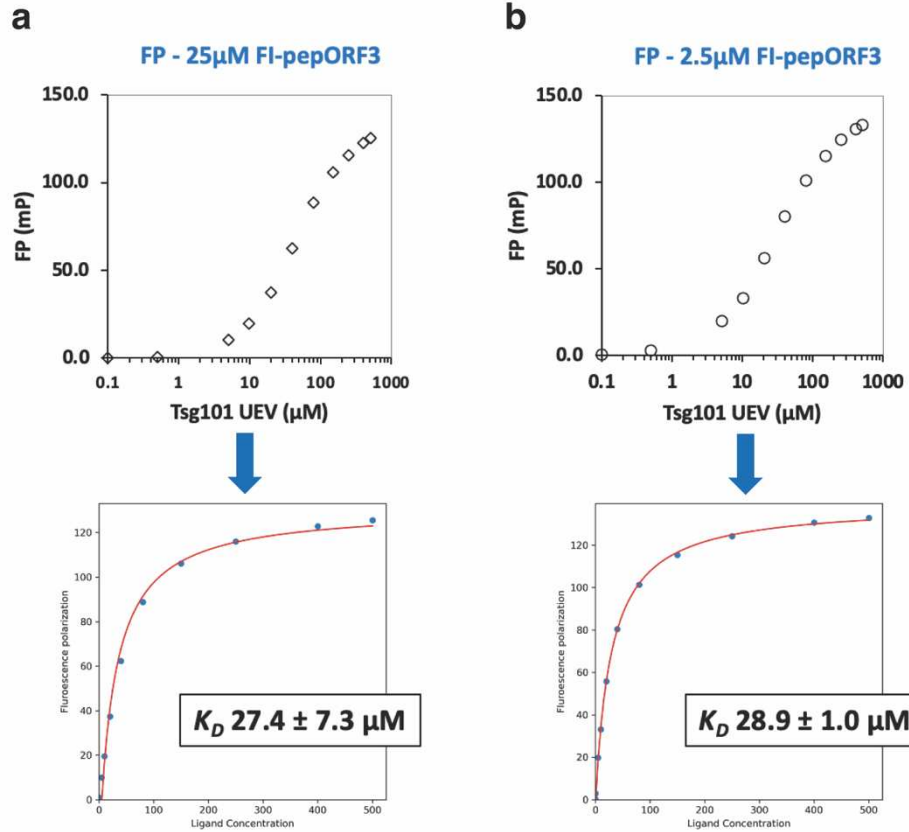
Figure 137. Fluorescence polarization plots and their analysis of (a) 25 µM and (b) 2.5 µM Fl-pepORF3 peptide with increasing concentration of Tsg101 UEV protein performed using PHERAstar microplate-reader (BMG labtech).

Next, in order to reduce the amount of the Fl-pepORF3 peptide used, we repeated the former titration experiment with lower Fl-pepORF3 concentrations, from 2 µM to 1.25 nM. To keep a good signal to noise ratio the concentration of 100 nM is found to be suitable. The titration curve for 100 nM Fl-pepORF3 peptide with increasing concentration of Tsg101 UEV domain and same titration points is shown in Figure 138.
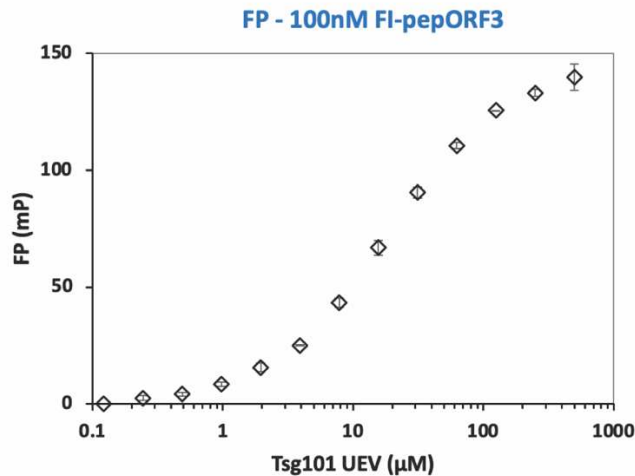
Figure 138. Fluorescence polarization plot of 100 nM Fl-pepORF3 peptide with increasing concentration of Tsg101 UEV protein performed using PHERAstar microplate-reader (BMG labtech).

Proving that the FP assay is working and obtained the same affinity data with Fl-pepORF3 peptide in several concentrations we started the competition assay. We choose a concentration of 100 nM for the fluorescent Fl-pepORF3 peptide and a concentration of 150 µM for Tsg101 UEV protein. This corresponds to roughly 80% of the peptide in the bound state. Using this setting, we performed several competition assays. To validate the competition assay, we first used the unlabeled pepORF3 peptide to compete with the Fl-pepORF3 peptide.

Having the concentration of Tsg101 UEV domain and Fl-pepORF3 peptide constant, the unlabeled pepORF3 peptide is added in the reaction with initial concentration at 500 µM and 2-fold dilution with 14 points are prepared and the polarization of the fluorescent peptide is recorded. Figure 139 depicts the decrease of the fluorescent polarization upon the addition of the unlabeled peptide signifying its competition with the FITC peptide and it scoops the Fl-pepORF3 out of the Tsg101 UEV groove. This result proves that the fluorescence polarization assay is suitable for a screening of compounds that can interfere with the interaction between Tsg101 UEV and ORF3 proteins.
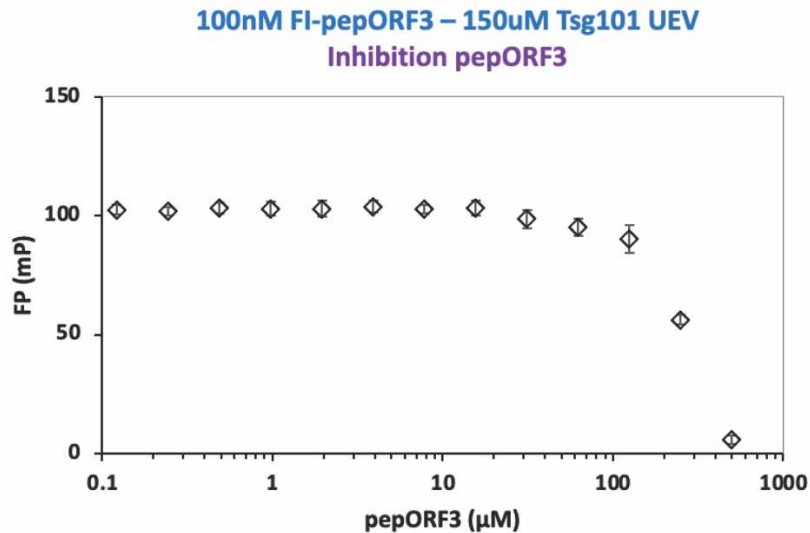
Figure 139. Fluorescence polarization plot of 100 nM Fl-pepORF3 peptide with 150 μM Tsg101 UEV protein and increasing concentration of unlabeled peptide pepORF3 performed using PHERAstar microplate-reader (BMG labtech).

In order to screen the Fr-PPIChem library, a library of compounds which is dedicated to target protein-protein interactions and contains 10.3 thousand compounds, a collaboration with Carine Derviaux and Xavier Morelli, leaders of the HiTS/IPCdd platform in Marseille Cancer Research Center, is started. Whereas the FP assay has been validated and is suitable for high-throughput screening purpose it suffers from potential false positive identifications in case of autofluorescence of the compounds tested. We thus agree to test another assay, the Homogeneous Time Resolved Fluorescence (HTRF) technology.

3.6.2 Homogeneous Time Resolved Fluorescence (HTRF) technology

The Homogeneous Time Resolved Fluorescence (HTRF) technology is based on the fluorescence resonance energy transfer (FRET) principle which involves a fluorescent donor and a fluorescent acceptor and measures the transferred energy between them based on their spatial proximity (Figure 140)[212].
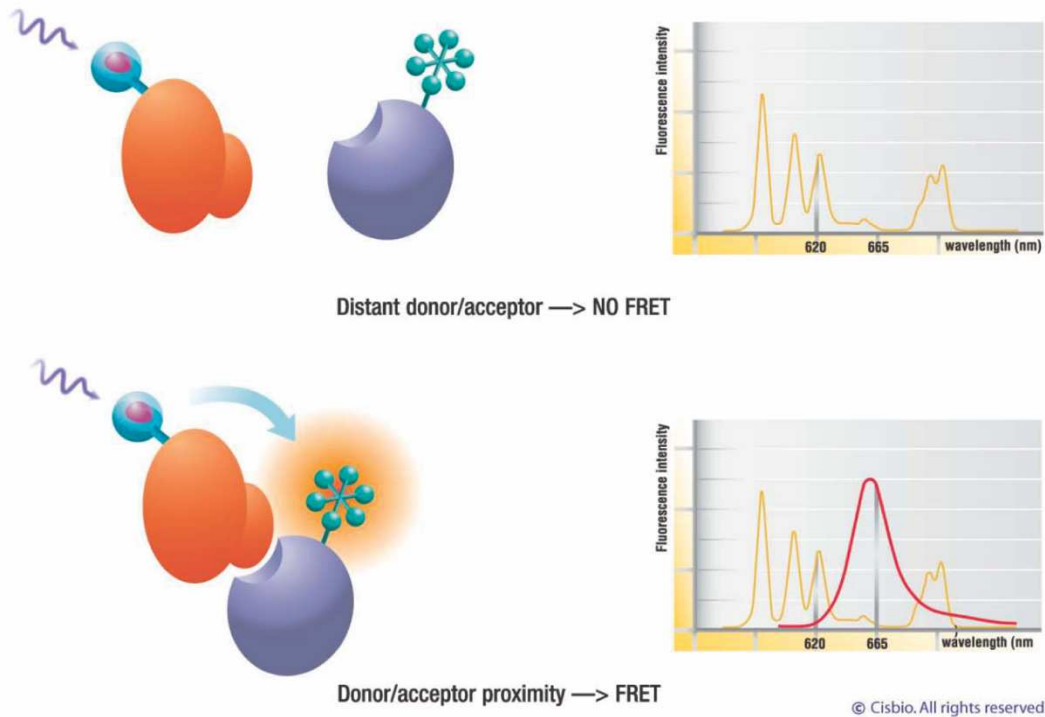
Figure 140. Principle of fluorescence resonance energy transfer (FRET) technology. The spatial proximity of fluorescent donor and acceptor generates a FRET signal (in red). From Degorce et al.[212], available under a Creative Commons Attribution License (CC BY 2.5).

For this purpose, the Tsg101 UEV FLAG-tag construct with an extra Flag-tag (DYKDDDDK) in the N-terminal is designed and an ORF3 peptide with a biotin in N-terminal, Biotin-GSTSPSAPPLPP, biotin-pepORF3, is designed and then synthesized by GeneCust.

The Tsg101 UEV FLAG-tag protein is first purified using Ni-affinity chromatography, the 6xHis-tag is cleaved by TEV protease, both the cleaved his-tag and the TEV protease are removed from Tsg101 UEV FLAG-tag using a HisTrap column. The latter is further purified by HiLoad 16/600 Superdex 75 pg (Cytiva) size exclusion column. Figure 141 illustrates the SEC chromatogram

monitoring using the 280 nm absorbance curve (a) and the 4-20% SDS-PAGE with the SEC fractions (b).
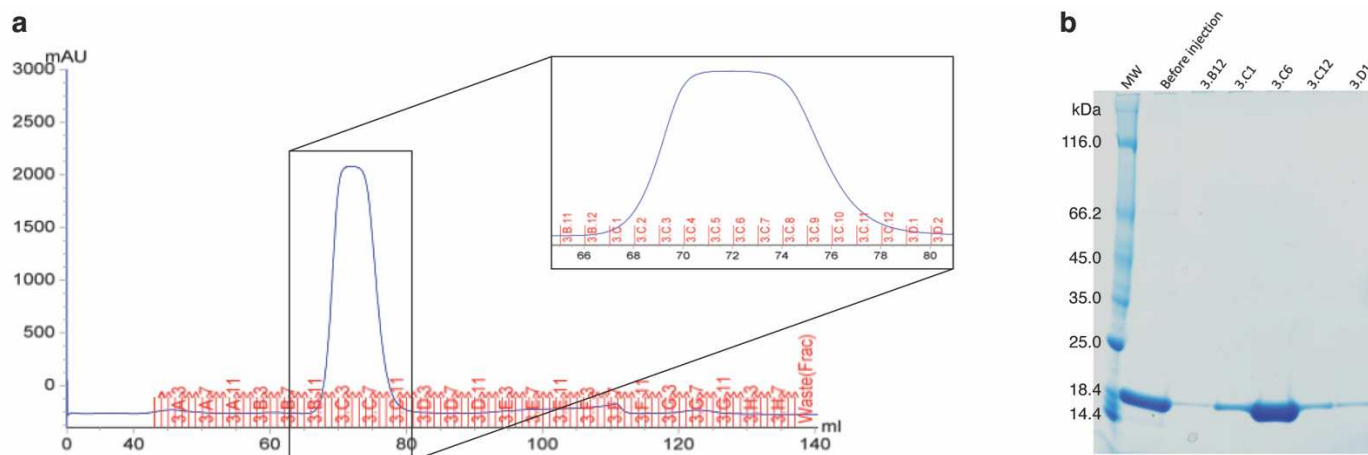
Figure 141. (a) Chromatogram of SEC purification of Tsg101 UEV FLAG-tag using HiLoad 16/600 Superdex 75 pg column monitoring using the 280 nm absorbance curve. (b) 4-20% SDS-PAGE with fractions collected in the SEC purifications (a) with Coomassie staining.

The concentrated Tsg101 UEV FLAG-tag protein and the biotin-pepORF3 are then sent to the HiTS/IPCdd platform in Marseille. Using streptavidin for the peptide and an antibody for the Flag-tag which both carry a fluorescence probe, they first test the system, but unfortunately, they could not detect any signal. In order to exclude the possibility that either the Flag-tag in Tsg101 UEV or the biotin moiety on the ORF3 peptide interfere with the interaction a four-point titration experiment using NMR Spectroscopy is performed using these 2 molecules. A 100 µM ^{15}N Tsg101 UEV FLAG-tag labeled protein in NMR buffer containing DMSO (50 mM Sodium Phosphate pH 6.1, 50 mM NaCl, 0.1 mM EDTA, 2.67% DMSO) is mixed with 20, 50, 100 and 400 µM of biotin-pepORF3 peptide. For each point, a sample with the appropriate amount of protein and peptide is placed in a 3 mm tube with 200 µL total volume and a 2D ^{1}H, ^{15}N HSQC spectrum is recorded at 293K on 900 MHz Spectrometer. In Figure 142, the overlay of the 2D HSQC spectra of ^{15}N Tsg101 UEV FLAG-tag with and without the biotin-pepORF3 peptide in ratio 1:1 (a) is compared with the corresponding ones of ^{15}N Tsg101 UEV domain with and without unlabeled ORF3 C20 protein in ratio 1:1 (b). The blue arrows indicate the affected peaks which are the same in both experiments. Therefore, the new constructs do not disrupt the interaction.
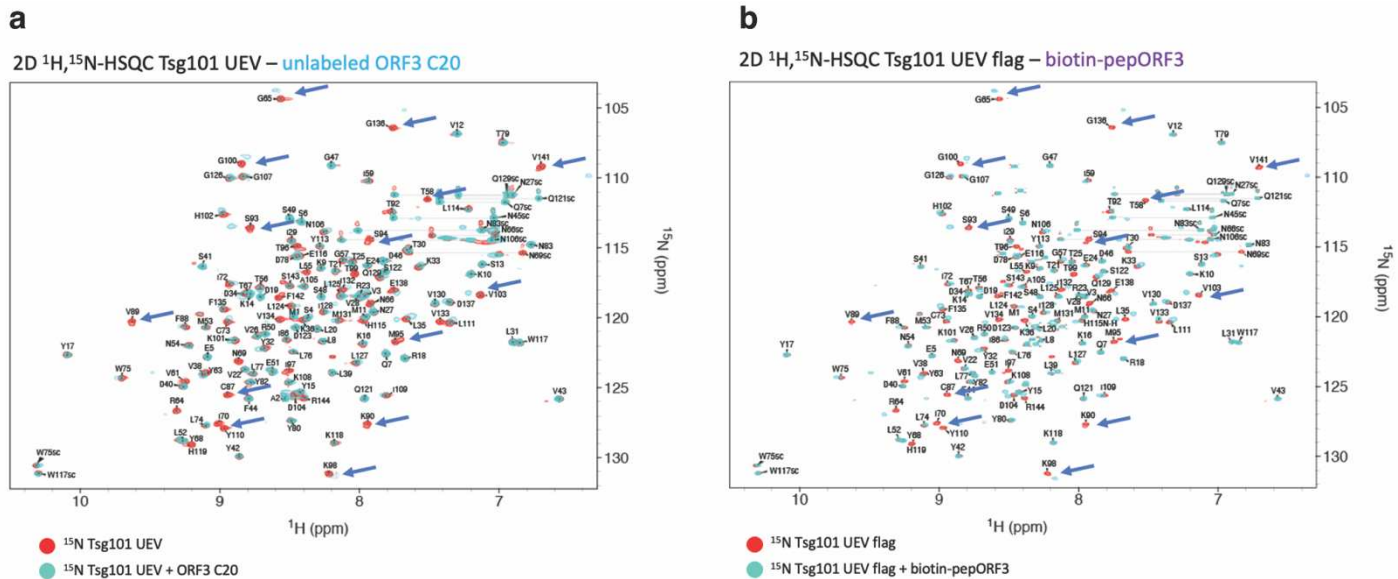
Figure 142. (a) Overlay of 2D ^1H, ^{15}N HSQC spectra of ^{15}N Tsg101 UEV domain without (in red) and with (in cyan) unlabeled ORF3 C20 protein in ratio 1:1. (b) Overlay of 2D ^1H, ^{15}N HSQC spectra of ^{15}N Tsg101 UEV FLAG-tag domain without (in red) and with (in cyan) biotin-pepORF3 peptide in ratio 1:1. The blue arrows indicate the affected peaks which are the same in both (a) and (b) experiments.

A possible explanation for the negative HTRF results is that the anti-Flag antibody, once bound to the N-terminal Flag sequence in Tsg101 UEV FLAG-tag, may interfere with the interaction with the biotin-pepORF3 peptide, as a consequence of steric hinderance. To overcome this, we plan to use another Tsg101 UEV protein that this time is fused, at its N-terminus, to a GST protein. The GST being much larger than the Flag-tag peptide, an anti-GST antibody may not interfere with the interaction studied. As mentioned in the Material and Methods section, the protein construct does not contain any cleavage site and therefore it is purified in two steps using an affinity GSTrap (Cytiva) column and a HiLoad 16/600 Superdex 75 pg (Cytiva) size exclusion column. The concentrated protein has been recently sent to Marseille and the HTRF experiments have to be done.

3.7 Study of interference of the prazole drugs with the ORF3-Tsg101 UEV interaction

As mentioned above, recent data in the literature have shown that prazole-based compounds, tenatoprazole and ilaprazole, are bound in the ubiquitin binding site of Tsg101 UEV domain resulting the blockade of the release of infectious HIV-1 from cells in culture without disruption of PTAP binding activity[154] and infectious Herpes Simplex Virus (HSV)-1/2 release from Vero cells in culture[155]. A solution-state NMR structural ensemble of Tsg101 UEV domain in complex with tenatoprazole compound is also available (PBD ID: 5vkg)[154].

Therefore, we wondered if prazole-based compounds could interfere with the Tsg101 UEV domain-ORF3 interaction. For this purpose, we conducted Thermal Shift Assay (TSA), NMR Spectroscopy and fluorescence polarization (FP) experiments, and the results are described below.

3.7.1 Thermal Shift Assay (TSA) of Tsg101 UEV domain with pepORF3 peptide and prazole drugs, ilaprazole sodium and tenatoprazole

The Thermal Shift Assay (TSA) technique was used to detect any difference in the thermal stability of the Tsg101 UEV domain first in presence of pepORF3 peptide and using this information to study the effect with prazole drugs, tenatoprazole and ilaprazole sodium.

This technique is based on the differential scanning fluorimetry (DSF) measuring the fluorescence of SYPRO Orange dye (Thermo Fisher Scientific) during the denaturation of the protein while the fluorophore binds to the exposed hydrophobic surfaces. The melting temperature of the protein (Tm) is determined from the fitted curve and is the temperature at which 50% of the protein denaturation is occurred[213]. The principal of Thermal Shift Assay is shown in Figure 143.
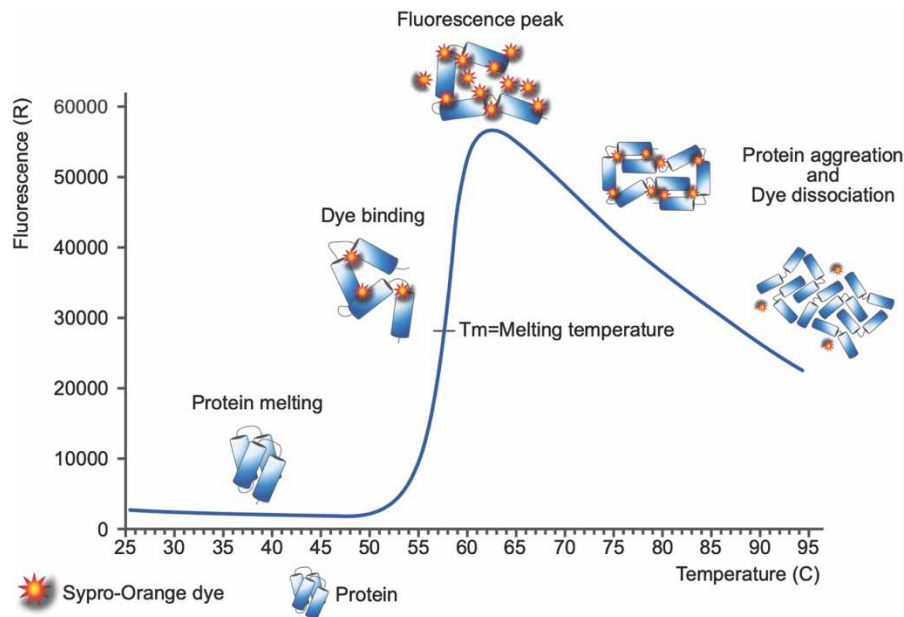
Figure 143. Principal of Thermal Shift Assay (TSA) using SYPRO orange dye. (Figure derived from website[214]).

The TSA experiments are performed with the LightCycler® 480 white Multiwell Plate 96 (Roche) device located at Pasteur Institute under the guidance of Adrien Herledan. First, the optimal conditions, the concentration of the protein and the SYPRO orange dye as well the buffer, were determined to be 10 µM, 10X and TSA Buffer containing 20 mM Tris pH 7.5, 50 mM NaCl, 1mM EDTA, respectively. The melting temperature of Tsg101 UEV protein alone is calculated around 69°C using the LightCycler Thermal Shift Analysis software v2.0[215] (Figure 144).
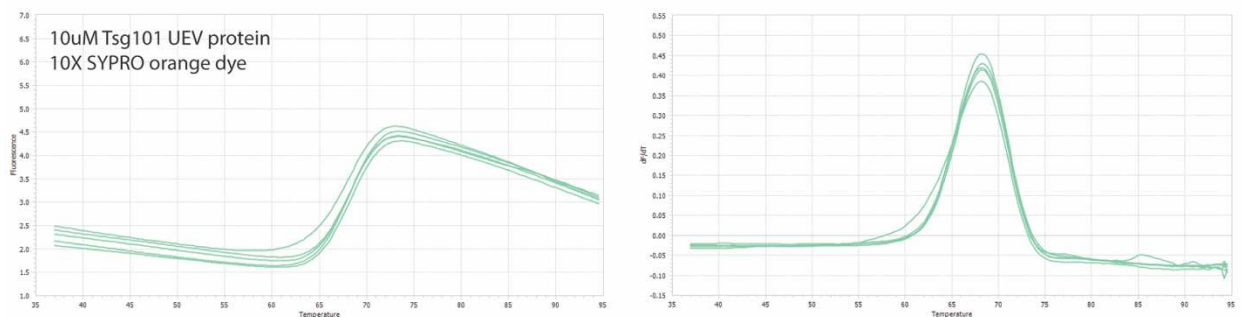
Figure 144. Thermal Shift Assay (TSA) curves for 10 µM Tsg101 UEV protein with 10X SYPRO orange dye in TSA Buffer (20 mM Tris pH 7.5, 50 mM NaCl, 1mM EDTA).

After determination of the melting temperature of Tsg101 UEV protein, the effect of HEV pepORF3 peptide, tenatoprazole and ilaprazole sodium in protein stability is studied. In general, an increase of the melting temperature means the increase of the stability of the protein bound to the molecule while the shift in lower temperatures means the protein destabilization. Before

studying the interference of prazole compounds with ORF3 protein, in this case with HEV pepORF3 peptide because of the disordered nature of the protein, the melting temperature of Tsg101 UEV protein in presence of each molecule is determined. A TSA experiment of 10 µM Tsg101 UEV protein with 10X SYPRO orange dye and 0, 7.5 and 20 µM pepORF3 peptide, ilaprazole sodium and tenatoprazole in TSA Buffer (20 mM Tris pH 7.5, 50 mM NaCl, 1mM EDTA) is performed as shown in Figure 145. The presence of pepORF3 peptide does not cause any change to the TSA curves and thus in the melting temperature of Tsg101 UEV protein. For ilaprazole sodium and tenatoprazole, there is a clear shift of the Tm peak at ~56°C and ~60°C, respectively, and the presence of two peaks at 20 µM of the molecule, as previously described in literature[155]. Due to the non-effect of melting temperature of Tsg101 UEV protein in presence of pepORF3 peptide, the interference with prazole compounds could not be further studied using this technique.
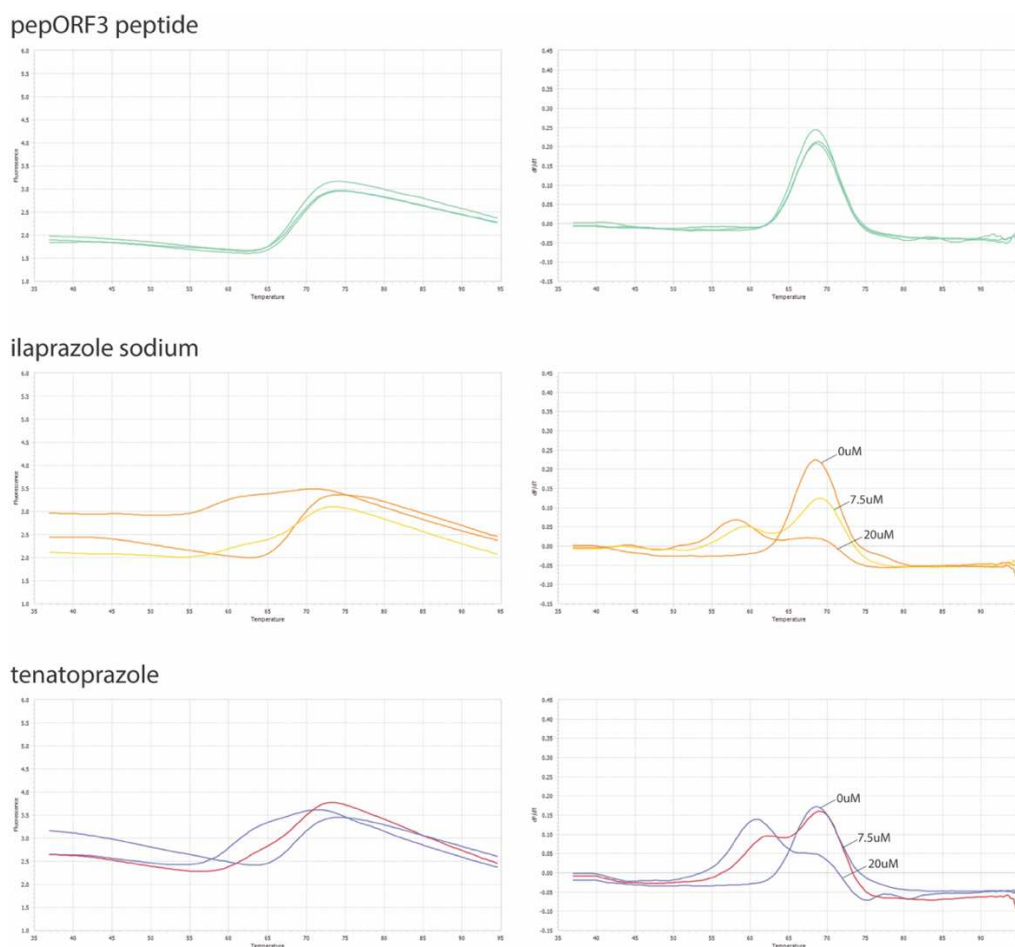
Figure 145. Thermal Shift Assay (TSA) curves for 10 µM Tsg101 UEV protein with 10X SYPRO orange dye and 0, 7.5 and 20 µM pepORF3 peptide, ilaprazole sodium and tenatoprazole in TSA Buffer (20 mM Tris pH 7.5, 50 mM NaCl, 1mM EDTA).

3.7.2 NMR study of ^{15}N Tsg101 UEV domain in presence of unlabeled ORF3 C20 protein, ilaprazole sodium and both.

Due to a lack of useful TSA results, the next technique used for the detection of the interference of the ilaprazole to the Tsg101 UEV domain-ORF3 protein interaction is solution-state NMR Spectroscopy. The ilaprazole compound is selected for NMR study because it is more potent than tenatoprazole regarding the *in vitro* binding properties[155].

A 100 μM ^{15}N, ^{13}C Tsg101 UEV protein sample in NMR Buffer (50 mM Sodium Phosphate pH 6.1, 50 mM NaCl, 0.1 mM EDTA) placed in 5 mm tube is used to record 2D ^{1}H, ^{15}N HSQC spectra at 293K on 900 MHz Spectrometer. Apart from the control spectrum (in red), 2D HSQC spectra are recorded for the labeled Tsg101 UEV domain with 100 μM unlabeled ORF3 C20 protein (in cyan), with 500 μM ilaprazole sodium (in green) and with both 100 μM unlabeled ORF3 C20 protein and 500 μM ilaprazole sodium (in blue). The overlay of the four 2D spectra is shown in Figure 146.
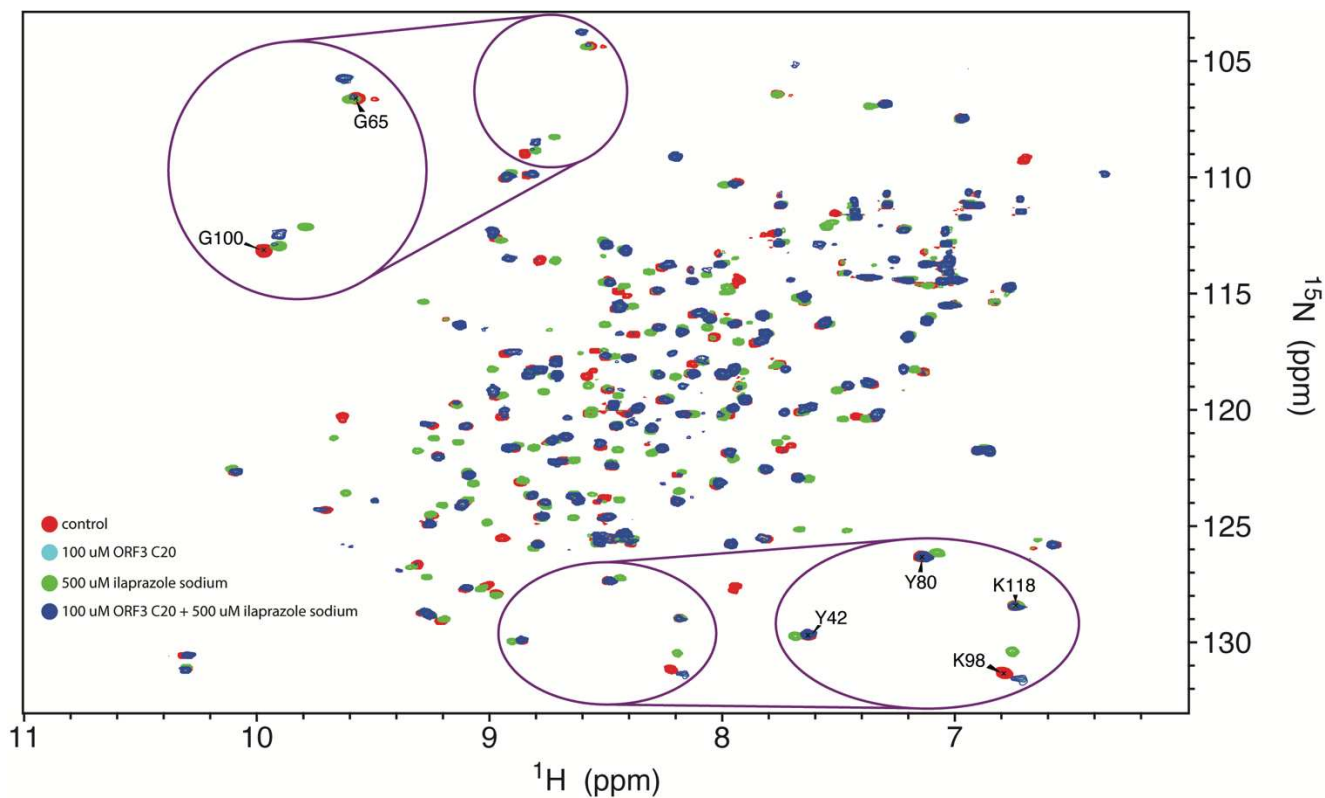
Figure 146. Overlay of 2D ^{1}H, ^{15}N HSQC spectra of 100 μM ^{15}N, ^{13}C Tsg101 UEV domain (in red), with 100 μM unlabeled ORF3 C20 protein (in cyan), with 500 μM ilaprazole sodium (in green) and with both 100 μM unlabeled ORF3 C20 protein and 500 μM ilaprazole sodium (in blue).

The analysis of the Chemical Shift Perturbations (CSP) values for all assigned non-Proline residues of Tsg101 UEV domain induced by ORF3 C20 protein (panel a), by ilaprazole sodium (panel b) and ORF3 C20 protein and ilaprazole sodium (panel c) are shown in Figure 147.
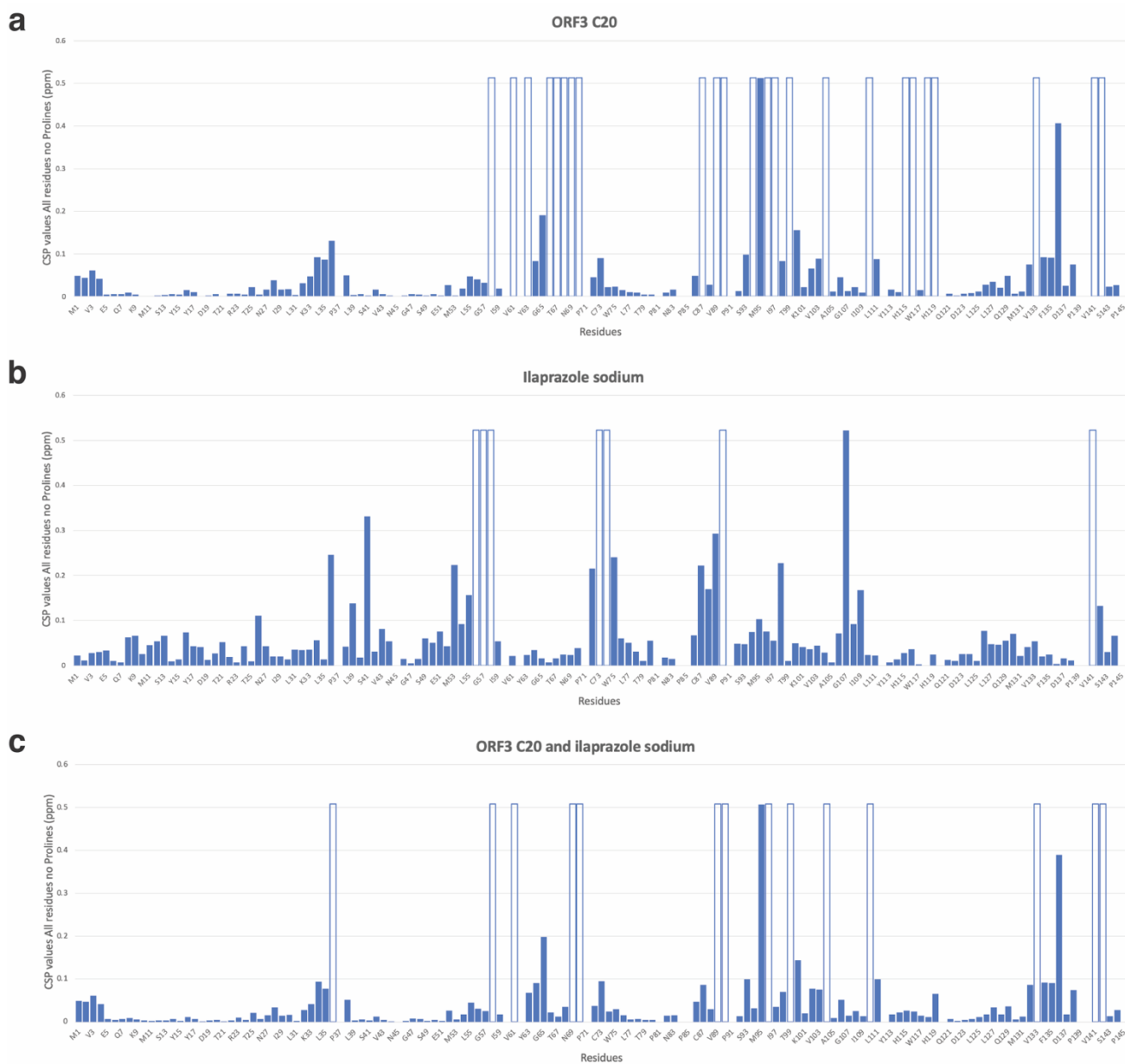
Figure 147. Chemical Shift Perturbation (CSP) values of all residues non-Prolines of Tsg101 UEV domain induced by ORF3 C20 (a) protein, by ilaprazole sodium (b) and by ORF3 C20 protein and ilaprazole sodium (c). The "empty" bars correspond to the residues that are disappeared. The Proline residues and unassigned Asn45 were not considered in the analysis.

Based on the HSQC spectra and CSP analysis for all experiments, both ORF3 C20 protein (as shown in the previous chapter) and ilaprazole sodium interact directly with Tsg101 UEV domain. The

overlay of HSQC spectra shows that some peaks, such as Lys98 and Gly100 in the zoomed panels, are affected with the addition of either ORF3 C20 protein or ilaprazole sodium, but the direction of the these shifted peaks is completely different. Comparing the CSP values, the affected residues when ORF3 C20 is added (Figure 147, panel a) and the ones in presence of ilaprazole sodium (Figure 147, panel b) are different as expected because they bind in different binding sites. In the experiment in which both components are added, the affected peaks are a combination of the ones in the previous experiments when only one component is present. Therefore, the ilaprazole sodium seems that it does not interfere with the Tsg101 UEV-ORF3 interaction.

3.7.3 Fluorescence Polarization of Tsg101 UEV domain with FITC-pepORF3 peptide – Competition Assays

Using the developed fluorescence polarization assay described above, competition assays with ilaprazole and tenatoprazole which both bind to the ubiquitin binding site of Tsg101 UEV in order to clarify if these molecules interfere with the Tsg101 UEV domain-ORF3 interaction are performed. The increasing concentration of the prazole-based compounds does not affect the fluorescence polarization of the FITC peptide as shown in Figure 148.
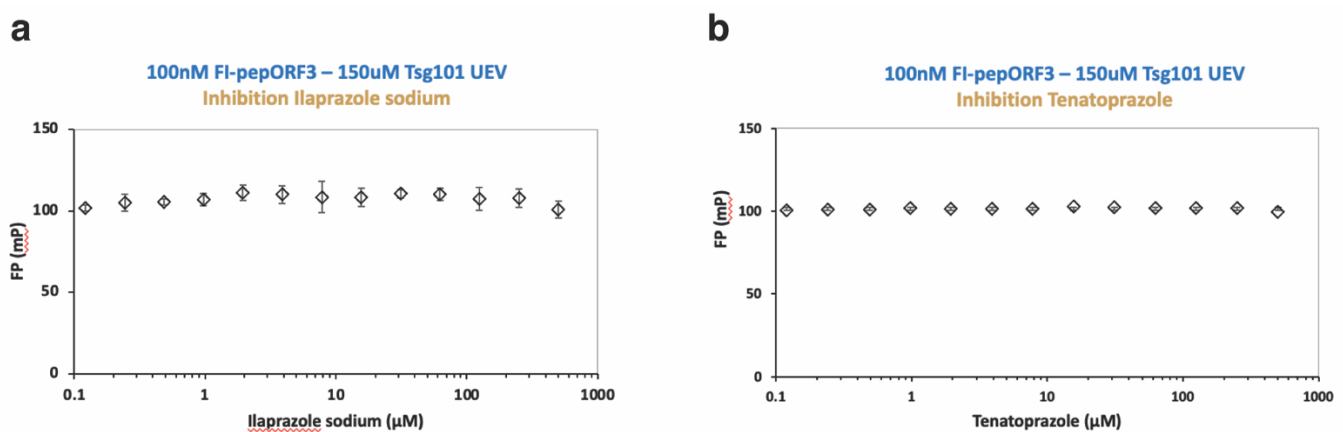
Figure 148. Fluorescence polarization plot of 100 nM Fl-pepORF3 peptide with 150 μM Tsg101 UEV protein and increasing concentration of (a) ilaprazole sodium and (b) tenatoprazole performed using PHERAstar microplate-reader (BMG labtech).

To conclude, the prazole-based drugs do not disrupt the Tsg101 UEV interaction with ORF3 protein.

Therefore, if it is shown that prazole drugs may be efficient against HEV infection it is highly probable that it will not due to the disruption of the Tsg101 UEV-ORF3 interaction.

Conclusions & Perspectives

Hepatitis E Virus (HEV) is the most common cause of acute viral hepatitis and represents a public health problem which is constantly growing around the world. It is a small, icosahedral virus which contains a ~7.2 kb positive-sense, single-stranded RNA genome which comprises three open reading frames, ORF1, ORF2 and ORF3. ORF1 encodes a non-structural polyprotein that includes multiple functional domains responsible for the replication of the viral genome. ORF2 encodes the viral capsid protein that assembles to make the viral particles. ORF3 encodes a small regulatory multifunctional protein which is poorly characterized. Previous studies have shown that ORF3 is mainly involved in the release of the infectious viral particles and interacts with other viral and host proteins inside the cell, but also it is associated with the cellular membranes. This study focuses in the first detailed molecular characterization of the ORF3 protein in order to decipher its functional role(s) during the HEV life cycle using multiple biophysical techniques, including NMR spectroscopy, Isothermal Titration Calorimetry (ITC) and X-ray crystallography.

In this study, the HEV Genotype 3 ORF3 sequence, found mainly in developed countries and that infects humans by consuming of uncooked or undercooked meat from infected animals, is used. Analyzing the protein sequence, two hydrophobic domains in the N-terminal, one cysteine-rich and one predicted to be transmembrane using the online TMHMM v2.0 server[186], are identified. The palmitoylation of one or multiple cysteine residues located on the cysteine-rich region results to the protein anchoring to the membrane based on the Gouttenoire *et al*[82]. On the other hand, Ding *et al* proposed a transmembrane insertion of ORF3 protein and its oligomerization that forms an ion channel correlating the ORF3 function with a viroporin function[81]. In addition, the C-terminal region contains many proline residues and is predicted to be disordered using the online GeneSilico MetaDisorder server[187]. The C-terminal region also contains a PSAP motif which is closed to the PTAP motif found in other viruses and through this motif they interact with the Ubiquitin E2 Variant (UEV) domain of human tumor susceptibility gene 101 (Tsg101) protein, essential protein for the viral secretion. Although different constructs are designed and studied, the ORF3 C20 in which all Cysteines are mutated to Serine except of Cysteine in position 20 of the protein sequence is the most characterized protein in this study.

The final optimum purification protocol for ORF3 protein contains the following steps, the cell lysis, the pellet resuspension with denaturing buffer (6M Urea), the Ni^{2+}-affinity chromatography followed by Reverse Phase chromatography and finally the dialysis and concentration of the protein fractions. Because of the lack of the aromatic residues in the ORF3 protein sequence, the monitoring of the purification is done using the unspecific 215 nm wavelength and the concentration estimation in the final purification step is determined combining the results of a 4-20% SDS-PAGE with increasing volume of ORF3 sample and Bradford assay estimation using the calibration curve specifically built for ORF3 protein. Regarding the ORF3 constructs, the ORF3 C8A protein had good expression, but it was soluble only in presence of detergent (2% octyl-β-D-glucopyranoside, β-OG) and therefore it was not further studied. The ORF3 C8S protein had good expression, but it was used only as template for designing the mutant ORF3 C20. The latter, the His-ORF3 C20, ORF3 Cter and the ORF3 WT had high expression yields and are further characterized by NMR spectroscopy.

First, the 2D ^{1}H, ^{15}N HSQC spectrum of ORF3 Cter protein has a narrow dispersion in its proton dimension and it is the first experimental evidence that the C-terminal region of the protein is disordered. The backbone and proline assignments of ORF3 Cter protein were obtained by recording 2D and 3D NMR spectra.

Next, the 2D ^{1}H, ^{15}N HSQC spectrum of ORF3 C20 protein has also a narrow dispersion in its proton dimension and thus, it suggests that the full-length ORF3 C20 protein is mainly disordered without excluding the possibility of the presence of low-populated helical segments. The analysis of the experimental NMR data using the Secondary Structure Propensities (SSP) and the overall low score for all residues is further evidence of the disordered nature of the ORF3 protein. The same conclusion is obtained after recording the ^{15}N relaxation experiments for ORF3 C20 protein. Moreover, as the 2D ^{1}H, ^{15}N HSQC spectrum of ORF3 Cter overlaps with the one of ORF3-C20 it means that the presence of the N-terminal part of ORF3 does not affect the in-solution conformation of its C-terminal half.

The 2D ^{1}H, ^{15}N HSQC spectrum of ORF3 WT protein has a narrow dispersion in the proton dimension as the one of ORF3 C20 construct concluding that ORF3 WT protein is mainly

disordered as expected. The backbone assignments of this construct were obtained by recording two set of 3D data, 3D HNCO and HNCACB spectra, combining with the corresponding ones of ORF3 C20 protein and then transferring to the HSQC spectrum of ORF3 WT protein. The comparison of the 2D ^{1}H, ^{15}N HSQC spectra of ORF3 C20 and ORF3 WT showed that even if few chemical shift perturbations were observed, it is expected as 7 serine residues were changed into cysteine residues, most of the NMR resonances are similar. This emphasizes that ORF3 C20 is a good experimental model that mostly mimics the in-solution behavior of ORF3 WT.

The analysis of the experimental NMR ^{13}C chemical shifts confirms the disordered nature of ORF3. This small HEV protein thus corresponds to a dynamic ensemble of interconverting conformers. Based on the NMR data, no helical tendency was found for the ORF3 region (residues 30-52) that has been proposed (and predicted) to be transmembrane.

Regarding the membrane anchoring of ORF3 protein, we tried to get biophysical and biochemical data on the two anchoring modes that have been proposed in the literature: the ORF3 transmembrane oligomerization model[81] and the model of ORF3 protein association via post-translational modification (palmitoylation)[82]. To this end, we used the ORF3 C20 construct that contains a unique cysteine residue and Nanodiscs that mimic the membrane as it is a lipid bilayer stabilized by two copies of membrane-scaffolding proteins (MSP). Regarding the transmembrane hypothesis we fail to obtain Nanodiscs with ORF3 embedded. We had issues with the ORF3 solubility in the presence of the sodium cholate, of the amphiphilic MSP protein and even of the Bio-beads SM-2. This issue was the reason that the transmembrane insertion approach cannot be studied using the Nanodisc technology. An additional experiment to study this approach and simultaneously the oligomerization state was the preparation of liposomes using *E. coli* polar lipid extract, mix them with ORF3 C20 protein and check with native PAGE analysis. Unfortunately, no band could be detected in native PAGE without a reasonable explanation. These negative results blocked the investigation on the transmembrane oligomerization approach in this study.

Regarding the second hypothesis, the ORF3 membrane association via its palmitoylation on cysteine resdues[82], we used a two-step procedure: a formation of "empty" Nanodiscs and then the attachment of ORF3 C20 protein. A prerequisite for the second step is the existence of

modified lipids on the "empty" NDs which in this study correspond to the PE-MCC and DGS-NTA(Ni) modified lipids. PE-MCC is expected to make a covalent link with the ORC3 C20 cysteine side chain, whereas DGS-NTA(Ni) is intended to bind the 6xHis-tag of His-ORF3 C20 protein. Only with DGS-NTA(Ni) modified lipids we managed to get His-ORF3 C20 protein bound to Nanodiscs, but the NMR experiment could not provide a reliable conclusion due to the existence of extra unknown peaks on the 2D HSQC spectrum. This point has to be solved before we can get reliable NMR results on the membrane attachment of ORF3. In addition, it is worth to try again the attachment of ORF3 C20 in the PE-MCC modified lipids testing different buffer conditions and lipid ratio which could help the interaction of the maleimide group of the lipid with the thiol group of the Cys20 of the protein.

Finally, we bring experimental evidence that HEV ORF3 binds the human Tsg101 UEV domain in a direct manner. Moreover, we performed an in-depth molecular characterization of this interaction using different biophysical tools. The backbone (including proline) assignments of the human Tsg101 UEV domain in apo-state are obtained and deposited in the Biological Magnetic Resonance Data Bank (BMRB) under accession code 50765. Using the ORF3 and Tsg101 UEV NMR assignments we identified, at a per residue level, the binding site in each protein. In Tsg101 UEV the binding site is similar to the one previously described for the binding of late domains from HIV-1 or Ebola, whereas in ORF3 the binding site encompasses the ^{95}PSAP^{98} late domain sequence. Both the stoichiometry (1:1) and the affinity (K_D ~25 µM) of the interaction between ORF3 and Tsg101 UEV were obtained using ITC. Both NMR and ITC data allowed the identification of a 10-mers ORF3-derived peptide, pepORF3 (^{93}TSPSAPPLPP^{102} from ORF3 protein sequence), that binds Tsg101 UEV with the same characteristics. Further characterization of the interaction was achieved by solving the crystal structure of the complex between human Tsg101 UEV domain and pepORF3 at 1.4 Å resolution. pepORF3 binds in a groove at the surface of Tsg101 UEV which is located next to the ubiquitin binding pocket of the protein, in which also prazole-based compounds have been shown to bind. With the goal of finding small molecules compounds that target the ORF3-Tsg101 interaction and further block the release of the infectious virions from infected cells, we developed a fluorescence polarization (FP) assay, compatible with high-throughput screening, in which we can monitor the displacement of a fluorescent-ORF3-derived

peptide. Another experimental assay, relying on Homogeneous Time Resolved Fluorescence (HTRF), has also been considered in order to overcome potential issue with the autofluorescence of some compounds and has still to be further improved before starting the high-throughput screening of compounds.

Recent data in the literature have shown that prazole-based compounds, such as tenatoprazole and ilaprazole, are bound in the ubiquitin binding site of Tsg101 UEV domain resulting the blockade of the release of infectious HIV from cells in culture without disruption of PTAP binding activity[154] and infectious Herpes Simplex Virus (HSV)-1/2 release from Vero cells in culture[155]. In this study, we wondered if prazole-based compounds could interfere with the Tsg101 UEV domain-ORF3 interaction. Both NMR spectroscopy and fluorescence polarization (FP) experiments have shown that the prazole compounds do not affect the Tsg101 UEV domain-ORF3 interaction. Nevertheless, as the prazole and ORF3 binding sites in Tsg101 UEV are close to each other, this is an interesting feature that could be further exploited to design new anti-HEV antivirals.

The NMR and biochemical tools that have been developed along this work could now be used to study others interactions involving HEV ORF3 protein in order to get a broader view of its multifunctionality.

References

(1) *World Health Organization. Hepatitis E: fact sheet*. https://www.who.int/news-room/fact-sheets/detail/hepatitis-e.

(2) Nimgaonkar, I.; Ding, Q.; Schwartz, R. E.; Ploss, A. Hepatitis E Virus: Advances and Challenges. *Nat Rev Gastroenterol Hepatol* **2018**, *15* (2), 96–110. https://doi.org/10.1038/nrgastro.2017.150.

(3) Balayart, M. S.; Andjaparidze, A. G.; Savinskaya, S. S.; Ketiladze, E. S.; Braginsky, D. M.; Savinov, A. P.; Poleschuk, V. F. Evidence for a Virus in Non-A, Non-B Hepatitis Transmitted via the Fecal-Oral Route. *Intervirology* **1983**, *20* (1), 23–31. https://doi.org/10.1159/000149370.

(4) Reyes, G. R.; Purdy, M. A.; Kim, J.; Luk, K.-C.; Young, L. M.; Fry, K. E.; Bradley, D. W. Isolation of a CDNA from the Virus Responsible for Enterically Transmitted Non-A, Non-B Hepatitis. *Science* **1990**, *247* (4948), 1335–1339. https://doi.org/10.1126/science.2107574.

(5) Webb, G. W.; Kelly, S.; Dalton, H. R. Hepatitis A and Hepatitis E: Clinical and Epidemiological Features, Diagnosis, Treatment, and Prevention. *Clinical Microbiology Newsletter* **2020**, *42* (21), 171–179. https://doi.org/10.1016/j.clinmicnews.2020.10.001.

(6) Tam, A. W.; Smith, M. M.; Guerra, M. E.; Huang, C.-C.; Bradley, D. W.; Fry, K. E.; Reyes, G. R. Hepatitis E Virus (HEV): Molecular Cloning and Sequencing of the Full-Length Viral Genome. *Virology* **1991**, *185* (1), 120–131. https://doi.org/10.1016/0042-6822(91)90760-9.

(7) Donnelly, M. C.; Scobie, L.; Crossan, C. L.; Dalton, H.; Hayes, P. C.; Simpson, K. J. Review Article: Hepatitis E-a Concise Review of Virology, Epidemiology, Clinical Presentation and Therapy. *Aliment Pharmacol Ther* **2017**, *46* (2), 126–141. https://doi.org/10.1111/apt.14109.

(8) Yarbough, P. O. Hepatitis E Virus. *Intervirology* **1999**, *42* (2–3), 179–184. https://doi.org/10.1159/000024978.

(9) Krain, L. J.; Nelson, K. E.; Labrique, A. B. Host Immune Status and Response to Hepatitis E Virus Infection. *Clin Microbiol Rev* **2014**, *27* (1), 139–165. https://doi.org/10.1128/CMR.00062-13.

(10) Wang, K.; Wang, J.; Zhou, C.; Sun, X.; Liu, L.; Xu, X.; Wang, J. Rapid and Direct Detection of Hepatitis E Virus in Raw Pork Livers by Recombinase Polymerase Amplification Assays. *Front. Cell. Infect. Microbiol.* **2022**, *12*, 958990. https://doi.org/10.3389/fcimb.2022.958990.

(11) Hoofnagle, J. H.; Nelson, K. E.; Purcell, R. H. Hepatitis E. *N Engl J Med* **2012**, *367* (13), 1237–1244. https://doi.org/10.1056/NEJMra1204512.

(12) Kamar, N.; Bendall, R.; Legrand-Abravanel, F.; Xia, N.-S.; Ijaz, S.; Izopet, J.; Dalton, H. R. Hepatitis E. *The Lancet* **2012**, *379* (9835), 2477–2488. https://doi.org/10.1016/S0140-6736(11)61849-7.

(13) *Current ICTV Taxonomy Release | ICTV*. https://ictv.global/taxonomy (accessed 2022-10-22).

(14) Raji, Y. E.; Toung, O. P.; Taib, N. M.; Sekawi, Z. B. Hepatitis E Virus: An Emerging Enigmatic and Underestimated Pathogen. *Saudi Journal of Biological Sciences* **2022**, *29* (1), 499–512. https://doi.org/10.1016/j.sjbs.2021.09.003.

(15) Meng, X.-J. Expanding Host Range and Cross-Species Infection of Hepatitis E Virus. *PLoS Pathog* **2016**, *12* (8), e1005695. https://doi.org/10.1371/journal.ppat.1005695.

(16) González, M. M.; Sanabria, L. P.; Castaño-Osorio, J. C. Hepatitis E Virus: A Review of the Current Status and Perspectives. *Hepatitis E Virus* 8.

(17) Kamar, N.; Selves, J.; Mansuy, J.-M.; Ouezzani, L.; Péron, J.-M.; Guitard, J.; Cointault, O.; Esposito, L.; Abravanel, F.; Danjoux, M.; Durand, D.; Vinel, J.-P.; Izopet, J.; Rostaing, L. Hepatitis E Virus and Chronic Hepatitis in Organ-Transplant Recipients. *N Engl J Med* **2008**, *358* (8), 811–817. https://doi.org/10.1056/NEJMoa0706992.

(18) Oechslin, N.; Moradpour, D.; Gouttenoire, J. On the Host Side of the Hepatitis E Virus Life Cycle. *Cells* **2020**, *9* (5), 1294. https://doi.org/10.3390/cells9051294.

(19) Lee, G.-H.; Tan, B.-H.; Chi-Yuan Teo, E.; Lim, S.-G.; Dan, Y.-Y.; Wee, A.; Kim Aw, P. P.; Zhu, Y.; Hibberd, M. L.; Tan, C.-K.; Purdy, M. A.; Teo, C.-G. Chronic Infection With Camelid Hepatitis E Virus in a Liver Transplant Recipient Who Regularly Consumes Camel Meat and Milk. *Gastroenterology* **2016**, *150* (2), 355-357.e3. https://doi.org/10.1053/j.gastro.2015.10.048.

(20) Nicot, F.; Dimeglio, C.; Migueres, M.; Jeanne, N.; Latour, J.; Abravanel, F.; Ranger, N.; Harter, A.; Dubois, M.; Lameiras, S.; Baulande, S.; Chapuy-Regaud, S.; Kamar, N.; Lhomme, S.; Izopet, J. Classification of the Zoonotic Hepatitis E Virus Genotype 3 Into Distinct Subgenotypes. *Front. Microbiol.* **2021**, *11*, 634430. https://doi.org/10.3389/fmicb.2020.634430.

(21) Zachou, K. Acute Non-A, Non-B, Non-C Hepatitis Differences and Similarities between Hepatitis E Virus Infection and Autoimmune Hepatitis, with Phylogenetic Analysis of Hepatitis E Virus in Humans and Wild Boars. *aog* **2022**. https://doi.org/10.20524/aog.2022.0731.

(22) Adlhoch, C.; Avellon, A.; Baylis, S. A.; Ciccaglione, A. R.; Couturier, E.; de Sousa, R.; Epštein, J.; Ethelberg, S.; Faber, M.; Fehér, Á.; Ijaz, S.; Lange, H.; Manďáková, Z.; Mellou, K.; Mozalevskis, A.; Rimhanen-Finne, R.; Rizzi, V.; Said, B.; Sundqvist, L.; Thornton, L.; Tosti, M. E.; van Pelt, W.; Aspinall, E.; Domanovic, D.; Severi, E.; Takkinen, J.; Dalton, H. R. Hepatitis E Virus: Assessment of the Epidemiological Situation in Humans in Europe, 2014/15. *Journal of Clinical Virology* **2016**, *82*, 9–16. https://doi.org/10.1016/j.jcv.2016.06.010.

(23) *ECDC report: 10-fold increase of hepatitis E cases in the EU/EEA between 2005 and 2015*. https://www.ecdc.europa.eu/en/news-events/ecdc-report-10-fold-increase-hepatitis-e-cases-eueea-between-2005-and-2015.

(24) Kamar, N.; Izopet, J.; Pavio, N.; Aggarwal, R.; Labrique, A.; Wedemeyer, H.; Dalton, H. R. Hepatitis E Virus Infection. *Nat Rev Dis Primers* **2017**, *3* (1), 17086. https://doi.org/10.1038/nrdp.2017.86.

(25) Rawla, P.; Raj, J. P.; Kannemkuzhiyil, A. J.; Aluru, J. S.; Thandra, K. C.; Gajendran, M. A Systematic Review of the Extra-Hepatic Manifestations of Hepatitis E Virus Infection. *Medical Sciences* **2020**, *8* (1), 9. https://doi.org/10.3390/medsci8010009.

(26) Cancela, F.; Noceti, O.; Arbiza, J.; Mirazo, S. Structural Aspects of Hepatitis E Virus. *Arch Virol* **2022**. https://doi.org/10.1007/s00705-022-05575-8.

(27) Webb, G. W.; Dalton, H. R. Hepatitis E: An Underestimated Emerging Threat. *Therapeutic Advances in Infection* **2019**, *6*, 204993611983716. https://doi.org/10.1177/2049936119837162.

(28) Nagashima, S.; Takahashi, M.; Kobayashi, T.; Tanggis; Nishizawa, T.; Nishiyama, T.; Primadharsini, P. P.; Okamoto, H. Characterization of the Quasi-Enveloped Hepatitis E Virus

Particles Released by the Cellular Exosomal Pathway. *J Virol* **2017**, *91* (22), e00822-17. https://doi.org/10.1128/JVI.00822-17.

(29) Wang, B. Structural and Molecular Biology of Hepatitis E Virus. *Computational and Structural Biotechnology Journal* **2021**, 10.

(30) Nan, Y.; Zhang, Y.-J. Molecular Biology and Infection of Hepatitis E Virus. *Front. Microbiol.* **2016**, *7*. https://doi.org/10.3389/fmicb.2016.01419.

(31) Sooryanarain, H.; Heffron, C. L.; Meng, X.-J. The U-Rich Untranslated Region of the Hepatitis E Virus Induces Differential Type I and Type III Interferon Responses in a Host Cell-Dependent Manner. *mBio* **2020**, *11* (1), e03103-19. https://doi.org/10.1128/mBio.03103-19.

(32) Graff, J.; Torian, U.; Nguyen, H.; Emerson, S. U. A Bicistronic Subgenomic MRNA Encodes Both the ORF2 and ORF3 Proteins of Hepatitis E Virus. *J Virol* **2006**, *80* (12), 5919–5926. https://doi.org/10.1128/JVI.00046-06.

(33) Cao, D.; Meng, X.-J. Molecular Biology and Replication of Hepatitis E Virus. *Emerging Microbes & Infections* **2012**, *1* (1), 1–10. https://doi.org/10.1038/emi.2012.7.

(34) Metzger, K.; Bentaleb, C.; Hervouet, K.; Alexandre, V.; Montpellier, C.; Saliou, J.-M.; Ferrié, M.; Camuzet, C.; Rouillé, Y.; Lecoeur, C.; Dubuisson, J.; Cocquerel, L.; Aliouat-Denis, C.-M. Processing and Subcellular Localization of the Hepatitis E Virus Replicase: Identification of Candidate Viral Factories. *Front. Microbiol.* **2022**, *13*, 828636. https://doi.org/10.3389/fmicb.2022.828636.

(35) Ansari, I. H.; Nanda, S. K.; Durgapal, H.; Agrawal, S.; Mohanty, S. K.; Gupta, D.; Jameel, S.; Panda, S. K. Cloning, Sequencing, and Expression of the Hepatitis E Virus (HEV) Nonstructural Open Reading Frame 1 (ORF1). *J Med Virol* **2000**, *60* (3), 275–283.

(36) Ropp, S. L.; Tam, A. W.; Beames, B.; Purdy, M.; Frey, T. K. Expression of the Hepatitis E Virus ORF1. *Arch. Virol.* **2000**, *145* (7), 1321–1337. https://doi.org/10.1007/s007050070093.

(37) Chen, J. P.; Strauss, J. H.; Strauss, E. G.; Frey, T. K. Characterization of the Rubella Virus Nonstructural Protease Domain and Its Cleavage Site. *J Virol* **1996**, *70* (7), 4707–4713. https://doi.org/10.1128/JVI.70.7.4707-4713.1996.

(38) Golubtsov, A.; Kääriäinen, L.; Caldentey, J. Characterization of the Cysteine Protease Domain of Semliki Forest Virus Replicase Protein NsP2 by in Vitro Mutagenesis. *FEBS Lett* **2006**, *580* (5), 1502–1508. https://doi.org/10.1016/j.febslet.2006.01.071.

(39) Proudfoot, A.; Hyrina, A.; Holdorf, M.; Frank, A. O.; Bussiere, D. First Crystal Structure of a Nonstructural Hepatitis E Viral Protein Identifies a Putative Novel Zinc-Binding Protein. *J Virol* **2019**, *93* (13), e00170-19. https://doi.org/10.1128/JVI.00170-19.

(40) Montpellier, C.; Wychowski, C.; Sayed, I. M.; Meunier, J.-C.; Saliou, J.-M.; Ankavay, M.; Bull, A.; Pillez, A.; Abravanel, F.; Helle, F.; Brochot, E.; Drobecq, H.; Farhat, R.; Aliouat-Denis, C.-M.; Haddad, J. G.; Izopet, J.; Meuleman, P.; Goffard, A.; Dubuisson, J.; Cocquerel, L. Hepatitis E Virus Lifecycle and Identification of 3 Forms of the ORF2 Capsid Protein. *Gastroenterology* **2018**, *154* (1), 211-223.e8. https://doi.org/10.1053/j.gastro.2017.09.020.

(41) Yamashita, T.; Mori, Y.; Miyazaki, N.; Cheng, R. H.; Yoshimura, M.; Unno, H.; Shima, R.; Moriishi, K.; Tsukihara, T.; Li, T. C.; Takeda, N.; Miyamura, T.; Matsuura, Y. Biological and Immunological Characteristics of Hepatitis E Virus-like Particles Based on the Crystal Structure. *Proceedings of the National Academy of Sciences* **2009**, *106* (31), 12986–12991. https://doi.org/10.1073/pnas.0903699106.

(42) Holla, R.; Ahmad, I.; Ahmad, Z.; Jameel, S. Molecular Virology of Hepatitis E Virus. *Semin Liver Dis* **2013**, *33* (01), 003–014. https://doi.org/10.1055/s-0033-1338110.

(43) Yamada, K.; Takahashi, M.; Hoshino, Y.; Takahashi, H.; Ichiyama, K.; Nagashima, S.; Tanaka, T.; Okamoto, H. ORF3 Protein of Hepatitis E Virus Is Essential for Virion Release from Infected Cells. *Journal of General Virology* **2009**, *90* (8), 1880–1891. https://doi.org/10.1099/vir.0.010561-0.

(44) Chandra, V.; Kar-Roy, A.; Kumari, S.; Mayor, S.; Jameel, S. The Hepatitis E Virus ORF3 Protein Modulates Epidermal Growth Factor Receptor Trafficking, STAT3 Translocation, and the Acute-Phase Response. *J Virol* **2008**, *82* (14), 7100–7110. https://doi.org/10.1128/JVI.00403-08.

(45) Nagashima, S.; Takahashi, M.; Jirintai, S.; Tanggis; Kobayashi, T.; Nishizawa, T.; Okamoto, H. The Membrane on the Surface of Hepatitis E Virus Particles Is Derived from the Intracellular Membrane and Contains Trans-Golgi Network Protein 2. *Arch Virol* **2014**, *159* (5), 979–991. https://doi.org/10.1007/s00705-013-1912-3.

(46) Nair, V. P.; Anang, S.; Subramani, C.; Madhvi, A.; Bakshi, K.; Srivastava, A.; Shalimar; Nayak, B.; Ct, R. K.; Surjit, M. Endoplasmic Reticulum Stress Induced Synthesis of a Novel Viral Factor Mediates Efficient Replication of Genotype-1 Hepatitis E Virus. *PLoS Pathog* **2016**, *12* (4), e1005521. https://doi.org/10.1371/journal.ppat.1005521.

(47) Yin, X.; Ambardekar, C.; Lu, Y.; Feng, Z. Distinct Entry Mechanisms for Nonenveloped and Quasi-Enveloped Hepatitis E Viruses. *J Virol* **2016**, *90* (8), 4232–4242. https://doi.org/10.1128/JVI.02804-15.

(48) Montero, H.; Pérez-Gil, G.; Sampieri, C. L. Eukaryotic Initiation Factor 4A (EIF4A) during Viral Infections. *Virus Genes* **2019**, *55* (3), 267–273. https://doi.org/10.1007/s11262-019-01641-7.

(49) Perttilä, J.; Spuul, P.; Ahola, T. Early Secretory Pathway Localization and Lack of Processing for Hepatitis E Virus Replication Protein PORF1. *Journal of General Virology* **2013**, *94* (4), 807–816. https://doi.org/10.1099/vir.0.049577-0.

(50) Rehman, S.; Kapur, N.; Durgapal, H.; Panda, S. K. Subcellular Localization of Hepatitis E Virus (HEV) Replicase. *Virology* **2008**, *370* (1), 77–92. https://doi.org/10.1016/j.virol.2007.07.036.

(51) Martin-Serrano, J.; Zang, T.; Bieniasz, P. D. HIV-1 and Ebola Virus Encode Small Peptide Motifs That Recruit Tsg101 to Sites of Particle Assembly to Facilitate Egress. *Nat Med* **2001**, *7* (12), 1313–1319. https://doi.org/10.1038/nm1201-1313.

(52) Fraga, M.; Gouttenoire, J.; Sahli, R.; Chtioui, H.; Marcu, C.; Pascual, M.; Moradpour, D.; Vionnet, J. Sofosbuvir Add-on to Ribavirin for Chronic Hepatitis E in a Cirrhotic Liver Transplant Recipient: A Case Report. *BMC Gastroenterol* **2019**, *19* (1), 76. https://doi.org/10.1186/s12876-019-0995-z.

(53) Ahmad, T.; Haroon, H.; Ahmad, K.; Shah, S. M.; Shah, M. W.; Shah, S. M.; Hussain, A.; Jalal, S.; Ahmad, W.; Khan, M.; Khan, M.; Harapan, H.; Dhama, K.; Baig, M.; Hui, J. Hepatitis E Vaccines: A Mini Review. *Biomed. Res. Ther.* **2021**, *8* (9), 4514–4524. https://doi.org/10.15419/bmrat.v8i9.690.

(54) Riedmann, E. M. Human Vaccines & Immunotherapeutics: News. *Human Vaccines & Immunotherapeutics* **2012**, *8* (12), 1741–1744. https://doi.org/10.4161/hv.23373.

(55) Li, S. W.; Zhang, J.; Li, Y. M.; Ou, S. H.; Huang, G. Y.; He, Z. Q.; Ge, S. X.; Xian, Y. L.; Pang, S. Q.; Ng, M. H.; Xia, N. S. A Bacterially Expressed Particulate Hepatitis E Vaccine: Antigenicity,

Immunogenicity and Protectivity on Primates. *Vaccine* **2005**, *23* (22), 2893–2901. https://doi.org/10.1016/j.vaccine.2004.11.064.

(56) Zhu, F.-C.; Zhang, J.; Zhang, X.-F.; Zhou, C.; Wang, Z.-Z.; Huang, S.-J.; Wang, H.; Yang, C.-L.; Jiang, H.-M.; Cai, J.-P.; Wang, Y.-J.; Ai, X.; Hu, Y.-M.; Tang, Q.; Yao, X.; Yan, Q.; Xian, Y.-L.; Wu, T.; Li, Y.-M.; Miao, J.; Ng, M.-H.; Shih, J. W.-K.; Xia, N.-S. Efficacy and Safety of a Recombinant Hepatitis E Vaccine in Healthy Adults: A Large-Scale, Randomised, Double-Blind Placebo-Controlled, Phase 3 Trial. *The Lancet* **2010**, *376* (9744), 895–902. https://doi.org/10.1016/S0140-6736(10)61030-6.

(57) National Institute of Allergy and Infectious Diseases (NIAID). *A Phase 1, Double-Blinded, Placebo Controlled, Clinical Trial to Evaluate the Safety, Reactogenicity, and Immunogenicity of HEV-239 (Hecolin(R)) in a Healthy US Adult Population*; Clinical trial registration NCT03827395; clinicaltrials.gov, 2021. https://clinicaltrials.gov/ct2/show/NCT03827395 (accessed 2022-10-25).

(58) Norwegian Institute of Public Health. *An Effectiveness Trial (Phase IV) to Evaluate Protection of Pregnant Women by Hepatitis E Virus (HEV) Vaccine in Bangladesh and Risk Factors for Severe HEV Infection.*; Clinical trial registration NCT02759991; clinicaltrials.gov, 2020. https://clinicaltrials.gov/ct2/show/NCT02759991 (accessed 2022-10-25).

(59) Zaman, K.; Dudman, S.; Stene-Johansen, K.; Qadri, F.; Yunus, M.; Sandbu, S.; Gurley, E. S.; Overbo, J.; Julin, C. H.; Dembinski, J. L.; Nahar, Q.; Rahman, A.; Bhuiyan, T. R.; Rahman, M.; Haque, W.; Khan, J.; Aziz, A.; Khanam, M.; Streatfield, P. K.; Clemens, J. D. HEV Study Protocol : Design of a Cluster-Randomised, Blinded Trial to Assess the Safety, Immunogenicity and Effectiveness of the Hepatitis E Vaccine HEV 239 (Hecolin) in Women of Childbearing Age in Rural Bangladesh. *BMJ Open* **2020**, *10* (1), e033702. https://doi.org/10.1136/bmjopen-2019-033702.

(60) Shrestha, M. P.; Scott, R. M.; Joshi, D. M.; Mammen, M. P.; Thapa, G. B.; Thapa, N.; Myint, K. S. A.; Fourneau, M.; Kuschner, R. A.; Shrestha, S. K.; David, M. P.; Seriwatana, J.; Vaughn, D. W.; Safary, A.; Endy, T. P.; Innis, B. L. Safety and Efficacy of a Recombinant Hepatitis E Vaccine. *N Engl J Med* **2007**, *356* (9), 895–903. https://doi.org/10.1056/NEJMoa061847.

(61) U.S. Army Medical Research and Development Command. *A Phase II, Prospective, Randomized, Double-Blind, Placebo Controlled, Field Efficacy Trial of a Candidate Hepatitis E Vaccine in Nepal.*; Clinical trial registration NCT00287469; clinicaltrials.gov, 2019. https://clinicaltrials.gov/ct2/show/NCT00287469 (accessed 2022-10-25).

(62) Cao, Y.-F.; Tao, H.; Hu, Y.-M.; Shi, C.-B.; Wu, X.; Liang, Q.; Chi, C.-P.; Li, L.; Liang, Z.-L.; Meng, J.-H.; Zhu, F.-C.; Liu, Z.-H.; Wang, X.-P. A Phase 1 Randomized Open-Label Clinical Study to Evaluate the Safety and Tolerability of a Novel Recombinant Hepatitis E Vaccine. *Vaccine* **2017**, *35* (37), 5073–5080. https://doi.org/10.1016/j.vaccine.2017.05.072.

(63) Ahmad, T.; Nasir, S.; Musa, T. H.; AlRyalat, S. A. S.; Khan, M.; Hui, J. Epidemiology, Diagnosis, Vaccines, and Bibliometric Analysis of the 100 Top-Cited Studies on Hepatitis E Virus. *Human Vaccines & Immunotherapeutics* **2021**, *17* (3), 857–871. https://doi.org/10.1080/21645515.2020.1795458.

(64) Zafrullah, M.; Ozdener, M. H.; Panda, S. K.; Jameel, S. The ORF3 Protein of Hepatitis E Virus Is a Phosphoprotein That Associates with the Cytoskeleton. *J. VIROL.* **1997**, *71*, 9.

(65) Tyagi, S.; Korkaya, H.; Zafrullah, M.; Jameel, S.; Lal, S. K. The Phosphorylated Form of the ORF3 Protein of Hepatitis E Virus Interacts with Its Non-Glycosylated Form of the Major

Capsid Protein, ORF2. *Journal of Biological Chemistry* **2002**, *277* (25), 22759–22767. https://doi.org/10.1074/jbc.M200185200.

(66) Kannan, H.; Fan, S.; Patel, D.; Bossis, I.; Zhang, Y.-J. The Hepatitis E Virus Open Reading Frame 3 Product Interacts with Microtubules and Interferes with Their Dynamics. *JVI* **2009**, *83* (13), 6375–6382. https://doi.org/10.1128/JVI.02571-08.

(67) Moin, S. M.; Panteva, M.; Jameel, S. The Hepatitis E Virus Orf3 Protein Protects Cells from Mitochondrial Depolarization and Death. *Journal of Biological Chemistry* **2007**, *282* (29), 21124–21133. https://doi.org/10.1074/jbc.M701696200.

(68) Moin, S. M.; Chandra, V.; Arya, R.; Jameel, S. The Hepatitis E Virus ORF3 Protein Stabilizes HIF-1α and Enhances HIF-1-Mediated Transcriptional Activity through P300/CBP. *Cellular Microbiology* **2009**, *11* (9), 1409–1421. https://doi.org/10.1111/j.1462-5822.2009.01340.x.

(69) He, M.; Wang, M.; Huang, Y.; Peng, W.; Zheng, Z.; Xia, N.; Xu, J.; Tian, D. The ORF3 Protein of Genotype 1 Hepatitis E Virus Suppresses TLR3-Induced NF-KB Signaling via TRADD and RIP1. *Sci Rep* **2016**, *6* (1), 27597. https://doi.org/10.1038/srep27597.

(70) Nagashima, S.; Takahashi, M.; Jirintai; Tanaka, T.; Yamada, K.; Nishizawa, T.; Okamoto, H. A PSAP Motif in the ORF3 Protein of Hepatitis E Virus Is Necessary for Virion Release from Infected Cells. *Journal of General Virology* **2011**, *92* (2), 269–278. https://doi.org/10.1099/vir.0.025791-0.

(71) Kenney, S. P.; Wentworth, J. L.; Heffron, C. L.; Meng, X.-J. Replacement of the Hepatitis E Virus ORF3 Protein PxxP Motif with Heterologous Late Domain Motifs Affects Virus Release via Interaction with TSG101. *Virology* **2015**, *486*, 198–208. https://doi.org/10.1016/j.virol.2015.09.012.

(72) Zarrinpar, A.; Bhattacharyya, R. P.; Lim, W. A. The Structure and Function of Proline Recognition Domains. *Sci. STKE* **2003**, *2003* (179). https://doi.org/10.1126/stke.2003.179.re8.

(73) Chandra, V.; Kalia, M.; Hajela, K.; Jameel, S. The ORF3 Protein of Hepatitis E Virus Delays Degradation of Activated Growth Factor Receptors by Interacting with CIN85 and Blocking Formation of the Cbl-CIN85 Complex . *J. VIROL.* **2010**, *84*, 11.

(74) Nagashima, S.; Takahashi, M.; Jirintai, S.; Tanaka, T.; Nishizawa, T.; Yasuda, J.; Okamoto, H. Tumour Susceptibility Gene 101 and the Vacuolar Protein Sorting Pathway Are Required for the Release of Hepatitis E Virions. *Journal of General Virology* 11.

(75) Garrus, J. E.; von Schwedler, U. K.; Pornillos, O. W.; Morham, S. G.; Zavitz, K. H.; Wang, H. E.; Wettstein, D. A.; Stray, K. M.; Côté, M.; Rich, R. L.; Myszka, D. G.; Sundquist, W. I. Tsg101 and the Vacuolar Protein Sorting Pathway Are Essential for HIV-1 Budding. *Cell* **2001**, *107* (1), 55–65. https://doi.org/10.1016/S0092-8674(01)00506-2.

(76) Qi, Y.; Zhang, F.; Zhang, L.; Harrison, T. J.; Huang, W.; Zhao, C.; Kong, W.; Jiang, C.; Wang, Y. Hepatitis E Virus Produced from Cell Culture Has a Lipid Envelope. *PLoS ONE* **2015**, *10* (7), e0132503. https://doi.org/10.1371/journal.pone.0132503.

(77) Nan, Y.; Ma, Z.; Wang, R.; Yu, Y.; Kannan, H.; Fredericksen, B.; Zhang, Y.-J. Enhancement of Interferon Induction by ORF3 Product of Hepatitis E Virus. *J Virol* **2014**, *88* (15), 8696–8705. https://doi.org/10.1128/JVI.01228-14.

(78) Nan, Y.; Yu, Y.; Ma, Z.; Khattar, S. K.; Fredericksen, B.; Zhang, Y.-J. Hepatitis E Virus Inhibits Type I Interferon Induction by ORF1 Products. *J Virol* **2014**, *88* (20), 11924–11932. https://doi.org/10.1128/JVI.01935-14.

(79) Nan, Y.; Ma, Z.; Kannan, H.; Stein, D. A.; Iversen, P. I.; Meng, X.-J.; Zhang, Y.-J. Inhibition of Hepatitis E Virus Replication by Peptide-Conjugated Morpholino Oligomers. *Antiviral Research* **2015**, *120*, 134–139. https://doi.org/10.1016/j.antiviral.2015.06.006.

(80) Osterman, A.; Stellberger, T.; Gebhardt, A.; Kurz, M.; Friedel, C. C.; Uetz, P.; Nitschko, H.; Baiker, A.; Vizoso-Pinto, M. G. The Hepatitis E Virus Intraviral Interactome. *Sci Rep* **2015**, *5* (1), 13872. https://doi.org/10.1038/srep13872.

(81) Ding, Q.; Heller, B.; Capuccino, J. M. V.; Song, B.; Nimgaonkar, I.; Hrebikova, G.; Contreras, J. E.; Ploss, A. Hepatitis E Virus ORF3 Is a Functional Ion Channel Required for Release of Infectious Particles. *Proc Natl Acad Sci USA* **2017**, *114* (5), 1147–1152. https://doi.org/10.1073/pnas.1614955114.

(82) Gouttenoire, J.; Pollán, A.; Abrami, L.; Oechslin, N.; Mauron, J.; Matter, M.; Oppliger, J.; Szkolnicka, D.; Dao Thi, V. L.; van der Goot, F. G.; Moradpour, D. Palmitoylation Mediates Membrane Association of Hepatitis E Virus ORF3 Protein and Is Required for Infectious Particle Secretion. *PLoS Pathog* **2018**, *14* (12), e1007471. https://doi.org/10.1371/journal.ppat.1007471.

(83) Nieva, J. L.; Madan, V.; Carrasco, L. Viroporins: Structure and Biological Functions. *Nat Rev Microbiol* **2012**, *10* (8), 563–574. https://doi.org/10.1038/nrmicro2820.

(84) Sze, C.; Tan, Y.-J. Viral Membrane Channels: Role and Function in the Virus Life Cycle. *Viruses* **2015**, *7* (6), 3261–3284. https://doi.org/10.3390/v7062771.

(85) Veit, M. Palmitoylation of Virus Proteins. *Biology of the Cell* **2012**, *104* (9), 493–515. https://doi.org/10.1111/boc.201200006.

(86) Rothman, J. H.; Stevens, T. H. Protein Sorting in Yeast: Mutants Defective in Vacuole Biogenesis Mislocalize Vacuolar Proteins into the Late Secretory Pathway. *Cell* **1986**, *47* (6), 1041–1051. https://doi.org/10.1016/0092-8674(86)90819-6.

(87) Henne, W. M.; Buchkovich, N. J.; Emr, S. D. The ESCRT Pathway. *Developmental Cell* **2011**, *21* (1), 77–91. https://doi.org/10.1016/j.devcel.2011.05.015.

(88) Christ, L.; Raiborg, C.; Wenzel, E. M.; Campsteijn, C.; Stenmark, H. Cellular Functions and Molecular Mechanisms of the ESCRT Membrane-Scission Machinery. *Trends in Biochemical Sciences* **2017**, *42* (1), 42–56. https://doi.org/10.1016/j.tibs.2016.08.016.

(89) Williams, R. L.; Urbé, S. The Emerging Shape of the ESCRT Machinery. *Nat Rev Mol Cell Biol* **2007**, *8* (5), 355–368. https://doi.org/10.1038/nrm2162.

(90) Flower, T. G.; Takahashi, Y.; Hudait, A.; Rose, K.; Tjahjono, N.; Pak, A. J.; Yokom, A. L.; Liang, X.; Wang, H.-G.; Bouamr, F.; Voth, G. A.; Hurley, J. H. A Helical Assembly of Human ESCRT-I Scaffolds Reverse-Topology Membrane Scission. *Nat Struct Mol Biol* **2020**, *27* (6), 570–580. https://doi.org/10.1038/s41594-020-0426-4.

(91) Ferraiuolo, R.-M.; Manthey, K. C.; Stanton, M. J.; Triplett, A. A.; Wagner, K.-U. The Multifaceted Roles of the Tumor Susceptibility Gene 101 (TSG101) in Normal Development and Disease. *Cancers* **2020**, *12* (2), 450. https://doi.org/10.3390/cancers12020450.

(92) Demirov, D. G.; Ono, A.; Orenstein, J. M.; Freed, E. O. Overexpression of the N-Terminal Domain of TSG101 Inhibits HIV-1 Budding by Blocking Late Domain Function. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (2), 955–960. https://doi.org/10.1073/pnas.032511899.

(93) VerPlank, L.; Bouamr, F.; LaGrassa, T. J.; Agresta, B.; Kikonyogo, A.; Leis, J.; Carter, C. A. Tsg101, a Homologue of Ubiquitin-Conjugating (E2) Enzymes, Binds the L Domain in HIV

Type 1 Pr55 ^{Gag}. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98* (14), 7724–7729. https://doi.org/10.1073/pnas.131059198.

(94) Schmidt, O.; Teis, D. The ESCRT Machinery. *Current Biology* **2012**, *22* (4), R116–R120. https://doi.org/10.1016/j.cub.2012.01.028.

(95) Raiborg, C.; Stenmark, H. The ESCRT Machinery in Endosomal Sorting of Ubiquitylated Membrane Proteins. *Nature* **2009**, *458* (7237), 445–452. https://doi.org/10.1038/nature07961.

(96) Ren, X.; Kloer, D. P.; Kim, Y. C.; Ghirlando, R.; Saidi, L. F.; Hummer, G.; Hurley, J. H. Hybrid Structural Model of the Complete Human ESCRT-0 Complex. *Structure* **2009**, *17* (3), 406–416. https://doi.org/10.1016/j.str.2009.01.012.

(97) Hierro, A.; Sun, J.; Rusnak, A. S.; Kim, J.; Prag, G.; Emr, S. D.; Hurley, J. H. Structure of the ESCRT-II Endosomal Trafficking Complex. *Nature* **2004**, *431* (7005), 221–225. https://doi.org/10.1038/nature02914.

(98) Teo, H.; Perisic, O.; González, B.; Williams, R. L. ESCRT-II, an Endosome-Associated Complex Required for Protein Sorting. *Developmental Cell* **2004**, *7* (4), 559–569. https://doi.org/10.1016/j.devcel.2004.09.003.

(99) Teo, H.; Gill, D. J.; Sun, J.; Perisic, O.; Veprintsev, D. B.; Vallis, Y.; Emr, S. D.; Williams, R. L. ESCRT-I Core and ESCRT-II GLUE Domain Structures Reveal Role for GLUE in Linking to ESCRT-I and Membranes. *Cell* **2006**, *125* (1), 99–111. https://doi.org/10.1016/j.cell.2006.01.047.

(100) Gill, D. J.; Teo, H.; Sun, J.; Perisic, O.; Veprintsev, D. B.; Emr, S. D.; Williams, R. L. Structural Insight into the ESCRT-I/-II Link and Its Role in MVB Trafficking. *EMBO J* **2007**, *26* (2), 600–612. https://doi.org/10.1038/sj.emboj.7601501.

(101) Im, Y. J.; Hurley, J. H. Integrated Structural Model and Membrane Targeting Mechanism of the Human ESCRT-II Complex. *Developmental Cell* **2008**, *14* (6), 902–913. https://doi.org/10.1016/j.devcel.2008.04.004.

(102) Vietri, M.; Radulovic, M.; Stenmark, H. The Many Functions of ESCRTs. *Nat Rev Mol Cell Biol* **2020**, *21* (1), 25–42. https://doi.org/10.1038/s41580-019-0177-4.

(103) Muzioł, T.; Pineda-Molina, E.; Ravelli, R. B.; Zamborlini, A.; Usami, Y.; Göttlinger, H.; Weissenhorn, W. Structural Basis for Budding by the ESCRT-III Factor CHMP3. *Developmental Cell* **2006**, *10* (6), 821–830. https://doi.org/10.1016/j.devcel.2006.03.013.

(104) Bajorek, M.; Schubert, H. L.; McCullough, J.; Langelier, C.; Eckert, D. M.; Stubblefield, W.-M. B.; Uter, N. T.; Myszka, D. G.; Hill, C. P.; Sundquist, W. I. Structural Basis for ESCRT-III Protein Autoinhibition. *Nat Struct Mol Biol* **2009**, *16* (7), 754–762. https://doi.org/10.1038/nsmb.1621.

(105) Zamborlini, A.; Usami, Y.; Radoshitzky, S. R.; Popova, E.; Palu, G.; Göttlinger, H. Release of Autoinhibition Converts ESCRT-III Components into Potent Inhibitors of HIV-1 Budding. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (50), 19140–19145. https://doi.org/10.1073/pnas.0603788103.

(106) Shim, S.; Kimpler, L. A.; Hanson, P. I. Structure/Function Analysis of Four Core ESCRT-III Proteins Reveals Common Regulatory Role for Extreme C-Terminal Domain. *Traffic* **2007**, *8* (8), 1068–1079. https://doi.org/10.1111/j.1600-0854.2007.00584.x.

(107) Teis, D.; Saksena, S.; Emr, S. D. Ordered Assembly of the ESCRT-III Complex on Endosomes Is Required to Sequester Cargo during MVB Formation. *Developmental Cell* **2008**, *15* (4), 578–589. https://doi.org/10.1016/j.devcel.2008.08.013.

(108) Saksena, S.; Wahlman, J.; Teis, D.; Johnson, A. E.; Emr, S. D. Functional Reconstitution of ESCRT-III Assembly and Disassembly. *Cell* **2009**, *136* (1), 97–109. https://doi.org/10.1016/j.cell.2008.11.013.

(109) Shiflett, S. L.; Ward, D. M.; Huynh, D.; Vaughn, M. B.; Simmons, J. C.; Kaplan, J. Characterization of Vta1p, a Class E Vps Protein in Saccharomyces Cerevisiae. *Journal of Biological Chemistry* **2004**, *279* (12), 10982–10990. https://doi.org/10.1074/jbc.M312669200.

(110) Azmi, I. F.; Davies, B. A.; Xiao, J.; Babst, M.; Xu, Z.; Katzmann, D. J. ESCRT-III Family Members Stimulate Vps4 ATPase Activity Directly or via Vta1. *Developmental Cell* **2008**, *14* (1), 50–61. https://doi.org/10.1016/j.devcel.2007.10.021.

(111) Babst, M. Endosomal Transport Function in Yeast Requires a Novel AAA-Type ATPase, Vps4p. *The EMBO Journal* **1997**, *16* (8), 1820–1831. https://doi.org/10.1093/emboj/16.8.1820.

(112) Babst, M. The Vps4p AAA ATPase Regulates Membrane Association of a Vps Protein Complex Required for Normal Endosome Function. *The EMBO Journal* **1998**, *17* (11), 2982–2993. https://doi.org/10.1093/emboj/17.11.2982.

(113) Scott, A.; Gaspar, J.; Stuchell-Brereton, M. D.; Alam, S. L.; Skalicky, J. J.; Sundquist, W. I. Structure and ESCRT-III Protein Interactions of the MIT Domain of Human VPS4A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (39), 13813–13818. https://doi.org/10.1073/pnas.0502165102.

(114) Katzmann, D. J.; Babst, M.; Emr, S. D. Ubiquitin-Dependent Sorting into the Multivesicular Body Pathway Requires the Function of a Conserved Endosomal Protein Sorting Complex, ESCRT-I. *Cell* **2001**, *106* (2), 145–155. https://doi.org/10.1016/S0092-8674(01)00434-2.

(115) Chu, T.; Sun, J.; Saksena, S.; Emr, S. D. New Component of ESCRT-I Regulates Endosomal Sorting Complex Assembly. *Journal of Cell Biology* **2006**, *175* (5), 815–823. https://doi.org/10.1083/jcb.200608053.

(116) Morita, E.; Sandrin, V.; Alam, S. L.; Eckert, D. M.; Gygi, S. P.; Sundquist, W. I. Identification of Human MVB12 Proteins as ESCRT-I Subunits That Function in HIV Budding. *Cell Host & Microbe* **2007**, *2* (1), 41–53. https://doi.org/10.1016/j.chom.2007.06.003.

(117) Kostelansky, M. S.; Schluter, C.; Tam, Y. Y. C.; Lee, S.; Ghirlando, R.; Beach, B.; Conibear, E.; Hurley, J. H. Molecular Architecture and Functional Model of the Complete Yeast ESCRT-I Heterotetramer. *Cell* **2007**, *129* (3), 485–498. https://doi.org/10.1016/j.cell.2007.03.016.

(118) Pornillos, O.; Higginson, D. S.; Stray, K. M.; Fisher, R. D.; Garrus, J. E.; Payne, M.; He, G.-P.; Wang, H. E.; Morham, S. G.; Sundquist, W. I. HIV Gag Mimics the Tsg101-Recruiting Activity of the Human Hrs Protein. *Journal of Cell Biology* **2003**, *162* (3), 425–434. https://doi.org/10.1083/jcb.200302138.

(119) Bouamr, F.; Melillo, J. A.; Wang, M. Q.; Nagashima, K.; de Los Santos, M.; Rein, A.; Goff, S. P. PPPYEPTAP Motif Is the Late Domain of Human T-Cell Leukemia Virus Type 1 Gag and Mediates Its Functional Interaction with Cellular Proteins Nedd4 and Tsg101. *J Virol* **2003**, *77* (22), 11882–11895. https://doi.org/10.1128/JVI.77.22.11882-11895.2003.

(120) Théry, C.; Boussac, M.; Véron, P.; Ricciardi-Castagnoli, P.; Raposo, G.; Garin, J.; Amigorena, S. Proteomic Analysis of Dendritic Cell-Derived Exosomes: A Secreted Subcellular Compartment Distinct from Apoptotic Vesicles. *J Immunol* **2001**, *166* (12), 7309–7318. https://doi.org/10.4049/jimmunol.166.12.7309.

(121) Théry, C.; Witwer, K. W.; Aikawa, E.; Alcaraz, M. J.; Anderson, J. D.; Andriantsitohaina, R.; Antoniou, A.; Arab, T.; Archer, F.; Atkin-Smith, G. K.; Ayre, D. C.; Bach, J.-M.; Bachurski, D.; Baharvand, H.; Balaj, L.; Baldacchino, S.; Bauer, N. N.; Baxter, A. A.; Bebawy, M.; Beckham, C.; Bedina Zavec, A.; Benmoussa, A.; Berardi, A. C.; Bergese, P.; Bielska, E.; Blenkiron, C.; Bobis-Wozowicz, S.; Boilard, E.; Boireau, W.; Bongiovanni, A.; Borràs, F. E.; Bosch, S.; Boulanger, C. M.; Breakefield, X.; Breglio, A. M.; Brennan, M. Á.; Brigstock, D. R.; Brisson, A.; Broekman, M. L.; Bromberg, J. F.; Bryl-Górecka, P.; Buch, S.; Buck, A. H.; Burger, D.; Busatto, S.; Buschmann, D.; Bussolati, B.; Buzás, E. I.; Byrd, J. B.; Camussi, G.; Carter, D. R.; Caruso, S.; Chamley, L. W.; Chang, Y.-T.; Chen, C.; Chen, S.; Cheng, L.; Chin, A. R.; Clayton, A.; Clerici, S. P.; Cocks, A.; Cocucci, E.; Coffey, R. J.; Cordeiro-da-Silva, A.; Couch, Y.; Coumans, F. A.; Coyle, B.; Crescitelli, R.; Criado, M. F.; D'Souza-Schorey, C.; Das, S.; Datta Chaudhuri, A.; de Candia, P.; De Santana, E. F.; De Wever, O.; del Portillo, H. A.; Demaret, T.; Deville, S.; Devitt, A.; Dhondt, B.; Di Vizio, D.; Dieterich, L. C.; Dolo, V.; Dominguez Rubio, A. P.; Dominici, M.; Dourado, M. R.; Driedonks, T. A.; Duarte, F. V.; Duncan, H. M.; Eichenberger, R. M.; Ekström, K.; EL Andaloussi, S.; Elie-Caille, C.; Erdbrügger, U.; Falcón-Pérez, J. M.; Fatima, F.; Fish, J. E.; Flores-Bellver, M.; Försönits, A.; Frelet-Barrand, A.; Fricke, F.; Fuhrmann, G.; Gabrielsson, S.; Gámez-Valero, A.; Gardiner, C.; Gärtner, K.; Gaudin, R.; Gho, Y. S.; Giebel, B.; Gilbert, C.; Gimona, M.; Giusti, I.; Goberdhan, D. C.; Görgens, A.; Gorski, S. M.; Greening, D. W.; Gross, J. C.; Gualerzi, A.; Gupta, G. N.; Gustafson, D.; Handberg, A.; Haraszti, R. A.; Harrison, P.; Hegyesi, H.; Hendrix, A.; Hill, A. F.; Hochberg, F. H.; Hoffmann, K. F.; Holder, B.; Holthofer, H.; Hosseinkhani, B.; Hu, G.; Huang, Y.; Huber, V.; Hunt, S.; Ibrahim, A. G.-E.; Ikezu, T.; Inal, J. M.; Isin, M.; Ivanova, A.; Jackson, H. K.; Jacobsen, S.; Jay, S. M.; Jayachandran, M.; Jenster, G.; Jiang, L.; Johnson, S. M.; Jones, J. C.; Jong, A.; Jovanovic-Talisman, T.; Jung, S.; Kalluri, R.; Kano, S.; Kaur, S.; Kawamura, Y.; Keller, E. T.; Khamari, D.; Khomyakova, E.; Khvorova, A.; Kierulf, P.; Kim, K. P.; Kislinger, T.; Klingeborn, M.; Klinke, D. J.; Kornek, M.; Kosanović, M. M.; Kovács, Á. F.; Krämer-Albers, E.-M.; Krasemann, S.; Krause, M.; Kurochkin, I. V.; Kusuma, G. D.; Kuypers, S.; Laitinen, S.; Langevin, S. M.; Languino, L. R.; Lannigan, J.; Lässer, C.; Laurent, L. C.; Lavieu, G.; Lázaro-Ibáñez, E.; Le Lay, S.; Lee, M.-S.; Lee, Y. X. F.; Lemos, D. S.; Lenassi, M.; Leszczynska, A.; Li, I. T.; Liao, K.; Libregts, S. F.; Ligeti, E.; Lim, R.; Lim, S. K.; Linē, A.; Linnemannstöns, K.; Llorente, A.; Lombard, C. A.; Lorenowicz, M. J.; Lörincz, Á. M.; Lötvall, J.; Lovett, J.; Lowry, M. C.; Loyer, X.; Lu, Q.; Lukomska, B.; Lunavat, T. R.; Maas, S. L.; Malhi, H.; Marcilla, A.; Mariani, J.; Mariscal, J.; Martens-Uzunova, E. S.; Martin-Jaular, L.; Martinez, M. C.; Martins, V. R.; Mathieu, M.; Mathivanan, S.; Maugeri, M.; McGinnis, L. K.; McVey, M. J.; Meckes, D. G.; Meehan, K. L.; Mertens, I.; Minciacchi, V. R.; Möller, A.; Møller Jørgensen, M.; Morales-Kastresana, A.; Morhayim, J.; Mullier, F.; Muraca, M.; Musante, L.; Mussack, V.; Muth, D. C.; Myburgh, K. H.; Najrana, T.; Nawaz, M.; Nazarenko, I.; Nejsum, P.; Neri, C.; Neri, T.; Nieuwland, R.; Nimrichter, L.; Nolan, J. P.; Nolte-'t Hoen, E. N.; Noren Hooten, N.; O'Driscoll, L.; O'Grady, T.; O'Loghlen, A.; Ochiya, T.; Olivier, M.; Ortiz, A.; Ortiz, L. A.; Osteikoetxea, X.; Østergaard, O.; Ostrowski, M.; Park, J.; Pegtel, D. M.; Peinado, H.; Perut, F.; Pfaffl, M. W.; Phinney, D. G.; Pieters, B. C.; Pink, R. C.; Pisetsky, D. S.; Pogge von Strandmann, E.; Polakovicova, I.; Poon, I. K.; Powell, B. H.; Prada, I.; Pulliam, L.; Quesenberry, P.; Radeghieri, A.; Raffai, R. L.; Raimondo, S.; Rak, J.; Ramirez, M. I.; Raposo, G.; Rayyan, M. S.; Regev-Rudzki, N.; Ricklefs, F. L.; Robbins, P. D.; Roberts, D. D.; Rodrigues, S. C.; Rohde, E.; Rome, S.; Rouschop, K. M.; Rughetti, A.; Russell, A. E.; Saá, P.; Sahoo, S.; Salas-Huenuleo, E.;

Sánchez, C.; Saugstad, J. A.; Saul, M. J.; Schiffelers, R. M.; Schneider, R.; Schøyen, T. H.; Scott, A.; Shahaj, E.; Sharma, S.; Shatnyeva, O.; Shekari, F.; Shelke, G. V.; Shetty, A. K.; Shiba, K.; Siljander, P. R.-M.; Silva, A. M.; Skowronek, A.; Snyder, O. L.; Soares, R. P.; Sódar, B. W.; Soekmadji, C.; Sotillo, J.; Stahl, P. D.; Stoorvogel, W.; Stott, S. L.; Strasser, E. F.; Swift, S.; Tahara, H.; Tewari, M.; Timms, K.; Tiwari, S.; Tixeira, R.; Tkach, M.; Toh, W. S.; Tomasini, R.; Torrecilhas, A. C.; Tosar, J. P.; Toxavidis, V.; Urbanelli, L.; Vader, P.; van Balkom, B. W.; van der Grein, S. G.; Van Deun, J.; van Herwijnen, M. J.; Van Keuren-Jensen, K.; van Niel, G.; van Royen, M. E.; van Wijnen, A. J.; Vasconcelos, M. H.; Vechetti, I. J.; Veit, T. D.; Vella, L. J.; Velot, É.; Verweij, F. J.; Vestad, B.; Viñas, J. L.; Visnovitz, T.; Vukman, K. V.; Wahlgren, J.; Watson, D. C.; Wauben, M. H.; Weaver, A.; Webber, J. P.; Weber, V.; Wehman, A. M.; Weiss, D. J.; Welsh, J. A.; Wendt, S.; Wheelock, A. M.; Wiener, Z.; Witte, L.; Wolfram, J.; Xagorari, A.; Xander, P.; Xu, J.; Yan, X.; Yáñez-Mó, M.; Yin, H.; Yuana, Y.; Zappulli, V.; Zarubova, J.; Žékas, V.; Zhang, J.; Zhao, Z.; Zheng, L.; Zheutlin, A. R.; Zickler, A. M.; Zimmermann, P.; Zivkovic, A. M.; Zocco, D.; Zuba-Surma, E. K. Minimal Information for Studies of Extracellular Vesicles 2018 (MISEV2018): A Position Statement of the International Society for Extracellular Vesicles and Update of the MISEV2014 Guidelines. *Journal of Extracellular Vesicles* **2018**, *7* (1), 1535750. https://doi.org/10.1080/20013078.2018.1535750.

(122) Maucuer, A.; Camonis, J. H.; Sobel, A. Stathmin Interaction with a Putative Kinase and Coiled-Coil-Forming Protein Domains. *Proceedings of the National Academy of Sciences* **1995**, *92* (8), 3100–3104. https://doi.org/10.1073/pnas.92.8.3100.

(123) Li, L.; Cohen, S. N. Tsg101: A Novel Tumor Susceptibility Gene Isolated by Controlled Homozygous Functional Knockout of Allelic Loci in Mammalian Cells. *Cell* **1996**, *85* (3), 319–329. https://doi.org/10.1016/S0092-8674(00)81111-3.

(124) Cheng, X. Role of TSG101 in Cancer. *Front Biosci* **2013**, *18* (1), 279. https://doi.org/10.2741/4099.

(125) Wagner, K.-U.; Dierisseau, P.; B Rucker, E.; Robinson, G. W.; Hennighausen, L. Genomic Architecture and Transcriptional Activation of the Mouse and Human Tumor Susceptibility Gene TSG101: Common Types of Shorter Transcripts Are True Alternative Splice Variants. *Oncogene* **1998**, *17* (21), 2761–2770. https://doi.org/10.1038/sj.onc.1202529.

(126) *Multiple Sequence Alignment: Methods and Protocols*; Katoh, K., Ed.; Methods in Molecular Biology; Springer US: New York, NY, 2021; Vol. 2231. https://doi.org/10.1007/978-1-0716-1036-7.

(127) Madeira, F.; Pearce, M.; Tivey, A. R. N.; Basutkar, P.; Lee, J.; Edbali, O.; Madhusoodanan, N.; Kolesnikov, A.; Lopez, R. Search and Sequence Analysis Tools Services from EMBL-EBI in 2022. *Nucleic Acids Research* **2022**, *50* (W1), W276–W279. https://doi.org/10.1093/nar/gkac240.

(128) *AlphaFold Protein Structure Database*. https://alphafold.ebi.ac.uk/entry/Q99816 (accessed 2022-11-03).

(129) Fabbro, M.; Zhou, B.-B.; Takahashi, M.; Sarcevic, B.; Lal, P.; Graham, M. E.; Gabrielli, B. G.; Robinson, P. J.; Nigg, E. A.; Ono, Y.; Khanna, K. K. Cdk1/Erk2- and Plk1-Dependent Phosphorylation of a Centrosome Protein, Cep55, Is Required for Its Recruitment to Midbody and Cytokinesis. *Developmental Cell* **2005**, *9* (4), 477–488. https://doi.org/10.1016/j.devcel.2005.09.003.

(130)   Morita, E.; Sandrin, V.; Chung, H.-Y.; Morham, S. G.; Gygi, S. P.; Rodesch, C. K.; Sundquist, W. I. Human ESCRT and ALIX Proteins Interact with Proteins of the Midbody and Function in Cytokinesis. *EMBO J* **2007**, *26* (19), 4215–4227. https://doi.org/10.1038/sj.emboj.7601850.

(131)   Lee, H. H.; Elia, N.; Ghirlando, R.; Lippincott-Schwartz, J.; Hurley, J. H. Midbody Targeting of the ESCRT Machinery by a Noncanonical Coiled Coil in CEP55. *Science* **2008**, *322* (5901), 576–580. https://doi.org/10.1126/science.1162042.

(132)   Elia, N.; Sougrat, R.; Spurlin, T. A.; Hurley, J. H.; Lippincott-Schwartz, J. Dynamics of Endosomal Sorting Complex Required for Transport (ESCRT) Machinery during Cytokinesis and Its Role in Abscission. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108* (12), 4846–4851. https://doi.org/10.1073/pnas.1102714108.

(133)   Katoh, K.; Suzuki, H.; Terasawa, Y.; Mizuno, T.; Yasuda, J.; Shibata, H.; Maki, M. The Penta-EF-Hand Protein ALG-2 Interacts Directly with the ESCRT-I Component TSG101, and Ca2+-Dependently Co-Localizes to Aberrant Endosomes with Dominant-Negative AAA ATPase SKD1/Vps4B. *Biochemical Journal* **2005**, *391* (3), 677–685. https://doi.org/10.1042/BJ20050398.

(134)   Suzuki, H.; Kawasaki, M.; Inuzuka, T.; Okumura, M.; Kakiuchi, T.; Shibata, H.; Wakatsuki, S.; Maki, M. Structural Basis for Ca2+-Dependent Formation of ALG-2/Alix Peptide Complex: Ca2+/EF3-Driven Arginine Switch Mechanism. *Structure* **2008**, *16* (10), 1562–1573. https://doi.org/10.1016/j.str.2008.07.012.

(135)   Boura, E.; Różycki, B.; Herrick, D. Z.; Chung, H. S.; Vecer, J.; Eaton, W. A.; Cafiso, D. S.; Hummer, G.; Hurley, J. H. Solution Structure of the ESCRT-I Complex by Small-Angle X-Ray Scattering, EPR, and FRET Spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108* (23), 9437–9442. https://doi.org/10.1073/pnas.1101763108.

(136)   Watanabe, M.; Yanagi, Y.; Masuhiro, Y.; Yano, T.; Yoshikawa, H.; Yanagisawa, J.; Kato, S. A Putative Tumor Suppressor, TSG101, Acts as a Transcriptional Suppressor through Its Coiled-Coil Domain. *Biochemical and Biophysical Research Communications* **1998**, *245* (3), 900–905. https://doi.org/10.1006/bbrc.1998.8547.

(137)   Ferrer, M.; López-Borges, S.; Lazo, P. A. Expression of a New Isoform of the Tumor Susceptibility TSG101 Protein Lacking a Leucine Zipper Domain in Burkitt Lymphoma Cell Lines. *Oncogene* **1999**, *18* (13), 2253–2259. https://doi.org/10.1038/sj.onc.1202551.

(138)   Ferrer, M.; Hernández, S.; Campo, E.; Lazo, P. Loss of the TSG101 Leucine Zipper Domain in Aggressive Non-Hodgkin's Lymphomas. *Leukemia* **2000**, *14* (11), 2014–2016. https://doi.org/10.1038/sj.leu.2401920.

(139)   Lu, Q.; Hope, L. W.; Brasch, M.; Reinhard, C.; Cohen, S. N. TSG101 Interaction with HRS Mediates Endosomal Trafficking and Receptor Down-Regulation. *Proceedings of the National Academy of Sciences* **2003**, *100* (13), 7626–7631. https://doi.org/10.1073/pnas.0932599100.

(140)   Kostelansky, M. S.; Sun, J.; Lee, S.; Kim, J.; Ghirlando, R.; Hierro, A.; Emr, S. D.; Hurley, J. H. Structural and Functional Organization of the ESCRT-I Trafficking Complex. *Cell* **2006**, *125* (1), 113–126. https://doi.org/10.1016/j.cell.2006.01.049.

(141)   Feng, G. H.; Lih, C. J.; Cohen, S. N. TSG101 Protein Steady-State Level Is Regulated Posttranslationally by an Evolutionarily Conserved COOH-Terminal Sequence. *Cancer Res* **2000**, *60* (6), 1736–1741.

(142)   Pornillos, O. Structure and Functional Interactions of the Tsg101 UEV Domain. *The EMBO Journal* **2002**, *21* (10), 2397–2406. https://doi.org/10.1093/emboj/21.10.2397.

(143)   McDonald, B.; Martin-Serrano, J. Regulation of Tsg101 Expression by the Steadiness Box: A Role of Tsg101-Associated Ligase. *MBoC* **2008**, *19* (2), 754–763. https://doi.org/10.1091/mbc.e07-09-0957.

(144)   Palencia, A.; Martinez, J. C.; Mateo, P. L.; Luque, I.; Camara-Artigas, A. Structure of Human TSG101 UEV Domain. *Acta Crystallogr D Biol Crystallogr* **2006**, *62* (4), 458–464. https://doi.org/10.1107/S0907444906005221.

(145)   Sundquist, W. I.; Schubert, H. L.; Kelly, B. N.; Hill, G. C.; Holton, J. M.; Hill, C. P. Ubiquitin Recognition by the Human TSG101 Protein. *Molecular Cell* **2004**, *13* (6), 783–789. https://doi.org/10.1016/S1097-2765(04)00129-7.

(146)   Pornillos, O.; Alam, S. L.; Davis, D. R.; Sundquist, W. I. Structure of the Tsg101 UEV Domain in Complex with the PTAP Motif of the HIV-1 P6 Protein. *Nat Struct Biol* **2002**. https://doi.org/10.1038/nsb856.

(147)   Im, Y. J.; Kuo, L.; Ren, X.; Burgos, P. V.; Zhao, X. Z.; Liu, F.; Burke, T. R.; Bonifacino, J. S.; Freed, E. O.; Hurley, J. H. Crystallographic and Functional Analysis of the ESCRT-I /HIV-1 Gag PTAP Interaction. *Structure* **2010**, *18* (11), 1536–1547. https://doi.org/10.1016/j.str.2010.08.010.

(148)   Freed, E. O. Viral Late Domains. *J Virol* **2002**, *76* (10), 4679–4687. https://doi.org/10.1128/JVI.76.10.4679-4687.2002.

(149)   Dolnik, O.; Kolesnikova, L.; Welsch, S.; Strecker, T.; Schudt, G.; Becker, S. Interaction with Tsg101 Is Necessary for the Efficient Transport and Release of Nucleocapsids in Marburg Virus-Infected Cells. *PLoS Pathog* **2014**, *10* (10), e1004463. https://doi.org/10.1371/journal.ppat.1004463.

(150)   Kim, S.-E.; Liu, F.; Im, Y. J.; Stephen, A. G.; Fivash, M. J.; Waheed, A. A.; Freed, E. O.; Fisher, R. J.; Hurley, J. H.; Burke, T. R. Elucidation of New Binding Interactions with the Human Tsg101 Protein Using Modified HIV-1 Gag-P6 Derived Peptide Ligands. *ACS Med. Chem. Lett.* **2011**, *2* (5), 337–341. https://doi.org/10.1021/ml1002579.

(151)   Tavassoli, A.; Lu, Q.; Gam, J.; Pan, H.; Benkovic, S. J.; Cohen, S. N. Inhibition of HIV Budding by a Genetically Selected Cyclic Peptide Targeting the Gag–TSG101 Interaction. *ACS Chem. Biol.* **2008**, *3* (12), 757–764. https://doi.org/10.1021/cb800193n.

(152)   Lennard, K. R.; Gardner, R. M.; Doigneaux, C.; Castillo, F.; Tavassoli, A. Development of a Cyclic Peptide Inhibitor of the P6/UEV Protein–Protein Interaction. *ACS Chem. Biol.* **2019**, *14* (9), 1874–1878. https://doi.org/10.1021/acschembio.9b00627.

(153)   Siarot, L.; Chutiwitoonchai, N.; Sato, H.; Chang, H.; Sato, H.; Fujino, M.; Murakami, T.; Aono, T.; Kodama, E.; Kuroda, K.; Takei, M.; Aida, Y. Identification of Human Immunodeficiency Virus Type-1 Gag-TSG101 Interaction Inhibitors by High-Throughput Screening. *Biochemical and Biophysical Research Communications* **2018**, *503* (4), 2970–2976. https://doi.org/10.1016/j.bbrc.2018.08.079.

(154)   Strickland, M.; Ehrlich, L. S.; Watanabe, S.; Khan, M.; Strub, M.-P.; Luan, C.-H.; Powell, M. D.; Leis, J.; Tjandra, N.; Carter, C. A. Tsg101 Chaperone Function Revealed by HIV-1 Assembly Inhibitors. *Nat Commun* **2017**, *8* (1), 1391. https://doi.org/10.1038/s41467-017-01426-2.

(155)   Leis, J.; Luan, C.-H.; Audia, J. E.; Dunne, S. F.; Heath, C. M. Ilaprazole and Other Novel Prazole-Based Compounds That Bind Tsg101 Inhibit Viral Budding of Herpes Simplex Virus 1

and 2 and Human Immunodeficiency Virus from Cells. *J Virol* **2021**, *95* (11). https://doi.org/10.1128/JVI.00190-21.

(156) Rabi, I. I.; Zacharias, J. R.; Millman, S.; Kusch, P. A New Method of Measuring Nuclear Magnetic Moment. *Phys. Rev.* **1938**, *53* (4), 318–318. https://doi.org/10.1103/PhysRev.53.318.

(157) *The Nobel Prize in Physics 1944*. NobelPrize.org. https://www.nobelprize.org/prizes/physics/1944/summary/ (accessed 2022-11-06).

(158) Becker, E. D. A BRIEF HISTORY OF NUCLEAR MAGNETIC RESONANCE. *Anal. Chem.* **1993**, *65* (6), 295A-302A. https://doi.org/10.1021/ac00054a716.

(159) Purcell, E. M.; Torrey, H. C.; Pound, R. V. Resonance Absorption by Nuclear Magnetic Moments in a Solid. *Phys. Rev.* **1946**, *69* (1–2), 37–38. https://doi.org/10.1103/PhysRev.69.37.

(160) Bloch, F.; Hansen, W. W.; Packard, M. Nuclear Induction. *Phys. Rev.* **1946**, *69* (3–4), 127–127. https://doi.org/10.1103/PhysRev.69.127.

(161) *The Nobel Prize in Physics 1952*. NobelPrize.org. https://www.nobelprize.org/prizes/physics/1952/summary/ (accessed 2022-11-06).

(162) Ernst, R. R.; Anderson, W. A. Application of Fourier Transform Spectroscopy to Magnetic Resonance. *Review of Scientific Instruments* **1966**, *37* (1), 93–102. https://doi.org/10.1063/1.1719961.

(163) *The Nobel Prize in Chemistry 1991*. NobelPrize.org. https://www.nobelprize.org/prizes/chemistry/1991/summary/ (accessed 2022-11-06).

(164) *The Nobel Prize in Chemistry 2002*. NobelPrize.org. https://www.nobelprize.org/prizes/chemistry/2002/summary/ (accessed 2022-11-06).

(165) *The Nobel Prize in Physiology or Medicine 2003*. NobelPrize.org. https://www.nobelprize.org/prizes/medicine/2003/summary/ (accessed 2022-11-06).

(166) Raja, P. M. V.; Barron, A. R. Physical Methods in Chemistry and Nano Science. 665.

(167) Rule, G. S.; Hitchens, T. K. *Fundamentals of Protein NMR Spectroscopy*; Focus on structural biology; Springer: Dordrecht, 2006.

(168) *NMR Spectroscopy Principles, Interpreting an NMR Spectrum and Common Problems*. Analysis & Separations from Technology Networks. http://www.technologynetworks.com/analysis/articles/nmr-spectroscopy-principles-interpreting-an-nmr-spectrum-and-common-problems-355891 (accessed 2022-11-07).

(169) Wright, P. E.; Dyson, H. J. Intrinsically Unstructured Proteins: Re-Assessing the Protein Structure-Function Paradigm. *Journal of Molecular Biology* **1999**, *293* (2), 321–331. https://doi.org/10.1006/jmbi.1999.3110.

(170) Deiana, A.; Forcelloni, S.; Porrello, A.; Giansanti, A. Intrinsically Disordered Proteins and Structured Proteins with Intrinsically Disordered Regions Have Different Functional Roles in the Cell. *PLoS ONE* **2019**, *14* (8), e0217889. https://doi.org/10.1371/journal.pone.0217889.

(171) Kumar, N.; Kaushik, R.; Tennakoon, C.; Uversky, V. N.; Longhi, S.; Zhang, K. Y. J.; Bhatia, S. Comprehensive Intrinsic Disorder Analysis of 6108 Viral Proteomes: From the Extent of Intrinsic Disorder Penetrance to Functional Annotation of Disordered Viral Proteins. *J. Proteome Res.* **2021**, *20* (5), 2704–2713. https://doi.org/10.1021/acs.jproteome.1c00011.

(172)   Dobrev, V. S.; Fred, L. M.; Gerhart, K. P.; Metallo, S. J. Characterization of the Binding of Small Molecules to Intrinsically Disordered Proteins. In *Methods in Enzymology*; Elsevier, 2018; Vol. 611, pp 677–702. https://doi.org/10.1016/bs.mie.2018.09.033.

(173)   Schanda, P. Development and Application of Fast NMR Methods for the Study of Protein Structure and Dynamics. 241.

(174)   garren. *NMR Spectroscopy*. SlideServe. https://www.slideserve.com/garren/nmr-spectroscopy (accessed 2022-11-16).

(175)   *1H-15N HSQC | Protein NMR*. https://www.protein-nmr.org.uk/solution-nmr/spectrum-descriptions/1h-15n-hsqc/ (accessed 2022-11-09).

(176)   Breukels, V.; Konijnenberg, A.; Nabuurs, S. M.; Doreleijers, J. F.; Kovalevskaya, N. V.; Vuister, G. W. Overview on the Use of NMR to Examine Protein Structure. In *Current Protocols in Protein Science*; Coligan, J. E., Dunn, B. M., Speicher, D. W., Wingfield, P. T., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2011; p 17.5.1-17.5.44. https://doi.org/10.1002/0471140864.ps1705s64.

(177)   *Triple Resonance Backbone Assignment | Protein NMR*. https://www.protein-nmr.org.uk/solution-nmr/assignment-theory/triple-resonance-backbone-assignment/ (accessed 2022-11-10).

(178)   *NCO | Protein NMR*. https://www.protein-nmr.org.uk/solid-state-mas-nmr/spectrum-descriptions/nco/ (accessed 2022-11-10).

(179)   Kleckner, I. R.; Foster, M. P. An Introduction to NMR-Based Approaches for Measuring Protein Dynamics. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2011**, *1814* (8), 942–968. https://doi.org/10.1016/j.bbapap.2010.10.012.

(180)   Ziarek, J. J.; Baptista, D.; Wagner, G. Recent Developments in Solution Nuclear Magnetic Resonance (NMR)-Based Molecular Biology. *J Mol Med* **2018**, *96* (1), 1–8. https://doi.org/10.1007/s00109-017-1560-2.

(181)   Alcaraz, L. A.; Gómez, J.; Ramírez, P.; Calvente, J. J.; Andreu, R.; Donaire, A. Folding and Unfolding in the Blue Copper Protein Rusticyanin: Role of the Oxidation State. *Bioinorganic Chemistry and Applications* **2007**, *2007*, 1–9. https://doi.org/10.1155/2007/54232.

(182)   Kasem, A.; Azeem, K.; Vlčková, J.; Zatloukalová, S.; Štěpánek, L.; Kyselý, Z.; Kollárová, H. Epidemiology of Hepatitis E Virus Infection. *Epidemiol Mikrobiol Imunol* **2019**, *68* (4), 176–182.

(183)   Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S.; Wilkins, M. R.; Appel, R. D.; Bairoch, A. Protein Identification and Analysis Tools on the ExPASy Server. In *The Proteomics Protocols Handbook*; Walker, J. M., Ed.; Humana Press: Totowa, NJ, 2005; pp 571–607. https://doi.org/10.1385/1-59259-890-0:571.

(184)   Hagn, F.; Nasr, M. L.; Wagner, G. Assembly of Phospholipid Nanodiscs of Controlled Size for Structural Studies of Membrane Proteins by NMR. *Nat Protoc* **2018**, *13* (1), 79–98. https://doi.org/10.1038/nprot.2017.094.

(185)   Lee, W.; Tonelli, M.; Markley, J. L. NMRFAM-SPARKY: Enhanced Software for Biomolecular NMR Spectroscopy. *Bioinformatics* **2015**, *31* (8), 1325–1327. https://doi.org/10.1093/bioinformatics/btu830.

(186)   *TMHMM v2.0 - Prediction of transmembrane helices in proteins*. https://services.healthtech.dtu.dk/service.php?TMHMM-2.0.

(187)   Kozlowski, L. P.; Bujnicki, J. M. MetaDisorder: A Meta-Server for the Prediction of Intrinsic Disorder in Proteins. *BMC Bioinformatics* **2012**, *13* (1), 111. https://doi.org/10.1186/1471-2105-13-111.

(188)   Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: Making Protein Folding Accessible to All. *Nat Methods* **2022**, *19* (6), 679–682. https://doi.org/10.1038/s41592-022-01488-1.

(189)   Mirdita, M.; Steinegger, M.; Söding, J. MMseqs2 Desktop and Local Web Server App for Fast, Interactive Sequence Searches. *Bioinformatics* **2019**, *35* (16), 2856–2858. https://doi.org/10.1093/bioinformatics/bty1057.

(190)   Mirdita, M.; von den Driesch, L.; Galiez, C.; Martin, M. J.; Söding, J.; Steinegger, M. Uniclust Databases of Clustered and Deeply Annotated Protein Sequences and Alignments. *Nucleic Acids Res* **2017**, *45* (D1), D170–D176. https://doi.org/10.1093/nar/gkw1081.

(191)   Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(192)   Mitchell, A. L.; Almeida, A.; Beracochea, M.; Boland, M.; Burgin, J.; Cochrane, G.; Crusoe, M. R.; Kale, V.; Potter, S. C.; Richardson, L. J.; Sakharova, E.; Scheremetjew, M.; Korobeynikov, A.; Shlemov, A.; Kunyavskaya, O.; Lapidus, A.; Finn, R. D. MGnify: The Microbiome Analysis Resource in 2020. *Nucleic Acids Res.* **2019**. https://doi.org/10.1093/nar/gkz1035.

(193)   Rutherfurd, S. M.; Gilani, G. S. Amino Acid Analysis. *Current Protocols in Protein Science* **2009**, *58* (1). https://doi.org/10.1002/0471140864.ps1109s58.

(194)   *29836-26-8   Octyl-beta-D-glucopyranoside   Formula,NMR,Boiling   Point,Density,Flash Point*. https://www.guidechem.com/dictionary/en/29836-26-8.html (accessed 2022-10-01).

(195)   Marsh, J. A.; Singh, V. K.; Jia, Z.; Forman-Kay, J. D. Sensitivity of Secondary Structure Propensities to Sequence Differences between Alpha- and Gamma-Synuclein: Implications for Fibrillation. *Protein Sci* **2006**, *15* (12), 2795–2804. https://doi.org/10.1110/ps.062465306.

(196)   Yousefi-Salakdeh, E.; Johansson, J.; Strömberg, R. A Method for S- and O-Palmitoylation of Peptides: Synthesis of Pulmonary Surfactant Protein-C Models. *Biochemical Journal* **1999**, *343* (3), 557–562. https://doi.org/10.1042/bj3430557.

(197)   *18:1 PE MCC*. Avanti Polar Lipids. https://avantilipids.com/product/780201 (accessed 2022-11-20).

(198)   Zare, F.; Potenza, A.; Greschner, A. A.; Gauthier, M. A. Consecutive Alkylation, "Click", and "Clip" Reactions for the Traceless Methionine-Based Conjugation and Release of Methionine-Containing Peptides. *Biomacromolecules* **2022**, acs.biomac.2c00357. https://doi.org/10.1021/acs.biomac.2c00357.

(199) Liu, J.; Zhu, L.; Zhang, X.; Wu, B.; Zhu, P.; Zhao, H.; Wang, J. Peptide-Based NTA(Ni)-Nanodiscs for Studying Membrane Enhanced FGFR1 Kinase Activities. *PeerJ* **2019**, *7*, e7234. https://doi.org/10.7717/peerj.7234.

(200) *18:1 DGS-NTA(Ni)*. https://avantilipids.com/product/790404 (accessed 2022-11-20).

(201) Anthis, N. J.; Clore, G. M. Sequence-Specific Determination of Protein and Peptide Concentrations by Absorbance at 205 Nm: Sequence-Specific Protein Concentration at 205 Nm. *Protein Science* **2013**, *22* (6), 851–858. https://doi.org/10.1002/pro.2253.

(202) Hafsa, N. E.; Arndt, D.; Wishart, D. S. CSI 3.0: A Web Server for Identifying Secondary and Super-Secondary Structure in Proteins Using NMR Chemical Shifts. *Nucleic Acids Res* **2015**, *43* (W1), W370–W377. https://doi.org/10.1093/nar/gkv494.

(203) Shen, Y.; Delaglio, F.; Cornilescu, G.; Bax, A. TALOS+: A Hybrid Method for Predicting Protein Backbone Torsion Angles from NMR Chemical Shifts. *J Biomol NMR* **2009**, *44* (4), 213–223. https://doi.org/10.1007/s10858-009-9333-z.

(204) Golovanov, A. P.; Blankley, R. T.; Avis, J. M.; Bermel, W. Isotopically Discriminated NMR Spectroscopy: A Tool for Investigating Complex Protein Interactions in Vitro. *J. Am. Chem. Soc.* **2007**, *129* (20), 6528–6535. https://doi.org/10.1021/ja070505q.

(205) He, C.; Lin, G.; Upton, K. T.; Imanaka, H.; Smith, M. A. Structural Investigation of Titan Tholins by Solution-State ^{1}H, ^{13}C, and ^{15}N NMR: One-Dimensional and Decoupling Experiments. *J. Phys. Chem. A* **2012**, *116* (19), 4760–4767. https://doi.org/10.1021/jp3016062.

(206) Freyer, M. W.; Lewis, E. A. Isothermal Titration Calorimetry: Experimental Design, Data Analysis, and Probing Macromolecule/Ligand Binding and Kinetic Interactions. In *Methods in Cell Biology*; Elsevier, 2008; Vol. 84, pp 79–113. https://doi.org/10.1016/S0091-679X(07)84004-0.

(207) *ITC • Isothermal Titration Calorimetry • Thermodynamics, Affinity, Stoichiometry*. 2bind. https://2bind.com/itc/ (accessed 2022-10-13).

(208) Frasca, V. Biophysical Characterization of Antibodies with Isothermal Titration Calorimetry. *J Appl Bioanal* **2016**, *2* (3), 90–102. https://doi.org/10.17145/jab.16.013.

(209) Potterton, L.; Agirre, J.; Ballard, C.; Cowtan, K.; Dodson, E.; Evans, P. R.; Jenkins, H. T.; Keegan, R.; Krissinel, E.; Stevenson, K.; Lebedev, A.; McNicholas, S. J.; Nicholls, R. A.; Noble, M.; Pannu, N. S.; Roth, C.; Sheldrick, G.; Skubak, P.; Turkenburg, J.; Uski, V.; von Delft, F.; Waterman, D.; Wilson, K.; Winn, M.; Wojdyr, M. *CCP* 4 *i* 2: The New Graphical User Interface to the *CCP* 4 Program Suite. *Acta Crystallogr D Struct Biol* **2018**, *74* (2), 68–84. https://doi.org/10.1107/S2059798317016035.

(210) Winn, M. D.; Ballard, C. C.; Cowtan, K. D.; Dodson, E. J.; Emsley, P.; Evans, P. R.; Keegan, R. M.; Krissinel, E. B.; Leslie, A. G. W.; McCoy, A.; McNicholas, S. J.; Murshudov, G. N.; Pannu, N. S.; Potterton, E. A.; Powell, H. R.; Read, R. J.; Vagin, A.; Wilson, K. S. Overview of the *CCP* 4 Suite and Current Developments. *Acta Crystallogr D Biol Crystallogr* **2011**, *67* (4), 235–242. https://doi.org/10.1107/S0907444910045749.

(211) Moerke, N. J. Fluorescence Polarization (FP) Assays for Monitoring Peptide-Protein or Nucleic Acid-Protein Binding. *Current Protocols in Chemical Biology* **2009**, *1* (1), 1–15. https://doi.org/10.1002/9780470559277.ch090102.

(212) Degorce, F. HTRF: A Technology Tailored for Drug Discovery - A Review of Theoretical Aspects and Recent Applications. *TOCHGENJ* **2009**, *3* (1), 22–32. https://doi.org/10.2174/1875397300903010022.

(213) Huynh, K.; Partch, C. L. Analysis of Protein Stability and Ligand Interactions by Thermal Shift Assay. *Current Protocols in Protein Science* **2015**, *79* (1). https://doi.org/10.1002/0471140864.ps2809s79.

(214) Thermal Shift Assay. *Wikipedia*; 2021.

(215) LightCycler® 480 Real-Time PCR System-Technical-Note.Pdf. https://www.gene-quantification.de/LC480-Technical-Note-01-HRM.pdf (accessed 2022-10-15).

3CLpro SARS-CoV-2 project

Apart from the HEV ORF3 project presented in this study, I also contributed to another scientific project. Since the beginning of the Covid19 pandemic, a Task-force including several research laboratories from the Pasteur Institute of Lille has been created to help finding potent solutions to fight against the newly emerging coronavirus SARS-CoV-2. Our Integrative Structural Biology group, in collaboration with the U1177 unit "Drugs & Molecules for Living Systems" (Prof. B. Deprez and Dr J. Charton) and the U1019-UMR9017 unit "Center for Infection and Immunity of Lille" (Dr. J. Dubuisson and Dr. S. Belouzard), is involved in a project that aim to find potent inhibitor(s) of the Main Protease (3CLpro) of the SARS-Cov-2, a viral enzyme that is essential to the virus life cycle. Our laboratory received a financial support from I-SITE ULNE (project 3CLPROSCREEN), the CPER CTRL (Transdisciplinary Research Center on Longevity) program, and the Pasteur Institute of Lille. The project is still ongoing and its scope has been expanded toward the discovery of pan-coronaviruses inhibitors.

In our laboratory, the initial objectives were: 1/ to express in a recombinant way and purify a unique batch of the SARS-CoV-2 3CLp enzyme that should be sufficient to perform a high-throughput screening with ~90,000 molecules; 2/ to use solution-state NMR spectroscopy to get information on 3CLp (that was never achieved before on any coronavirus); 3/ to perform an NMR-based fragment screening on SARS-CoV-2 3CLp to start the *de novo* design of an inhibitor; 4/ to use X-ray crystallography to support the medicinal chemistry.

From May 2020, in this project, I was involved in the 3CLp protein production as well the crystallization assays of 3CLp with different molecules. Especially, I prepared unlabeled and isotopically labeled protein samples of 3CLp protease WT and two point-mutations (G11A and R298A) monomeric mutants. The labeled samples were used for NMR Spectroscopy studies, particularly for obtaining the chemical shift backbone assignments for 3CLp WT protein and performing interaction experiments with different molecules. In addition, using a two-step NMR-based fragment screening, we identified 38 molecular fragments that binds SARS-CoV-2 in three different binding sites. The best fragment (F01), without optimization, has been shown to have

some antiviral properties in an infected cells culture model. The high-resolution crystal structure of the complex 3CLp:F01 has been solved to get the molecular details of its binding in the active site of the protease. This information is still use to optimize the potency of F01 derivatives.

Actually, we mainly support the lab of medicinal chemistry by solving crystal structures of the SARS-CoV-2 3CLp in complex with either F01-derived molecules or compounds that have been identified in the HTS enzymatic screening. We also express and purify Main Protease from other coronaviruses, such as MERS-CoV and h229E-CoV, to help the development of pan-coronavirus inhibitors.

Our results are included in the two scientific articles presented in the Annex 2. The first one is published on Angewandte Chemie International Edition, titled "NMR spectroscopy of the main protease of SARS-CoV-2 and fragment-based screening identify three protein hotspots and an antiviral fragment.", by Cantrelle FX, Boll E, Brier L, Moschidi D, Belouzard S, Landry V, Leroux F, Dewitte F, Landrieu I, Dubuisson J, Deprez B, Charton J and Hanoulle X, in which I am second author. The second one published on European Journal of Medicinal Chemistry journal with title "Novel dithiocarbamates selectively inhibit 3CL protease of SARS-CoV-2 and other coronaviruses", by Brier L, Hassan H, Hanoulle X, Landry V, Moschidi D, Desmarets L, Rouillé Y, Dumont J, Herledan A, Warenghem S, Piveteau C, Carré P, Ikherbane S, Cantrelle FX, Dupré E, Dubuisson J, Belouzard S, Leroux F, Deprez B and Charton J.

Annex

1. HEV ORF3 project

The following article titled "Backbone NMR resonance assignment of the apo human Tsg101-UEV domain" is already published on "Biomolecular NMR assignments" journal. In addition, the results presented in this study about HEV ORF3 protein is included in a journal paper which is in preparation.

Backbone NMR resonance assignment of the apo human Tsg101-UEV domain

Danai Moschidi, François-Xavier Cantrelle, Emmanuelle Boll & Xavier Hanoulle[*]

[1] CNRS EMR9002 Integrative Structural Biology, F-59000, Lille, France
[2] Univ. Lille, Inserm, CHU Lille, Institut Pasteur de Lille, U1167 - RID-AGE - Risk Factors and Molecular Determinants of Aging-Related Diseases, F-59000, Lille, France

Corresponding author:
Xavier Hanoulle
xavier.hanoulle@univ-lille.fr
ORCID: 0000-0002-3755-2680

Abstract

The Endosomal Sorting Complex Required for Transport (ESCRT) pathway, through inverse topology membrane remodeling, is involved in many biological functions, such as ubiquitinated membrane receptor trafficking and degradation, multivesicular bodies (MVB) formation and cytokinesis. Dysfunctions in ESCRT pathway have been associated to several human pathologies, such as cancers and neurodegenerative diseases. The ESCRT machinery is also hijacked by many enveloped viruses to bud away from the plasma membrane of infected cells. Human tumor susceptibility gene 101 (Tsg101) protein is an important ESCRT-I complex component. The structure of the N-terminal ubiquitin E2 variant (UEV) domain of Tsg101 (Tsg101-UEV) comprises an ubiquitin binding pocket next to a late domain [P(S/T)AP] binding groove. These two binding sites have been shown to be involved both in the physiological roles of ESCRT-I and in the release of the viral particles, and thus are attractive targets for antivirals. The structure of the Tsg101-UEV domain has been characterized, using X-ray crystallography or NMR spectroscopy, either in its apo-state or bound to ubiquitin or late domains. In this study, we report the backbone NMR resonance assignments, including the proline signals, of the apo human Tsg101-UEV domain, that so far was not publicly available. These data, that are in good agreement with the crystallographic structure of Tsg101-UEV domain, can therefore be used for further NMR studies, including protein-protein interaction studies and drug discovery.

Biological context

Human tumor susceptibility gene 101 protein (Tsg101 protein) is one of the components of the Endosomal Sorting Complex Required for Transport (ESCRT) machinery that is highly conserved in eukaryotes. The ESCRT pathway is involved in the sorting, trafficking and lysosomal degradation of ubiquitinated proteins though the multivesicular bodies (MVB), but it is also involved in many others biological processes, such as membrane recycling, cytokinesis, autophagy or exosome secretion (Williams and Urbé 2007; Henne et al. 2011). In these physiological processes, ESCRTs mediate inverse membrane remodeling with the formation of vesicles which contain cytosol and bud away from it, either at cell surface or inside cellular organelles (Flower et al. 2020).

The ESCRT machinery comprises five different multi-subunit complexes (ESCRT-0, -I, -II, -III and the Vps4 complex) that assemble on the cytosolic side of the membrane (Schmidt and Teis 2012; Vietri et al. 2020). ESCRT-0 initiates the recognition of the ubiquitinated proteins and established interactions with both lipids (phosphatidylinositol3-phosphate, PI3P) and proteins. The hepatocyte growth factor (HGF)-regulated Tyrosine kinase substrate (HRS) from ESCRT-0 establishes a direct protein-protein interaction with Tsg101 that belongs to ESCRT-I. The assembly of Tsg101 with Vps28, Vps37 (A, B, C, D) and Mvb12 (A, B) or ubiquitin-associated protein 1 (UBAP1) constitutes the hetero-tetrameric ESCRT-I. The latter one recruits ESCRT-II that is a hetero-terameric complex with a GLUE domain that can simultaneously bind to Vps28 from ESCRT-I, ubiquitin and PI3P. Then, ESCRT-II recruits ESCRT-III, a multimeric complex made of charged multivesicular body proteins (CHMP) and accessory proteins. This complex, in contrast to the others one, will transiently assemble and constitute the main driving force with the AAA-ATPase Vps4 for membrane constriction and scission (Vietri et al. 2020).

The ESCRT machinery, by controlling the membrane proteins recycling, allows the tight regulation of cell receptors signaling. As a consequence of this central regulation mechanism, altered ESCRT functions have been associated with many human diseases, such as cancers and neurodegenerative diseases (Ferraiuolo et al. 2020). Moreover, it has been shown that the ESCRT machinery is a host factor that is required for the replication cycle of some viruses. Indeed, many enveloped RNA viruses, such as HIV-1 (VerPlank et al. 2001; Garrus et al. 2001), Ebola (Martin-Serrano et al. 2001), human T-lymphotropic virus (Bouamr et al. 2003) or HEV (Nagashima et al. 2011), hijack the ESCRT pathway to bud away from the plasma membrane of infected cells, a process that is similar to vesicles budding into intracellular organelles from the cytosol. Whereas ESCRT-0 and -II seem dispensable, ESCRT-I and -III have been shown to be involved in viral particles maturation and release. A key player in this process is the Tsg101 protein from ESCRT-I.

The Tsg101 protein (Vps23 in yeast) is a ~44 kDa protein that is composed of several domains. From the N-terminus to the C-terminus, there are a ubiquitin E2 variant (UEV) domain, a proline-rich region (PRR), a Stalk domain and a Head domain (Flower et al. 2020). Both the Stalk and Head domains of Tsg101 are components of the ESCRT-I core

that displays an elongated rod-like shape structure (Boura et al. 2011). The N-terminal UEV domain of Tsg101 (residues 2-145) protrudes from the ESCRT-I core to which it is connected via a disordered and flexible PRR. The Tsg101 UEV domain (Tsg101-UEV) binds ubiquitin, but is devoid of enzymatic E2 ligase activity as the catalytic cysteine residue is absent (Pornillos et al. 2002b). This interaction allows ESCRT-I to bind ubiquitinated cargo that have to be processed through the MVB pathway. Next to this ubiquitin-binding pocket, there is a groove at the surface of Tsg101-UEV that interacts with P(S/T)AP peptide sequence. This is how ESCRT-I interacts with ESCRT-0, in which HRS displays a PSAP peptide sequence. Tsg101 biological functions are auto-regulated by an interaction between an internal PTAP motif and its Tsg101-UEV domain (Lu et al. 2003; McDonald and Martin-Serrano 2008). Several viruses, such as HIV-1, Ebola, HEV and Marburg virus (MARV), have been shown to recruit ESCRT-I at their assembly/budding sites through interaction of Tsg101-UEV with viral late domains [P(T/S)AP] (Freed 2002; Pornillos et al. 2002a; Dolnik et al. 2014). These latter ones are located in the HIV-1 Gag, Ebola VP40, HEV ORF3 and MARV NP proteins. This protein-protein interaction thus constitutes an attractive drug target to develop antivirals. Peptidomimetics, including cyclic peptides (Tavassoli et al. 2008; Lennard et al. 2019) and small molecules (Siarot et al. 2018), have indeed been shown to abolish the interaction between HIV-1 Gag and human Tsg101-UEV and to interfere with viral particles release. In addition, prazole-based drugs abolishing the ubiquitin binding function of Tsg101-UEV proved to interfere with the early HIV-1 assembly, independently of its interaction with the HIV-1 Gag PTAP late domain (Strickland et al. 2017). More recently, tenatoprazole and ilaprazole have been shown to inhibit the release of HIV-1 and Herpes Simplex Virus (HSV) 1/2 infectious particles from infected cells (Leis et al. 2021). Two contiguous binding pockets in Tsg101-UEV could thus be targeted to develop broad-spectrum antivirals against enveloped viruses.

Here, we report the backbone ($^{13}C_\alpha$, $^{13}C_\beta$, ^{13}CO, $^{15}N^H$, $^1H^N$, $^1H_\alpha$) assignments, including the proline ($^{13}C_\alpha$, $^{13}C_\beta$, ^{13}CO, ^{15}N ,$^1H_\alpha$) assignments, of human Tsg101-UEV domain in its apo-state. These data could be further used to study the molecular details of Tsg101-UEV interactions with protein partners, as well as its structural/conformational consequences, or for drug screening purpose. In this case, the assignments give the possibility to distinguish the P(S/T)AP and ubiquitin binding pockets.

Protein expression and purification

The DNA sequence coding for the human UEV domain of Tsg101 (residues 1-145) (Uniprot accession number Q99816) was synthesized, with codon optimization for *E. coli*, by GeneCust and inserted into pET28a(+) plasmid between the *Nco*I and *Bam*HI restriction sites to give the pET28a-Tsg101-UEV plasmid. The coding sequence has been designed to produce the Tsg101-UEV domain with an His$_6$-tag at its N-terminus followed by a Tobacco Etch Virus (TEV) cleavage site. Thus, the amino-acid sequence of the recombinant protein corresponds to GSSHHHHHHSSG*ENLYFQ/GA*-(Tsg101-UEV residue 1-145). The pET28a-Tsg101-UEV plasmid was transformed into *E. coli* BL21(DE3) cells for overexpression.

Few colonies from a *Luria-Bertani* (LB)-agar plate supplemented with kanamycin (25 µg/mL) were used to inoculate 25 mL of LB medium. Bacteria were grown overnight at 37°C with shaking. Then, 20 mL of this pre-culture were added in 1 L of an optimized M9 minimal medium containing 1 g/L $^{15}NH_4Cl$, 3 g/L $^{13}C_6$-D-glucose and 0.5 g/L ISOGRO-^{15}N,^{13}C (Sigma-Aldrich) supplemented with kanamycin (25 µg/mL). The cells were grown at 37°C, at 160 rpm, until the OD$_{600}$ reached ~1.0 and then the protein expression was induced by addition of 0.4 mM isopropyl-β-D-thiogalactopyranoside (IPTG). After 4-5 hours at 37°C, the cells were harvested by centrifugation at 5,000 x *g* for 20 min at 4°C. The cell pellet was resuspended in 40 mL of Buffer R (50 mM Na$_2$HPO$_4$/NaH$_2$PO$_4$ (NaPi) pH 7.8, 500 mM NaCl, 10 mM Imidazole) supplemented with EDTA-free protease inhibitor cocktail (cOmplete, Roche, 1 tablet), DNaseI (150 µL at 6 mg/mL) and RNaseA (15 µL at 40 mg/mL). The cell lysis was done using an homogenizer (EmulsiFlexC3, Avestin), with 5 passages at 20,000 psi at 4°C. The cell extract was centrifuged at 39,000 x *g* for 40 min at 4°C. After filtration (0.45 µm), the supernatant was loaded on a HisTrapHP column (1 mL, Cytiva) to purify the isotopically labeled UEV domain of Tsg101 protein, thanks to the N-terminal His$_6$-tag. The protein bound to the affinity column was eluted using a linear gradient with increasing concentration of Imidazole (from 10 to 300 mM Imidazole in 15 column volume) while 0.8 mL elution fractions were collected. The elution fractions were analyzed by SDS-PAGE (4-20%) and Coomassie staining. The fractions containing Tsg101-UEV were pooled, supplemented with

TEV protease (500 µL at 5 mg/mL) and dialyzed (6-8 kDa cut-off) overnight at 15°C against Buffer D (50 mM Tris-HCl pH 6.4, 250 mM NaCl, 5 mM β-mercaptoethanol). This allowed for simultaneous His$_6$-tag cleavage and Imidazole removal. Then, the cleaved protein was loaded on the HisTrapHP column to remove both the cleaved His$_6$-tag and the TEV protease. The Tsg101-UEV domain was then dialyzed (6-8 kDa cut-off) overnight at 4°C against NMR Buffer (50 mM NaPi pH 6.1, 50 mM NaCl, 0.1 mM EDTA) and finally concentrated to 330 µM. The doubly labeled ^{15}N,^{13}C-Tsg101-UEV protein was flash frozen in liquid Nitrogen and stored at -80°C until used.

NMR sample and NMR data acquisition and processing

A doubly labeled ^{15}N,^{13}C Tsg101-UEV sample at 330 µM concentration was placed into a 5 mm Shigemi tube (300 µL) to record all the NMR experiments, at 298 K, using a 600 MHz Bruker spectrometer (Avance III HD) equipped with a CPQCI cryoprobe. The sample was in NMR Buffer (50 mM NaPi pH 6.1, 50 mM NaCl, 0.1 mM EDTA) and was supplemented with 5% (*v/v*) D$_2$O and Trimethyl Silyl Propionate (TMSP).

The classical backbone assignments were obtained from 3D experiments CBCANH, HNCO, CBCACONH, HN(CA)CO, HNHA and HN(CA)NNH, while the assignments of prolines resonances were based on the carbon-detected 3D hCACON and HCAN and the 2D NCO and CACO experiments. All 3D spectra were recorded using non-uniform sampling (Table 1). The data were acquired and processed using Topspin 3.5 (Bruker Biospin). The TMSP signal was used as the ^{1}H chemical shift reference, and the ^{15}N and ^{13}C chemical shifts were indirectly referenced based on ^{1}H chemical shifts. The NMR data were analyzed using NMRFAM-Sparky (Lee et al. 2015) and checked by I-PINE web server (Lee et al. 2019).

Table 1 List of NMR experiments acquired at 600 MHz spectrometer and corresponding parameters that have been used for Tsg101-UEV assignment.

	Time domain data size (points)			Spectral width/Carrier frequency (ppm)			NS	Delay time (s)	NUS points	NUS (%)
	t1	t2	t3	F1	F2	F3				
^{1}H,^{15}N HSQC	2048	256		14/4.7 (^{1}H)	28/117.5 (^{15}N)		16	1	64	50
CBCANH	2048	82	120	14/4.7 (^{1}H)	28/117.5 (^{15}N)	60/42 (^{13}C)	48	1	688	28
CBCACONH	2048	82	120	14/4.7 (^{1}H)	28/117.5 (^{15}N)	60/42 (^{13}C)	48	1	590	24
HNCO	1432	82	128	12/4.7 (^{1}H)	30/118 (^{15}N)	11/173 (^{13}C)	16	0.25	419	16
HNCACO	1426	82	128	12/4.7 (^{1}H)	28/ 117.5 (^{15}N)	14/173.5 (^{13}C)	64	0.25	524	20
HN(CA)NNH	2048	128	128	14/4.7 (^{1}H)	28/117.5 (^{15}N)	28/117.5 (^{15}N)	48	1	491	12
HACAN	2048	128	128	14/4.7 (^{1}H)	40/52 (^{13}C)	28/117.5 (^{15}N)	32	1	737	18
HNHA	2048	256	128	14/4.7 (^{1}H)	12/4.7 (^{1}H)	28/117.5 (^{15}N)	32	1	491	6
hCACON	1024	128	64	40/173.5 (^{13}C)	43/123 (^{15}N)	40/173.5 (^{13}C)	32	1	409	20
^{13}C,^{15}N NCO	1024	160		40/173.5 (^{13}C)	47/123 (^{15}N)		160	1	40	50
^{13}C,^{13}C CACO	724	256		20/173.5 (^{13}C)	50/173.5 (^{13}C)		160	1	32	50

Extent of assignments and data deposition

After TEV cleavage, two extra residues from the purification tag were present at the N-terminus of the protein. These were not taken into account for both the assignment and the sequence numbering purpose. Therefore, the first residue (residue 1) is the methionine (Met) of human Tsg101-UEV domain. The advantage of the presence of these extra residues was that even the resonances corresponding to the first N-terminal residues of the Tsg101-UEV were observed in the ^{1}H,^{15}N HSQC spectrum.

The 2D ^{1}H,^{15}N HSQC spectrum of the apo form of Tsg101-UEV domain (residues 1-145 of the full-length protein) displays well spread resonances with almost no overlap. All backbone ^{1}H-^{15}N resonances in the ^{1}H,^{15}N HSQC spectrum were successfully assigned except the one corresponding to residue Asn45 (N45) that was not visible (Fig. 1, 99% completeness). We also successfully assigned all backbone ^{15}N-^{13}CO resonances in the 2D ^{15}N,^{13}C-NCO spectrum in which a correlation can be observed for each protein residue, including the 13 proline residues with their corresponding resonances located in the ^{15}N 131-143 ppm region of the spectrum (Fig. 2, completeness 100%). Overall, the analysis of our full NMR dataset resulted in the assignment of 131 out of 132 backbone ^{1}H-^{15}N correlations (99%), 138 out of 145 ^{1}H$_\alpha$ resonances (95%), 145 out of 145 backbone ^{15}N^H and ^{13}C$_\alpha$ resonances (100%), 133 out of 138 ^{13}C$_\beta$ resonances (96%), 144 out of 145 ^{13}CO resonances (99%) and 145 out of 145 ^{1}H$_\alpha$ resonances (100%) [2 of the 7 glycine residues have only one ^{1}H$_\alpha$ resonance]. In addition, the ^{1}H-^{15}N correlations, in the ^{1}H,^{15}N HSQC spectrum, arising from side chains of Asn, Gln and Trp residues were also assigned (**Error! Reference source not found.**). Only for Asn54 we could not determine the ^{13}C$_\gamma$, ^{15}N$_{\delta2}$ and ^{1}H$_{\delta2}$ resonances. Regarding the 13 prolines residues, all the ^{15}N, ^{13}C$_\alpha$, ^{1}H$_\alpha$, ^{13}C$_\beta$ and ^{13}CO (except the one from Pro145) signals were assigned and we also assigned 7 out of 13 ^{13}C$_\delta$ and ^{1}H$_{\delta2/3}$ resonances (54%) using the 3D HCAN spectrum. The analysis of the ^{13}C$_\beta$ chemical shift of prolines showed that two proline residues, Pro81 and Pro120, are in *cis* conformation with their ^{13}C$_\beta$ value being close to ~34 ppm. This observation is in agreement with the crystal structure of the Tsg101-UEV domain (PDB entry 2F0R) in which these two prolines are referred as *cis*-peptides (Palencia et al. 2006). Two proline residues, Pro81 and Pro84, display unusual chemical shifts. Based on the crystallographic structure, we attributed this to the close packing with surrounding aromatic residues (Trp117, Tyr80, Tyr82 and His119). Moreover, Pro84 is part of a Pro-Pro motif in the sequence. All the assigned chemical shifts have been deposited in the Biological Magnetic Resonance Bank (https://bmrb.io/) under the accession number 50765.

Fig. 1 Annotated 2D ^{1}H,^{15}N HSQC spectrum of apo human Tsg101-UEV domain. The resonance assignments are shown with black labels. Resonances corresponding to Asn, Gln and Trp side-chains are annotated with a "sc" label

Fig. 2 Annotated 2D carbon-detected $^{15}N,^{13}C$-NCO spectrum of apo human Tsg101-UEV domain. The spectrum is centered on the proline region (132-143 ppm for the ^{15}N dimension). The resonance assignments are shown with black labels

Using the backbone NMR chemical shifts, we analyzed the secondary structure content in the apo Tsg101-UEV domain using both the CSI 3.0 (Hafsa et al. 2015) web server and the Secondary Structure Propensities (SSP) program (Marsh et al. 2006). Both the CSI and SSP methods gave similar results. The results were compared with the secondary structures from the crystal structure of the Tsg101-UEV domain (PDB entry 2F0R) (Palencia et al. 2006) and proved to be well correlated (Fig. 3). A short α-helix (residues 111-116) was not identified using CSI 3.0 because of its lower propensity highlighted by the ~0.4 SSP score (Fig. 3b,c). Moreover, the flexibility of the apo Tsg101-UEV domain in solution was predicted from the experimental NMR chemical shifts using TALOS+ server (Shen et al. 2009). The Random Coil Index (RCI) derived S^2 values for all residues are shown in Fig.3c (blue dots) and matched with the increase flexibility observed for the N- and C-termini and the loops connecting the secondary structures of Tsg101-UEV domain.

The assignments of the apo Tsg101-UEV domain may be used for drug screening purpose or to perform in-depth study of any molecular interaction with others molecular partners, such as peptides or proteins.
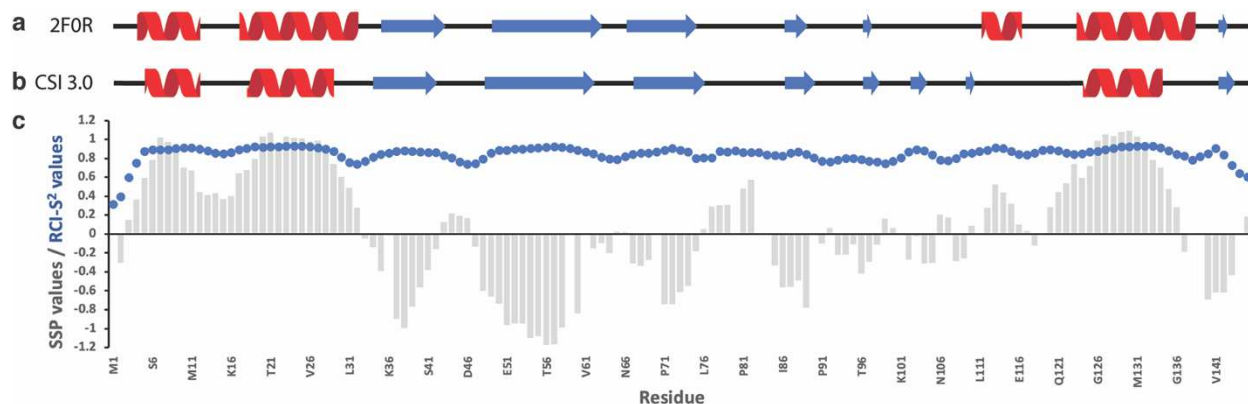
Fig. 3 Secondary structure analysis and flexibility of Tsg101-UEV domain. (a) Secondary structure analysis from the crystal structure of Tsg101-UEV domain (PDB ID: 2F0R) (Palencia et al. 2006; Hafsa et al. 2015). (b) Secondary structure analysis based on the experimental backbone NMR chemical shifts ($^1H^N$, $^{15}N^H$, $^{13}C_\alpha$, $^1H_\alpha$, $^{13}C_\beta$ and ^{13}CO) of the protein using the CSI 3.0 web server (Hafsa et al. 2015). Red cartoon represents the α-helix and blue arrow the β-strand. (c) Secondary Structure Propensities (SSP) score values (grey bars) based on $^{13}C_\alpha$, $^{13}C_\beta$ and $^1H_\alpha$ chemical shifts of human Tsg101-UEV domain (Marsh et al. 2006). Positive and negative values indicate α-helix and extended structure propensities, respectively. The flexibility of the protein is highlighted by the predicted order parameter (Random Coil Index S^2, RCI-S^2, blue dots) based on the backbone chemical shifts ($^1H^N$, $^{15}N^H$, $^{13}C_\alpha$, $^1H_\alpha$, $^{13}C_\beta$ and ^{13}CO) using the TALOS+ server (Shen et al. 2009)

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
All authors have agreed to the publication of the manuscript.

Availability of data and material
All the chemical shift values of apo human Tsg101-UEV domain were deposited in the Biological Magnetic Resonance Data Bank (BMRB) under accession code 50765.

Competing interests
The authors declare that they have no conflict of interest.

Authors' contributions
D.M. and X.H performed protein expression and purification. F-X.C. and E.B. collected the NMR data. D.M. analysed the data and wrote the first draft of the manuscript. D.M. prepared figures 1-3. All authors commented on previous versions of the manuscript. All authors reviewed the manuscript. Funding acquisition and supervision were done by X.H.

Authors' information (optional)
Not applicable

References

Bouamr F, Melillo JA, Wang MQ, et al (2003) PPPYEPTAP Motif Is the Late Domain of Human T-Cell Leukemia Virus Type 1 Gag and Mediates Its Functional Interaction with Cellular Proteins Nedd4 and Tsg101. J Virol 77:11882–11895. https://doi.org/10.1128/JVI.77.22.11882-11895.2003

Boura E, Różycki B, Herrick DZ, et al (2011) Solution structure of the ESCRT-I complex by small-angle X-ray scattering, EPR, and FRET spectroscopy. Proc Natl Acad Sci 108:9437–9442. https://doi.org/10.1073/pnas.1101763108

Dolnik O, Kolesnikova L, Welsch S, et al (2014) Interaction with Tsg101 Is Necessary for the Efficient Transport and Release of Nucleocapsids in Marburg Virus-Infected Cells. PLoS Pathog 10:e1004463. https://doi.org/10.1371/journal.ppat.1004463

Ferraiuolo R-M, Manthey KC, Stanton MJ, et al (2020) The Multifaceted Roles of the Tumor Susceptibility Gene 101 (TSG101) in Normal Development and Disease. Cancers 12:450. https://doi.org/10.3390/cancers12020450

Flower TG, Takahashi Y, Hudait A, et al (2020) A helical assembly of human ESCRT-I scaffolds reverse-topology membrane scission. Nat Struct Mol Biol 27:570–580. https://doi.org/10.1038/s41594-020-0426-4

Freed EO (2002) Viral Late Domains. J Virol 76:4679–4687. https://doi.org/10.1128/JVI.76.10.4679-4687.2002

Garrus JE, von Schwedler UK, Pornillos OW, et al (2001) Tsg101 and the Vacuolar Protein Sorting Pathway Are Essential for HIV-1 Budding. Cell 107:55–65. https://doi.org/10.1016/S0092-8674(01)00506-2

Hafsa NE, Arndt D, Wishart DS (2015) CSI 3.0: a web server for identifying secondary and super-secondary structure in proteins using NMR chemical shifts. Nucleic Acids Res 43:W370–W377. https://doi.org/10.1093/nar/gkv494

Henne WM, Buchkovich NJ, Emr SD (2011) The ESCRT Pathway. Dev Cell 21:77–91. https://doi.org/10.1016/j.devcel.2011.05.015

Lee W, Bahrami A, Dashti HT, et al (2019) I-PINE web server: an integrative probabilistic NMR assignment system for proteins. J Biomol NMR 73:213–222. https://doi.org/10.1007/s10858-019-00255-3

Lee W, Tonelli M, Markley JL (2015) NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. Bioinforma Oxf Engl 31:1325–1327. https://doi.org/10.1093/bioinformatics/btu830

Leis J, Luan C-H, Audia JE, et al (2021) Ilaprazole and Other Novel Prazole-Based Compounds That Bind Tsg101 Inhibit Viral Budding of Herpes Simplex Virus 1 and 2 and Human Immunodeficiency Virus from Cells. J Virol 95:e00190-21. https://doi.org/10.1128/JVI.00190-21

Lennard KR, Gardner RM, Doigneaux C, et al (2019) Development of a Cyclic Peptide Inhibitor of the p6/UEV Protein–Protein Interaction. ACS Chem Biol 14:1874–1878. https://doi.org/10.1021/acschembio.9b00627

Lu Q, Hope LW, Brasch M, et al (2003) TSG101 interaction with HRS mediates endosomal trafficking and receptor down-regulation. Proc Natl Acad Sci 100:7626–7631. https://doi.org/10.1073/pnas.0932599100

Marsh JA, Singh VK, Jia Z, Forman-Kay JD (2006) Sensitivity of secondary structure propensities to sequence differences between α- and γ-synuclein: Implications for fibrillation. Protein Sci 15:2795–2804. https://doi.org/10.1110/ps.062465306

Martin-Serrano J, Zang T, Bieniasz PD (2001) HIV-1 and Ebola virus encode small peptide motifs that recruit Tsg101 to sites of particle assembly to facilitate egress. Nat Med 7:1313–1319. https://doi.org/10.1038/nm1201-1313

McDonald B, Martin-Serrano J (2008) Regulation of Tsg101 Expression by the Steadiness Box: A Role of Tsg101-associated Ligase. Mol Biol Cell 19:754–763. https://doi.org/10.1091/mbc.e07-09-0957

Nagashima S, Takahashi M, Jirintai S, et al (2011) Tumour susceptibility gene 101 and the vacuolar protein sorting pathway are required for the release of hepatitis E virions. J Gen Virol 92:2838–2848. https://doi.org/10.1099/vir.0.035378-0

Palencia A, Martinez JC, Mateo PL, et al (2006) Structure of human TSG101 UEV domain. Acta Crystallogr D Biol Crystallogr 62:458–464. https://doi.org/10.1107/S0907444906005221

Pornillos O, Alam SL, Davis DR, Sundquist WI (2002a) Structure of the Tsg101 UEV domain in complex with the PTAP motif of the HIV-1 p6 protein. Nat Struct Mol Biol 9:812–817. https://doi.org/10.1038/nsb856

Pornillos O, Alam SL, Rich RL, et al (2002b) Structure and functional interactions of the Tsg101 UEV domain. EMBO J 21:2397–2406. https://doi.org/10.1093/emboj/21.10.2397

Schmidt O, Teis D (2012) The ESCRT machinery. Curr Biol 22:R116–R120. https://doi.org/10.1016/j.cub.2012.01.028

Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44:213–223. https://doi.org/10.1007/s10858-009-9333-z

Siarot L, Chutiwitoonchai N, Sato H, et al (2018) Identification of human immunodeficiency virus type-1 Gag-TSG101 interaction inhibitors by high-throughput screening. Biochem Biophys Res Commun 503:2970–2976. https://doi.org/10.1016/j.bbrc.2018.08.079

Strickland M, Ehrlich LS, Watanabe S, et al (2017) Tsg101 chaperone function revealed by HIV-1 assembly inhibitors. Nat Commun 8:1391. https://doi.org/10.1038/s41467-017-01426-2

Tavassoli A, Lu Q, Gam J, et al (2008) Inhibition of HIV Budding by a Genetically Selected Cyclic Peptide Targeting the Gag−TSG101 Interaction. ACS Chem Biol 3:757–764. https://doi.org/10.1021/cb800193n

VerPlank L, Bouamr F, LaGrassa TJ, et al (2001) Tsg101, a homologue of ubiquitin-conjugating (E2) enzymes, binds the L domain in HIV type 1 Pr55Gag. Proc Natl Acad Sci 98:7724–7729. https://doi.org/10.1073/pnas.131059198

Vietri M, Radulovic M, Stenmark H (2020) The many functions of ESCRTs. Nat Rev Mol Cell Biol 21:25–42. https://doi.org/10.1038/s41580-019-0177-4

Williams RL, Urbé S (2007) The emerging shape of the ESCRT machinery. Nat Rev Mol Cell Biol 8:355–368. https://doi.org/10.1038/nrm2162

2. 3CLpro SARS-CoV-2 project

Two research articles on main protease (3CLp) of SARS-CoV-2 project, one titled "NMR spectroscopy of the main protease of SARS-CoV-2 and fragment-based screening identify three protein hotspots and an antiviral fragment." published on "Angewandte Chemie International Edition" journal and one titled "Novel dithiocarbamates selectively inhibit 3CL protease of SARS-CoV-2 and other coronaviruses." published on "European Journal of Medicinal Chemistry" journal.

NMR spectroscopy of the main protease of SARS-CoV-2 and fragment-based screening identify three protein hotspots and an antiviral fragment

François-Xavier Cantrelle, Emmanuelle Boll, Lucile Brier, Danai Moschidi, Sandrine Belouzard, Valérie Landry, Florence B Leroux, Frédérique Dewitte, Isabelle Landrieu, Jean Dubuisson, et al.

Accepted Article

Title: NMR spectroscopy of the main protease of SARS-CoV-2 and fragment-based screening identify three protein hotspots and an antiviral fragment

Authors: François-Xavier Cantrelle, Emmanuelle Boll, Lucile Brier, Danai Moschidi, Sandrine Belouzard, Valérie Landry, Florence Leroux, Frédérique Dewitte, Isabelle Landrieu, Jean Dubuisson, Benoit Deprez, Julie Charton, and Xavier Hanoulle

WILEY-VCH

RESEARCH ARTICLE

NMR spectroscopy of the main protease of SARS-CoV-2 and fragment-based screening identify three protein hotspots and an antiviral fragment

François-Xavier Cantrelle[a][b]#, Emmanuelle Boll[a][b]#, Lucile Brier[c]#, Danai Moschidi[a][b], Sandrine Belouzard[d], Valérie Landry[c], Florence Leroux[c], Frédérique Dewitte[a][b], Isabelle Landrieu[a][b], Jean Dubuisson[d], Benoit Deprez*[c], Julie Charton[c], and Xavier Hanoulle*[a][b]

[a]   Dr.F-X. Cantrelle, E. Boll, D. Moschidi, F. Dewitte, Dr. I. Landrieu, Dr. X. Hanoulle
         CNRS ERL9002 - BSI - Integrative Structural Biology
         50 avenue Halley, F-59658 Villeneuve d'Ascq, Lille, France
         E-mail: xavier.hanoulle@univ-lille.fr
[b]   Dr.F-X. Cantrelle, E. Boll, D. Moschidi, F. Dewitte, Dr. I. Landrieu, Dr. X. Hanoulle
         Univ. Lille, INSERM, CHU Lille, Institut Pasteur de Lille, U1167 - RID-AGE - Risk Factors and Molecular Determinants of Aging-Related Diseases
         1 rue du Professeur Calmette, F-59019, Lille, France
         E-mail: xavier.hanoulle@univ-lille.fr
[c]   L. Brier, V. Landry, Dr. F. Leroux, Pr. B. Deprez, Dr. J. Charton
         Univ. Lille, INSERM, Institut Pasteur de Lille, U1177 - Drugs and Molecules for Living Systems, F-59000, Lille, France; European Genomic Institute for
         Diabetes, EGID, University of Lille
         3 rue du Professeur Laguesse, F-59006, Lille, France
         E-mail: benoit.deprez@univ-lille.fr
[d]   Dr. S. Belouzard, Dr. J. Dubuisson
         Univ. Lille, CNRS, INSERM, CHU Lille, Institut Pasteur de Lille, U1019-UMR 9017 - CIIL - Center for Infection and Immunity of Lille
         1 rue du Professeur Calmette, F-59019, Lille, France
#   The authors contributed equally

         Supporting information for this article is given via a link at the end of the document.

Abstract: The main protease (3CLp) of the SARS-CoV-2, the causative agent for the COVID-19 pandemic, is one of the main targets for drug development. To be active, 3CLp relies on a complex interplay between dimerization, active site flexibility, and allosteric regulation. The deciphering of these mechanisms is a crucial step to enable the search for inhibitors. In this context, using NMR spectroscopy, we studied the conformation of dimeric 3CLp from the SARS-CoV-2 and monitored ligand binding, based on NMR signal assignments. We performed a fragment-based screening that led to the identification of 38 fragment hits. Their binding sites showed three hotspots on 3CLp, two in the substrate binding pocket and one at the dimer interface. **F01** is a non-covalent inhibitor of the 3CLp and has antiviral activity in SARS-CoV-2 infected cells. This study sheds light on the complex structure-function relationships of 3CLp, and constitutes a strong basis to assist in developing potent 3CLp inhibitors.

Introduction

Since the end of 2019, the world faces the global COVID-19 pandemic that represents a major health burden worldwide with strong societal and economic impacts. The etiological agent is the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) with a case fatality rate of ~2%[1]. This virus represents the seventh coronavirus that infects humans and causes the third β-coronavirus outbreak that emerged in the 21st century. Even

though, both vaccines[2–5] and neutralizing antibodies[6–8] are now available to fight against SARS-CoV-2, specific and efficient antivirals against β-coronaviruses are urgently needed to overcome the limited vaccine coverage, variant escapes from antibodies and the future outbreaks.

The RNA genome of SARS-CoV-2 encodes for up to 27 different proteins[9,10]: the structural proteins, the nonstructural proteins (Nsp) and finally several accessory proteins. The Nsp, corresponding to the replicase-transcriptase, are first translated in two polyproteins, pp1a and pp1ab, which are then cleaved by two viral proteases, the main protease (Mpro or 3CLp) and papain-like protease to release 16 functional proteins. 3CLp cleaves at 11 sites (Nsp4-Nsp16), including its own release. Native 3CLp (306 aa) is composed of three domains[11]. Domains I and II are chymotrypsin-like domains with a β-barrel fold and domain III is a 5 α-helices globular domain that is involved in the regulation of 3CLp dimerization. A long linker (L3)[12] connects the domains II and III whereas the N-ter and C-ter (N-terminal and C-terminal) ends are located at the interface between the protomers (Fig. S1). The functional and active SARS-CoV-2 3CLp corresponds to a homodimeric[13] cysteine protease with an unusual catalytic dyad (Cys145, His41). These are buried in a cleft between the domains I and II that is highly conserved among coronaviruses. The recognition sequence, (L,F)Q↓(S,A,G)[14], for the proteolytic cleavage (↓) requires a Gln at position P1 that is a hallmark feature shared by 3CLp of others coronaviruses[15,16], and which in contrast is not present in human proteases[17]. The substrate binding site is made by 4 pockets named S1', S1, S2

RESEARCH ARTICLE

and S4[11,18] formed residues from domains I and II and also by residues from the linker (L3). The active conformation of the active site is further stabilized by Ser1 from the second protomer, which stresses the functional importance of 3CLp dimerization.

The function, conservation, substrate specificity, and the absence of human homologue all contribute to make 3CLp an attractive drug target. Structural biology here plays a tremendous role as it helps to select the ligands, to get the molecular details of their interactions, and to find the proper ways to improve their potency. So far, the 3CLp inhibitors correspond either to compounds that covalently react with the catalytic Cys145, or to non-covalent molecules that bind either in the active site or at several allosteric sites. Among the first category, there are boceprevir[19,20], GC376[21], inhibitors 11a[22] and 13b[11]; and more recently PF-00835231[23,24]. In the second, molecules that bind either in its active site (MUT056399, 23R) or elsewhere on its molecular surface, including two allosteric sites (Pelitinib, AT7519), have been identified[25,26]. Moreover, fragment screening has also been performed to identify fragments[27] that can be grown, linked or merged to develop potent 3CLp inhibitors[28]. Numerous structural biology methods, including crystallography (X-ray[11,18,22,25] or neutrons[29]), mass spectrometry[13,19,27], computational analyses[30,31,31,32], have been used to get a better understanding of the complex structure-function relationships in 3CLp, including the conformational flexibility of its active site[11,12,18,19,25,30,33–36], and then to find or conceive inhibitors.

In this work, we used NMR spectroscopy to study the dimeric SARS-CoV-2 3CLp. We obtained its NMR chemical shift backbone assignment and used these data in a fragment-based screening that led to the identification of 38 fragment hits. The deciphering of their binding sites and the conformational consequences they induced in 3CLp led to the identification of 3 protein hotspots, two located in the active site of the protease, with two different NMR signatures, and one at the dimerization interface. We further show that the fragment lead **F01** binds in the active site and is, without optimization, a reversible 3CLp inhibitor

with antiviral activity in SARS-CoV-2 infected Vero-81 cells. The crystal structure of F01-bound 3CLp that we have solved will help its optimization. These NMR data should help to get a better understanding of the complex interplays between the active site plasticity, the dimerization and the enzymatic activity of 3CLp. This also constitutes a new tool to assist the development of potent 3CLp inhibitors for the present or future outbreaks.

Results and Discussion

NMR spectroscopy of SARS-CoV-2 3CLp dimer. We produced SARS-CoV-2 3CLp samples with different isotopic labeling schemes to study by liquid-state NMR spectroscopy. The purified protease (306 aa, 67.6 kDa) has both native N- and C-terminal ends (SI and Fig. S1), which is crucial for both its enzymatic activity and its proper dimerization. We obtained good quality ^1H,^{15}N-TROSY HSQC spectrum, with ~280 resonances (Figure 1) and then recorded 3D ^1H,^{15}N,^{13}C TROSY- HNCACB, -HN(CO)CACB, -HNCO, -HN(CA)CO, -HN(CO)CA spectra. Due to unfavorable magnetic relaxation properties some ^{13}C signals were not observed and thus we had to record additional data on other samples, including 3CLp bound to boceprevir, and a monomeric 3CLp R298A mutant, in order to reduce the protein dynamics and the molecular weight, respectively. To perform the NMR backbone assignments of SARS-CoV-2 3CLp, we used a combined and integrated strategy that includes classical sequential assignment, analyses of chemical shift perturbations (CSPs) upon boceprevir binding, CS predictions and previous NMR assignments for the isolated N-ter and C-ter domains of SARS-CoV 3CLp[37] (see SI). We assigned 183 proton amide correlations (183/293, 63%) and further obtained 239, 207 and 234 chemical shifts for Cα, Cβ and C', respectively (Figure 1, SI, BMRB entry 50780).
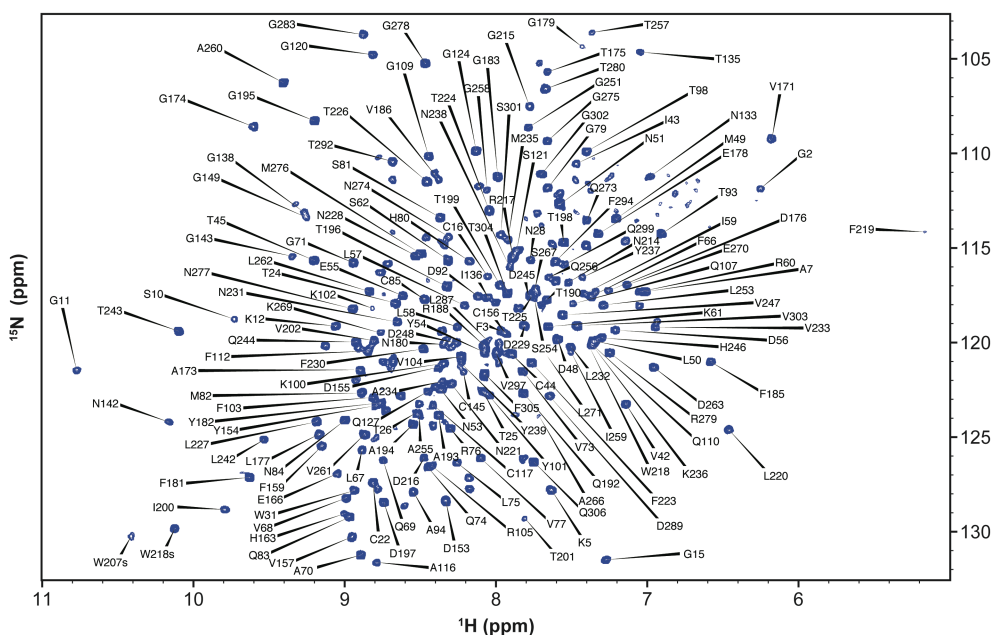
Figure 1. 2D ^1H,^{15}N-TROSY-HSQC NMR spectrum of SARS-CoV-2 3CLp dimer. The assignments are annotated with black labels.

2

RESEARCH ARTICLE

Most of the unassigned proton amides lie in the first two β-barrel domains or at the dimerization interface (Fig. S2). Whereas previous attempts to record multidimensional NMR data on SARS-CoV[38] and SARS-CoV-2[39] 3CLp have failed, these new NMR data open the field to a large range of future studies of the dimeric 3CLp in solution and at temperature close to physiological, an important parameter when considering dynamics. To assess the potential of our experimental system, we analyzed the 3CLp spectral perturbations upon binding of either boceprevir or GC376 (Figs. S3-S4). In both cases, the perturbations induced are highest in the active site but also propagate further in the two catalytic domains, and even toward its C-terminal end with GC376. NMR perturbations may arise from ligand binding but also from the subsequent conformational changes. GC376 indeed induces perturbations both at the active site and at the dimerization interface, the two regions of the protease that are targeted to develop inhibitors[13,25,27,40]. Moreover, in the presence of GC376, a few 3CLp NMR resonances split into two new ones (Fig. S5), probably highlighting the two conformations of the P3 moiety of the bound inhibitor[20]. The split resonances notably match with Val42, Asn142, Gln192 and Gly2. The later one showing that we can detect inter-protomer conformational consequences. Interestingly, when using a R298A 3CLp monomeric mutant, we observed ~115 additional resonances in the 2D ^{1}H,^{15}N NMR spectrum that is ~100 more than expected. This could be due to the two orientations of the domain III that have been described for SARS-CoV 3CLp R298A[41]. These data highlight the potential for in-solution studies of the 3CLp. Based on the NMR assignments we are able to not only detect ligand binding and map the binding site(s), but also to analyze the conformational rearrangement(s) throughout the dimer, providing essential molecular detail for medicinal chemistry.
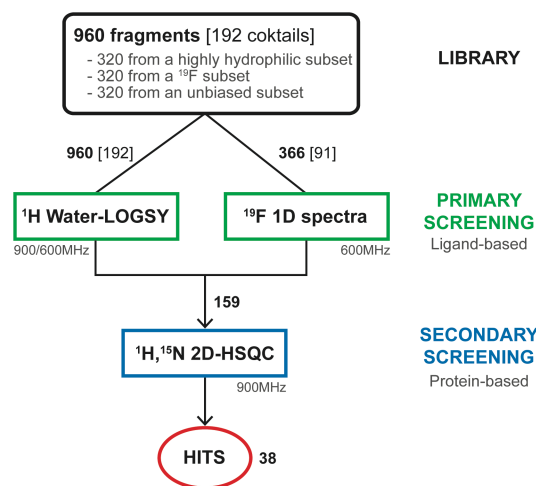
Figure 2. NMR fragment screening.

NMR Fragment-based screening set up. Fragment screening is widely used in drug discovery as it allows to efficiently probe the chemical space while keeping reasonable the numbers of molecule that have to be assessed[28]. The fragment hits identified (low MW) that bind to the target are then optimized to give lead compounds. We used a library of 960 commercially available fragments with physio-chemical properties that mostly fulfill the

'rule of three' criteria[42] (Fig. S7a-d). We designed a strategy with a primary and a secondary screening using ligand- and protein-observed NMR methods, respectively (Figure 2). The screening steps were performed in the presence of DTT, a nucleophile and reducing agent, to minimize the selection of highly electrophilic and nonspecific compounds that would covalently bind to the protease.

^{1}H and ^{19}F NMR ligand-based primary screening. The 960 fragments were split into 192 cocktails of 5 fragments, as this strategy already proved efficient[43]. All the cocktails have been analyzed with ^{1}H Water-LOGSY[44] and additionally with ^{19}F spectroscopy for 91 of them (Figure 2), as our library contains 427 fluorine fragments in total. With Water-LOGSY, the detection of the hits is straightforward since their signals have opposite phase (Figure 3a). When using ^{19}F spectroscopy, the spectra only contain one NMR signal for each ^{19}F-fragment present in the cocktail and we monitored both CSPs and signal broadening (Figure 3b). The primary screening led to the identification of 159 binders (Scheme S1), corresponding to a 16.6 % hit rate (Figure 2).
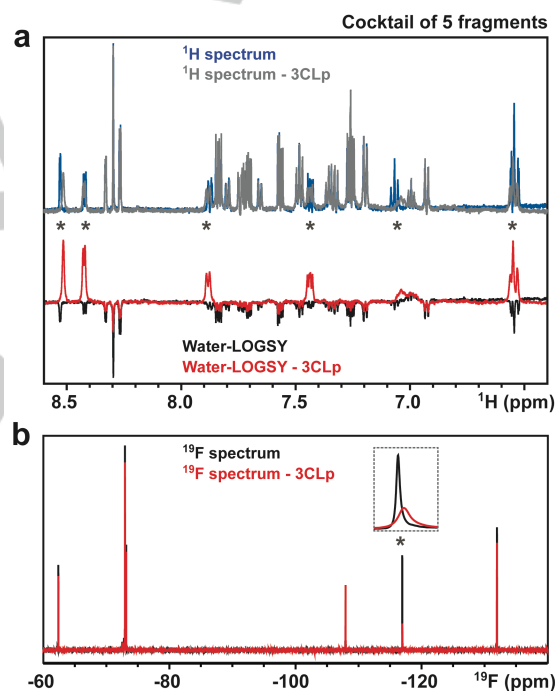
Figure 3. Ligand-based NMR primary screening. Analyses of a 5-fragment cocktail in the absence and in the presence of unlabeled 3CLp. (a) 1D ^{1}H and ^{1}H Water-LOGSY spectra. (b) 1D ^{19}F-NMR of the same cocktail. Signals, annotated with an asterisk, correspond to the F04 fragment that is a direct binder. See Scheme S1 for other cocktails.

Secondary screening using NMR spectra of 3CLp. We performed the secondary screening using 2D ^{1}H,^{15}N TROSY-HSQC spectra that have been acquired on SARS-CoV-2 3CLp in the presence of each of the 159 binders identified in the primary screening. Using both CSPs and signal broadening (Figure 4), we confirmed 38 fragments as direct binders of 3CLp, corresponding to an overall ~4% hit rate (Figures 2 and 4 and Scheme S2, Tables S1-S2). This value can be compared with the ~6% obtained in a combined MS and X-ray approach[27]. The ratio of

3

RESEARCH ARTICLE
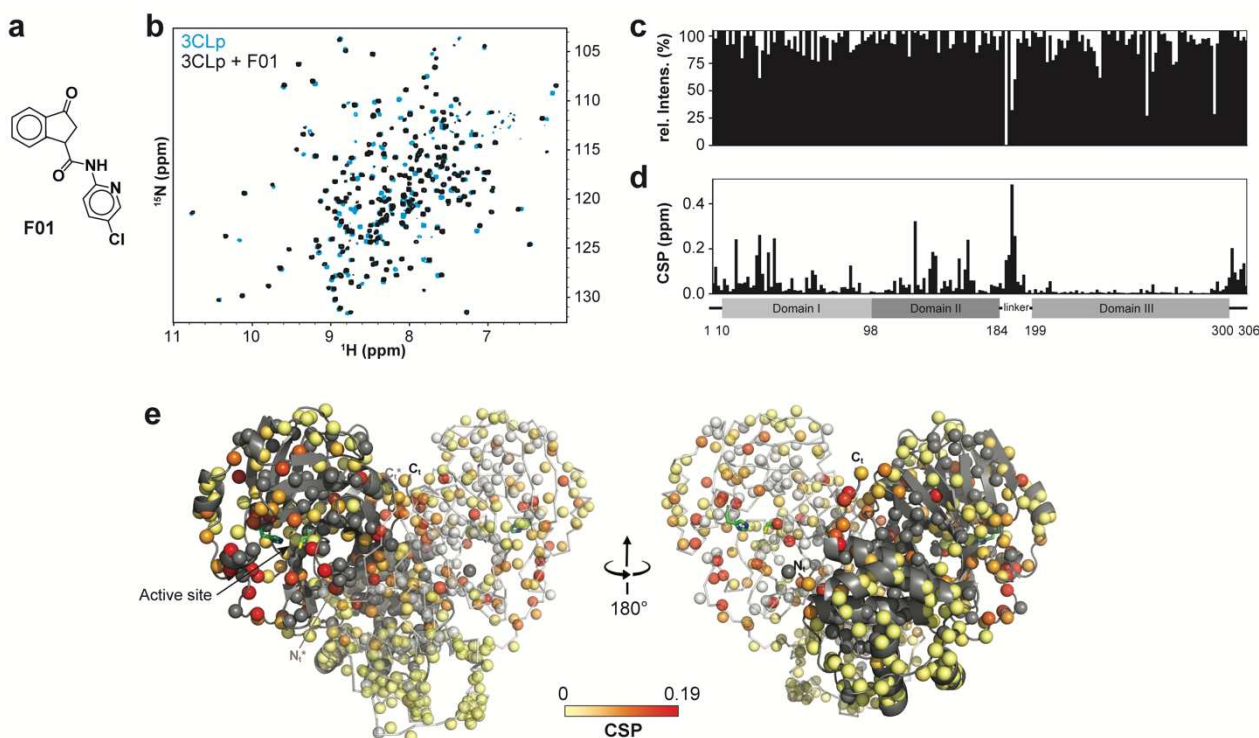
Figure 4. Protein-based NMR secondary screening. (a) Fragment **F01** structure. (b) Overlay of two 2D ^{1}H,^{15}N-TROSY-HSQC spectra acquired on 3CLp in the absence (in light blue) and in the presence (in black) of fragment **F01**. The broadening of the resonances and ^{1}H and ^{15}N-combined CSPs induced upon fragment binding are shown along the 3CLp sequence in (c) and (d), respectively. (e) Structure of the 3CLp dimer (PDB: 7k3t), with protomers A and B shown in grey and white, respectively. Each small ball represents a proton amide and thus should correspond to a resonance in the ^{1}H,^{15}N 2D spectrum. The CSPs, shown in (d), have been color coded (from light yellow to red) and are displayed on these balls. Unassigned residues were kept in the original color of the protomer. Catalytic His41 and Cys145, are shown in green. See Scheme S2 and Table S2 for other hits. See the SI for a color-blind-friendly version of this figure.

^{19}F-containing fragments in the hits (~40%) is close to the ratio in the library used. In contrast, both the average MW and lipophilicity of the fragment hits are higher than those in the entire library (Fig. S7a-d)

Identification of three different classes of binders corresponding to three protein hotspots. Using the backbone assignments, the analysis of the CSPs induced by the 38 hits shows that they can be grouped into three classes corresponding to three 3CLp hotspots (Figure 5; Fig. S8). In Class I (24 hits), CSPs are observed for resonances assigned to residues distributed in the active site cleft, in the loop L3, and in the C-ter end, whereas residues from the N-ter end are only moderately affected. Class II is made by 8 hits that induce CSPs for only a restricted set of residues, in the substrate binding site, that belong exclusively to either the domain I or the tip of the loop L3 and that corresponds to the S2 and S3 binding sites. Class III (5 hits) is defined by CSPs for residues located at the dimerization interface of 3CLp (N-ter and C-ter ends). As to the fragment F27, it induces a strong reduction in the signal intensity all along the 3CLp sequence (Fig. S8 and Table S2), and may correspond to a false positive.

Analysis of the 3CLp binding hotspots. The CSPs pattern in Class I, illustrated by fragment **F01**, is similar to the ones observed in 3CLp upon binding of either boceprevir or GC376 (Fig. S8), two potent inhibitors. The NMR CSPs induced upon binding of **F01** (see Figure 4) correspond to residues distributed all along the 3CLp active site cleft (S1-S4 pockets) and indeed match with

the residues involved in the binding of GC376 (Fig. S9a). Moreover, the CSPs propagate toward the 3CLp dimerization interface, as with GC376 (Fig. S4).

These NMR data are fully supported by the crystal structure of fragment **F01**-bound 3CLp that we solved (Figure 6 and Fig. S10 and Table S3; PDB: 7p51). The 3-oxo-2,3-dihydro-indene ring and 5-chloro-2-pyridyl group of **F01** occupy the S1 and S2 pockets of 3CLp, respectively. Three hydrogen bonds are formed between **F01** and 3CLp. One of them involves the ketone in the indene ring of **F01** that is electrophilic and could covalently react with the catalytic Cys145. This group, located in a key position of the active site, rather behaves as a H-bond acceptor and interacts with His163 (see SI). The binding of **F01** induces conformational changes in all the active site of 3CLp (see SI, Figure 6 and Fig. S10b). It induces the displacement of: the α-helix (Ser46-Leu50) around the S2 pocket, the loop L3 and of Asn142 and Glu166 residues around the S1 pocket. This last movement propagates to the 3CLp dimeric interface with Ser1 of protomer B being slightly displaced. It has been shown that in the 3CLp dimer, Ser1 from protomer B interacts with Glu166 of protomer A and stabilizes the active conformation of the S1 pocket[11,18]. Thus, the CSPs observed in 3CLp spectrum upon **F01** binding both match with its binding site and the induced conformational changes through allosteric pathways (Fig. S10c).

Our data show that conformational plasticity[29,36] and allosteric regulations [13,25,35] within 3CLp can be studied using NMR

4
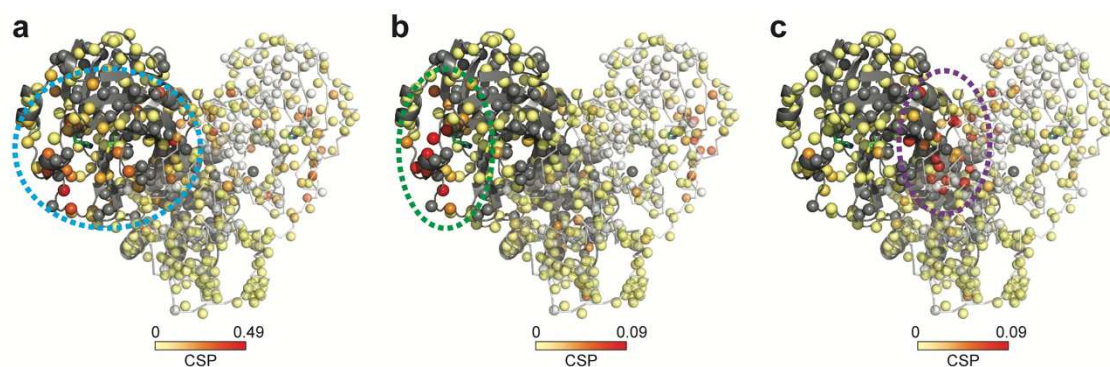
Figure 5. The 38 hits identified in the NMR screening can be grouped into three classes according to the CSPs they induced on the 2D NMR spectrum of ^{2}H,^{15}N-3CLp upon binding. The representation is similar to that in Figure 4e. (a) Class I (**F01**) - The CSPs are distributed in all the active site cleft, including the S1-S4 substrate pockets, and extend toward the dimerization interface of the protease. (b) Class II (**F30**) - The CSPs induced correspond to a binding of the fragments in the S2 and S3 pockets, with the highest perturbations observed for residues located in a short α-helix (Ser46-Leu50). (c) Class III (**F15**) - Upon binding these fragments induce CSPs at the dimerization interface of 3CLp. See Figs. S8-S9. See the SI for a color-blind-friendly version of this figure.

spectroscopy, especially the tight interplay between substrate binding, active site conformation and dimerization.

The hits from Class II, such as **F30**, induced CSPs that would correspond to their binding into the S2 and S3 pockets located in the domain I-side of the 3CLp substrate binding site, as SEN1269[25] (Fig. S9b). This molecule binds to S2 and induced the displacement of the short α-helix (Ser46-Leu50), for which we observed the highest CSPs upon binding of Class II hits (Fig. S8). The NMR CSPs induced upon binding of the Class III hits, which includes **F15**, are localized at the 3CLp dimeric interface and could be predicted to resemble the binding of x1086 and x1187[13,27] in the hydrophobic pocket made by residues both in the N-ter (Met6, Phe8) and C-ter (Arg298, Gln299, Val303) ends (Fig. S9c). With **F15**, we also observed a high CSP for the resonance corresponding to Gln127, which is at the dimeric interface, and that has been shown to make a hydrogen bond with x1086.

Interestingly, no NMR perturbations observed in our screening match with fragment binding into the allosteric sites 1 and 2 that

have been identified by Günther *et al*[25]. It could be that the binding in these two sites requires bigger and more complex molecule structures, or simply that the fragment library used did not allow to probe all the possible binding sites.

Looking at the chemical properties of the fragment hits on the basis of their Class I, II or III belonging, we found that in average Class II hits are smaller than Class I hits (avg. 233.3 Da *vs* 245.7 Da), and that Class III hits are even smaller (avg. 206.85 Da) and are also in more hydrophobic (80% with 2<AlogP<3) (Fig. S7).

Among the 38 hits identified in this work, **F01** induced the highest CSPs in the NMR spectrum of SARS-CoV-2 3CLp (Table S2 and Scheme S2).

F01 is a reversible inhibitor of 3CLp and has antiviral activity against SARS-CoV-2. We further characterized **F01**, the main hit of our screening. First, using NMR titration experiments, we determined a dissociation constant K_D = 73±14 µM for the interaction between **F01** and 3CLp (Figure 7a and Fig. S11).
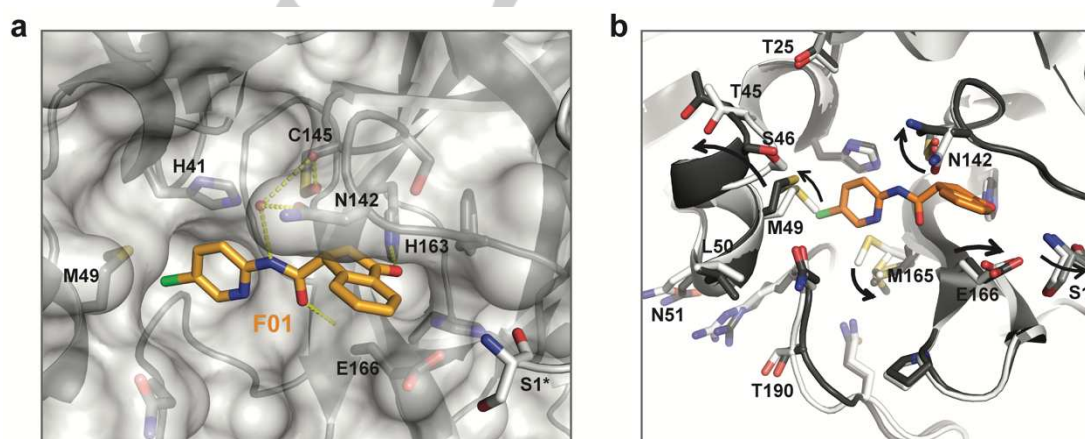
Figure 6. Crystal structure of the fragment **F01**-bound 3CLp. (a) Close-up view of the **F01** binding in the active site. Protomer A is shown in grey and with surface representation, whereas protomer B is displayed in white and in cartoon representation. Three hydrogen bonds between **F01** and 3CLp are displayed as yellow dashes. Residue from protomer A are labeled in black and residue S1 from protomer B is marked with an asterisk. (b) Conformational changes in the **F01**-bound 3CLp structure (PDB: 7p51) compared to the apo 3CLp structure (PDB: 7nts). See Fig. S10.

5

RESEARCH ARTICLE
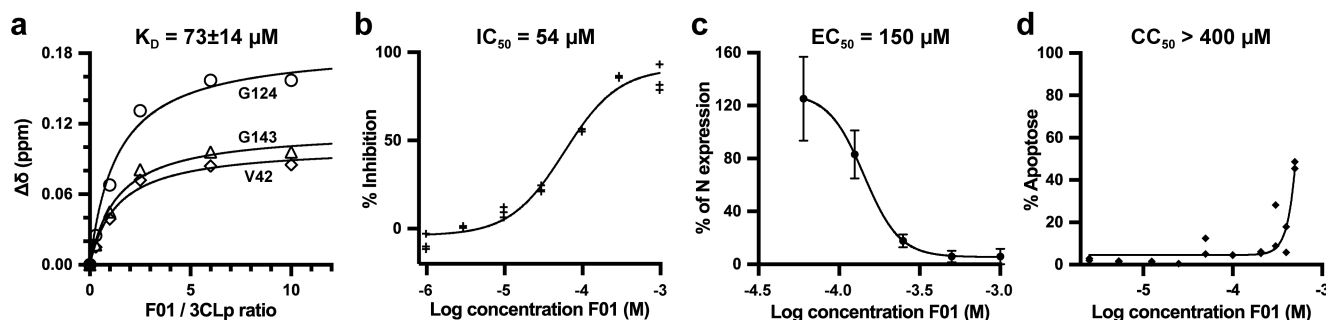
Figure 7. **F01** is an inhibitor of 3CLp and is active against SARS-CoV-2 in Vero-81 cells. (a) Affinity of the Interaction between **F01** and 3CLp. NMR titration curves where the ^{1}H,^{15}N-combined CSPs ($\Delta\delta$, ppm) were plotted as a function of the F01/3CLp ratios. The K_D value (µM) corresponds to the mean (±SD) calculated over 18 3CLp resonances (Fig. S11). (b) **F01** inhibits the *in vitro* enzymatic activity of 3CLp. The half-maximal inhibitory concentration (IC$_{50}$) has been calculated using the initial velocities of the reactions. (c) The antiviral activity of **F01** against SARS-CoV-2 has been tested on Vero-81 infected cells. After infection in the presence of increasing **F01** concentrations, the cells were lysed (t=16h) and the viral N-protein content was quantified and was used to determine the half-maximal effective concentration (EC$_{50}$). Viral titers were also measured in the cell supernatants (Fig. S13). (d) The 50% cytotoxic concentration (CC$_{50}$) of **F01** has been assayed on Vero-81 cells (t=20h).

This affinity is higher than expected, as initial hits from fragment-based screening usually bind to their target with a low affinity, in the 1-10 mM range[45]. Second, using an *in-vitro* enzymatic assay, we showed that **F01** is an inhibitor of 3CLp with a moderate potency (IC$_{50}$ = 54 µM) (Figure 7b). Third, using jump dilution assay, we showed that **F01** is a reversible inhibitor of the protease (Fig. S12), which agrees with the crystal structure (see Figure 6). Finally, Vero-81 cells were infected with SARS-CoV-2 in the presence of increasing concentrations of **F01** and then both the viral N protein cellular content was assayed and the number of infectious viral particles was determined in the cell supernatants. The results showed that **F01** has antiviral activity (EC$_{50}$ = 150 µM) against SARS-CoV-2 (Figure 7c and Fig. S13) while displaying a low cytotoxicity (CC$_{50}$ > 400 µM) (Figure 7d).

Usually, the initial fragment hits have neither *in vitro* nor biological activity, as they often are too small and bind to their target with very low affinity. In this work, we identified the fragment **F01** that even without optimization has antiviral activity against SARS-CoV-2. Very recently, Bajusz *et al*. have reported a fragment, SX013, that blocks the SARS-CoV-2 replication in Vero E6 cells with an EC$_{50}$ of 304 µM[46], which is double of that for **F01** in Vero-81 cells. The ligand efficiency of **F01** is 0.29-0.30 kcal.mol^{-1}.heavy atom^{-1} showing that **F01** is a good fragment lead and deserved to be optimized in order to increase its potency and other drug related properties[28,47].

Conclusion

Whereas structural biology plays a central role in drug discovery and drug development, up to date, NMR spectroscopy has not successfully been pushed forward to study the 3CLp from coronaviruses[37–39]. In this work, we used solution-state NMR spectroscopy to study the dimeric 3CLp protease of the SARS-CoV-2, which is one of the main targets to develop efficient antivirals to fight against the COVID-19 pandemic. Considering the high sequence conservation between the 3CLps[20,40], our data will also be valuable for others β-coronaviruses, such as MERS-CoV and SARS-CoV (67% and 98% sequence similarity, respectively), and possibly for future emerging β-coronaviruses. Even being incomplete, the 3CLp backbone chemical shift

assignment, obtained at pH and temperature close to physiological ones, has proved to be highly valuable in a medicinal chemistry project as these new NMR data allowed the study of both the structure and conformation of the dimeric protease. As a complement to the molecular dynamics[12,30,35,48], these data also provide, for future studies, an experimental mean to assess the 3CLp dynamics in solution, an important point to consider in drug development.

Since mid-2020, the world faces the apparition of SARS-CoV-2 variants that may, at least partially, escape to currents vaccines. This stresses that there is a need for direct acting antiviral(s) and also that there is a high risk for emergence of resistance mutations in 3CLp if targeted. To help resolve this common issue in drug development, a promising strategy consists in the combination of both orthosteric and allosteric drugs[49,50]. In this way, our NMR data could be valuable to identify both the allosteric sites of SARS-CoV-2 3CLp and the molecules that bind into, and to identify the allosteric pathways along which resistance mutations may also occur.

Using a two-step fragment screening, we identified 38 hits, including the promising fragment **F01**, and three binding sites, or hotspots, located in the active site and at the dimerization interface of 3CLp. It has been shown that 3CLp can indeed be efficiently targeted at its active site, at its dimerization interface and even at different allosteric sites[11,13,14,18,22,24,25,27,29,51]. We showed that **F01** binds to 3CLp active site with a rather good affinity (K_D = 73 µM), is a non-covalent reversible inhibitor of the protease (IC$_{50}$ = 54 µM) and demonstrates antiviral activity against SARS-CoV-2 (EC$_{50}$ = 150 µM), despite no optimization. Our results indicates that **F01** is a promising fragment lead that deserved to be optimized to give more potent compounds[28,52]. Structure-activity relationship studies, guided by the the crystal structure, will help this process and two approaches could be considered: first, **F01** (Class I) could be linked or merged to Class II hits, and second, **F01** could be studied in combination with fragments from Class III that bind at the dimerization interface.

This work and our NMR results will benefit to the better understanding of the complex structure-function relationships in the dimer of 3CLp and assist the rational design of potent 3CLp inhibitors, that may both block its active site and interfere with its dimerization, in order to tackle current, or even future, coronaviruses pandemics.

6

RESEARCH ARTICLE

Accession codes

Backbone NMR assignments of SARS-CoV-2 3CLp have been deposited in the Biological Magnetic Resonance Data Bank (Entry 50780).

Crystal structures of apo 3CLp and 3CLp in complex with fragment F01 have been deposited in the Protein Data Bank as entries 7NTS and 7P51, respectively.

Acknowledgements

Keywords: NMR spectroscopy • Viruses • Protein structure • Fragment screening • Drug discovery

[1] World Health Organization, "WHO Coronavirus (COVID-19) Dashboard," can be found under https://covid19.who.int, **2021**.

[2] A. Mullard, *Lancet* **2020**, *395*, 1751–1752.

[3] F. P. Polack, S. J. Thomas, N. Kitchin, J. Absalon, A. Gurtman, S. Lockhart, J. L. Perez, G. Pérez Marc, E. D. Moreira, C. Zerbini, R. Bailey, K. A. Swanson, S. Roychoudhury, K. Koury, P. Li, W. V. Kalina, D. Cooper, R. W. Frenck, L. L. Hammitt, Ö. Türeci, H. Nell, A. Schaefer, S. Ünal, D. B. Tresnan, S. Mather, P. R. Dormitzer, U. Şahin, K. U. Jansen, W. C. Gruber, C4591001 Clinical Trial Group, *N Engl J Med* **2020**, *383*, 2603–2615.

[4] L. R. Baden, H. M. El Sahly, B. Essink, K. Kotloff, S. Frey, R. Novak, D. Diemert, S. A. Spector, N. Rouphael, C. B. Creech, J. McGettigan, S. Khetan, N. Segall, J. Solis, A. Brosz, C. Fierro, H. Schwartz, K. Neuzil, L. Corey, P. Gilbert, H. Janes, D. Follmann, M. Marovich, J. Mascola, L. Polakowski, J. Ledgerwood, B. S. Graham, H. Bennett, R. Pajon, C. Knightly, B. Leav, W. Deng, H. Zhou, S. Han, M. Ivarsson, J. Miller, T. Zaks, COVE Study Group, *N Engl J Med* **2021**, *384*, 403–416.

[5] M. Voysey, S. A. C. Clemens, S. A. Madhi, L. Y. Weckx, P. M. Folegatti, P. K. Aley, B. Angus, V. L. Baillie, S. L. Barnabas, Q. E. Bhorat, S. Bibi, C. Briner, P. Cicconi, A. M. Collins, R. Colin-Jones, C. L. Cutland, T. C. Darton, K. Dheda, C. J. A. Duncan, K. R. W. Emary, K. J. Ewer, L. Fairlie, S. N. Faust, S. Feng, D. M. Ferreira, A. Finn, A. L. Goodman, C. M. Green, C. A. Green, P. T. Heath, C. Hill, H. Hill, I. Hirsch, S. H. C. Hodgson, A. Izu, S. Jackson, D. Jenkin, C. C. D. Joe, S. Kerridge, A. Koen, G. Kwatra, R. Lazarus, A. M. Lawrie, A. Lelliott, V. Libri, P. J. Lillie, R. Mallory, A. V. A. Mendes, E. P. Milan, A. M. Minassian, A. McGregor, H. Morrison, Y. F. Mujadidi, A. Nana, P. J. O'Reilly, S. D. Padayachee, A. Pittella, E. Plested, K. M. Pollock, M. N. Ramasamy, S. Rhead, A. V. Schwarzbold, N. Singh, A. Smith, R. Song, M. D. Snape, E. Sprinz, R. K. Sutherland, R. Tarrant, E. C. Thomson, M. E. Török, M. Toshner, D. P. J. Turner, J. Vekemans, T. L. Villafana, M. E. E. Watson, C. J. Williams, A. D. Douglas, A. V. S. Hill, T. Lambe, S. C. Gilbert, A. J. Pollard, Oxford COVID Vaccine Trial Group, *Lancet* **2021**, *397*, 99–111.

[6] C. Wang, W. Li, D. Drabek, N. M. A. Okba, R. van Haperen, A. D. M. E. Osterhaus, F. J. M. van Kuppeveld, B. L. Haagmans, F. Grosveld, B.-J. Bosch, *Nat Commun* **2020**, *11*, 2251.

[7] X. Chen, R. Li, Z. Pan, C. Qian, Y. Yang, R. You, J. Zhao, P. Liu, L. Gao, Z. Li, Q. Huang, L. Xu, J. Tang, Q. Tian, W. Yao, L. Hu, X. Yan, X. Zhou, Y. Wu, K. Deng, Z. Zhang, Z. Qian, Y. Chen, L. Ye, *Cell Mol Immunol* **2020**, *17*, 647–649.

[8] Y. Cao, B. Su, X. Guo, W. Sun, Y. Deng, L. Bao, Q. Zhu, X. Zhang, Y. Zheng, C. Geng, X. Chai, R. He, X. Li, Q. Lv, H. Zhu, W. Deng, Y. Xu, Y. Wang, L. Qiao, Y. Tan, L. Song, G. Wang, X. Du, N. Gao, J. Liu, J. Xiao, X.-D. Su, Z. Du, Y. Feng, C. Qin, C. Qin, R. Jin, X. S. Xie, *Cell* **2020**, *182*, 73-84.e16.

[9] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, Y.-Z. Zhang, *Nature* **2020**, *579*, 265–269.

[10] Y. Chen, Q. Liu, D. Guo, *J Med Virol* **2020**, *92*, 418–423.

[11] L. Zhang, D. Lin, X. Sun, U. Curth, C. Drosten, L. Sauerhering, S. Becker, K. Rox, R. Hilgenfeld, *Science* **2020**, *368*, 409–412.

[12] N. Verma, J. A. Henderson, J. Shen, *J. Am. Chem. Soc.* **2020**, *142*, 21883–21890.

[13] T. J. El-Baba, C. A. Lutomski, A. L. Kantsadi, T. R. Malla, T. John, V. Mikhailov, J. R. Bolla, C. J. Schofield, N. Zitzmann, I. Vakonakis, C. V. Robinson, *Angewandte Chemie International Edition* **2020**, *59*, 23544–23548.

[14] W. Rut, K. Groborz, L. Zhang, X. Sun, M. Zmudzinski, B. Pawlik, X. Wang, D. Jochmans, J. Neyts, W. Młynarski, R. Hilgenfeld, M. Drag, *Nature Chemical Biology* **2020**, DOI 10.1038/s41589-020-00689-z.

[15] K. Fan, P. Wei, Q. Feng, S. Chen, C. Huang, L. Ma, B. Lai, J. Pei, Y. Liu, J. Chen, L. Lai, *Journal of Biological Chemistry* **2004**, *279*, 1637–1642.

[16] C.-P. Chuck, H.-F. Chow, D. C.-C. Wan, K.-B. Wong, *PLoS One* **2011**, *6*, e27228.

[17] L. Zhang, D. Lin, Y. Kusov, Y. Nian, Q. Ma, J. Wang, A. von Brunn, P. Leyssen, K. Lanko, J. Neyts, A. de Wilde, E. J. Snijder, H. Liu, R. Hilgenfeld, *J. Med. Chem.* **2020**, *63*, 4562–4578.

[18] Z. Jin, X. Du, Y. Xu, Y. Deng, M. Liu, Y. Zhao, B. Zhang, X. Li, L. Zhang, C. Peng, Y. Duan, J. Yu, L. Wang, K. Yang, F. Liu, R. Jiang, X. Yang, T. You, X. Liu, X. Yang, F. Bai, H. Liu, X. Liu, L. W. Guddat, W. Xu, G. Xiao, C. Qin, Z. Shi, H. Jiang, Z. Rao, H. Yang, *Nature* **2020**, *582*, 289–293.

[19] C. Ma, M. D. Sacco, B. Hurst, J. A. Townsend, Y. Hu, T. Szeto, X. Zhang, B. Tarbet, M. T. Marty, Y. Chen, J. Wang, *Cell Res* **2020**, *30*, 678–692.

[20] L. Fu, F. Ye, Y. Feng, F. Yu, Q. Wang, Y. Wu, C. Zhao, H. Sun, B. Huang, P. Niu, H. Song, Y. Shi, X. Li, W. Tan, J. Qi, G. F. Gao, *Nat Commun* **2020**, *11*, 4417.

[21] W. Vuong, M. B. Khan, C. Fischer, E. Arutyunova, T. Lamer, J. Shields, H. A. Saffran, R. T. McKay, M. J. van Belkum, M. A. Joyce, H. S. Young, D. L. Tyrrell, J. C. Vederas, M. J. Lemieux, *Nature Communications* **2020**, *11*, DOI 10.1038/s41467-020-18096-2.

[22] W. Dai, B. Zhang, X.-M. Jiang, H. Su, J. Li, Y. Zhao, X. Xie, Z. Jin, J. Peng, F. Liu, C. Li, Y. Li, F. Bai, H. Wang, X. Cheng, X. Cen, S. Hu, X. Yang, J. Wang, X. Liu, G. Xiao, H. Jiang, Z. Rao, L.-K. Zhang, Y. Xu, H. Yang, H. Liu, *Science* **2020**, *368*, 1331–1335.

[23] R. L. Hoffman, R. S. Kania, M. A. Brothers, J. F. Davies, R. A. Ferre, K. S. Gajiwala, M. He, R. J. Hogan, K. Kozminski, L. Y. Li, J. W. Lockner, J. Lou, M. T. Marra, L. J. Mitchell, B. W. Murray, J. A. Nieman, S. Noell, S. P. Planken, T. Rowe, K. Ryan, G. J. Smith, J. E. Solowiej, C. M. Steppan, B. Taggart, *J Med Chem* **2020**, DOI 10.1021/acs.jmedchem.0c01063.

[24] B. Boras, R. M. Jones, B. J. Anson, D. Arenson, M. A. Bakowski, N. Beutler, J. Binder, E. Chen, H. Eng, J. Hammond, R. Hoffman, E. P. Kadar, R. Kania, M. G. Kirkpatrick, L. Lanyon, E. K. Lendy, J. R. Lillis, S. A. Luthra, C. Ma, S. Noell, R. S. Obach, M. N. O. Brien, R. O'Connor, K. Ogilvie, D. Owen, M. Pettersson, M. R. Reese, T. F. Rogers, M. I. Rossulek, J. G. Sathish, C. Steppan, L. W. Updyke, Y. Zhu, J. Wang, A. K. Chatterjee, A. S. Anderson, A. Allerton, **n.d.**, 32.

[25] S. Günther, P. Y. A. Reinke, Y. Fernández-García, J. Lieske, T. J. Lane, H. M. Ginn, F. H. M. Koua, C. Ehrt, W. Ewert, D. Oberthuer, O. Yefanov, S. Meier, K. Lorenzen, B. Krichel, J.-D. Kopicki, G. Gelisio, W. Brehm, I. Dunkel, B. Seychell, H. Gieseler, B. Norton-Baker, B. Escudero-Pérez, M. Domaracky, S. Saouane, A. Tolstikova, T. A. White, A. Hänle, M. Groessler, H. Fleckenstein, F. Trost, M. Galchenkova, Y. Gevorkov, C. Li, S. Awel, A. Peck, M. Barthelmess, F. Schluenzen, P. L. Xavier, N. Werner, H. Andaleeb, N. Ullah, S. Falke, V. Srinivasan, B. A. França, M. Schwinzer, H. Brognaro, C. Rogers, D. Melo, J. J. Zaitseva-Doyle, J. Knoska, G. E. Peña-Murillo, A. R. Mashhour, V. Hennicke, P. Fischer, J. Hakanpää, J. Meyer, P. Gribbon, B. Ellinger, M. Kuzikov, M. Wolf, A. R. Beccari, G. Bourenkov, D. von Stetten, G. Pompidor, I. Bento, S. Panneerselvam, I. Karpics, T. R. Schneider, M. M. Garcia-Alai, S. Niebling, C. Günther, C. Schmidt, R. Schubert, H. Han, J. Boger, D. C. F. Monteiro, L. Zhang, X. Sun, J. Pletzer-Zelgert, J. Wollenhaupt, C. G. Feiler, M. S. Weiss, E.-C. Schulz, P. Mehrabi, K. Karničar, A. Usenik, J. Loboda, H. Tidow, A. Chari, R. Hilgenfeld, C. Uetrecht, R. Cox, A. Zaliani, T. Beck, M. Rarey, S.

Günther, D. Turk, W. Hinrichs, H. N. Chapman, A. R. Pearson, C. Betzel, A. Meents, *Science* **2021**, DOI 10.1126/science.abf7945.

[26] N. Kitamura, M. D. Sacco, C. Ma, Y. Hu, J. A. Townsend, X. Meng, F. Zhang, X. Zhang, M. Ba, T. Szeto, A. Kukuljac, M. T. Marty, D. Schultz, S. Cherry, Y. Xiang, Y. Chen, J. Wang, *J. Med. Chem.* **2021**, DOI 10.1021/acs.jmedchem.1c00509.

[27] A. Douangamath, D. Fearon, P. Gehrtz, T. Krojer, P. Lukacik, C. D. Owen, E. Resnick, C. Strain-Damerell, A. Aimon, P. Ábrányi-Balogh, J. Brandão-Neto, A. Carbery, G. Davison, A. Dias, T. D. Downes, L. Dunnett, M. Fairhead, J. D. Firth, S. P. Jones, A. Keeley, G. M. Keserü, H. F. Klein, M. P. Martin, M. E. M. Noble, P. O'Brien, A. Powell, R. N. Reddi, R. Skyner, M. Snee, M. J. Waring, C. Wild, N. London, F. von Delft, M. A. Walsh, *Nature Communications* **2020**, *11*, DOI 10.1038/s41467-020-18709-w.

[28] D. A. Erlanson, S. W. Fesik, R. E. Hubbard, W. Jahnke, H. Jhoti, *Nat Rev Drug Discov* **2016**, *15*, 605–619.

[29] D. W. Kneller, G. Phillips, K. L. Weiss, S. Pant, Q. Zhang, H. M. O'Neill, L. Coates, A. Kovalevsky, *Journal of Biological Chemistry* **2020**, *295*, 17365–17373.

[30] V. Mody, J. Ho, S. Wills, A. Mawri, L. Lawson, M. C. C. J. C. Ebert, G. M. Fortin, S. Rayalam, S. Taval, *Commun Biol* **2021**, *4*, 93.

[31] M. Sencanski, V. Perovic, S. B. Pajovic, M. Adzic, S. Paessler, S. Glisic, *Molecules* **2020**, *25*, DOI 10.3390/molecules25173830.

[32] J. Breidenbach, C. Lemke, T. Pillaiyar, L. Schäkel, G. A. Hamwi, M. Diett, R. Gedschold, N. Geiger, V. Lopez, S. Mirza, V. Namasivayam, A. C. Schiedel, K. Sylvester, D. Thimm, C. Vielmuth, L. P. Vu, M. Zyulina, J. Bodem, M. Gütschow, C. E. Müller, *Angewandte Chemie International Edition* **2021**, *60*, 10423–10429.

[33] M. A. Walsh, J. M. Grimes, D. I. Stuart, *Biochemical and Biophysical Research Communications* **2021**, *538*, 40–46.

[34] J. Qiao, Y.-S. Li, R. Zeng, F.-L. Liu, R.-H. Luo, C. Huang, Y.-F. Wang, J. Zhang, B. Quan, C. Shen, X. Mao, X. Liu, W. Sun, W. Yang, X. Ni, K. Wang, L. Xu, Z.-L. Duan, Q.-C. Zou, H.-L. Zhang, W. Qu, Y.-H.-P. Long, M.-H. Li, R.-C. Yang, X. Liu, J. You, Y. Zhou, R. Yao, W.-P. Li, J.-M. Liu, P. Chen, Y. Liu, G.-F. Lin, X. Yang, J. Zou, L. Li, Y. Hu, G.-W. Lu, W.-M. Li, Y.-Q. Wei, Y.-T. Zheng, J. Lei, S. Yang, *Science* **2021**, *371*, 1374–1378.

[35] D. Suárez, N. Díaz, *J. Chem. Inf. Model.* **2020**, *60*, 5815–5831.

[36] D. W. Kneller, G. Phillips, H. M. O'Neill, R. Jedrzejczak, L. Stols, P. Langan, A. Joachimiak, L. Coates, A. Kovalevsky, *Nature Communications* **2020**, *11*, DOI 10.1038/s41467-020-16954-7.

[37] S. Zhang, N. Zhong, X. Ren, C. Jin, B. Xia, *Biomolecular NMR Assignments* **2011**, *5*, 143–145.

[38] J. Shi, J. Song, *FEBS Journal* **2006**, *273*, 1035–1045.

[39] A. L. Kantsadi, E. Cattermole, M.-T. Matsoukas, G. A. Spyroulias, I. Vakonakis, *J Biomol NMR* **2021**, *75*, 167–178.

[40] B. Goyal, D. Goyal, *ACS Combinatorial Science* **2020**, *22*, 297–305.

[41] J. Shi, J. Sivaraman, J. Song, *Journal of Virology* **2008**, *82*, 4620–4629.

[42] M. Congreve, R. Carr, C. Murray, H. Jhoti, *Drug Discovery Today* **2003**, *8*, 876–877.

[43] D. Valenti, J. F. Neves, F.-X. Cantrelle, S. Hristeva, D. L. Santo, T. Obšil, X. Hanoulle, L. M. Levy, D. Tzalis, I. Landrieu, C. Ottmann, *Med. Chem. Commun.* **2019**, DOI 10.1039/C9MD00215D.

[44] C. Dalvit, G. Fogliatto, A. Stewart, M. Veronesi, B. Stockman, *Journal of biomolecular NMR* **2001**, *21*, 349–359.

[45] **N.d.**

[46] D. Bajusz, W. S. Wade, G. Satała, A. J. Bojarski, J. Ilaš, J. Ebner, F. Grebien, H. Papp, F. Jakab, A. Douangamath, D. Fearon, F. von Delft, M. Schuller, I. Ahel, A. Wakefield, S. Vajda, J. Gerencsér, P. Pallai, G. M. Keserű, *Nat Commun* **2021**, *12*, 3201.

[47] C. W. Murray, D. C. Rees, *Nature Chem* **2009**, *1*, 187–192.

[48] M. Macchiagodena, M. Pagliai, P. Procacci, *Chemical Physics Letters* **2020**, *750*, 137489.

[49] D. Ni, Y. Li, Y. Qiu, J. Pu, S. Lu, J. Zhang, *Trends in Pharmacological Sciences* **2020**, *41*, 336–348.

[50] S. Lu, Y. Qiu, D. Ni, X. He, J. Pu, J. Zhang, *Drug Discovery Today* **2020**, *25*, 177–184.

[51] C. Liu, S. Boland, M. D. Scholle, D. Bardiot, A. Marchand, P. Chaltin, L. M. Blatt, L. Beigelman, J. A. Symons, P. Raboisson, Z. A. Gurard-Levin, K. Vandyck, J. Deval, *Antiviral Res* **2021**, *187*, 105020.

[52] A. L. Hopkins, G. M. Keserü, P. D. Leeson, D. C. Rees, C. H. Reynolds, *Nat Rev Drug Discov* **2014**, *13*, 105–121.

WILEY-VCH

RESEARCH ARTICLE

Entry for the Table of Contents
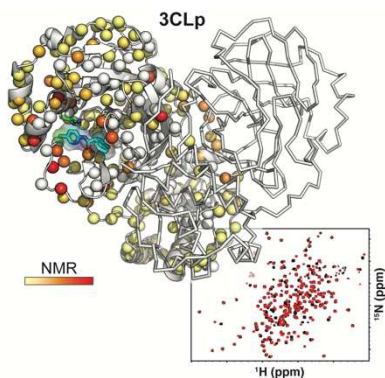
We herein report the liquid-sate NMR spectroscopy analysis of the dimeric SARS-CoV-2 main protease (3CLp), including its backbone assignments, to study its complex conformational regulation. Using fragment-based NMR screening, we highlighted three hotspots on the protein, two in the substrate binding pocket and one at the dimer interface, and we identified a non-covalent reversible inhibitor of 3CLp that has antiviral activity in infected cells.

Novel dithiocarbamates selectively inhibit 3CL protease of SARS-CoV-2 and other coronaviruses

Lucile Brier[1,‡], Haitham Hassan[1,‡,†], Xavier Hanoulle[4,5,‡], Valerie Landry[1,3], Danai Moschidi[4,5],

Lowiese Desmarets[6], Yves Rouillé[6], Julie Dumont[1,3], Adrien Herledan[1,3], Sandrine Warenghem[1,3],

Catherine Piveteau[1], Paul Carré[1,3], Sarah Ikherbane[1,3], François-Xavier Cantrelle[4,5], Elian

Dupré[4,5], Jean Dubuisson[6], Sandrine Belouzard[6], Florence Leroux[1,2], Benoit Deprez[2,3,*,+], Julie

Charton[2,+]

1. Univ. Lille, Inserm, Institut Pasteur de Lille, U1177 - Drugs and Molecules for Living Systems, F-59000 Lille, France
2. Univ. Lille, Inserm, Institut Pasteur de Lille, U1177 - Drugs and Molecules for Living Systems, EGID, F-59000 Lille, France
3. Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, US 41 - UMS 2014 - PLBS, F-59000 Lille, France
4. CNRS, ERL9002 - BSI - Integrative Structural Biology, F-59000, Lille, France.
5. Univ. Lille, INSERM, CHU Lille University Hospital, Institut Pasteur de Lille, UMR1167 - RID-AGE - Risk factors and molecular determinants of aging-related, F-59000, Lille, France.
6. Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur Lille, U1019 - UMR 9017 - CIIL - Center for Infection and Immunity of Lille, F-59000 Lille, France

‡ LB, HH and XH are equally contributing first authors.

+ Joint last authors: B.D. and J.C. contributed equally.

ABSTRACT. Since end of 2019, the global and unprecedented outbreak caused by the coronavirus

SARS-CoV-2 led to dramatic numbers of infections and deaths worldwide. SARS-CoV-2 produces

two large viral polyproteins which are cleaved by two cysteine proteases encoded by the virus, the

3CL protease (3CL^{pro}) and the papain-like protease, to generate non-structural proteins essential

for the virus life cycle. Both proteases are recognized as promising drug targets for the development of anti-coronavirus chemotherapy. Aiming at identifying broad spectrum agents for the treatment of COVID-19 but also to fight emergent coronaviruses, we focused on 3CL^{pro} that is well conserved within this viral family. Here we present a high-throughput screening of more than 89,000 small molecules that led to the identification of a new chemotype, potent inhibitor of the SARS-CoV-2 3CL^{pro}. The mechanism of inhibition, the interaction with the protease using NMR and X-Ray, the specificity against host cysteine proteases and promising *in cellulo* antiviral properties are reported.

INTRODUCTION

The outbreak of severe acute respiratory syndrome (SARS)[1] in 2002, as well as the Middle-East respiratory syndrome (MERS)[2] in 2012 and COVID-19[3] since 2019 demonstrates the potential of coronaviruses to cross boundaries between species and highlights the importance of urgently developing efficient antiviral compounds with broad spectrum activity against this virus family.

The newly emerged coronavirus SARS-CoV-2 is the seventh reported coronavirus having the potential to infect human. It is an enveloped, positive single-stranded RNA virus that can infect both humans and animals. Coronaviruses contain the largest known RNA genome encoding, in addition to the structural and accessory proteins, two large viral polyproteins, pp1a and pp1ab. These polyproteins are processed by two viral proteases, the papain-like protease (PLP) and the 3C-like protease (3CL^{pro}, also known as the main protease M^{pro}) to generate a series of functional non-structural proteins essential for virus replication and transcription.

The viral main protease (3CL^{pro}), a 33.8 kDa cysteine protease with a non-classical Cys-His dyad (Cys145-His41) is active as a homodimer. As an essential component for the formation of the coronavirus replication complex, this protease is an attractive target for the development of anti-

coronavirus therapeutics. Moreover, the remarkable degree of conservation of this protease among the viruses of this family (96% sequence identity between the SARS-CoV-2 and SARS-CoV) is a strong asset to design pan-anti-coronavirus drugs to tackle not only the current SARS-CoV-2 but also potential future outbreaks of emergent human coronaviruses.[4,5] Interestingly, 3CL[pro] cleaves the polyprotein 1ab at 11 cleavage sites with a common cleavage sequence LQ↓(S/A/G) that is unusual for human proteases[6,7]. This may allow for identifying drug candidates with high specificity for the viral protease reducing unwanted polypharmacology and potential side effects.

Several 3CL[pro] inhibitors are currently in preclinical and clinical development for the treatment of COVID-19[8,9,10,11,12,13,14,15]. Amongst the most potent compounds, GC-376, a peptidomimetic broad-spectrum antiviral agent developed for feline coronavirus infections, inhibits the replication of noroviruses, picornaviruses, and coronaviruses[16]. Recently two other peptidomimetic broad-spectrum coronavirus 3CL[pro] inhibitors with high selectivity over human proteases; PF-07304814, a phosphate prodrug metabolized *in vivo* to the active moiety PF-00835231 was developed[17] and PF-7321332 (nirmatrelvir) specifically designed to be administered orally, is currently approved[18]. However, the promising clinical results reported by Pfizer for nirmatrelvir are dependent upon coadministration with ritonavir as a PK enhancer. In addition to peptidomimetic inhibitors that can be challenging in terms of selectivity and PK profile, novel non peptidomimetic small molecules pan-inhibitors of 3CL[pro] are also highly desirable to fight COVID-19 and potential emerging coronaviruses.

Here we report the high-throughput screening of a large library of more than 89,000 small compounds on the 3CL[pro] of SARS-CoV-2 and the identification of a novel class of non-peptidomimetic small inhibitors with promising anti-coronaviral activity.

RESULTS AND DISCUSSION

Development of the enzymatic assay and high-throughput screening.

SARS-CoV-2 3CL^pro expression. It has been shown that extra amino-acid residues or affinity tags at either the N-terminus or C-terminus significantly decrease 3CL^pro enzymatic activity via disturbing its dimerization and/or active site conformation. The native SARS-CoV-2 3CL^pro was thus recombinantly expressed in *Escherichia coli* with a cleavable tag and purified to homogeneity. The N-terminal 6xHis-SUMO tag that was later cleaved using SENP2 protease (His-tagged) to release the native SARS-CoV-2 3CL^pro without any extra residue at its extremities. The high yield for expression and purification (~100mg/L culture) allowed us to work with a single batch of the protease leading to a high reproducibility of our enzymatic high throughput screening.

Enzyme characterization, assay development, validation of the assay with known inhibitors. Aiming at rapidly developing a Förster Resonance Energy Transfer (FRET)-based assay and on the basis of the strong identity between SARS-CoV and SARS-CoV-2 main proteases, we have turned to a fluorogenic substrate previously reported to monitor the catalytic activity of the SARS-CoV 3CL^pro [19]. The Dabcyl-KTSAVLQ/SGFRKME(Edans) substrate bearing the sequence between polypeptide nsp4-nsp5 junction of SARS-CoV-2 was selected. The fluorescence resulting from the cleavage of this substrate by the protease was measured with a Victor 3V instrument (Perkin-Elmer) using excitation and emission wavelengths of 340(25) nm and 535(25) nm (Figure 1d). Using this fluorogenic substrate, proteolytic activity of the 3CL^pro was evaluated in a pH 7.4 buffer containing 50 mM HEPES, 0.1 mg/mL BSA, 0.01% Triton X-100 and 2 mM GSH. Due to the presence of a catalytic cysteine in the 3CL^pro active site, glutathione (GSH) was added in the assay buffer to both maintain the catalytic cysteine in its active state and discard highly electrophilic compounds from the hit list ($[GSH] = 100000 \times [3CL^{pro}]$). Triton X-100 has also been added to avoid non-specific inhibitory binding of the compounds with the protease by forming aggregate. We characterized the enzymatic activity of the recombinant SARS-CoV-2 3CL^pro by determining

the Michaelis-Menten constant (K_M) value. Fluorogenic substrate concentrations ranging from 6

to 400 μM were used in this kinetic study with a fixed enzyme concentration of 15 nM. The initial

velocity was measured and plotted against substrate concentration. Curve fitting with Michaelis-

Menton equation gave the best-fit values of K_M as 54 μM (Figure 1b). Previously we had checked

that the enzymatic reaction was linear for enzyme concentration between 3.1 nM and 25 nM (Figure

1c).
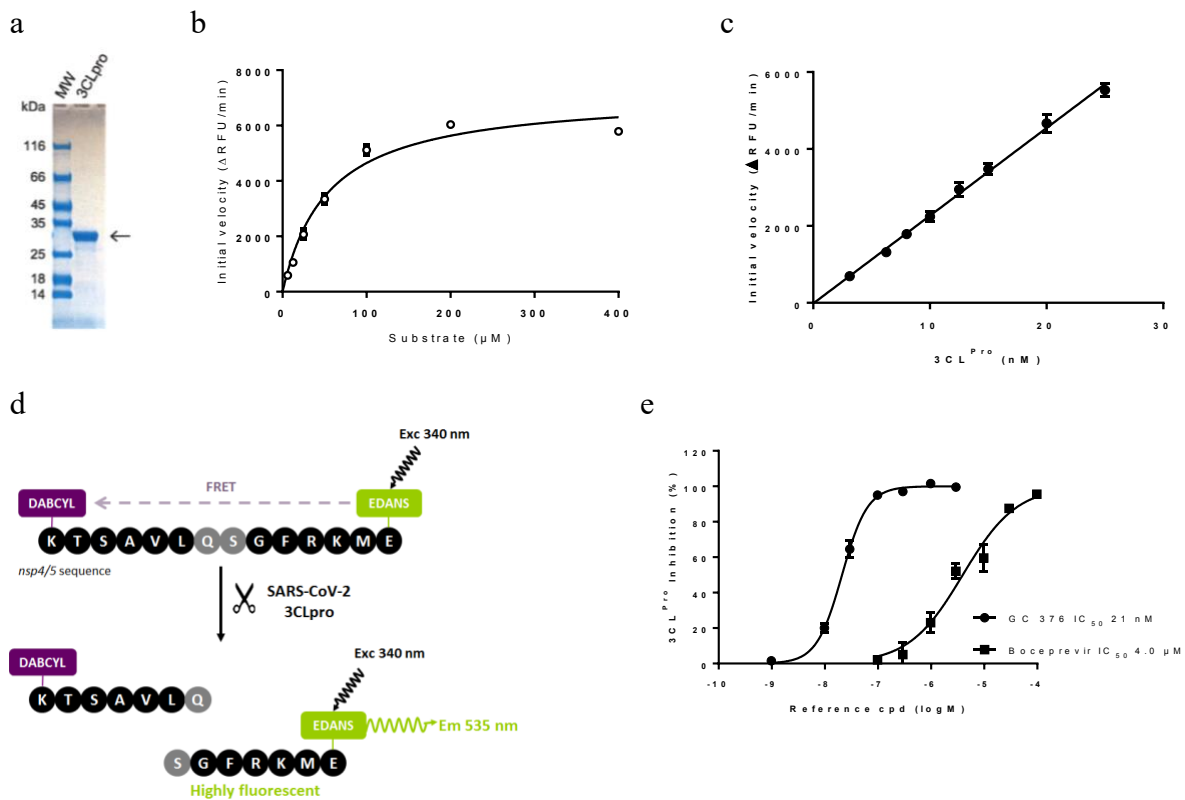
Figure 1. SARS-CoV-2 3CL^pro expression and characterization. (a) SDS-PAGE of 3CL^pro; the

calculated molecular weight of the 3CLpro is 33,796 Da. (b) Michaelis-Menten plot of 15 nM

3CL^pro with various concentrations of FRET substrate (c) Initial velocity plot of enzyme reaction

with various concentrations of 3CL^pro. (d) Principle of the enzymatic FRET-based assay used to

monitor 3CL^pro activity. Once the fluorogenic substrate is cleaved by the enzyme, the fluorophore

(Edans) and the fluorescence quencher (Dabcyl) are spatially separated, resulting in an increase in

fluorescence which is proportional to the enzyme activity. (e) SARS-CoV-2 3CL^pro inhibition by

reference compounds. 3CL^{pro} was pre-incubated 30 minutes with various concentration of GC-376 or boceprevir before FRET substrate addition and initial velocity measurement. Data are shown as mean ± SD of duplicates from representative experiments.

As a too high substrate concentration could be detrimental for the purposes of identification of inhibitors that usually compete with substrate for the enzyme active site, we selected the lowest enzyme (15 nM) and substrate concentration (10 μM; $0.2×K_M$) that yielded a strong, reliable and reproducible signal in the 384-well plate screening format (0.8<Z'<0.9). Finally, tolerance to DMSO was evaluated and a final maximal DMSO concentration of 1% was used. To validate the enzyme assay, two known inhibitors of the SARS-CoV-2 3CL^{pro}; GC-376 and boceprevir[16] were evaluated in the screening assay. Both compounds dose-dependently inhibited the enzyme activity with IC_{50} values of 4 μM for boceprevir and 0.02 μM for GC-376, in line with the values reported in the literature (Figure 1e).

Primary high-throughput screening. A library of 89,193 compounds, selected from commercial vendors or prepared by our chemists using state-of-the-art selection and design criteria, in terms of diversity and "drug / lead-likeness" properties, was screened at 30 μM against the SARS-CoV-2 3CL protease by enzymatic end-point fluorescence intensity assay in 384-well microplate format on a semi-automated system (Figure 2). To allow detection of slow/tight enzyme binders, a 30-minute pre-incubation of compounds with the enzyme was applied before starting the reaction with substrate. Fluorescence was measured after 30 minutes. Boceprevir at 4 μM (IC_{50} value) and 40 μM (10 x IC_{50}) was used as a positive reference compound in each plate. The HTS demonstrated robust performance with an average Z' factor of 0.85.
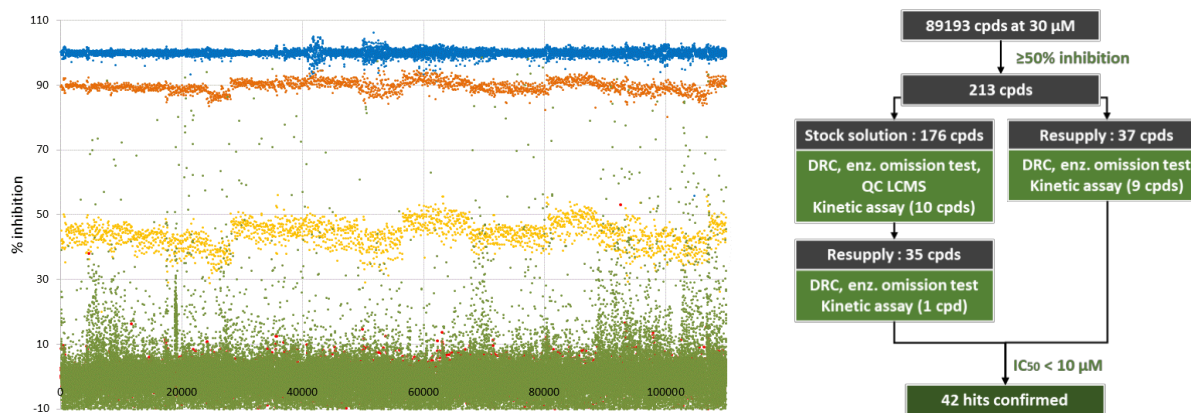
Figure 2. Left. Overview of the screening. SARS-CoV-2 3CL^{pro} inhibition was reported (%) for each incubate. Green dots: tested compounds at 30µM, Red dots: positive controls (incubations with vehicle), Blue dots: negative controls (incubations w/o enzyme), Orange dots: reference compound Boceprevir at 40 µM, Light orange dots: Boceprevir at 4 µM. Right. Workflow of the screening campaign.

The cut-off to select hits was set to 50 % inhibition of the enzyme activity in order to focus on the most potent inhibitors. 213 compounds were thus cherry-picked from our inventories or repurchased. They were retested in the same assay in dose–response experiments to determine their IC$_{50}$ values. In parallel the compounds were tested at the highest concentration (100 µM) without enzyme in order to test for fluorescence signal interference. Following the enzyme omission test, 20 compounds were retested in a continuous kinetic assay that allows rate calculation, for confirmation. The DMSO stock solutions of the 176 compounds were controlled for purity and identity by LCMS. From these 176 compounds, 35 were then selected and repurchased based on chemical structures, synthetic access and IP criteria. Finally, 42 confirmed hits were selected using a cut-off of maximal inhibition greater than 50% and IC$_{50}$ less than 10 µM (Figure 2).

We report here the identification and characterization of a dithiocarbamate 3CL^{pro} inhibitor that represents a novel chemical series with promising anti-coronavirus activity (compound **1**, Figure 3)

Cpd 1
SARS-CoV-2 3CLpro
IC_{50} = 1.8 µM (GSH+)
IC_{50} = 0.021 µM (GSH-)
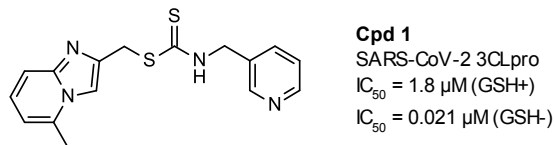
Figure 3. Structure of hit compound **1**.

Hit compound 1 characterization.

Slow binding inhibitor. Compound **1** was retested with or without a 60-minute pre-incubation step before substrate addition. As can be seen in the concentration-effect curve (Figure 4a), observed inhibition is shifted to lower concentrations with pre-incubation, revealing higher compound potency as expected for slow binding inhibitor.

Effect of reducing agents on inhibitory potency. Enzymatic assays for the SARS-CoV-2 3CL[pro] inhibition were performed with different reducing conditions classically used with cysteine protease (Figure 4b). We obtained highly decreased potencies for compound **1** in reducing conditions (IC_{50}=0.008 µM without reducing agent, 1.75 µM with 2 mM GSH and 1.94 µM with 2mM THP (Figure 4b)). To know if the loss of potency was due to a degradation of the compound by the reducing agents, compound **1** was incubated at 100 µM in the enzymatic assay buffer in presence of 2 mM of GSH. After 1h30 incubation, the very low disappearance of the compound in both conditions (less than 10%) cannot explain the 2-log decrease in potency observed with the reducing agents suggesting an impact of the reducing condition on the enzyme and rather than on the compound. Others reported this effect but failed to bring mechanistic description[20].
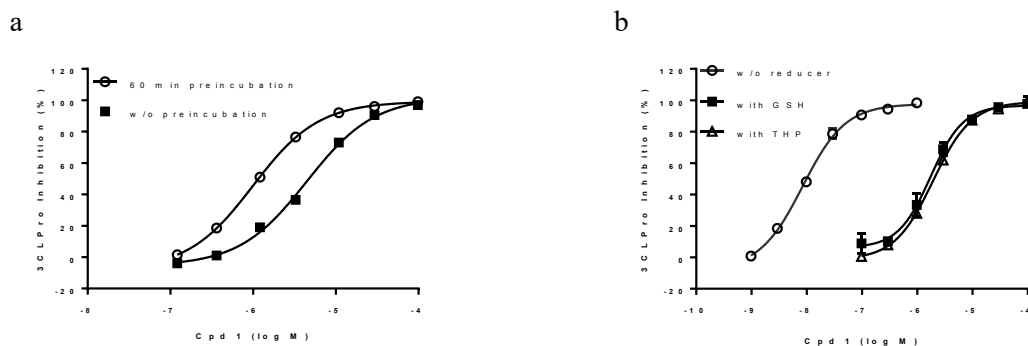
Figure 4. (a) SARS-CoV-2 3CL^{pro} inhibition by hit compounds **1**. 3CL^{pro} was pre-incubated 60 minutes or not with various concentration of compound before FRET substrate addition and initial velocity measurement. Inhibitions (%) are shown as mean ± SD of duplicate from a representative experiment. (b) SARS-CoV-2 3CL^{pro} inhibition by hit compounds in buffer containing (2mM) or not GSH and THP. 3CL^{pro} was pre-incubated 60 minutes with various concentration of compound before FRET substrate addition and initial velocity measurement. Inhibitions (%) are shown as mean ± SD of duplicate from a representative experiment.

Reversibility of inhibition. To elucidate the binding mode of compound **1** with the 3CL^{pro}, the reversibility of inhibition was evaluated. Jump dilution assay is commonly used to evaluate the reversibility of inhibition. To that aim, the compound was pre-incubated at 10 times its IC$_{50}$ with 3CL^{pro} in presence of GSH (2mM) as reducing agent. Then, incubates were quickly diluted to 1/100 with substrate solution before measuring the fluorescence kinetics. Inhibition is compared to control standard incubations at 10*IC$_{50}$ and 0.1*IC$_{50}$ final concentrations of compounds (Figure 5). Final enzyme and substrate concentrations are 15 nM and 10 μM, respectively. As can be seen in Figure 5, an irreversible inhibition was observed for compound **1**. During the first 10 minutes after jump dilution, the enzyme was inhibited at around 80%, a value similar to the 10*IC$_{50}$ control value, that is consistent with the presence of an electrophilic center (dithiocarbamate function) that

may form covalent bond with the nucleophilic Cys145 of the catalytic site. Later, the enzyme activity was progressively recovered and no more inhibition was observed after 2 hours.
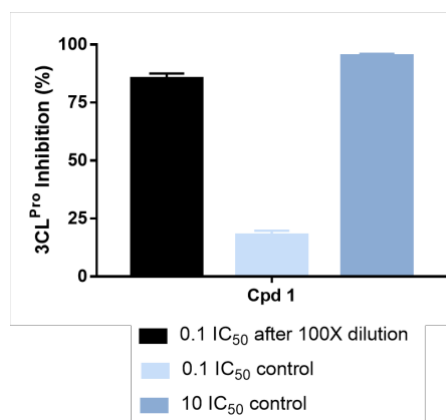
Figure 5. Assessment of the irreversible/reversible mechanism of inhibition with compound **1**. SARS-CoV-2 3CL^{pro} was incubated with compound **1** in a jump dilution assay. Initial rates are measured just after the compound dilution with substrate and enzymatic inhibitions (%) are calculated as mean ± SD of triplicate. Data are representative of three other experiments.

Thermal Shift Assay (TSA). The direct binding of compound **1** to the 3CL protease was evaluated in the TSA assay. The melting temperature of SARS-CoV-2 3CL^{pro} was shifted by 14.17°C upon binding of compound **1** supporting the direct binding of compound **1** with the 3CL^{pro} (Figure 6). In agreement with values reported in the literature, Boceprevir and GC-376 shifted the melting curve of 3CL^{pro} by 3.82 and 19.36°C respectively upon binding[16]. Moreover, the binding of compound **1** to the 3CL protease is maintained when adding GSH 2mM in the buffer (ΔTm=11.33°C).
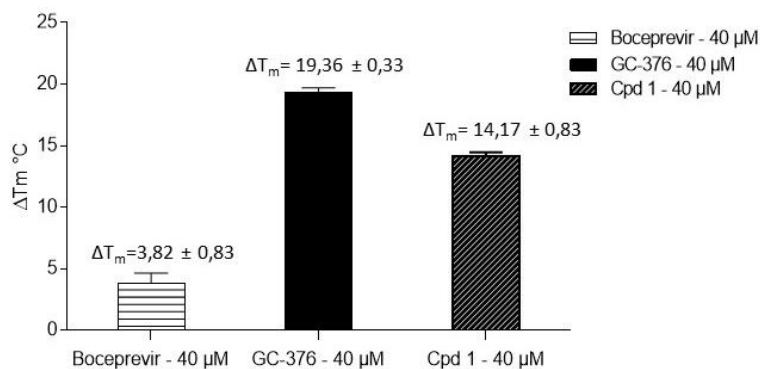
Figure 6. Evaluation of thermal stabilization of SARS-CoV-2 3CL^{pro} by Thermal Shift Assay (TSA) in presence of Boceprevir, GC376 or Cpd **1** (40 µM). Thermal shift (ΔTm) was calculated by subtracting reference melting temperature of the protease from the Tm in presence of the compound. Values presented are the means of ΔTm±SD of the eight independent TSA experiments.

Structural data. The direct binding of compound **1** to the 3CL^{pro} was also investigated using solution NMR spectroscopy. Using a ^{2}H,^{15}N-labelled 3CL^{pro} we compared the 2D ^{1}H,^{15}N-TROSY-HSQC NMR spectra of the 3CL^{pro} acquired in the absence or the presence of an excess of molecule[21]. Spectral perturbations (chemical shift perturbations and/or signal broadening) were observed on the 3CL^{pro} spectrum upon addition of compound **1** (Figure 7). These chemical shift perturbations show the direct binding of this compound in the 3CL^{pro} active site, for which the highest CSP are observed.
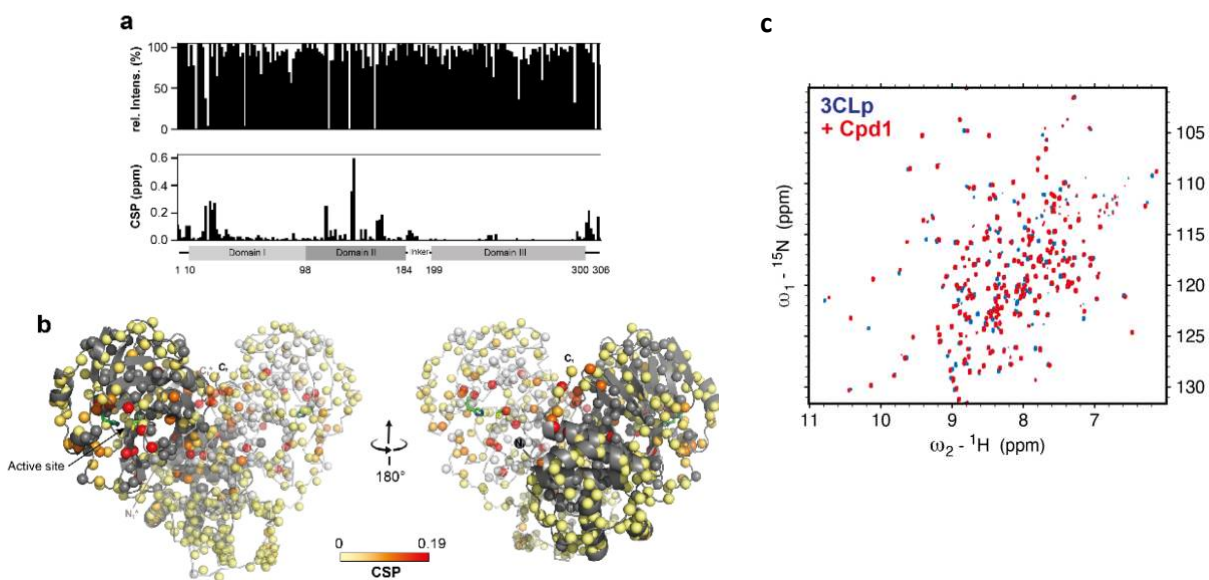
Figure 7. Interaction of SARS-CoV-2 3CL[pro] with compound **1** assessed by NMR spectroscopy. a) The variations in NMR resonance intensities (top) and chemical shift perturbations (CSP) (bottom) induced upon cpd **1** addition are displayed along the main protease sequence. b) The CSPs, shown in (a, bottom), have been color coded (from light yellow to red) and are displayed on the structure of the dimeric 3CL[pro], with the two protomers shown in dark grey and white, respectively. The side chains of both catalytic His41 and Cys145 are shown in green. c) Overlaid 2D ^{1}H,^{15}N-TROSY-HSQC spectra acquired on ^{2}H,^{15}N-labelled 3CL[pro] (100µM) in the absence (in blue) or in the presence (in red) of cpd **1** (target concentration of 2 mM). The final DMSO-d6 concentration was 3%. The spectra were acquired at 305K on a 900MHz NMR spectrometer.

To get additional structural data on the binding mode of compound **1**, we solved the crystal structure of the 3CL[pro]: compound **1** complex. We crystallized the native 3CL[pro] and obtained the complex with compound **1** using a soaking procedure. The 3CL[pro] crystals were soaked for 1 hour in the crystallization solution containing 10 mM of the molecule (10% DMSO), and then briefly soaked into a cryoprotective solution containing 10% glycerol before flash-freezing in liquid

nitrogen. The structure of the complex, solved at 1.49Å resolution, reveals that the Sulfur of active site cysteine 145 undergoes a transthiocarbamoylation by reacting with the electrophilic carbon of the dithiocarbamate function of compound **1**, triggering the loss of (5-methylimidazo[1,2-a]pyridin-2-yl)methanethiol (Figure 8a). This binding mode is fully consistent with our NMR data where the highest CSP were observed for residues surrounding the S1 pocket. The pyridine ring of the newly formed dithiocarbamate adduct occupies the S1 pocket that usually binds the P1 residue of the substrate. The 3CL[pro] S1 pocket is defined by the residues Phe140, Leu141, Asn142, Gly143, Ser144, His163, Met165, Glu166, His172. The ligand establishes two hydrogen bonds with the protease, one with His163 and one with Asn142 through a water molecule, as well as many hydrophobic contacts with the surrounding 3CL[pro] residues (Figure 8b). The pyridinyl substituent also present in ML188 and calpain inhibitor XII forms the same hydrogen bond with the H163 imidazole in S1 pocket as for compound **1**[22,23]. The His163 residue at the S1 pocket was described as a binding hot spot for 3CL[pro] inhibitors and the pyridinyl substituent appears as a suitable scaffold to finely fit into the S1 pocket. The 3CL[pro] S1 pocket is occupied by a cyclobutyl or a glutamine surrogate γ-lactam ring of boceprevir and GC-376 respectively[16,24]. Upon binding of compound **1** the loops (Glu166-His172 and Phe185-Ala194) and the helical segment (Val42-Pro52), forming the walls of the enzymatic cleft, were slightly displaced compared to the structure of the apo-3CL[pro], making the cleft slightly wider. These observations are consistent with the location of the observed NMR chemical shift perturbations. In addition, we also observed NMR CSPs for residues that are outside of the active site (Gly2, Phe3, Ser10, Gly11, Ala116, Cys117, Ser121, Ala124, Glu166, Gln299, Ser301, Gly303, Val303 and Phe305) (Figure 7). This NMR data show that compound **1** while binding at the active site induces structural and/or dynamical perturbations up to the dimerization interface of 3CL[pro], where both the N-terminus and C-terminus of the protease, including Ser10, Gly11, R298, have been shown to be essential for its

dimerization[25,26]. This path, highlighted by NMR CSPs, clearly shows an allosteric regulation pathway in 3CL[pro]. The later one could correspond to the conserved communication network (His163, Ser147, Leu115 and Ser10) that have been recently proposed in 3CL[pro] based on mutational and functional analyses[26]. The formation of the dithiocarbamate function with the Cys145 also explains the irreversible inhibition observed in the jump dilution assay for this compound. However, the new dithiocarbamate remains electrophilic and hydrolysable. Indeed, upon dilution the enzyme recovers slowly its catalytic activity. We have previously observed the ability of compound **1** to react with thiols like GSH derivatives but in the case of 3CL[pro], this reactivity is greatly improved as **1** yields complete reaction with 3CL[pro] and only a low reaction conversion with GSH was observed after 90 minutes (<10%) despite high GSH concentration (2 mM).
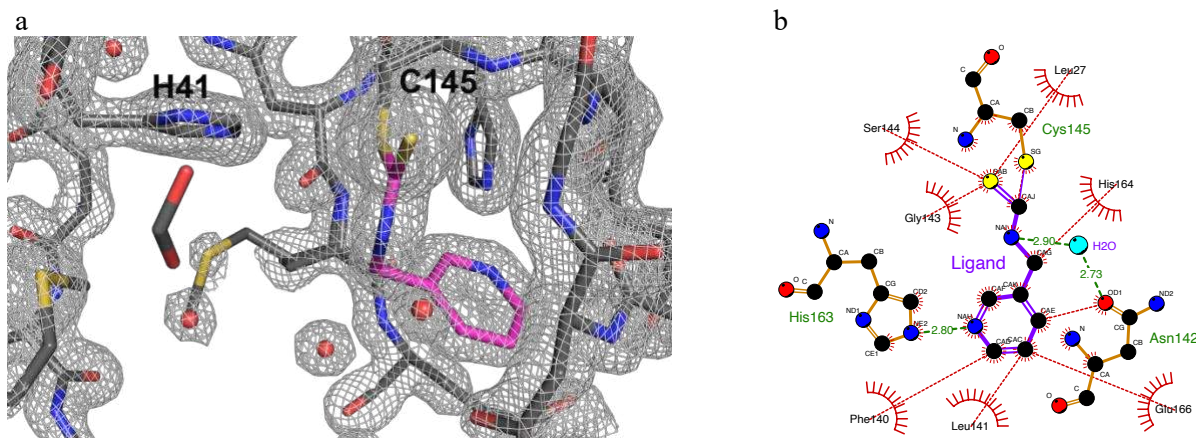
Figure 8. Crystallographic structure of the SARS-CoV-2 3CL[pro] bound to the N-(pyridin-3-ylmethyl)thioformamide moiety (in pink) from the compound **1** (PDB ID: 7NTQ) (a). The 2*Fo-Fc* electron-density map, contoured at 1.5σ, is shown as light grey mesh. (b) Interaction of the ligand (cpd **1** after Cys145 binding) with the 3CL[pro] S1 pocket. The analysis of the interaction was made using LigPlot+ (v2.2.4). 3CL[pro] and ligand bonds are shown in brown and purple, respectively. Water molecules are displayed as cyan spheres. Dashed green lines and dashed red lines represent hydrogen bonds and hydrophobic interactions, respectively. The 3CL[pro] residues that are involved

in hydrogen bonds have their name displayed in green whereas the residues making hydrophobic contact(s) are indicated with their name surrounded by red spikes. Atoms involved in hydrophobic contact(s) are surrounded by red spikes.

*Analogs synthesis of compound **1** and enzymatic activity on 3CL^{pro}.* To gain molecular insight into the binding properties of the dithiocarbamate-based inhibitor **1** within the catalytic site, four analogs were synthesized. In compound **1A**, the electrophilic warhead dithiocarbamate was replaced by a thiourea moiety. After covalent binding of compound **1** with the catalytic Cys145 of the 3CL^{pro}, only the pyridyl part remains in the catalytic site. To evaluate the importance of the leaving moiety (5-methylimidazo[1,2-a]pyridin-2-yl)methanethiol for the binding of **1**, this part was replaced by either an ethyl (**1B**) or a benzyl (**1C**) moiety while keeping the dithiocarbamate warhead. Finally, the dithiocarbamate in **1C** was replaced by a thiocarbamate in **1D** (Figure 9).
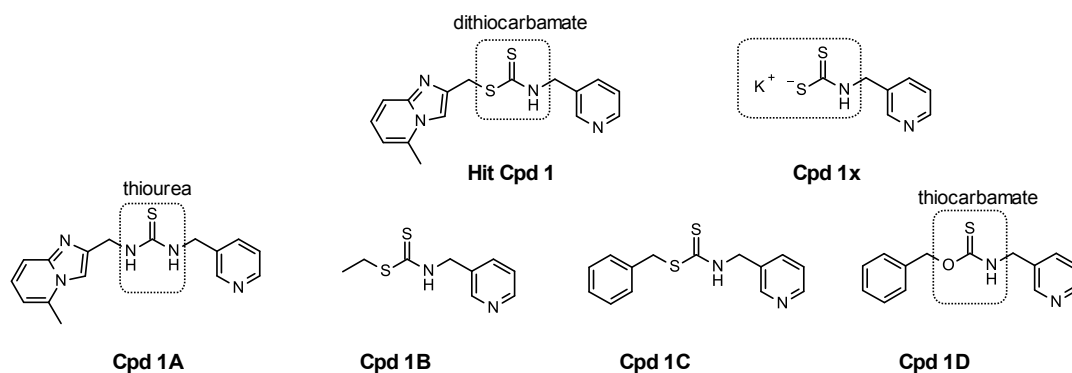
Figure 9. Structures of analogs **1x** and **1A-1D** of hit compound **1**

Hit compound **1** was resynthesized in three steps described in scheme 1. 3-(aminomethyl)pyridine was reacted with carbon disulfide in presence of potassium hydroxide to obtain intermediate **1x**. 1,3-dichloroacetone and 2-amino-6-methylpyridine were refluxed in

ethanol to lead to chlorine intermediate **1y**. Finally, compound **1** was obtained by nucleophilic

substitution between **1x** and **1y**.

Scheme 1. Synthetic Route to hit compound 1[a]
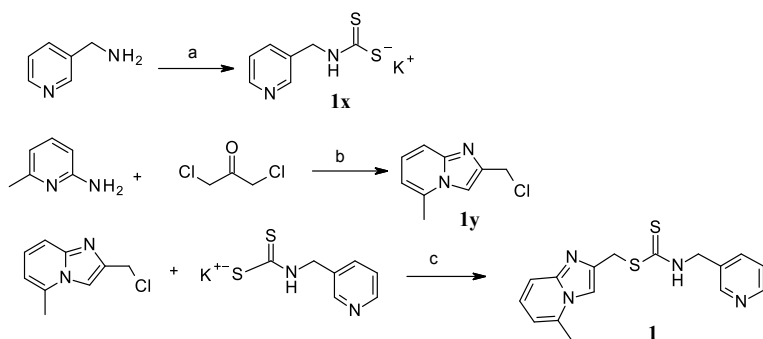
[a]Reagents and conditions: (a) CS₂, KOH, MeOH, 2 h, 0°C, 43% (b) EtOH, reflux, overnight, 30

% (c) NEt₃, MeOH, 3 h, room temp., 85%

1A-1D were obtained from isothiocyanate **1i** (scheme 2) and different nucleophiles (amine, thiol

and alcohol). **1i** was prepared according a reported procedure using 3-picolylamine and bis(2-

pyridyloxy)methanethione (DPT).[27]

Scheme 2. Synthetic Routes to 1A-1D[a]
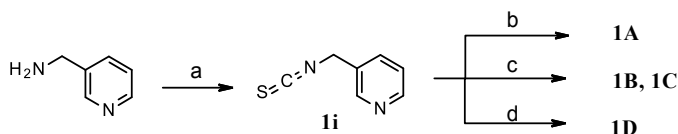
[a]Reagents and conditions: (a) NaH (60% dispersion in mineral oil) in dry THF, 0°C, 1 h then

bis(2-pyridyloxy)methanethione, room temp., overnight, 67 % (b) R-NH₂, NEt₃, DCM, 1 h, room

temp., 96% (c) R-SH, NEt$_3$, DCM, 1 h, room temp., 90-96 % (d) R-OH, NaH (60% dispersion in mineral oil), dry THF, 0°C then **1i**, room temp., 1 h, 85%.

The activities presented in Table 1 show that the dithiocarbamate function present in **1**, **1B** and **1C** is essential for the inhibitory effect as replacement by thiourea (**1A**) or thiocarbamate (**1D**) led to inactive compounds. The replacement of the 5-methylimidazo[1,2-a]pyridin-2-yl moiety of compound **1** by a phenyl in compound **1C** led to a similar inhibitory activity while the replacement by an ethyl group in **1B** resulted in 8-fold lower potency. Interestingly, the potassium dithiocarbamate intermediate **1x** devoid of thiol leaving group is as potent as compound **1**. Salts of dithiocarbamate as Ditiocarb, the sodium salt of diethyldithiocarbamate, are known to be powerful metal chelating agent. As zinc complexes are known to inhibit 3CL^{pro} [28], we hypothesized that compound **1x** could inhibit the 3CL^{pro} as a zinc complex. The compound was thus tested in presence of the zinc chelator EDTA (500 µM) but **1x** revealed as potent in presence or in absence of EDTA in contrast to zinc acetate that completely loose its inhibitory activity in presence of EDTA. To get insight into the binding mode of compound **1x**, we solved the crystal structure of the 3CL^{pro}:compound **1x** complex. As previously for compound **1**, we crystallized the native 3CL^{pro} and obtained the complex with compound **1x** using a soaking procedure. The structure of the complex reveals that the active site cysteine 145 makes a covalent bound with the electrophilic carbon of the dithiocarbamate function of compound **1x** leading to the same fragment bound to the protease as with compound **1** (Figure S1a). The superimposition of the structures of the SARS-CoV-2 3CL^{pro} bound to the compound **1** and to the compound **1x** (Figure S2b) shows that the N-(pyridin-3-ylmethyl)thioformamide moiety bound to the protease has the same binding mode (Figure S1b).

Table 1. SARS-CoV-2 3CL^{Pro} inhibition of hit compound 1 and analogues 1A-D, 1x

Cpd	Assay without GSH		Assay with GSH	
	IC$_{50}$ (µM)	pIC$_{50}$ ± SD	IC$_{50}$ (µM)	pIC$_{50}$ ± SD
1	0.021	7.67 ± 0.30	1.46	5.85 ± 0.05
1A	35.0	4.46 ± 0.19	>100	< 4
1B	0.178	6.75 ± 0.63	22.40	4.65 ± 0.07
1C	0.023	7.65 ± 0.32	9.36	5.03 ± 0.10
1D	>100	< 4	>100	< 4
1x	0.027	7.57 ± 0.22	1.36	5.87 ± 0.04

Enzymatic assays were performed with or without GSH 2mM in the buffer. The enzyme was pre-incubated for 1h with compound at increasing concentrations before starting the reaction with the substrate. Initial rates were recorded to calculate % inhibition and IC$_{50}$ were obtained from concentration-response curves by a nonlinear regression analysis of the data. IC$_{50}$ values are averages of three independent experiments.

Inhibition of the 3CL protease of another human coronavirus. Since the catalytic sites of 3CL proteases are highly conserved among coronaviruses family, we hypothesized that compound **1** could act as a pan-inhibitor of coronaviruses 3CL proteases. Consistent with this hypothesis, compound **1** was shown to be inhibitor of the 3CL protease of coronaviruses of alpha (HCoV-229E) and beta (SARS-CoV-2, MERS-CoV) groups of *Coronaviridae* (Table 2) and could represent a good starting point to develop a broad spectrum anti-coronavirus compound to fight emerging coronaviruses. A weaker potency has been obtained for all tested compounds on the 3CL^{pro} of MERS-CoV. Moreover, an increase in enzymatic activity was observed in the presence of low concentrations of compounds while inhibition of enzymatic activity was observed at higher inhibitor concentrations (Figure S2). The activation of MERS-CoV 3CL^{pro} by ligands at low concentration was previously described as a result of dimerization induced upon partial occupation

of the substrate binding pocket. Indeed, MERS-CoV 3CL^{pro} is a weakly associated dimer requiring ligand binding for dimer formation and enzymatic activity. [29]

Table 2. Activity of compounds against 3CL^{pro} of different human coronaviruses

Cpd	hCoV-229E (Alpha-CoV) 3CL^{pro} IC$_{50}$ μM (pIC$_{50}$)	SARS-CoV-2 (Beta-CoV) 3CL^{pro} IC$_{50}$ μM (pIC$_{50}$)	MERS-CoV (Beta-CoV) 3CL^{pro} IC$_{50}$ μM (pIC$_{50}$)
1	0.016 (7.79 ± 0.14)	0.021 (7.67 ± 0.30)	2.00 (5.70 ± 0.10)
1x	0.019 (7.73 ± 0.10)	0.027 (7.57 ± 0.22)	1.36 (5.87 ± 0.21)
GC-376	0.028 (7.56 ± 0.09)	0.012 (7.92 ± 0.02)	0.43 (6.37 ± 0.06)

Enzymatic assays were performed without GSH in the buffer. The enzymes were pre-incubated for 1h with compound at increasing concentrations before starting the reaction with the substrate. Initial rates were used to calculate % inhibitions and IC$_{50}$ were obtained from concentration-response curves by a nonlinear regression analysis of the data. IC$_{50}$ values are averages of at least three independent experiments.

Selectivity against human cysteine proteases (Human Calpain 1, Cathepsin L). Compound **1** and **1x** were tested on two host-cell proteases, the human cysteine proteases calpain 1 and cathepsin L. Both compounds showed no inhibitory activity against Calpain 1 (IC$_{50}$ >300 μM) and a 80 and 30-fold decrease of potency on Cathepsin L compared to SARS-CoV-2 3CL^{pro} was obtained respectively for compound **1** and **1x**, suggesting promising selectivity for coronavirus 3CL proteases. In contrast, GC-376 is highly potent on SARS-CoV-2 3CL^{pro}, human Calpain 1 and Cathepsin L (Table 3).

Table 3. Selectivity profile of compounds against human cysteine proteases

Cpd	3CL^{pro} SARS-CoV-2 IC$_{50}$ μM (pIC$_{50}$) [a]	Human Calpain 1 IC$_{50}$ μM (pIC$_{50}$) [b]	Cathepsin L IC$_{50}$ μM (pIC$_{50}$) [b]
1	1.46 (5.85 ± 0.05)	> 300 (>3.52)	122 (3.91 ± 0.19)
1x	1.36 (5.87 ± 0.04)	> 300 (>3.52)	42.8 (4.37 ± 0.10)
GC-376	0.02 (7.64 ± 0.27)	0.016 (7.79 ± 0.05)	0.0007 (9.16 ± 0.11)

Enzymatic assays were performed with GSH (2mM) as reducing agent. The enzymes were pre-incubated with compound at increasing concentrations before starting the reaction with the substrate. Initial rates were used to calculate % inhibitions and IC_{50} were obtained from concentration-response curves by a nonlinear regression analysis of the data.

[a]IC_{50} values are averages of three independent experiments; enzyme and compound pre-incubation of 60 min

[b]IC_{50} values are averages of two independent experiments; enzyme and compound pre-incubation of 30 min.

Antiviral activity of 3CL[pro] inhibitor 1 in SARS-CoV-2 and HCoV-229E live viruses assay

Antiviral effect on SARS-CoV-2. To evaluate the antiviral activity of our 3CL[pro] inhibitors against SARS-CoV-2, compounds **1** and **1x** were tested in a cellular assay in Vero-81 cells stably expressing a fluorescent reporter probe to detect SARS-CoV-2 infection (F1G cells)[30]. As Vero cells are known to express high levels of the efflux transporter P-glycoprotein (P-gp), the assay was performed in presence of the P-gp inhibitor CP-100356 (0.5 µM) that had no antiviral or cytotoxic activity at the concentration used. As can be seen in figure 12, compound **1** demonstrated promising antiviral with micromolar potency (IC_{50} = 1.06 µM) without cytotoxicity at the active doses. For compound **1x**, it was unfortunately toxic at the doses that gave antiviral activity.
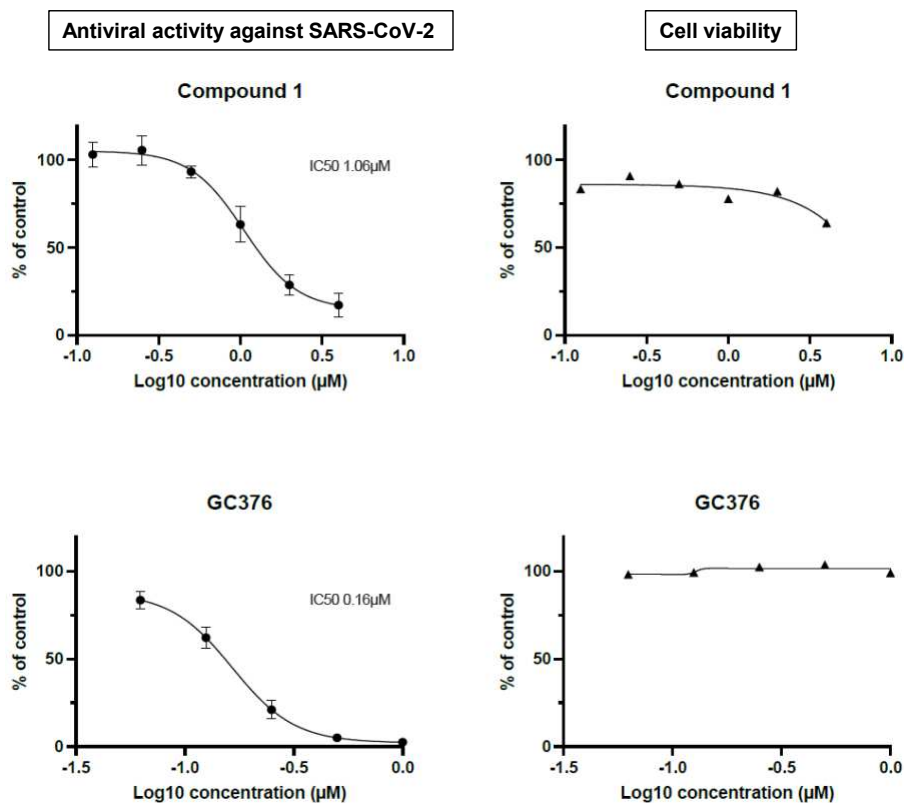
Figure 12. F1G cells were infected with SARS-CoV-2 in presence of P-gp inhibitor (CP-100356) and increasing concentrations of compound **1** (top) or **GC-376** (bottom). 16 h later, cells were fixed in presence of Hoechst 33342. Infected cells (GFP positive nuclei) and the total number of cells (number of nuclei, Hoechst staining) were determined. Results are presented as the percentage of the control and are the average of four independent experiments. Error bars represent the SEM.

HCoV-229E antiviral assay. As compound **1** is a potent inhibitor of the 3CL^{pro} of the human coronavirus 229E, its antiviral activity *in cellulo* on this coronavirus was also evaluated. Compound toxicity was first evaluated in a high-content screening apoptosis assay to determine the range of concentrations devoid of toxicity. This assay allows the rapid detection and quantification of apoptotic Huh-7 cells by detection of Caspase-3/7 activity in real-time imaging using NucView™ 488 substrate containing peptide sequence DEVD attached to a nucleic acid dye

(data not shown). Then, *in cellulo* (Huh-7 cells) antiviral activity on the human coronavirus 229E was assessed using recombinant HCoV-229E expressing the *Renilla* luciferase (Rluc) reporter gene. Rluc activity was measured using the Renilla-Glo luciferase assay system (Promega). As can be seen in Figure 13 compound **1** inhibits viral replication of this coronavirus with a micromolar potency showing that compound **1** is a good starting point to develop a new broad spectrum anti-coronavirus candidate.
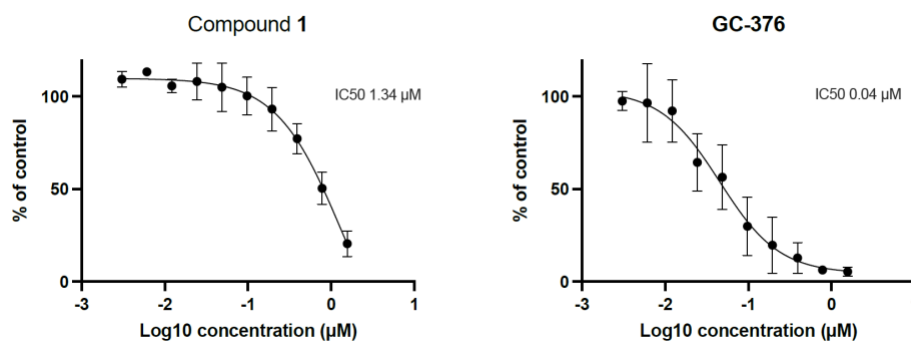
Figure 13. Antiviral activity on the human coronavirus 229E assessed using recombinant HCoV-229E expressing the *Renilla* luciferase (Rluc) reporter gene for compounds **1** (left) and GC-376 (right). Results are the mean from at least 3 independent experiments. Error bars represent the SEM.

CONCLUSIONS. Compound **1** bearing a dithiocarbamate warhead represents a novel class of covalent, 3CL^{pro} inhibitor with potency in the nanomolar range on the 3CL^{pro} of SARS-CoV-2 (IC$_{50}$ = 21 nM). This compound also inhibited the 3CL protease of two other coronaviruses; HCoV-229E (Alpha-CoV) and MERS-CoV (Beta-CoV). The selectivity over host proteases, often challenging with electrophilic covalent inhibitors, was evaluated on two human cysteine proteases and revealed a more than 80-fold selectivity. We also showed that inhibition of the enzyme, performed through the formation of a labile dithiocarbamate with the sulfur atom of the catalytic cysteine, is reversible

within 2 hours, a blockade time likely sufficient to cover the protein turnover time in infected cells. Indeed, a robust cellular antiviral activity in the low micromolar range on both SARS-CoV-2 and HCoV-229E was obtained. Therefore, these results collectively support compound **1**, a selective 3CL protease inhibitor with antiviral activity, as a promising starting point for further development of this series as broad-spectrum anti-coronavirus agents.

EXPERIMENTAL SECTION

3CL^{pro} substrate {Dabcyl}-KTSAVLQSGFRKM-{Glu(Edans)} was purchased from LifeTein (purity >95%). Boceprevir was purchased from MedChemExpress, GC376 from Carbosynth (BG167367), compound **1** from enamine (T5466440).

Expression and purification of the 3CL^{pro} of SARS-CoV-2, MERS-CoV and hCoV-229E. The gene coding for the SARS-CoV-2 main protease (3CL^{pro}) was synthesized, with codon optimization, and inserted into an in-house modified pET24a plasmid in order to produce the 3CL^{pro} fused to a N-terminal 6xHis-SUMO tag in *Escherichia coli* BL21(DE3) (pHis-SUMO-3CL^{pro}). The bacteria were grown at 37°C in Luria-Bertani (LB) medium supplemented with Kanamycin (25µg/mL). When A$_{600nm}$ reached ~0.9, the temperature was lowered to 21°C and induction was carried out with 0.3 mM isopropyl-β-D-galactopyranoside for 12 hours. Harvested cells were lysed using homogenizer (Emulsiflex C-3) in lysis buffer (50 mM Tris.Cl pH 8.0, 300 mM NaCl) supplemented with both DNAseI and RNAseA. The fusion protein was first purified using a HisTrap HP column (Cytiva) and eluted with an elution buffer containing 400 mM imidazole. The 6xHis-SUMO-3CL^{pro} fractions were selected from SDS-PAGE analysis, then pooled and dialyzed (cut-off 6-8kDa) 2 hours at 4°C against 2 L of cleavage buffer (40 mM Tris-Cl pH 7.5, 100 mM NaCl, 5 mM β-mercaptoethanol). The 6xHis-SUMO tag was cleaved by adding

SENP2 protease (His tagged) into the dialysis bag and by dialyzing the sample over-night at 4°C against 2 L of fresh cleavage buffer. The sample was then passed through a HisTrap HP column (Cytiva) to eliminate the SENP2 protease and the 6xHis-SUMO tag. The flow-through containing the native 3CL^{pro} was dialyzed (cut-off 6-8kDa) at 4°C against 2 times 3 L of storage buffer (50 mM Tris-Cl pH7.5, 20 mM NaCl, 1 mM EDTA, 1mM DTT). The 3CL^{pro} was concentrated using a stirred cell with (Amicon) a 10 kDa membrane. A 15.8 mg/mL batch was kept at 4°C whereas a 12.6 mg/ml batch containing 20% glycerol was flash frozen in liquid nitrogen and then stored at -80°C until used. The final yield was about 100 mg per L of culture. The native sequence of the purified 3CL^{pro} was checked by MALDI-tof analysis (Axima Assurance, Shimadzu).

The expression of native main protease from both 229E and MERS-CoV was performed following the same strategy as above. The proteins were stored at -80°C, in storage buffer (20 mM Tris-Cl pH 7.5, 50 mM NaCl, 1 mM EDTA, 2 mM DTT) at 6 mg/mL and 12.6 mg/mL, respectively. The final yields were about 80 mg and 35 mg per L of culture, respectively.

Expression and purification of ^{2}H,^{15}N labelled SARS-CoV-2 3CL^{pro}. The protocol is similar to the one for unlabeled 3CL^{pro} (see above) but with the following modifications. Bacteria were grown in a M9-based semi-rich medium (M9 medium supplemented with ^{15}NH$_4$Cl (1 g/L), D-glucose-C-d7 (3 g/L), Isogro N,D-powder growth medium (0.5 g/L; Sigma-Aldrich), and kanamycin (25 µg/mL). The final storage buffer was (50 mM NaPi pH 6.8, 40 mM NaCl, 0.2 mM EDTA, 3 mM THP). The ^{2}H,^{15}N-labelled 3CL^{pro} was concentrated up to 8.37 mg/mL, flash frozen in liquid nitrogen and then stored at -80°C until used.

NMR analysis of SARS-CoV-2 3CL^{pro} with ligands. All NMR experiments were performed at 305 K using Bruker Neo 900 MHz NMR spectrometer equipped with a cryogenic triple resonance probe (Bruker, Karlsruhe, Germany). The proton chemical shifts were referenced using the methyl signal of TMSP (sodium 3-trimethylsillyl-[2,2,3,3-d4]-propionate) at 0 ppm. Spectra were

processed with the Bruker TopSpin software package 4.0.6. Data analysis was done with Sparky software[31].

NMR data were acquired on 200 µL samples in 3 mm tubes containing 100 µM of ^{2}H,^{15}N-doubly labelled 3CL^{pro} sample in NMR buffer (50mM NaPi pH 6.8, 40 mM NaCl, 3 mM THP, 3% DMSO-d6, 5% D$_2$O) and 2 mM of the ligands. 2D ^{1}H,^{15}N-TROSY-HSQC spectra were acquired with 64 scans and 2048 and 128 complex points in the ^{1}H and ^{15}N dimensions respectively.

Crystallization of SARS-CoV-2 3CL^{pro}. A 3CL^{pro} sample at 5 mg/mL in storage buffer (50 mM Tris-Cl pH 7.5, 20 mM NaCl, 1 mM EDTA, 1 mM DTT) was used for crystallogenesis. Crystals with flower-shape were obtained in 0.2 M sodium formate, 20% PEG 3350 at room temperature. These crystals were crushed with a micro-tool to make a seed stock and new crystals were grown in the same condition using the microseeding technique with a cat whisker.

For the complexes with compounds **1** and **1x**, the 3CL^{pro} crystals were soaked for 1 hour in a solution containing 10 mM compound **1**, 10% DMSO, 0.2 M sodium formate, 20% PEG 3350 and then briefly soaked into 0.2 M sodium formate, 20% PEG 3350, 10% glycerol before freezing.

X-ray data collection and processing. X-Ray data were collected on the Proxima1 or Proxima2a beamlines [32] of the SOLEIL synchrotron facility (Paris, France). The data collection was done remotely using the MXCuBE2[33] software and the crystals were handled by a Staubli sample changer. The data were collected at 100K using an Eiger-X 16M or 9M (Dectris) detector. The data were processed with XDS[34] (xdsme scripts from the synchrotron facility, https://github.com/legrandp/xdsme). The molecular replacement (using the PDB entry 7K3T, DOI: 10.2210/pdb7K3T/pdb) and the refinement steps were done using the CCP4i2 interface[35] of the CCP4 program suite[36]. The statistics for data collection and refinement are summarized in the Table 1 in supporting Information.

The final models and the structure factors corresponding to the 3CL^{pro} bound to N-(pyridin-3-ylmethyl)thioformamide from compounds **1** and **1x** have been deposited in the Protein Data Bank as entries 7NTQ and 8AEB respectively.

The figure 8a has been generated using Pymol (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC

SARS-CoV-2 3CL^{pro}enzymatic assays. For HTS, the FRET-based assay was optimized and miniaturized in Corning 384-wells plates (dark, low-binding, low volume). The assay was conducted at room temperature using a reaction volume of 20 µL and a buffer containing 50 mM HEPES, 0.1 mg/mL BSA, 0.01% Triton and 2 mM GSH at pH 7.5. Compounds stored as 10 mM stock solutions in DMSO were dispensed by acoustic nanodispensing with an Echo Liquid Handler (Labcyte). 10 µL 3CL^{pro} was added to the tested compounds and preincubated for 30 min before addition of 10 µL FRET-based substrate {Dabcyl}-KTSAVLQSGFRKM-{Glu(Edans)}. Final concentrations were 15 nM, 10 µM and 30µM, for enzyme, substrate and compound library respectively. Final concentration of DMSO did not exceed 1%. After 30 min, the fluorescence intensity was monitored with a Victor 3V instrument (Perkin-Elmer) using excitation and emission wavelengths of 340(25) nm and 535(25) nm. Boceprevir was used as a positive reference compound at 4µM (close to the IC$_{50}$ value) and 40µM. Minimal and maximal fluorescence values were obtained in each test plates by incubating the substrate alone (negative control) or with the 3CL^{pro} enzyme (positive control), respectively. Data were normalized intra-plate with these controls: the average of the negative control values was subtracted from all raw data and percentages of inhibition were calculated by the following equation:

$$Inhibition \ (\%) = 100 - \left(\frac{X*100}{positiveControlMean} \right)$$

Z' factors were calculated according to Zhang et al.[37] using mean and standard deviations from positive and negative controls. Plates were validated if their respective Z' factors were ≥ 0.5 and if enzyme inhibition with 40 µM or 4 µM boceprevir were $\geq 80\%$ and $\geq 40\%$, respectively.

For determination of Km or compound IC_{50}, the reaction progress was monitored for 30 minutes to measure the initial velocities used in calculations. Km was obtained by curve fitting with Michaelis-Menton equation using GraphPad Prism 7. IC_{50} values were obtained from concentration-response curves by a nonlinear regression analysis of the data using an equation at four parameters (using XL fit^{TM} 5.2.0.0. from IDBS (Guilford, United Kingdom) or GraphPad Prism 7 (San Diego, USA).

$$y = A + \frac{B - A}{1 + \left(\dfrac{10^C}{x}\right)^D}$$

A, minimum y value; B, maximum y value; C, $LogIC_{50}$ value; D, slope factor.

A jump dilution assay was used to evaluate inhibition reversibility. Compounds are pre-incubated for 1 hour at 10 times their IC_{50} with 100X 3CL^{pro}. Then, the incubates are quickly diluted to 100th with substrate solution before measuring the fluorescence kinetics. Final concentrations after dilution are 0.1X IC_{50}, 15 nM and 10 µM for compound, enzyme and substrate, respectively. In the same experiment, control pre-incubations of 3CL^{pro} with compounds are performed (60 min) with just, as usual, a two-fold dilution with the substrate to obtain 10X IC_{50} and 0.1X IC_{50} final concentrations of compounds. Enzyme and substrate are at 15 nM and 10 µM respectively. All incubates are performed in triplicate.

hCov-229E and MERS-CoV 3CL^{pro} Enzymatic Assays. Incubations are performed similarly to the SARS-Cov-2 3CL^{pro} assay, using the same substrate and the same buffer. Enzyme is at 15 nM and 400 nM for hCov-229E and MERS-CoV 3CL^{pro} assay, respectively. Substrate is at 30 µM

($0.2 \times K_M$) and 60 μM ($0.2 \times K_M$), for hCoV-229E and MERS-CoV 3CL^{pro} assay, respectively. The reaction progress was monitored for 30 minutes to measure the initial velocities and inhibition data expressed as % inhibition, in the same way that for the SARS-Cov-2 3CL^{pro} enzymatic assay.

Calpain Assay. Effect on human Calpain 1 activity was measured using the InnoZyme™ Calpain 1/2 Activity kit purchased by Calbiochem®. Assay was performed in Corning 96-wells plates according to the kit instructions with 1/400 diluted enzyme and 0.5 mM (DABCYL)-TPLKSPPPSPR-(EDANS) substrate in the presence of calcium ions and reducing agent (GSH 2mM). Compounds were pre-incubated for 30 min at room temperature with Calpain 1. The fluorescence (excitation at 320 nm and emission at 480 nm) was measured for 30 min at room temperature with a Victor 3V (Perkin-Elmer) and initial velocities used for inhibition (%) calculations.

Cathepsin Assay. Effect on human Cathepsin L activity was measured using the SensoLyte® 520 Cathepsin L Assay Kit purchased by Anaspec®. Assay was performed in 96-wells plates according to the kit instructions with 1/1000 diluted enzyme and QXL™ 520/HiLyte Fluor™ 488 substrate. GSH (2 mM) was used as reducing agent. Compounds were pre-incubated for 60 min at room temperature with Cathepsin L. The fluorescence (excitation at 490 nm and emission at 520 nm) was measured for 30 min at room temperature with a Victor 3V (Perkin-Elmer) and initial velocities used for inhibition (%) calculations.

Thermal Shift Assay. Fluorescent dye SYPRO® Orange Protein Gel Stain was produced by Sigma-aldrich (S692-500UL). Two compounds have been used as positive shifter of the 3CL^{pro} melting temperature protein: Boceprevir (HY-10237, Medchemexpress), GC-376 (BG167367, Biosynth Carbosynth®). The assay was performed in 50 mM Tris/HCl, pH 7.5, 150 mM NaCl, 2 mM MgCl$_2$ and 1% DMSO in LightCycler® 480 white Multiwell Plate 96 (Roche, 04729692001). First, 5 μL of reference and compound **1** have been transferred in assay microplates. Then, 10 μL

of a 3CL^{pro} of SARS-Cov2 and Sypro Orange Dye mix were loaded to achieve final concentrations of 40 or 100 µM for tested and reference compounds, 5X for Sypro orange dye and 2.5 µM for 3CL^{pro}. After 30 min of incubation at room temperature, assay microplate was launched in LightCycler480 device (Roche) for thermal denaturation of the protein under a temperature gradient range from 37 to 95 °C with 0.05 °C/s incremental step. The melting temperature (Tm) was calculated as the mid log of the transition phase from the native to the unfolded protein using a Boltzmann model in LightCycler_Thermal_Shif_Analysis software v2.0. ΔTm was obtained by subtracting reference Tm of proteins in the presence of DMSO from the Tm in the presence of compounds.

Cellular antiviral activity. *Cells and viruses.* Vero-81 cells stably expressing a fluorescent reporter probe to detect SARS-CoV-2 infection (F1G cells) and Huh-7 cells were grown at 37°C with 5% CO2 in Dulbecco's modified eagle medium (DMEM, Gibco) containing Glutamax and supplemented with 10% FBS (Life technologies).

SARS-CoV-2 virus (hCoV-19_IPL_France strain; NCBI MW575140) was propagated on Vero-81 cells expressing TMPRSS2 at 37°C and recombinant HCoV-229E expressing the Renilla luciferase (kindly provided by Dr Volker Thiel, University of Bern, Switzerland) was propagated on Huh-7 cells at 32°C. After complete lysis of the cells, supernatants containing the viruses were centrifuged, aliquoted and stored at -80°C.

Dose-response Experiments. For SARS-CoV-2 infection assay, F1G cells were plated on coverslips in 24 well-plates. The next day, cells were infected at an MOI of 0.1 in presence of 0.5 µM CP-100356 and increasing concentrations of GC376 or compound 1. Sixteen hours later, cells were fixed with 4 % PFA for 30 min containing 10 µg/ml Hoechst 33342 (life technologies). Cells were rinsed with PBS and mounted on glass slides in Mowiol 4-88 containing medium. Images acquisitions were performed with an EVOS M5000 imaging system (Thermo Fischer Scientific)

equipped with a 10X objective and light cubes for DAPI and GFP. The total number of cells was determined by counting the number of nuclei and the number of infected cells was determined by counting the number of GFP positive nuclei. The experiment was performed four times.

For HCoV-229E infection assays, Huh-7 cells were plated in 96-well plates and infected 24h later in presence of increasing concentration of GC376 or compound **1**. Cells were incubated for 6h and lysed. Luciferase activity was measured by using the renilla luciferase assay system (Promega) and a Berthold luminometer.

Chemistry.

All commercial reagents and solvents were used without further purification. Flash chromatography was performed using a Puriflash®430 with prepacked silica columns. UV detection was used to collect the desired product.

NMR spectra were recorded on a Bruker DRX-300 spectrometer. The assignments were made using one-dimensional (1D) ^{1}H and ^{13}C spectra and two-dimensional (2D) HSQC, HMBC and COSY spectra. Chemical shifts are in parts per million (ppm).

LC-MS Waters system was equipped with a 2747 sample manager, a 2695 separations module, a 2996 photodiode array detector (200-400 nm) and a Micromass ZQ2000 detector. XBridge C18 column (50 mm x 4.6 mm, 3.5 mm, Waters) was used. The injection volume was 20 μL. A mixture of water and acetonitrile was used as mobile phase in gradient-elution. The pH of the mobile phase was adjusted with HCOOH and NH$_4$OH to form a buffer solution at pH 3.8. The analysis time was 5 min (at a flow rate at 2 mL/ min). Purity (%) was determined by reversed phase HPLC using UV detection (215 nm), and all isolated compounds showed purity greater than 95%. HRMS analysis was performed on a LC-MS system equipped with a LCT Premier XE mass spectrometer (Waters),

using a XBridge C18 column (50 mm_ 4.6 mm, 3.5 mm,Waters). A gradient starting from 98% H_2O 5mM Ammonium Formate pH 3.8 and reaching 100% CH_3CN 5 mM Ammonium Formate pH 3.8 within 3 min at a flow rate of 1 mL/min was used.

Potassium 3-pyridylmethylimino(thioxo)methanethiolate (1x). To a solution of 3-(aminomethyl) pyridine (400 mg, 3.70 mmol) in methanol (15 ml) was added CS_2 (1.55 ml, 25.09 mmol, 7.0 equiv) and KOH (137 mg, 3.70 mmol, 1.0 equiv) at 0°C. The reaction mixture was stirred for 2 h at 0°C. Then the reaction mixture was concentrated under reduced pressure and recrystallized in ethanol to afford the desired compound as white amorphous solid (350 mg, 43%). HRMS (ESI): $[M+H]^+$ $C_7H_9N_2S_2$: calcd. 185.0207 found 185.0192. ^1H NMR (300 MHz, CD_3OD): δ (ppm) 8.55-8.51 (m, 1H), 8.41-8.36 (m, 1H), 7.88-7.82 (m, 1H), 7.37 (ddd, 1H, $J = 0.6, 4.8, 7.8$ Hz), 4.88 (s, 2H). ^{13}C NMR (75 MHz, CD_3OD): δ (ppm) 216.3, 149.4, 148.3, 137.6, 137.0, 125.0, 49.1.

2-(chloromethyl)-5-methylimidazo[1,2-a]pyridine (1y). A solution of 1,3-dichloroacetone (593 mg, 4.67 mmol) and 2-amino-6- methylpyridine (500 mg, 4.62 mmol) in ethanol (5.0 mL) was heated to reflux overnight. Then, the reaction mixture was cooled to room temperature and concentrated under reduced pressure. The residue was diluted with aqueous saturated solution of $NaHCO_3$ (5.0 mL) and extracted with ethyl acetate (3 x 5.0 mL). The combined organic layers were dried over $MgSO_4$ and concentrated under reduced pressure. The crude product was purified on silica gel column chromatography eluting by (Hexane/EtOAc 5/5) to give 2- (chloromethyl)-5-methylimidazo[1,2-a]pyridine (250 mg, 30%). LC-MS: $t_R = 1.75$ min. MS (ESI): $[M+H]^+$ 181.00. ^1H NMR (300 MHz, $CDCl_3$): δ (ppm) 7.53-7.47 (m, 2H), 7.16 (dd, 1H, $J = 6.9, 9.0$ Hz), 6.66-6.60 (m, 1H), 4.81-4.78 (m, 2H), 2.57 (s, 3H). ^{13}C NMR (75 MHz, $CDCl_3$): δ (ppm) 145.9, 143.1, 134.8, 125.5, 115.2, 111.9, 108.2, 39.9, 18.8.

(5-methylimidazo[1,2-a]pyridin-2-yl)methyl (pyridin-3-ylmethyl)carbamodithioate (1). To a solution of potassium N-(3-pyridylmethyl)carbamodithioate (100 mg, 0.45 mmol, 1.0 equiv) and

2-(chloromethyl)-5-methyl-imidazo[1,2-a]pyridine (81 mg, 0.45 mmol, 1.0 equiv) in MeOH (2.0 mL) was added Et$_3$N (0.01 mL, 0.67 mmol, 1.5 equiv). The reaction was stirred at room temperature for 3 h. Then, the solvent was evaporated under reduced pressure and the residue was diluted with saturated solution of NaHCO$_3$ and EtOAc. The organic layer was separated and the aqueous layer was extracted for 3 times with EtOAc. The combined organic layers were dried over MgSO$_4$ and concentrated under reduced pressure to give the crude product which was purified though column chromatography eluting by (EtOAc/MeOH 100/0 to 100/10) to give the desired product as yellow solid which upon washing with Et$_2$O gave white solid (130 mg, 85%). LC-MS: t$_R$ = 2.08 min. MS (ESI) m/z = 329.02 [M+H]+. HRMS (ESI): [M+H]$^+$ C$_{16}$H$_{17}$N$_4$S$_2$: calcd. 329.0895 found 329.0896. ^{1}H NMR (300 MHz, CDCl$_3$): δ (ppm) 11.82 (br, 1H). 8.66 (d, J = 1.7 Hz , 1H), 8.49 (dd, 1H, J = 1.5, 4.8 Hz), 7.78-7.72 (m, 1H), 7.33 (s, 1H), 7.22 (dd, 1H, J = 4.8, 13 Hz), 7.13-7.06 (m, 1H), 7.04-7.98 (m, 1H), 6.61-6.56 (m, 1H), 4.97 (d, 2H, J = 4.8 Hz), 4.27 (s, 2H), 2.50 (s, 3H). ^{13}C NMR (75 MHz, CDCl$_3$): δ (ppm) 197.0, 149.8, 148.9, 144.8, 142.8, 136.4, 134.8, 132.4, 126.0, 123.5, 113.8, 112.1, 107.1, 48.3, 33.3, 18.7.

3-(isothiocyanatomethyl)pyridine (1i). To a suspension of NaH (60% dispersion in mineral oil) (163 mg, 4.07 mmol, 1.1 equiv.) in dry THF (10.0 mL), 3-Picolylamine (400 mg, 3.70 mmol) dissolved in dry THF (15 mL) was added dropwise at 0°C. After 1 h, bis(2-pyridyloxy)methanethione (DPT) (868 mg, 3.74 mmol, 1.0 equiv.) was added and the mixture gradually allowed to reach to room temperature and stirred overnight. Then, the reaction was quenched with drops of water and the volatiles were removed under reduced pressure. The residue was diluted with EtOAc and washed with water for 3 times. The combined organic layers were dried over MgSO$_4$ and concentrated under reduced pressure to give dark yellow crude product. The crude product was purified through column chromatography eluting by (Hexane/EtOAc 50/50) to give the desired product as pale-yellow oil (370 mg, 67%). LC-MS: t$_R$ = 2.22 min. MS (ESI$^+$): m/z

= 151 [M+ H]$^+$. ^{1}H NMR (300 MHz, CDCl$_3$): δ (ppm) 8.56 (dd, 1H, J = 1.3, 4.8 Hz), 8.53 (d, 1H, J = 1.8 Hz), 7.68-7.61 (m, 1H), 7.30 (dd, 1H, J = 4.8, 7.8 Hz), 4.71 (s, 2H). ^{13}C NMR (75 MHz, CDCl$_3$): δ (ppm) 149.8, 148.3, 134.6, 134.0, 130.2, 123.8, 46.4.

General procedure A for the synthesis of 1A-1C

To a solution of 3-(isothiocyanatomethyl)pyridine **1i** and thiol or primary amine(1.10 equiv.) in DCM was added Et$_3$N (1.2 equiv.). The reaction mixture was stirred at room temperature for 1 hour. Then, the volatiles were removed under reduced pressure and the residue was diluted by saturated solution of NaHCO$_3$ and EtOAc. The organic layer was separated and the aqueous layer was extracted with EtOAc for more 3 times. The combined organic layers were dried over MgSO$_4$ and concentrated under reduced pressure. The crude product was purified through column chromatography.

1-((5-methylimidazo[1,2-a]pyridin-2-yl)methyl)-3-(pyridin-3-ylmethyl)thiourea (1A). The titled compound was synthesized according to the general procedure A using 3-(isothiocyanatomethyl)pyridine (70 mg, 0.46 mmol), (8-Methylimidazo[1,2-a]pyridin-2-yl)methanamine dihydrochloride (120 mg, 0.51 mmol, 1.10 equiv.) and Et$_3$N (156 mg, 1.54 mmol, 3.30 equiv.) in DCM (5.0 mL). The crude product was purified through column chromatography eluting by (DCM/MeOH 95/5) to give **1A** as a white solid (140 mg, 96%). LC-MS: t$_R$ = 1.70 min. HRMS (ESI): [M+H]$^+$ C$_{16}$H$_{18}$N$_5$S: calcd. 312.1283 found 312.1294. ^{1}H NMR (300 MHz, CDCl$_3$): δ (ppm) 8.45 (s, 1H), 8.38 (d, 1H, J = 4.2 Hz), 7.63-7.55 (m, 1H), 7.39 (s, 1H), 7.21-7.14 (m, 1H), 7.13-7.02 (m, 2H), 6.58 (d, 1H, J = 6.9 Hz), 4.89-4.67 (m, 4H), 3.10 (br, 2H), 2.49 (s, 3H). ^{13}C NMR (75 MHz, CDCl$_3$): δ (ppm) 183.7, 149.2, 148.4, 145.2, 142.8, 135.8, 135.0, 134.0, 125.8, 123.4, 113.7, 112.1, 107.8, 46.1, 42.1, 18.8.

ethyl (pyridin-3-ylmethyl)carbamodithioate (1B). The titled compound was synthesized according the general procedure A using 3-(isothiocyanatomethyl)pyridine (70 mg, 0.46 mmol), ethanethiol

(31.9 mg, 0.513 mmol, 1.10 equiv.) and Et$_3$N (56.6 mg, 0.55 mmol, 1.20 equiv.) in DCM (5.0 mL) then. The crude product was purified through column chromatography eluting by (EtOAc 100%) to give the **1B** as a white solid (95 mg, 96%). LC-MS: t$_R$ = 2.28 min. HRMS (ESI): [M+H]$^+$ C$_9$H$_{13}$N$_2$S$_2$: calcd. 213.0520 found 213.0516. ^{1}H NMR (300 MHz, CDCl$_3$): δ (ppm) 8.50 (d, 2H, *J* = 1.6 Hz), 8.42 (dd, 1H, *J* = 1.5, 4.9 Hz), 7.84-7.77 (m, 1H), 7.38 (dd, 1H, *J* = 4.9, 7.7 Hz), 4.93 (s, 2H), 3.23 (q, 2H, *J* = 12.0 Hz), 1.29 (t, 3H, *J* = 12.5 Hz). ^{13}C NMR (75 MHz, CDCl$_3$): δ (ppm) 200.5, 149.6, 148.8, 137.9, 135.6, 125.1, 48.0, 30.0, 14.8.

benzyl (pyridin-3-ylmethyl)carbamodithioate (1C). The titled compound was synthesized according the general procedure A using 3-(isothiocyanatomethyl)pyridine (70 mg, 0.46 mmol), benzylthiol (63.7 mg, 0.513 mmol, 1.10 equiv.) and Et$_3$N (56.6 mg, 0.55 mmol, 1.20 equiv.) in DCM (5.0 mL) . The crude product was purified through column chromatography eluting by (EtOAc 100%) to give **1C** as a white solid (115 mg, 90%). LC-MS: tR = 2.73 min. HRMS (ESI): [M+H]$^+$ C$_{14}$H$_{15}$N$_2$S$_2$: calcd. 275.0677 found 275.0667. ^{1}H NMR (300 MHz, CD$_3$OD): δ (ppm) 8.50 (d, 1H, *J* = 1.6 Hz), 8.41 (dd, 1H, *J* = 1.5, 4.9 Hz), 7.80-7.74 (m, 1H), 7.40-7.31 (m, 3H), 7.30-7.17 (m, 3H), 4.93 (s, 2H), 4.54 (s, 2H). ^{13}C NMR (75 MHz, CD$_3$OD): δ (ppm) 199.9, 149.6, 148.4, 138.4, 137.9, 135.4, 130.1, 129.5, 128.3, 125.1, 48.3, 40.3.

O-benzyl (pyridin-3-ylmethyl)carbamothioate (1D). To a suspension of NaH (60% dispersion in mineral oil) (15.9 mg, 0.41 mmol, 1.25equiv.) in dry THF (2.0 mL), benzylalcohol (39.6 mg, 0.36 mmol, 1.10 equiv.) dissolved in dry THF (1.0 mL) was added dropwise at 0°C. Then, 3-(isothiocyanatomethyl)pyridine (50 mg, 0.33 mmol) was added and the mixture gradually allowed to reach to room temperature and stirred for further 1 h. The reaction was quenched with drops of water and the volatiles were removed under reduced pressure. The residue was diluted by saturated solution of NaHCO$_3$ and EtOAc. The organic layer was separated and the aqueous layer was extracted with EtOAc for more 3 times. The combined organic layers were dried over MgSO$_4$ and

concentrated under reduced pressure. The crude product was purified through column chromatography eluting by (Hexane/ EtOAc 50/50) to give **1D** as a white solid (73 mg, 85%). LC-MS: t_R = 2.60 min. HRMS (ESI): $[M+H]^+$ $C_{14}H_{15}N_2OS$: calcd. 259.0905 found 259.0898. ^{1}H NMR (300 MHz, CDCl$_3$): δ (ppm) 8.53-8.38 (m, 2H), 7.88-7.67 (m, 1H), 7.57-7.47 (m, 1H), 7.41-7.17 (m, 6H), 5.55-5.48 (m, 2H), 4.80-4.37 (m, 2H). ^{13}C NMR (75 MHz, CDCl$_3$): δ (ppm) 190.8, 189.7, 149.1, 149.0, 148.9, 136.0, 135.7, 135.5, 135.2, 132.9, 132.5, 128.6, 128.5, 128.4, 123.7, 73.5, 72.4, 46.6, 44.6 (mixture of rotamers).

ASSOCIATED CONTENT

Supporting Information

Additional experimental details and materials, including Crystallography data collection and refinement statistics; crystallographic structure of the SARS-CoV-2 3CL[pro] after binding to **1x,** dose-response curves in the enzymatic assay on the MERS-CoV 3CL[pro]; ^{1}H NMR and ^{13}C NMR spectra of synthetic compounds (PDF).

Accession Codes

Coordinates for the crystal structures have been deposited in the Protein Data Bank with ID 7NTQ (compound **1**) and 8AEB (compound **1x**).

AUTHOR INFORMATION

Corresponding Authors

Benoit Deprez

E-mail: benoit.deprez@univ-lille.fr.

[2] Univ. Lille, Inserm, Institut Pasteur de Lille, U1177 - Drugs and Molecules for Living Systems, EGID, F-59000 Lille, France
[3] Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, US 41 - UMS 2014 - PLBS, F-59000 Lille, France

Authors

Lucile Brier

[1] Univ. Lille, Inserm, Institut Pasteur de Lille, U1177 - Drugs and Molecules for Living Systems, F-59000 Lille, France

Haitham Hassan

[1] Univ. Lille, Inserm, Institut Pasteur de Lille, U1177 - Drugs and Molecules for Living Systems, F-59000 Lille, France

Xavier Hanoulle

[4] CNRS, ERL9002 - BSI - Integrative Structural Biology, F-59000, Lille, France.
[5] Univ. Lille, INSERM, CHU Lille University Hospital, Institut Pasteur de Lille, UMR1167 - RID-AGE - Risk factors and molecular determinants of aging-related, F-59000, Lille, France.

Valerie Landry

[1] Univ. Lille, Inserm, Institut Pasteur de Lille, U1177 - Drugs and Molecules for Living Systems, F-59000 Lille, France
[3] Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, US 41 - UMS 2014 - PLBS, F-59000 Lille, France

Danai Moschidi

[4] CNRS, ERL9002 - BSI - Integrative Structural Biology, F-59000, Lille, France.
[5] Univ. Lille, INSERM, CHU Lille University Hospital, Institut Pasteur de Lille, UMR1167 - RID-AGE - Risk factors and molecular determinants of aging-related, F-59000, Lille, France.

Lowiese Desmarets

[6] Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur Lille, U1019 - UMR 9017 - CIIL - Center for Infection and Immunity of Lille, F-59000 Lille, France

Yves Rouillé

[6] Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur Lille, U1019 - UMR 9017 - CIIL - Center for Infection and Immunity of Lille, F-59000 Lille, France

Julie Dumont

[1] Univ. Lille, Inserm, Institut Pasteur de Lille, U1177 - Drugs and Molecules for Living Systems, F-59000 Lille, France
[3] Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, US 41 - UMS 2014 - PLBS, F-59000 Lille, France

Adrien Herledan

[1] Univ. Lille, Inserm, Institut Pasteur de Lille, U1177 - Drugs and Molecules for Living Systems, F-59000 Lille, France
[3] Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, US 41 - UMS 2014 - PLBS, F-59000 Lille, France

Sandrine Warenghem

[1] Univ. Lille, Inserm, Institut Pasteur de Lille, U1177 - Drugs and Molecules for Living Systems, F-59000 Lille, France
[3] Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, US 41 - UMS 2014 - PLBS, F-59000 Lille, France

Catherine Piveteau

[1] Univ. Lille, Inserm, Institut Pasteur de Lille, U1177 - Drugs and Molecules for Living Systems, F-59000 Lille, France

Paul Carré

[1] Univ. Lille, Inserm, Institut Pasteur de Lille, U1177 - Drugs and Molecules for Living Systems, F-59000 Lille, France
[3] Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, US 41 - UMS 2014 - PLBS, F-59000 Lille, France

Sarah Ikherbane

[1] Univ. Lille, Inserm, Institut Pasteur de Lille, U1177 - Drugs and Molecules for Living Systems, F-59000 Lille, France
[3] Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, US 41 - UMS 2014 - PLBS, F-59000 Lille, France

François-Xavier Cantrelle

[4] CNRS, ERL9002 - BSI - Integrative Structural Biology, F-59000, Lille, France.
[5] Univ. Lille, INSERM, CHU Lille University Hospital, Institut Pasteur de Lille, UMR1167 - RID-AGE - Risk factors and molecular determinants of aging-related, F-59000, Lille, France.

Elian Dupré

[4] CNRS, ERL9002 - BSI - Integrative Structural Biology, F-59000, Lille, France.
[5] Univ. Lille, INSERM, CHU Lille University Hospital, Institut Pasteur de Lille, UMR1167 - RID-AGE - Risk factors and molecular determinants of aging-related, F-59000, Lille, France.

Jean Dubuisson

[6] Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur Lille, U1019 - UMR 9017 - CIIL - Center for Infection and Immunity of Lille, F-59000 Lille, France

Sandrine Belouzard

[6] Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur Lille, U1019 - UMR 9017 - CIIL - Center for Infection and Immunity of Lille, F-59000 Lille, France

Florence Leroux

[1] Univ. Lille, Inserm, Institut Pasteur de Lille, U1177 - Drugs and Molecules for Living Systems, F-59000 Lille, France
[3] Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, US 41 - UMS 2014 - PLBS, F-59000 Lille, France

Julie Charton

[2] Univ. Lille, Inserm, Institut Pasteur de Lille, U1177 - Drugs and Molecules for Living Systems, EGID, F-59000 Lille, France

Author Contributions

[‡] LB, HH and XH are equally contributing first authors.

+ Joint last authors: B.D. and J.C. contributed equally.

All authors approved the final manuscript.

Notes

The authors declare no competing financial interest.

ABBREVIATIONS USED

DCM, dichloromethane; DMSO, dimethylsulfoxide; DPT, bis(2-pyridyloxy)methanethione; MERS, Middle East respiratory syndrome; SARS, severe acute respiratory Syndrome; TCEP, Tris(2-carboxyethyl)phosphine; THF, tetrahydrofuran ; 3CL^{pro}, chymotrypsin-like cysteine protease; Mpro, main protease; HPLC, high performance liquid chromatography; HCoV-229E, human coronavirus 229E

REFERENCES

(1) Zhong, N. S.; Zheng, B. J.; Li, Y. M.; Xie, Z. H.; Chan, K. H.; Li, P. H.; Tan, S. Y.; Chang, Q.; Xie, J. P.; Liu, X. Q.; Xu, J.; Li, D. X.; Yuen, K. Y.; Peiris, J. S. M.; Guan, Y. Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, people's Republic of China, in February, 2003. *Lancet* **2003**, *362*, 1353–1358.

(2) Zaki, A. M.; van Boheemen, S.; Bestebroer, T. M.; Osterhaus, A. D. M. E.; Fouchier, R. A. M. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med* **2012**, *367*, 1814–1820.

(3) Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Huang, B.; Shi, W.; Lu, R.; Niu, P.; Zhan, F.; Ma, X.; Wang, D.; Xu, W.; Wu, G.; Gao, G. F.; Tan, W., A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* **2020**, 382, 727−733.

(4) Shitrit, A.; Zaidman, D.; Kalid, O.; Bloch, I.; Doron, D.; Yarnizky, T.; Buch, I.; Segev, I.; Ben-Zeev, E.; Segev, E.; Kobiler, O., Conserved interactions required for inhibition of the main protease of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). *Scientific Reports* **2020**, 10, 20808.

(5) Wang, H.; He, S.; Deng, W.; Zhang, Y.; Li, G.; Sun, J.; Zhao, W.; Guo, Y.; Yin, Z.; Li, D.; Shang, L., Comprehensive Insights into the Catalytic Mechanism of Middle East Respiratory Syndrome 3C-Like Protease and Severe Acute Respiratory Syndrome 3C-Like Protease. *ACS Catalysis* **2020,** 10, 5871−5890.

(6) Fan, K., Ma, L., Han, X., Liang, H., Wei, P., Liu, Y., Lai, L. The substrate specificity of SARS coronavirus 3C-like proteinase. *Biochem. Biophys. Res. Commun.* **2005**, *329*, 934–940.

(7) Anand, K.; Ziebuhr, J.; Wadhwani, P.; Mesters, J. R.; Hilgenfeld, R., Coronavirus Main Proteinase (3CLpro) Structure: Basis for Design of Anti-SARS Drugs. *Science* **2003**, *300*, 1763−1767.

(8) Pillaiyar, T.; Manickam, M.; Namasivayam, V.; Hayashi, Y.; Jung, S.-H. An Overview of Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) 3CL Protease Inhibitors: Peptidomimetics and Small Molecule Chemotherapy. *J. Med. Chem.* **2016**, *59*, 6595−6628.

(9) Cannalire, R.; Cerchia, C.; Beccari, A. R.; Di Leva, F. S.; Summa, V., Targeting SARS-CoV-2 Proteases and Polymerase for COVID-19 Treatment: State of the Art and Future Opportunities. *J. Med. Chem.* **2022**, *65*, 2716−2746.

(10) Konwar, M.; Sarma, D., Advances in developing small molecule SARS 3CLpro inhibitors as potential remedy for corona virus infection. *Tetrahedron* **2021**, *77*, 131761.

(11) Liu, Y.; Liang, C.; Xin, L.; Ren, X.; Tian, L.; Ju, X.; Li, H.; Wang, Y.; Zhao, Q.; Liu, H.; Cao, W.; Xie, X.; Zhang, D.; Wang, Y.; Jian, Y., The development of Coronavirus 3C-Like protease (3CLpro) inhibitors from 2010 to 2020. *Eur. J. Med. Chem.* **2020**, *206*, 112711.

(12) Gao, K.; Wang, R.; Chen, J.; Tepe, J. J.; Huang, F.; Wei, G.-W. Perspectives on SARS-CoV-2 Main Protease Inhibitors. *J. Med. Chem.* **2021**, *64*, 16922–16955.

(13) Kitamura, N.; Sacco, M. D.; Ma, C.; Hu, Y.; Townsend, J. A.; Meng, X.; Zhang, F.; Zhang, X.; Ba, M.; Szeto, T.; Kukuljac, A.; Marty, M. T.; Schultz, D.; Cherry, S.; Xiang, Y.; Chen, Y.; Wang, J. Expedited Approach toward the Rational Design of Noncovalent SARS-CoV-2 Main Protease Inhibitors. *J. Med. Chem.* **2022**, *65*, 2848−2865.

(14) Han, S. H.; Goins, C. M.; Arya, T.; Shin, W.-J.; Maw, J.; Hooper, A.; Sonawane, D. P.; Porter, M. R.; Bannister, B. E.; Crouch, R. D.; Lindsey, A. A.; Lakatos, G.; Martinez, S. R.; Alvarado, J.; Akers, W. S.; Wang, N. S.; Jung, J. U.; Macdonald, J. D.; Stauffer, S. R. Structure-Based Optimization of ML300-Derived, Noncovalent Inhibitors Targeting the Severe Acute Respiratory Syndrome Coronavirus 3CL Protease (SARS-CoV-2 3CLpro). *J. Med. Chem.* **2022**, *65*, 2880−2904.

(15) Zhang, C.-H.; Spasov, K. A.; Reilly, R. A.; Hollander, K.; Stone, E. A.; Ippolito, J. A.; Liosi, M.-E.; Deshmukh, M. G.; Tirado-Rives, J.; Zhang, S.; Liang, Z.; Miller, S. J.; Isaacs, F.; Lindenbach, B. D.; Anderson, K. S.; Jorgensen, W. L. Optimization of Triarylpyridinone Inhibitors of the Main Protease of SARS-CoV-2 to Low-Nanomolar Antiviral Potency. *ACS Med. Chem. Lett.* **2021**, *12*, 1325–1332.

(16) Ma, C. A.-O.; Sacco, M. D.; Hurst, B.; Townsend, J. A.; Hu, Y.; Szeto, T.; Zhang, X.; Tarbet, B.; Marty, M. T.; Chen, Y.; Wang, J., Boceprevir, GC-376, and calpain inhibitors II, XII inhibit SARS-CoV-2 viral replication by targeting the viral main protease. *Cell Res* **2020**, *30*, 678–692.

(17) Hoffman, R. L.; Kania, R. S.; Brothers, M. A.; Davies, J. F.; Ferre, R. A.; Gajiwala, K. S.; He, M.; Hogan, R. J.; Kozminski, K.; Li, L. Y.; Lockner, J. W.; Lou, J.; Marra, M. T.; Mitchell, L. J.; Murray, B. W.; Nieman, J. A.; Noell, S.; Planken, S. P.; Rowe, T.; Ryan, K.; Smith, G. J.; Solowiej, J. E.; Steppan, C. M.; Taggart, B., Discovery of Ketone-Based Covalent Inhibitors of Coronavirus 3CL Proteases for the Potential Therapeutic Treatment of COVID-19. *J. Med. Chem.* **2020**, *63*, 12725−12747.

(18) Hammond, J.; Leister-Tebbe, H.; Gardner, A.; Abreu, P.; Bao, W.; Wisemandle, W.; Baniecki, M.; Hendrick, V. M.; Damle, B.; Simón-Campos, A.; Pypstra, R.; Rusnak, J. M.; EPIC-HR Investigators. Oral Nirmatrelvir for High-Risk, Nonhospitalized Adults with Covid-19. *N. Engl. J. Med.* **2022**, *386*, 1397−1408.

(19) Lee, H.; Mittal, A.; Patel, K.; Gatuz, J. L.; Truong, L.; Torres, J.; Mulhearn, DC.; Johnson, M. E. Identification of novel drug scaffolds for inhibition of SARS-CoV 3-Chymotrypsin-like protease using virtual and high-throughput screenings. *Bioorg Med Chem.* **2014**, *22*, 167−177.

(20) Lee, H.; Torres, J.; Truong, L.; Chaudhuri, R.; Mittal, A.; Johnson, M. E. Reducing agents affect inhibitory activities of compounds: Results from multiple drug targets. *Anal. Biochem.* **2012**, *423*, 46−53.

(21) Cantrelle, F. X.; Boll, E.; Brier, L.; Moschidi, D.; Belouzard, S.; Landry, V.; Leroux, F.; Dewitte, F.; Landrieu, I.; Dubuisson, J.; Deprez, B.; Charton, J.; Hanoulle, X. NMR Spectroscopy of the Main Protease of SARS-CoV-2 and Fragment-Based Screening Identify Three Protein Hotspots and an Antiviral Fragment. *Angew. Chem. Int. Ed. Engl.* **2021**, *60*, 25428−25435.

(22) Sacco, M. D.; Ma, C.; Lagarias, P.; Gao, A.; Townsend, J. A.; Meng, X.; Dube, P.; Zhang, X.; Hu, Y.; Kitamura, N.; Hurst, B.; Tarbet, B.; Marty, M. T.; Kolocouris, A.; Xiang, Y.; Chen, Y.; Wang, J. Structure and inhibition of the SARS-CoV-2 main protease reveal strategy for developing dual inhibitors against M(pro) and cathepsin L. *Sci. Adv.* **2020**, *6*, eabe0751

(23) Jacobs, J.; Grum-Tokars, V.; Zhou, Y.; Turlington, M.; Saldanha, S. A.; Chase, P.; Eggler, A.; Dawson, E. S.; Baez-Santos, Y. M.; Tomar, S.; Mielech, A. M.; Baker, S. C.; Lindsley, C. W.; Hodder, P.; Mesecar, A.; Stauffer, S. R. Discovery, synthesis, and structure-based optimization of a series of N-(tert-butyl)-2-(Narylamido)-2-(pyridin-3-yl) acetamides (ML188) as potent noncovalent small molecule inhibitors of the severe acute respiratory syndrome coronavirus (SARS-CoV) 3CL protease. *J. Med. Chem*. **2013**, *56*, 534−546.

(24) Fu, L.; Ye, F.; Feng, Y.; Yu, F.; Wang, Q.; Wu, Y.; Zhao, C.; Sun, H.; Huang, B.; Niu, P.; Song, H.; Shi, Y.; Li, X.; Tan, W.; Qi, J.; Gao, G. F. Both Boceprevir and GC376 Efficaciously Inhibit SARS-CoV-2 by Targeting Its Main Protease. *Nat. Commun* **2020**, *11*, 4417.

(25) Chen, S.; Zhang, J.; Hu, T.; Chen, K.; Jiang, H.; Shen, X. Residues on the Dimer Interface of SARS Coronavirus 3C-like Protease: Dimer Stability Characterization and Enzyme Catalytic Activity Analysis, *J. Biochem.*, **2008**, *143*, 525−536.

(26) Flynn, J. M.; Samant, N.; Schneider-Nachum, G.; Barkan, D. T.; Yilmaz, N. K.; Schiffer, C. A.; Moquin, S. A.; Dovala, D.; Bolon, D. N. Comprehensive Fitness Landscape of SARS-CoV-2 Mpro Reveals Insights into Viral Resistance Mechanisms. *eLife* **2022**, *11*, e77433.

(27) Maingot, L.; Elbakali, J.; Dumont, J.; Bosc, D.; Cousaert, N.; Urban, A.; Deglane, G.; Villoutreix, B.; Nagase, H.; Sperandio, O.; Leroux, F.; Deprez, B.; Deprez-Poulain, R., Aggrecanase-2 inhibitors based on the acylthiosemicarbazide zinc-binding group. *Eur. J. Med. Chem.* **2013**, *69*, 244−261.

(28) Lee, C. C.; Kuo, C. J.; Hsu, M. F.; Liang, P. H.; Fang, J. M.; Shie, J. J.; Wang, A. H. Structural basis of mercury- and zinc-conjugated complexes as SARS-CoV 3C-like protease inhibitors. *FEBS Lett.* **2007**, *581*, 5454−5458.

(29) Tomar, S.; Johnston, M. L.; St John, S. E.; Osswald, H. L.; Nyalapatla, P. R.; Paul, L. N.; Ghosh, A. K.; Denison, M. R.; Mesecar, A. D. Ligand-induced Dimerization of Middle East

Respiratory Syndrome (MERS) Coronavirus nsp5 Protease (3CLpro): implications for nsp5 regulation and the development of antivirals. *J Biol Chem.* **2015**, *290*, 19403−19422.

(30) Desmarets, L.; Callens, Nr.; Hoffmann, E.; Danneels, A.; Lavie, M.; Couturier, C.; Dubuisson, J.; Belouzard, S.; Rouillé, Y. A Reporter Cell Line for the Automated Quantification of SARS-CoV-2 Infection in Living Cells. *Front Microbiol*, **2022**, *13*, 1031204

(31) Lee, W.; Tonelli, M.; Markley, J. L. NMRFAM-SPARKY: Enhanced Software for Biomolecular NMR Spectroscopy. *Bioinformatics* **2015**, *31*, 1325–1327.

(32) Coati, A.; Chavas, L. M. G.; Fontaine, P.; Foos, N.; Guimaraes, B.; Gourhant, P.; Legrand, P.; Itie, J.-P.; Fertey, P.; Shepard, W.; Isabet, T.; Sirigu, S.; Solari, P.-L.; Thiaudiere, D.; Thompson, A. Status of the Crystallography Beamlines at Synchrotron SOLEIL⋆. *Eur. Phys. J. Plus* **2017**, *132*, 174.

(33) Oscarsson, M.; Beteva, A.; Flot, D.; Gordon, E.; Guijarro, M.; Leonard, G.; McSweeney, S.; Monaco, S.; Mueller-Dieckmann, C.; Nanao, M.; Nurizzo, D.; Popov, A.; von Stetten, D.; Svensson, O.; Rey-Bakaikoa, V.; Chado, I.; Chavas, L.; Gadea, L.; Gourhant, P.; Isabet, T.; Legrand, P.; Savko, M.; Sirigu, S.; Shepard, W.; Thompson, A.; Mueller, U.; Nan, J.; Eguiraun, M.; Bolmsten, F.; Nardella, A.; Milan-Otero, A.; Thunnissen, M.; Hellmig, M.; Kastner, A.; Schmuckermaier, L.; Gerlach, M.; Feiler, C.; Weiss, M. S.; Bowler, M. W.; Gobbo, A.; Papp, G.; Sinoir, J.; McCarthy, A.; Karpics, I.; Nikolova, M.; Bourenkov, G.; Schneider, T.; Andreu, J.; Cuní, G.; Juanhuix, J.; Boer, R.; Fogh, R.; Keller, P.; Flensburg, C.; Paciorek, W.; Vonrhein, C.; Bricogne, G.; de Sanctis, D. MXCuBE2: The Dawn of MXCuBE Collaboration. *J Synchrotron Rad* **2019**, *26*, 393–405.

(34) Kabsch, W. XDS. *Acta Crystallogr D Biol Crystallogr* **2010**, *66*, 125–132.

(35) Potterton, L.; Agirre, J.; Ballard, C.; Cowtan, K.; Dodson, E.; Evans, P. R.; Jenkins, H. T.; Keegan, R.; Krissinel, E.; Stevenson, K.; Lebedev, A.; McNicholas, S. J.; Nicholls, R. A.; Noble,

M.; Pannu, N. S.; Roth, C.; Sheldrick, G.; Skubak, P.; Turkenburg, J.; Uski, V.; von Delft, F.; Waterman, D.; Wilson, K.; Winn, M.; Wojdyr, M. CCP4i2: The New Graphical User Interface to the CCP4 Program Suite. *Acta Crystallogr D Struct Biol.* **2018**, *74*, 68–84.

(36) Winn, M. D.; Ballard, C. C.; Cowtan, K. D.; Dodson, E. J.; Emsley, P.; Evans, P. R.; Keegan, R. M.; Krissinel, E. B.; Leslie, A. G. W.; McCoy, A.; McNicholas, S. J.; Murshudov, G. N.; Pannu, N. S.; Potterton, E. A.; Powell, H. R.; Read, R. J.; Vagin, A.; Wilson, K. S. Overview of the CCP4 Suite and Current Developments. *Acta Crystallogr D Biol Crystallogr* **2011**, *67*, 235–242.

(37) Zhang, J. H.; Chung, T. D.; Oldenburg, K. R. A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. *J. Biomol. Screen.* **1999**, 4, 67–73.

Table of Content (TOC)
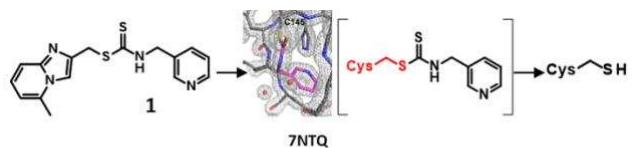
Coronavirus 3CL proteases IC$_{50}$ (µM)			Human proteases IC$_{50}$ (µM)	
SARS-CoV-2	HCoV-229E	MERS-CoV	Calpain 1	Cathepsin L
0.021	0.016	2.00	> 300	122