Novel aspects regarding structural variability in membrane protein families by the example of tetraspanins

kumulative Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Nikolas Reppert

(geb. Hochheimer)

aus

Recklinghausen

Bonn 2022

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

Gutachter: Herr Professor Dr. Thorsten Lang

Gutachter: Herr Professor Dr. Matthias Geyer

Tag der Promotion: 10.10.2022

Erscheinungsjahr: 2022

„Ich habe keine besondere Begabung,

sondern bin nur leidenschaftlich neugierig.“

-Albert Einstein-

Ich widme diese Arbeit meinem Vater

und seiner ansteckenden Neugierde, die mich geprägt hat.

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Dissertation eigenständig und ohne unerlaubte Hilfe angefertigt habe. Es wurden keine anderen als die angegebenen Hilfsmittel verwendet. Direkt oder indirekt übernommenes Gedankengut wurde nach bestem Wissen und Gewissen kenntlich gemacht. Die Arbeit liegt in dieser oder ähnlicher Form keiner anderen Prüfungsbehörde vor.

(Datum) (Unterschrift)

Anmerkung

Teile dieser Arbeit wurden bereits in folgenden Publikationen veröffentlicht:

Chapter 2:

Lang, T. & Hochheimer, N. Tetraspanins. *Curr. Biol. CB* **30**, R204–R206 (2020). DOI: https://doi.org/10.1016/j.cub.2020.01.007

Chapter 4:

Hochheimer, N., Sies, R., Aschenbrenner, A. C., Schneider, D. & Lang, T. Classes of non-conventional tetraspanins defined by alternative splicing. *Sci. Rep.* **9**, 14075 (2019). DOI: https://doi.org/10.1038/s41598-019-50267-0

Chapter 5:

Reppert, N. & Lang, T. A conserved sequence in the small intracellular loop of tetraspanins forms an M-shaped inter-helix turn. *Sci. Rep.* **12**, 4494 (2022). DOI: https://doi.org/10.1038/s41598-022-07243-y

Danksagung

Zuallererst möchte ich mich bei Herrn Prof. Thorsten Lang bedanken, der mir die Chance gegeben hat, in seiner Arbeitsgruppe zu forschen und mich dabei angeleitet und unterstützt hat. Ich bin ihm besonders dankbar für die Freiheiten, die er mir dabei eingeräumt hat, meine eigenen Forschungsinteressen zu verfolgen.

Zu Beginn meiner Zeit am LIMES Institut haben mich besonders Dr. Dennis de Coninck, aber auch Dr. Pascal Weber und Dr. Jérôme Finke betreut, angeleitet und gut aufgenommen. In den Jahren danach haben mich vorallem die fachlichen Diskussionen, die alltägliche Unterstützung im Labor und die Freundschaft von Jasmin Mertins, Lisa Hitschler, Jérôme Finke und Dominik Sons getragen.

Des Weiteren möchte ich mich bei Dr. Michael Tomlinson, Prof. Annemiek van Spriel und Prof. Waldemar Kolanus für die freundliche Bereitstellung von verschiedenen Materialien für meine Arbeit bedanken. Die Arbeit von Dr. Yahya Homsi in unserem Labor möchte ich hier auch erwähnen, da sie einen wichtigen Grundstein für mein Projekt gelegt hat. Zusätzlich habe ich im Rahmen meiner verschiedenen Veröffentlichungen große Unterstützung von Kollegen aus unserem Institut aber auch von außerhalb erhalten und möchte mich daher bei Ricarda Sies, Dr. Anna Aschenbrenner, Prof. Dirk Schneider, Prof. Shoshana Levy, Prof. María Yañéz-Mó, Dominik Sons und Sara C. Schmidt bedanken. Darüber hinaus haben mich Dr. Eileen C. Reppert und Elke Reppert bei der ermüdenden Aufgabe des Korrekturlesens unterstützt. Ein besonderer Dank geht an die Mitglieder meines Promotionskomitees Prof. Thorsten Lang, Prof. Matthias Geyer, Prof. Eva Kiermaier und Prof. Ulrich Kubitscheck.

Ich möchte mich auch bei allen Mitarbeitern des LIMES Institutes bedanken, besonders bei den Mitgliedern der AG Thiele und AG Burgdorf sowie bei Elvira und Jenny aus der Spülküche und allen Mitgliedern der AG Lang, die mich auf meiner Reise begleitet haben: Dennis de Coninck, Pascal Weber, Elisa Merklinger, Nora Karnowski, Vivien Averesch, Jérôme Finke, Lisa Hitschler, Ricarda Sies, Jasmin Mertins, Dominik Sons, Annika Massenberger und Sara C. Schmidt. Die gute Stimmung und positive Atmosphäre, die dank all dieser Menschen entstanden ist, haben meine Zeit am LIMES geprägt.

Da man nur so groß ist wie die Schultern, auf denen man steht, möchte ich hier meiner Familie und meinen Freunden danken. Ohne die Unterstützung und Liebe von meinen Familien- Mitgliedern Jochen, Renate, Fabian, Frederik, Marie, Jakob, Jelle, Jürgen, Elke, Eileen und Vivien sowie meinen guten Freunden Stefan und Hamudi wäre ich niemals so weit gekommen. Gerade die Fürsorge und Unterstützung durch meine Eltern in der Kindheit, Schulzeit und im Studium haben mir meinen Weg geebnet. Auch wenn meine Großeltern nicht mehr bei uns sind, möchte ich mich bei Opa Hans und Oma Liesel, Opa Joseph und Oma Ursel sowie meiner Großtante Anne bedanken. Zu guter Letzt geht mein größter Dank an meine unglaubliche Frau Eileen, die mich durch alle Stimmungshochs und -tiefs in allen Lebenslagen getragen hat.

Danke!

Table of Contents

List of Figures

Abbreviations

Å	Angstrom, 0.1 nm
aa	amino acid
ADAM10	A Disintegrin and metalloproteinase domain-containing protein 10
ADAm17	A Disintegrin and metalloproteinase domain-containing protein 17
ALIX	ALG-2-interacting protein X, part of ESCRT complex
APP	Amyloid-Precursor-Protein
APP-CTF	Amyloid-Precursor-Protein C-terminal fragment
Asx	Aspartate and Asparagine group
ATP	Adenosine triphosphate
BCR	B-cell receptor
CD151	Tspan24
CD19	B-lymphocyte antigen CD19
CD37	Tspan26
CD53	Tspan25
CD63	Tspan30
CD81	Tspan28
CD82	Tspan27
CD9	Tspan29
cryo-EM	Cryogenic electron microscopy
DNA	Deoxyribonucleic acid
ER	Endoplasmic reticulum
ESCRT	endosomal sorting complexes required for transport
EV	extracellular vesicle
EWI-2	Immunoglobulin superfamily member 8
EWI-F	Prostaglandin F2 receptor negative regulator
GCC	glycine-cysteine-cysteine motif
GFP	green fluorescent protein
GYEVM	Internalization motif of CD63
HepG2	immortalized hepatocellular carcinoma cell line
HIV	human immunodeficiency virus
HPV	Human papillomavirus
HSC70	Heat shock 70 kDa protein 8 (part of the ESCRT complex)
HSP90beta	heat shock protein 90kDa alpha family class B member 1, part of the ESCRT complex
LEL	Large extracellular loop
MVB	multi vesicular body
MERS	Middle East respiratory syndrome–related coronavirus
mRNA	messenger ribonucleic acid
NCBI	National Center for Biotechnology Information
nm	nanometer
NMD	nonsense mediated decay
OPM	Orientations of Proteins in Membranes database
PCR	polymerase chain reaction
PTC	premature termination codon
RNA	ribonucleic acid
SEL	small extracellular loop
SIL	small intracellular loop
snRNA	small nuclear ribonucleic acid
snRNP	small nuclear ribonucleoproteins
ST	serine threonine group
TEM	Tetraspanin-enriched microdomain

TMS	Transmembrane segment
tRNA	transfer ribonucleic acid
TSG101	Tumor susceptibility gene 101 (part of the ESCRT complex)
Tspan	Tetraspanin
VEGTR-2	vascular endothelial growth factor receptor 2, part of the ESCRT complex
Φ	phi, dihedral angle between amino acid residue, alpha carbon atom and amino group
Ψ	psi, dihedral angle between amino acid residue, alpha carbon atom and carboxyl group

Abstract

The work at hand focuses on structural variability in membrane protein families by the example of tetraspanins. The family of tetraspanins is found in almost all multicellular organisms and they play important roles in an exceptionally broad range of cellular processes. Tetraspanins are small membrane proteins that contain four transmembrane segments (TMSs). The TMSs are connected on the extracellular side by a small and a large extracellular loop between TMS1-TMS2 and TMS3-TMS4, respectively, and on the cytosolic side by a small intracellular loop between TMS2 and TMS3. The amino and carboxyl termini are also located at the cytosolic site. Two structural aspects of the tetraspanin family were analysed.

First, I addressed the question of alternative splicing and its implications on the family of tetraspanins. A comparative analysis of the known 33 human tetraspanins isoforms listed in the NCBI database revealed a large increase in the number of gene products due to alternative splicing. Surprisingly, these novel isoforms differ in their number of transmembrane domains and therefore represent in fact mono-, di- and trispanins. This allows for a classification into four classes of non-conventional tetraspanins. Expression of these non-conventional tetraspanins shows that alternative splicing can effect ER exit and non-conventional tetraspanins could modulate tetraspanin-enriched microdomain function. To my knowledge, this study is the first analysis of alternative splicing of a whole membrane protein family.

After exploring the realm of structural possibilities in the tetraspanin family that is associated with non-conventional tetraspanins, I refocused my work on segments modulating the structure of conventional tetraspanins. The center of this effort evolved around the small intracellular loop (SIL), the tetraspanins' least studied domain, which connects TMS2 and TMS3. In human and other animal tetraspanins, the sequence analysis of this domain yields a conserved five amino acid core sequence. The two available crystal structures that show the SIL reveal an M-shaped backbone of this loop. The M-motifs dihedral angles cement the turn character of this domain and enforce the inversion of the peptide chain orientation between TMS2 and TMS3. The same M-motif, albeit constituted by different amino acids, is found in many other inter-helix turns of other proteins and presents a new class of inter-helix turn of widespread importance. The glutamate at the second SIL core position is of particular interest, because it forms a salt-bridge in CD9 with a conserved lysine at the N-terminal juxta membrane region before TMS1. This salt-bridge regulates protein expression, localization and interaction. This can be explained by the hypothesis of a hinge-like

mechanism of the salt-bridge that stabilizes the conformational switch between the tetraspanins' helix bundle and funnel shape.

I hope that my findings will stimulate further research towards the structural aspects of tetraspanins. Future studies could investigate the expression patterns of the tetraspanin isoforms and their precise function. Triggered by this systematic analysis, a more detailed picture of the whole tetraspanin family will emerge. Furthermore, it could be a guideline for similar analyses of other membrane protein families. Additionally, my description of the tetraspanin hinge could be another piece of the puzzle that is the tetraspanins' conformation and could help to unravel its function. Along with a better understanding of tetraspanin structure and function, we might in future be able to harness their full potential in terms of therapeutics, diagnostics and biotechnological applications.

Chapter 1: Introduction

1.1 The cell and the cell membrane

The smallest unit of life is the cell, which is the building block of all known living organisms. The simplest organisms consist of only one cell (single-celled organism) but higher organisms can be built of many trillions of cells (multicellular organisms), such as humans that are built of thirty-seven trillion cells[1]. The four basic building materials that make up all living materials are proteins, nucleic acids, saccharides and lipids. A fundamental principle of life is compartmentation. This is reflected in the basic cellular organisation, which consists of an intracellular (cytoplasm) and extracellular space that is separated by the plasma membrane. The subdivision of the cytoplasm into different compartments forms the cellular organelles (s. Fig.1). The different storage principles for the genetic information divides the realms of life. There are organisms with the genome stored open in the cytoplasm (prokaryotes; bacteria and archaea) or tucked into the nucleus (eukaryotes). All higher organisms such as fungi, plants and animals including us humans are eukaryotes and share the same fundamental building plan of their cells.

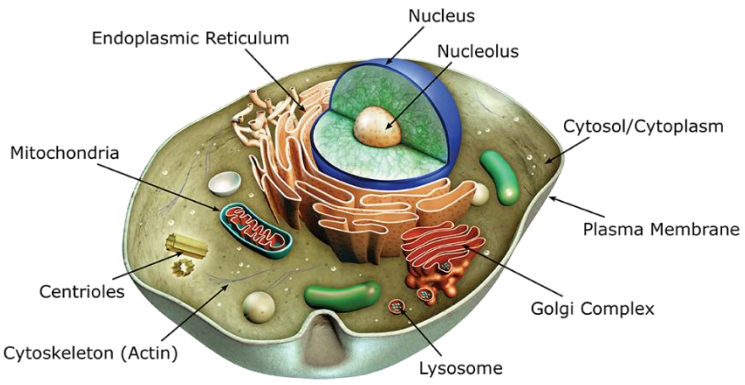


Figure 1: Illustration of a eukaryotic cell.

The key organelles of an eukaryotic cell are shown, such as the nucleus with its nucleolus, endoplasmic reticulum, Golgi apparatus, lysosomes together with some unlabelled endosomes, mitochondria, centrioles, cytoskeleton and the plasma membrane. Image taken from the open access illustration database https://www.aatbio.com/.

A fundamental attribute of all known life is the assembly of amino acids into polypeptides and proteins that can form supramolecular structures. The sequence of amino acids for each protein is based on the construction plan written down in the genetic code. The first step of this process is the transcription of the protein's construction plan from the deoxyribonucleic acid (DNA) into ribonucleic acid (RNA).

The protein coding RNA is also called messenger ribonucleic acid (mRNA) and, in eukaryotes, consists of introns and exons. The mature mRNA is generated by splicing, which results in the removal of the unwanted introns and sometimes also exons. From this process, a variety of possible mRNAs arise, which increase the number of possible gene products. In humans, for example, there are 20,000 genes but 80,000 different gene products created by alternative splicing[2]. The mature mRNA is equipped with a 5'-cap and a 3'-poly-adenosine tail and is transported out of the nucleus into the cytoplasm. Here, the genetic information is translated from the mRNA into an amino acid sequence via the ribosome, which is a macromolecular complex of RNAs and proteins. The growing polypeptide chain is injected into the endoplasmic reticulum (ER) during its manufacturing process. In cooperation with the Golgi apparatus, these two organelles play a major role in protein production, folding and refinement. A majority of post translational modifications such as glycosylation, palmitoylation, disulphide bond formation and many more take place in these compartments. At the end of a protein's lifetime it is unfolded and degraded by a proteasome within the cytosol or at the lysosome, a spherical membrane enclosed organelle that is specialised in breaking down biomolecules. The sorting and delivery of lipid vesicles and their contents is typically organised by endosomes and the endocytic pathway, which involves a conglomerate of intracellular vesicles that can be subdivided in early, late and recycling endosomes. The organelles as well as the cell itself are encased and delimited by biological membranes, which are formed by a lipid bilayer. The plasma membrane contains mainly phospholipids, glycolipids and cholesterol which are asymmetrically distributed between the intra- and extracellular side. The basic model to describe the arrangement of biological membrane components is the fluid mosaic model. It states a two-dimensional liquid-like bilayer formed by lipids with their hydrophobic parts at the membrane's core and the hydrophilic groups at the membrane's surface. The macromolecules like, for example, proteins within membranes, can be pictured as free floating icebergs[3] in a sea of lipids. The limitations of this free lateral movement by the membrane attached cytoskeleton was accounted for in the fences and picket model[4]. In addition, the model of protein clustering[5] and the lipid raft model[6] try to explain the variability in local protein concentrations in biological membranes. Membrane proteins mediate a variety of functions, such as: transport (signals and materials), interaction (with the cytoskeleton, other cells or surfaces), membrane reorganisation (membrane fusion and constriction) as well as enzymatic activity. These proteins can be simply associated to the membrane (peripheral membrane proteins) as well as partially or fully embedded into the membrane (integral membrane proteins).

1.2 Splicing, spliceosome and alternative splicing

The blueprint for the proteome of each cell is encoded in its DNA. The information from the DNA must be transcribed into RNA, which then is translated into an amino acid sequence and forms a protein in the end. Surprisingly, eukaryotic and archaea genes are discontinuous, with coding stretches of DNA (exons) interrupted by non-coding ones (introns). The only domain of life spared by those uncoding stretches is the one of bacteria[7]. As a logical consequence, introns must be removed from the pre-mRNA in order to create a nascent mRNA that encodes a functional protein. This process is called splicing and there are three known processes to catalyse it: the spliceosomal complex, self-splicing and transfer-RNA (tRNA) splicing. In the following, I will focus on the most common form of splicing, which is the spliceosome. It consists of five different ribonucleoprotein subunits (also called small nuclear ribonucleoproteins, snRNPs) and associated cofactors[8]. The small nuclear RNAs (snRNAs) that are part of these subunits associate to the exon-intron junctions. Due to base pairing of these RNAs to the target DNA, the conserved stepwise assembly of the spliceosome starts, which in the end leads to intron removal. The catalytically steps are believed to be based more on the snRNAs than on the associated proteins and are fuelled by the hydrolysis of ATP[8].

The tightly regulated and highly complex splicing process introduces another level of post transcriptional control of gene expression, because wrongly spliced mRNA is filtered out by nonsense-mediated decay (NMD). Additionally, splicing enables a broader spectrum of proteins based on the same number of coding genes by a process called alternative splicing. Alternative mature mRNAs, and consequently proteins, are created via different exon composition or alternative splicing site usage. This wide spread mechanism leads to an increased biodiversity in proteins and affects up to 95% of human multi-exonic genes[9]. However, not all alternative splicing products are increasing cell viability. Features like a premature termination codon (PTC) or intron retention lead to NMD-based degradation of one-third of humans alternatively spliced mRNAs[10] in humans.

1.3 Membrane proteins and their structures

There are 21 proteinogenic amino acids that are defined by their different side chains, which can be charged, polar, un-polar or hydrophobic. Proteins in polar media such as an aqueous solution, bury the hydrophobic residues within their core, which

stabilises the protein and hinders protein aggregation. This does not apply to proteins within hydrophobic media such as a biological membrane, which are able to expose their hydrophobic residues towards the lipophilic environment of the lipid bilayer.

Proteins are made of amino acids that are covalently linked by a peptide bond. The resulting sequence of linked amino acids is also called primary structure. The so-formed continued chain of peptide bonds and alpha carbon atoms is called the protein's backbone, with the side chains of the amino acids (also called residues) branching out to its sides. The backbone of certain sequences is prone to repetitive three-dimensional patterns that are stabilized by hydrogen bonding between the backbone's amino and carbonyl groups. In the hierarchy of protein structure, these elements are called secondary structure elements. The two most common ones are the alpha-helix and the beta-strand/sheet (s. Fig. 2A). The alpha-helix is a right-handed spiral with distinct spatial organisation and the residues within a beta-strand lie in a plane forming a zick-zack pattern. The interaction of neighbouring beta-strands forms a beta-sheet. These secondary structure elements display characteristic dihedral angles between the amino acid's alpha carbon atom and the amino group (Φ) or the carboxyl group (Ψ). The relation between those angles can be visualized in a Ramachandran plot, which simplifies the assignment of amino acids to certain secondary structure elements (s. Fig 2B). The three-dimensional condensation of those secondary structure elements together with their linking unstructured sequence stretches is called tertiary structure. The spatial organisation of multiple proteins within a complex is described as quaternary structure.

Most integral membrane proteins contain one of two structural features (s. Fig. 2C): the first feature is a hydrophobic alpha-helix spanning through the membrane, which is called transmembrane segment (TMS). There are proteins with only one TMS that are called single-spanning and there are examples with multiple TMSs that are called multi-spanning membrane proteins. The family of tetraspanins, which is the focus of this thesis, has four TMSs[11] and is therefore part of the multi-spanning group. The second structure element common for integral membrane proteins is the beta-barrel, which is formed by hydrophobic beta-strands. This conformation is often seen in porins (transporters for small molecules and ions), because the barrel forms a tunnel through the membrane. The types of cargo for such membrane transporters range from ions over small molecules to macro molecules. The outer surface of these beta barrel is exposed to the lipid bilayer and is hydrophobic, whereas the inner surface often displays polar residues enabling specific cargo to pass through.

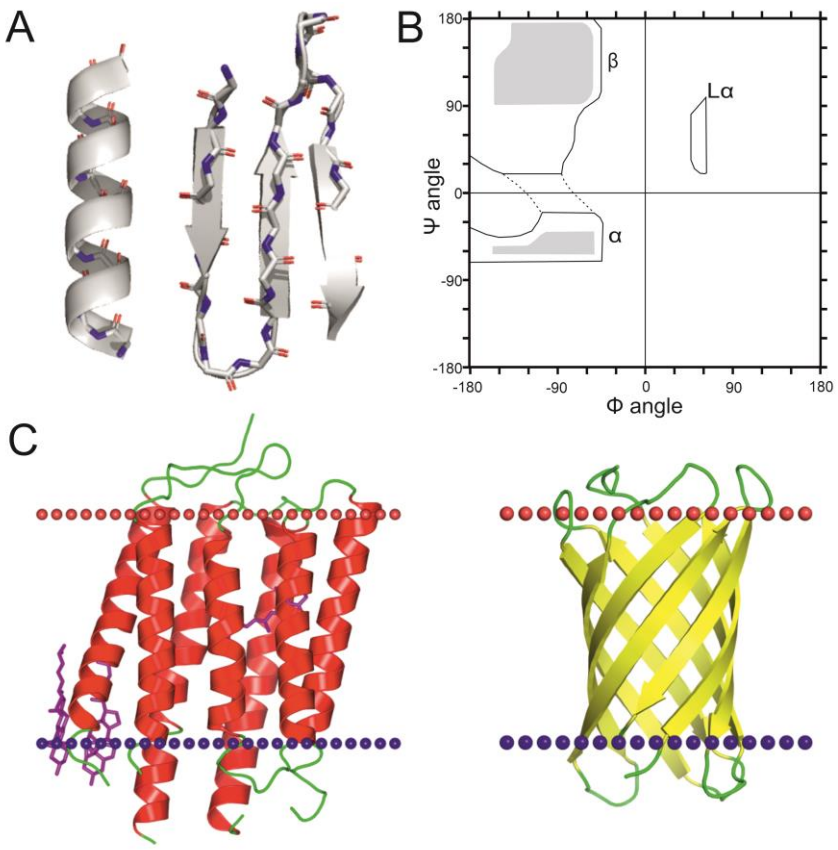


Figure 2: Depiction of secondary structure elements and their embedding into the biological membrane. (A) The most common secondary structure elements are alpha-helices (cartoon spiral, pdb: 3AM6) and multiple beta-strands forming a beta-sheet (cartoon arrows, pdb: 4P79). The backbone carbonyl (oxygen in red) and amino group (nitrogen in purple) of the peptide bonds interact and stabilize the structure along the helix axis or are perpendicular to the beta-strand tying together the beta-sheet. (B) The angles of the peptide chain are characteristic for each structure element (e.g.: alpha- helices, beta-strands/sheets, Lα- left-handed helix) and can be visualized in a Ramachandran plot. (C) Secondary structure elements specific for membrane spanning proteins. The alpha-helix (red) forms transmembrane segments (left, rhodopsin AR2, PDB: 3AM6) and beta sheets (yellow) are the basis for the beta-barrel (right, De novo transmembrane beta-barrel, PDB: 6X9Z). The connecting loops (green) and non-protein molecules such as cholesterol (purple) are shown, too. The membrane is indicated by the dotted lines (extracellular site by red, cytosolic site by blue). The illustration of protein orientation is based on the Orientations of Proteins in Membranes (OPM) database (https://opm.phar.umich.edu/).

The sequence stretches connecting the classical secondary structure elements can be called loops. In many proteins, these loops mediate functions such as protein–protein interactions, ligand binding, forming enzyme active sites and many more[12]. This is also true for tetraspanins, because the most important structure is believed to be the large extracellular loop (LEL) due to its role in protein-protein interaction (e.g. CD81 dimerization[13] and CD81-CD19 interaction[14]). A short loop (≤ 6 residues) that introduces an inversion of the peptide chain direction is also referred to as turn[15]. They are crucial for the protein architecture, because they bring regular secondary structure elements together during protein folding[16]. The turns are classified by the number of residues that are not included in the hydrogen bonding of the flanking

secondary structure elements. There are δ, γ, β , α, and π turns[17] with two to six residues forming the turn. The most common one is the β-turn formed by four amino acids and a characteristic short distance (7Å or less)[18] between the first and last alpha carbon atom of the turn. The β-turns are further subdivided based on the dihedral angles[19] of their backbone. In addition, there is a different category of turns which are formed between the interaction of a side chain with the backbone. They are named Asx or ST group depending on the side chain that forms this interaction (aspartate, asparagine or serine, threonine, respectively). In both groups, the polar side chain forms a hydrogen bond with the carboxyl group of the backbone and forces a bend of the peptide chain[20].

Chapter 2: Introduction to tetraspanins – an overview

In the context of the cluster of differentiation (CD) protocol in the late 1980s, the first tetraspanins were described which led to their names CD9, CD37, CD53, CD63, CD81, CD82 and CD151. The family of tetraspanins was first defined by a group of scientists around Shoshana Levy in 1990[23]. But it took until 2007 for the whole family of 33 human tetraspanins to be first mentioned[24] as such. However, the current tetraspanin research focuses mainly on single tetraspanins and their involvement in certain diseases or physiological processes. In general, little research exists on tetraspanins as a family.

The work at hand is focused on membrane protein families by the example of tetraspanins. Consequently, it resembles one of the few studies addressing the inherent structural and functional features of the tetraspanin family as a whole. Therefore, this chapter, and the published review that it is based on, gives an introduction to this protein family. The topics of tetraspanin structure, their names, subcellular location and function are addressed as well as the open questions in tetraspanin research.

2.1 Abstract

Tetraspanins are small membrane proteins that contain four characteristic TMSs, which are connected by a small (SEL) and a large extracellular loop (LEL) as well as a small intracellular loop (SIL). The N- and C-terminus are located at the intracellular site as well. Based on a crystal structure of the CD81-LEL[25], the first tetraspanin structure was proposed by Seigneuret[26] in 2006, which painted the picture of the TMSs in a compact four-helix bundle with the extracellular domains enthroned on top. In 2016, there was the first crystallographic structure of a full tetraspanin[27] showing a fundamentally different picture of a funnel-shaped helix bundle with the LEL tilted on top the funnel closing the resulting cavity. Both structures were integrated in the conformation switch model[28,29]. Furthermore, tetraspanins are subject to post-translational modification such as palmitoylation, glycosylation and bisulfide-bond formation, which affects their function and structure.

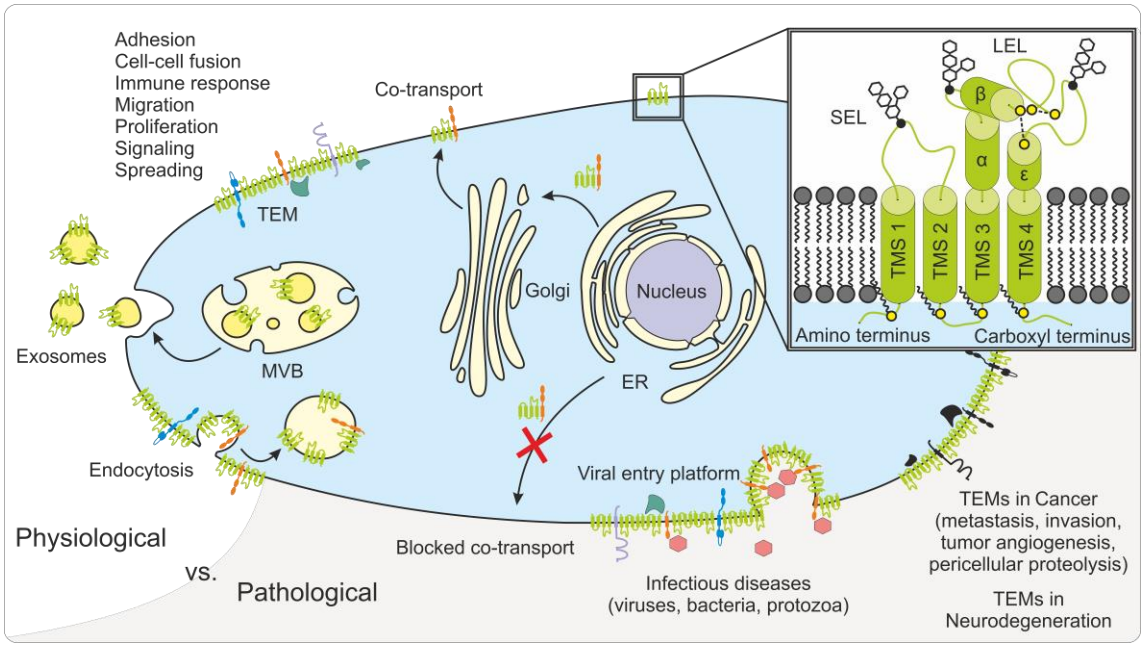


Figure 3: Illustration of different tetraspanin functions and locations in physiological and pathological contexts

On the one side, the physiological role of tetraspanins (green) in co-transport, tetraspanin enriched microdomains (TEMs, interaction partner proteins are depicted in different colours), exosome and endocytosis are shown. On the other side, the pathological processes of tetraspanins are shown, such as blocked co-transport, pathogen entry (viruses are shown as red hexagons) and malfunctioning TEMs (black). The zoomed in image shows a schematic illustration of the tetraspanins' structural features such as helices (barrels), cysteines (yellow) important for intracellular palmitoylation and disulphide bond formation in the LEL as well as potential glycosylation sites (black hexagons). For a more detailed description of this figure, please see the publication: Lang & Hochheimer (2020)[30] which is listed in Appendix A.

Members of the tetraspanin family can be found in all multicellular organisms, whereby some are expressed ubiquitously and others are specific to certain tissues or cell types. Because tetraspanins were discovered as surface antigens on leucocytes, they were first conceptualised as strictly located to the plasma membrane. However, they are also found within intracellular organelles (e.g. endosomes) and extracellular vesicles (s. Fig. 3). Tetraspanins are involved in an astonishingly broad set of cellular processes that can be both physiological and pathological. Their entanglement in many pathological processes, from cancer over neurodegeneration to infection and many more, make them a promising target for new therapeutic strategies. Interestingly, they do not possess catalytic or signaling functions on their own. Their main function is believed to be the spatial organization of a broad range of partner proteins within the plasma membrane. This function might be an explanation for their involvement in seemingly unrelated cellular processes. However, there are speculations on an unknown tetraspanin function that could link all these processes together. One of these hypotheses is the impact of tetraspanins on

membrane curvature. Additional open questions on the tetraspanin field arise from the publication in Chapter 4, such as the poorly described family members and their functions as well as the function of large intracellular domains found in only a hand full of tetraspanins. In general, there are still many open questions and aspects to learn about this family of membrane proteins.

2.2 Review: Tetraspanins

Tetraspanins

Thorsten Lang[1] and Nikolas Hochheimer[1]

[1]Department of Membrane Biochemistry, Life & Medical Sciences (LIMES) Institute, University of Bonn, Carl-Troll-Straße 31, 53115 Bonn, Germany

Curr. Biol. CB **30**, R204–R206 (2020).
DOI: https://doi.org/10.1016/j.cub.2020.01.007

My own contributions: figure preparation and writing the original draft together with Thorsten Lang

This publication is listed in Appendix A.

Chapter 3: Introduction to tetraspanins - in detail

The review in Chapter 2 gives a well-rounded overview of the tetraspanin family that provides the reader with enough information for Chapters 4 and 5. However, this review is meant to be a quick guide to the topic and naturally there are details missing. Consequently, Chapter 3 gives a more complete overview including the details which are missing from Chapter 2.

3.1 Tetraspanin topology and structure

The family of tetraspanins owes their name to the characteristic four TMSs that on average make up around one third of the protein (Fig. 4). They are rather small membrane proteins consisting of 204 – 393 amino acids in humans[31]. The N- and C-terminus as well as the small intracellular loop (SIL) connecting TMS2 and TMS3 are located near the cytosol. The small extracellular loop (SEL) and the large extracellular loop (LEL) are directed towards the extracellular or luminal side, connecting the TMSs 1 and 2 as well as the TMSs 3 and 4, respectively. The largest domain is the LEL, which can be subdivided into a constant and a variable domain and is structured by disulphide bridges[32]. A key feature of the LEL is a GCC (glycine-cysteine-cysteine) motif that marks the crossing of the constant to the variable domain of the LEL. There are three groups of tetraspanins differentiated by their number of disulphide bridge forming cysteines (4, 6 or 8) in the LEL. The extracellular domains of some family members can be glycosylated. An incorrect glycosylation leads to a retention within the endoplasmic reticulum (ER)[33]. Furthermore, the amino acid sequence of each TMS has a distinct heptad repeat motif which is involved in intra- and intermolecular interactions[34].

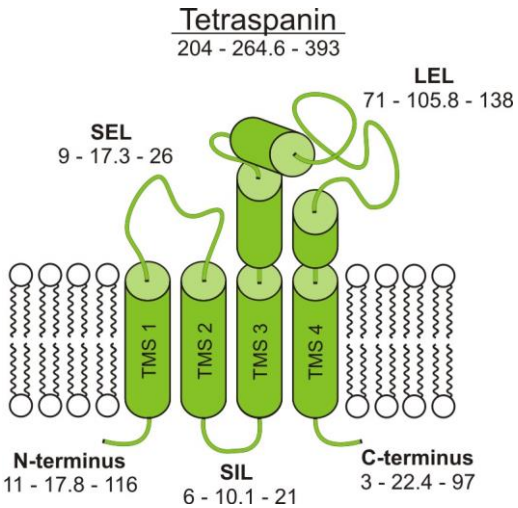


Figure 4: Schematic illustration of a human tetraspanin.

The size of each domain is shown with the numbers referring to the amino acid length of each domain (smallest – mean – largest). The three helices of the LEL that are within the constant part are indicated as well as the variable domain, which is illustrated by the winded loop. Schematized lipids depict the biological membrane. Illustration adopted from Hochheimer et al. 2019[31].

The first structural data were obtained via crystallography of the extracellular domain of CD81 by Kitadokoro in 2001[25]. This set the development of the first structural model of tetraspanins in motion, which was achieved by Seigneuret in 2006[35]. He used molecular modelling, based on the crystallographic structure mentioned before, to paint the picture of a compact four-helix bundle with the extracellular domains enthroned on top. This structure resembles a mushroom with the four helices forming the stalk and the LEL forming the cap. The helix bundle is organized in a left-handed coiled coil with the TMS3 and TMS4 protruding into the LEL helices. Additionally, the upright orientation of the LEL is stabilized by the SEL, which holds a small mostly hydrophobic beta-strand that locks into a hydrophobic groove of the LEL.

The first crystallographic structure of a full length tetraspanin was achieved for CD81 in 2016 by Zimmermann et al. [27]. This structure showed a fundamentally different picture of a funnel shaped helix bundle with the LEL tilted on top (Fig. 5), which seals the resulting cavity. Furthermore, there is an enclosed cholesterol in the tetraspanin bundle docked with its hydroxyl group to a conserved glutamate in the TMS4. The mutation of this glutamate resulted in a diminished cholesterol binding and altered interaction with CD19. The authors deduced a cholesterol induced conformational switch of CD81.

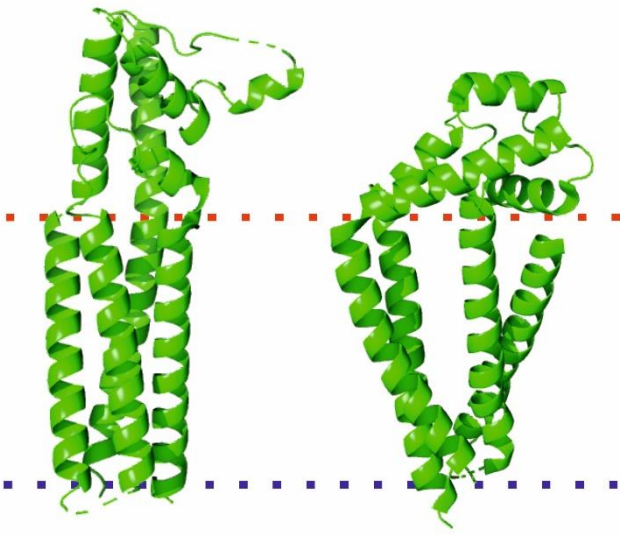


Figure 5: Illustration of the two tetraspanin conformations by the example of CD81.

The compact helix bundle (left, pdb: 7JIC) was found for CD81 in complex with CD19 and contrasts the funnel shaped conformation (right, pdb: 5TCX). The CD81/CD19 complex structure bases on a cryo-EM analysis[14] and the funnel shaped structure was achieved by crystallographic analysis of only CD81 using the lipidic cubic phase method.

There are two more crystal structures of full-length tetraspanins (CD9[36] and CD53[28]) that were solved in the last years and they also show the tetraspanins in the funnel shape, but without a bound cholesterol. The advances in cryogenic electron microscopy led to high resolution images of tetraspanins in complex with hetero proteins (CD81-CD19[14], CD9-EWI-F[37] and CD81/CD19[14]). These structures all show the tetraspanin in a conformation with an erected LEL and a more compact

arrangement of the helices (Fig. 5). The cryo-EM structure of CD81 in complex with CD19 shows a conformation that resembles the one predicted by Seigneuret the most. The current structural model of conformational switching bases on these two shapes that tetraspanins can adopt, which may enable them to participate in a much broader range of interactions[28,29].

The intracellular domains are comparably short and less structured. There are exceptions to that rule such as the comparably large C-terminus of peripherin-2 (Tspan22). This long domain was shown to contain an amphipathic helix. The pheripherin-2 is located at the outer segment of photoreceptor cells and the C-terminal helix interacts with the curved membrane[38]. Another characteristic feature of the intracellular domains of tetraspanins are conserved cysteines that can be palmityolated[39]. These post-translational modifications are believed to stabilize the homotypic (tetraspanins with tetraspanins) as well as the heterotypic (tetraspanins with other proteins) interactions[40,41].

3.2 Physiological functions of tetraspanins

Tetraspanins are associated with an astoundingly broad variety of physiological and pathological cellular processes, such as: adhesion, cell–cell fusion, endocytosis, exosome formation, immune response, migration, neurite navigation, pericellular proteolysis, proliferation, signaling, spreading, trafficking, vascular morphogenesis and remodelling, thrombosis, tumor progression and metastasis, viral and other pathogen entry and viral release[30] (s. Fig. 3).

The extraordinary high number of cellular processes that they are involved in raises the question of their function. It is well established that tetraspanins have no catalytic function or signalling properties themselves, but one of the main characteristics that distinguishes tetraspanins from other membrane proteins is their ability to interact with each other and non-tetraspanins/ heteroproteins. This was first shown for CD81 and its interaction partner CD19 on the surface of B cells[42,43]. It led to the discovery of a wide field of interactions between tetraspanins and heteroproteins such as integrins, members of the immunoglobulin superfamily and signalling receptors[44–46].

The different interactions of tetraspanins were first tested via immunoprecipitation assays, which resulted in three types of interactions that are ranked by their resilience towards detergents: (i) Primary interactions are the most stable ones and are direct interactions between a tetraspanin and one heteroprotein[47]. (ii) Secondary

interactions are formed between tetraspanins themselves and are supported by their palmitoylation[48]. (iii) The weakest interactions are the tertiary ones, which refer to indirect interactions between tetraspanins and heteroproteins within tetraspanin rich areas. These different levels of interaction inspired the hypothesis of the tetraspanin web (Fig. 6A) in which the interaction hierarchies orchestrate the interplay between tetraspanins and heteroproteins[49].

This model was recently expanded by homodimerization studies of Tspan8 within living cells[50], which showed an extraordinary weak Tspan8 dimer interaction. This represents one of the highest dissociation constants ever measured for membrane proteins in living cells. The authors of this study proposed a crucial function of the short dimer lifetime for the Tspan8 function.

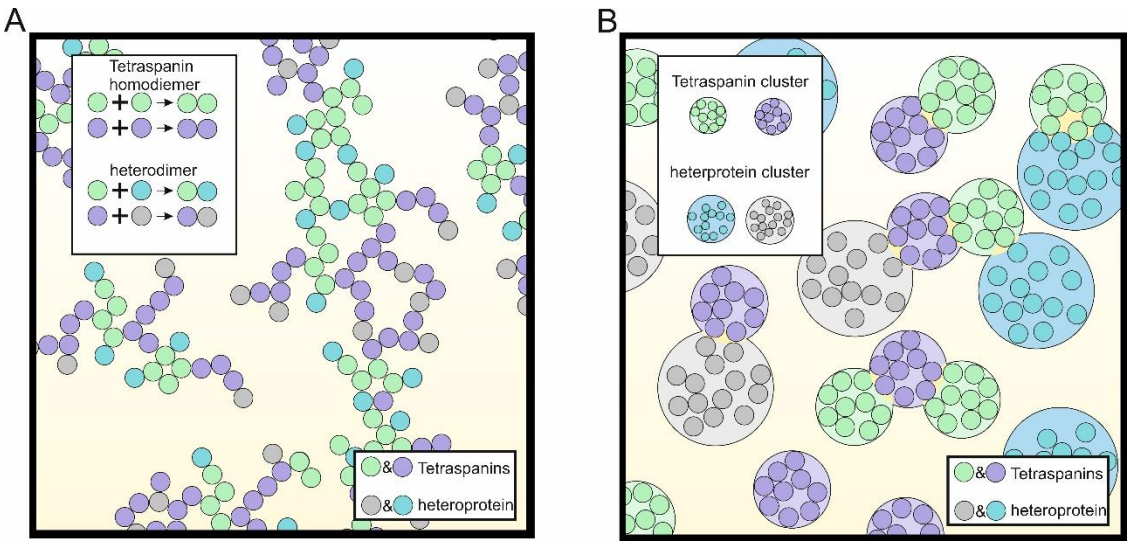


Figure 6: Comparison of the two models describing the tetraspanin arrangement in the plasma membrane.

(A) The tetraspanin web model describes an interaction network of single molecules based on the dimerization capacity of tetraspanins and their ability to bind to heteroproteins. (B) The contrasting model shows small clusters of tetraspanins (less than 10 molecules) and heteroproteins. The clusters are spatially separated or overlap slightly. The clusters (lighter coloured big circles) are shown to overlap (yellow area) and are formed by less than ten molecules of tetraspanins. Please note that within this illustration tetraspanins can interact with each other, but only with one heteroprotein each.

Furthermore, tetraspanins are known to cluster (Fig. 6B) with partner proteins and cholesterol in so called tetraspanin-enriched microdomains (TEMs)[51]. With the help of super-resolution microscopy it was shown that small clusters of tetraspanins (around 10 molecules in a cluster of 120nm diameter) are adjacently positioned in the membrane, but they hardly show any overlap. However, they revealed relatively closer proximity to their partner proteins[52]. These findings challenge the solely biochemically-based picture of the tetraspanin web. The most recent advance on this

topic is the molecular modelling of multiple CD81 molecules of the funnel shape in a lipid bilayer. This simulation analysed the cluster architecture of CD81 and revealed a cluster formation process that includes a surprising segregation of one CD81 molecule from the cluster. This excluded CD81 molecule adopted the closed helix bundle conformation[53] within this simulation.

The unifying idea behind these two models is the ability of tetraspanins to interact and organize heteroproteins and thereby to facilitate their functions. This is the reason for the above-mentioned nickname ‘master organisator of the plasma membrane’. The functions of the heteroproteins are believed to be facilitated and directed via spatial organisation in the tetraspanin network and by conformational rearrangement upon tetraspanin binding. To name a few examples, the tetraspanin network brings together components of signaling complexes and through this process enables their function (e.g. CD63 links VEGFR2 to beta1-integrins or CD81 enables CD19 and BCR cross talk)[44]. Furthermore, the substrate affinity of certain proteases can be altered via conformational changes induced by tetraspanin binding (e.g. scissor sister model, substrates affinity of ADAM10 differs upon binding to Tspan5, 10, 14, 15, 17 or Tspan33) [54]. A new aspect of tetraspanin function was proposed in a recent publication showing a computationally docked complex of the Angiotensin-converting enzyme 2 (ACE2) with A disintegrin and metalloprotease 17 (ADAM17) and CD9, which force a local membrane curvature[55].

The key domain that facilitates these interactions is the LEL (Fig. 7) with its variable domain that mediates the tetraspanin’s individual interaction properties[56,57]. The variable domain differs in sequence as well as in its secondary structure elements between tetraspanins, which resembles the unique fingerprint of each individual family member[58]. Additionally, the homotypic interaction and clustering between tetraspanin monomers is supported by the LEL’s delta-loop, which was shown for CD81[13,59]. However, the tetraspanins’ interaction capacity exceeds the extracellular domain and includes the TMSs and N- and C-termini as well.

TMSs mediate intra- and intermolecular interactions that are important for tetraspanin TMS organisation and dimerisation[34]. Furthermore, the interplay with heteroproteins is modulated by the transmembrane helices such as the ER clearance of CD19, which is based on the CD81-TMS1 association[49]. A similar mechanism is found for CD5 in association with CD9[61]. In addition, the integration of cholesterol into the four helix bundle of CD81 is mainly influenced by a glutamate in the TMS4[27], which is also

present in other tetraspanins[34]. The conformational influence of cholesterol on CD81 can be described as an allosteric mechanism[62].

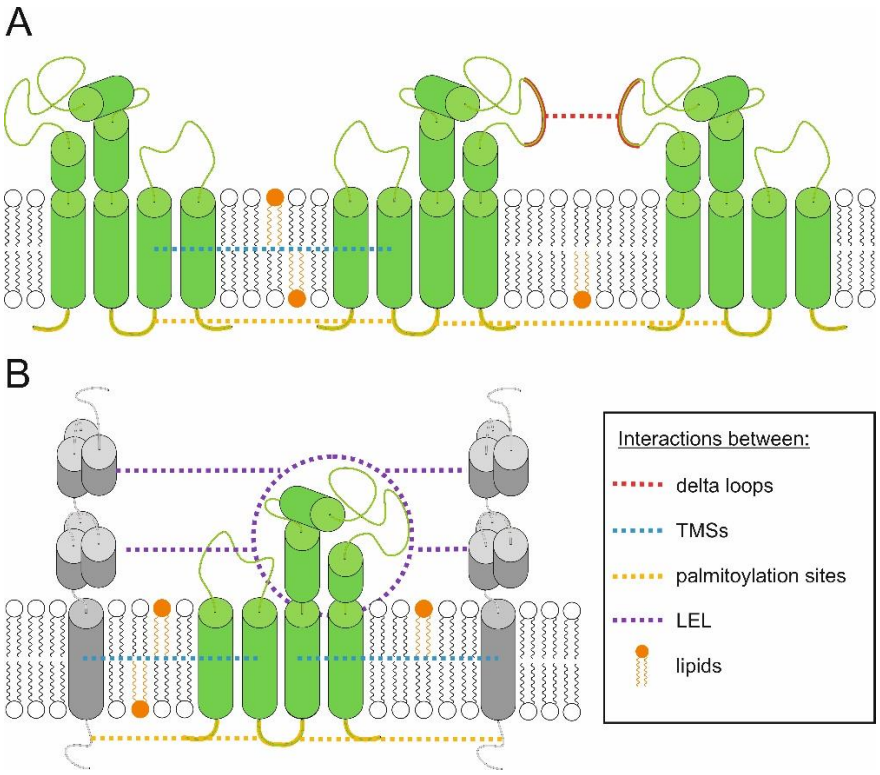


Figure 7: Schematic illustration of tetraspanin interactions.

(A) Interactions between tetraspanins (green) are shaped by direct interactions of the LEL (red dotted line, with focus on the delta-loop) and the TMSs (blue dotted line). Further connections are based on intracellular cysteine palmitoylations (yellow dotted line) and different lipids (orange) such as gangliosides and cholesterol. These interactions are also known as secondary interactions, because of their lower stability against detergents in immunoprecipitation experiments. Well-defined locations of interaction are colour-coded (e.g. delta loop and intracellular palmitoylation sites) and broader interaction surfaces (e.g. TMSs and LEL) are not. (B) Similar mechanisms affect the heteroprotein interactions (primary interactions). The direct connection between the interaction partners (grey) and tetraspanins is mediated by the LEL (purple dotted line). However, the above mentioned other mechanisms play a role as well. Image inspired by Deventer et al., 2021[60].

The intracellular domains are implemented in many functions, which is surprising for such short domains. The internalisation and following degradation of tetraspanins is linked to their ubiquitination of N-terminal lysines (e.g. CD81[63] and CD151[64]). The SIL is the shortest and least studied tetraspanin domain and it is only known that the SILs of CD81 and CD82 are involved in virus protein association[22]. The most divers intracellular domain is the C-terminus. As mentioned above, the long C-terminus of Peripherin-2 (Tspan22) holds an amphipathic alpha-helix that mediates membrane association and curvature[38]. Additionally, the most known internalisation sites of

tetraspanins are located here[65]. Moreover, direct interactions to heteroproteins are associated with this domain, influencing the tetraspanins' function and spatial organisation[66] of tetraspanins.

Post-translational modifications play a role in tetraspanin trafficking and interactions as well. On the extracellular site, a number of family members are N- and O-glycosylated. These modifications are important for the correct maturation and trafficking of the protein: as it was shown for Tspan1 and its four distinct glycosylation sites that impact the transition of the protein through the ER[33]. On the intracellular site, there are conserved cysteine residues at juxtamembrane positions that were shown to be palmitoylated[39]. These fatty acid additions are required for homotypic as well as heterotypic interactions, for example the association of CD9 with CD63 or integrin alpha(IIb)beta(3) on the surface of platelets[67].

The lateral association of tetraspanins is further determined by the lipid composition of the surrounding membrane. As mentioned before, cholesterol integrates into tetraspanins and furthermore seems to interact with the intracellular palmitoylations and thus contributes to the lateral association with other proteins[68]. Consequently, cholesterol impacts the organization of tetraspanins into TEMs as well[69]. However, there are more lipids with a connection to tetraspanins, for example gangliosides. The destabilising effect of ganglioside depletion on the formation of CD82 complexes and overall TEM integrity[70] was shown.

3.3 Where to find them?

The best-studied organism regarding tetraspanins is the human (*Homo sapiens*) but they are spread much more widely. Thus, they have been found in most multicellular organisms. The number of tetraspanin family members differs between species. For example, there are 33 in *Homo sapiens*[31], 31 in *Mus musculus*[31], 50 in *Danio rerio*[71], 34 in *Drosophila melanogaster*[72] and 17 in *Arabidopsis thaliana*[73]. It was even shown that the connection between multicellularity and tetraspanins is no coincidence, as tetraspanins coevolved with cell-cell interactions and they correlate with multicellularity[74].

Some family members are expressed ubiquitously (e.g. CD9, CD63 and CD151), but some are specific to certain cell types. For example, Tspan32 is only found on hematopoietic cells[75]. The family members found abundantly on the surface of leukocytes still represent the best-studied group. This is based on their close relation

to hematopoietic cells and regulation of cellular processes, including stem cell quiescence, migration, and niche adhesion[76].

First, tetraspanin research focused on the plasma membrane located protein pool. This was based on the history of tetraspanin discovery, but they are also found in intracellular membranes (e.g. CD63[77]). Naturally, they pass through the ER and Golgi apparatus as all integral membrane proteins do. Some tetraspanins show specific interactions in these compartments that enable their pathing. The analysis of Uroplakin-1a and -1b (Tspan21 and Tspan20) revealed their need of forming hetero dimers with Uroplakin-2 and -3 to be able to exit the ER[48].

Apart from the ER and Golgi apparatus, they are found in a variety of vesicles from endosomes to extracellular vesicles. The tetraspanin that is best known to locate into endosomes is CD63, which reflects in its alternative name of Lysosome-associated membrane glycoprotein 3 (Lamp-3). Hence, CD63 is not trapped in late endosomes and lysosomes, but it can escape and recycle towards other compartments[78]. Their endosomal location reflects in a variety of internalization motifs found on their intracellular domains, which is often a tyrosine-based sorting motif (YXXΦ, with X being any residue and Φ being a bulky hydrophobic one)[65]. It was shown for CD63 that a C-terminal GYEVM motif targets the protein towards the lysosome via the early endosome system in an adaptor complex AP-3 dependent manner[79]. The ability to reenter the cytosol from the plasma membrane and even carry heteroproteins along was shown for CD63[80] and CD151[81]. The role of tetraspanins in extracellular vesicles gained more and more attention over the last years and is addressed in the next chapter.

3.4 Tetraspanins and extracellular vesicles

More recently, the role of tetraspanins in extracellular vesicle (EV) formations and trafficking has come into focus. A variety of vesicles is known to facilitate intercellular communication by carrying functional membranes, cytosolic proteins, lipids and RNAs between cells. They are referred to as microvesicles, ectosomes or exosomes (and many more) depending on their origin, composition, size[82] and density[83]. Here, I will focus on exosomes, which are small extracellular vesicles with a typical diameter of around 30-150 nm[84]. They are formed by inward budding of early endosome membranes, which leads to the maturation into multicellular vesicles (MVBs). These organelles are mainly linked with endocytosis and trafficking and are involved in protein sorting, recycling, storage, transport and release[85].

There are two possible destinations of an MVB. The first one is the fusion with a lysosome and the resulting degradation of its cargo. The second path is the merger with the plasma membrane and the consequent release of its cargo (e.g. exosomes) into the extracellular space. The factors shaping the fate of an MVB are not well understood. It was shown that the cholesterol content of the MVBs influences their fate and cholesterol-rich MVBs are less likely to end up in a lysosome[86]. The formation of exosomes within the MVBs is orchestrated by the endosomal sorting complex required for transport (ESCRT) pathway[87]. The exosome cargo is versatile and consists of signal transducer proteins, RNAs, DNA, lipids and metabolites that are used for intercellular communication[88]. Furthermore, their proteome naturally consists of many ESCRT proteins such as Alix, TSG101, HSC70, and HSP90β. Apart from the components of this sorting machinery, tetraspanins such as CD63, CD9 and CD82[84] are abundant on exosomes as well. At first, tetraspanins were only seen as EV markers, but their function in EV formation and function was soon realized[83]. The pathway whereby tetraspanins are targeted towards exosomes is still unknown and may be different for each individual family member[89]. A possible targeting pathway might be the tetraspanins' well-defined internalization motifs[65]. Not only single tetraspanin molecules are involved in exosome targeting but TEMs can also act as sorting machinery towards exosomes[90]. The most basic function of exosomes is their ability to be selectively taken up by cells. Unfortunately, this is also a poorly understood aspect of their biology. The organization of adhesion molecules (e.g. integrins[46], intercellular adhesion molecule-1 and vascular cell adhesion molecule-1[91]) on the exosome's surface is believed to be crucial for exosome uptake and is influenced by tetraspanins. Interestingly, the proteins on the exosome surface are distributed unevenly, which was shown very recently[92].

3.5 Tetraspanins and disease

The attention of the scientific community was drawn towards tetraspanins because of their involvement in many pathological processes. They are involved in many cellular processes and consequently also in many pathological processes such as cancer, neurodegenerative diseases, autoimmune diseases, inflammation, viral and bacterial infections.

Their influence on adhesion, protease activity and signalling make them a key player in cancer development via regulating cell migration, invasion and metastasis[93]. This makes them a promising target for cancer therapeutics such as anti-tetraspanin

antibody treatments, which has shown an impact on tumor progression[94]. A prominent example is CD151, which modulates tumor cell activity via connecting to growth factors, matrix-metalloproteases and integrins[95]. However, not only the expression pattern of tetraspanins is connected with cancer, but their post-translational modifications such as the degree of CD63 glycosylation is also implicated in clinical outcomes of breast cancer patients[96].

Their link to the development of Alzheimer's disease is explained by their action in the APP processing/ degradation pathway (e.g. APP processing by alpha–secretases like ADAM10[54,97], Tspan6 determines APP-CTF degradation or secretion[98]), and their function in exosomes links them to the spreading and seeding[99]of the disease. Moreover, they play an important role in a number of different diseases such as: autoimmune diseases like diabetes type 1 (Tspan7 acts as an autoantigen and induces autoimmunity against beta cells in the pancreas)[100], pulmonary emphysema[101], osteopenia[101], leukemia[102] and defects in blood and lymphatic vessels[103].

Furthermore, they can be hijacked by pathogens like parasites, fungi, bacteria or viruses to gain access into the cell[104]. An interesting example for tetraspanin involvement in virus infections is the Zika virus. It was shown that Zika virus-infected cells secrete a characteristic subset of EVs that contain viral proteins and genomes that are able to infect other cells upon uptake[105]. As a result of Zika infection, the production of EVs is enhanced and their size and density distribution changes. This process is influenced by the EV enriched tetraspanin CD63. Other viruses connected to tetraspanins are HIV[106,107], HPV[108], MERS[109] and other Corona viruses[110] as well as influenza viruses[110]. A very recent publication highlights the role of tetraspanins in SARS-CoV-2 infection[55]. This bioinformatic analysis links the shedding of the ACE2 receptor, which is the binding partner of the viral spike protein and therefore crucial for infection, with the metalloprotease ADAM17. The author further states that ADAM17 can be regulated by tetraspanin interaction. Consequently, tetraspanins are involved in the SARS-CoV-2 infection. An example for the involvement of tetraspanins in bacterial infection is the formation of multinucleated giant cell (MNGC) in mouse macrophages during a *Burkholderia thailandensis* infection[111]. The tetraspanins CD9 and CD81 are both involved in the process of MNGC formation and CD9 turned out to be a negative regulator of this process[111].

3.6 Applications in diagnostics, therapy and biotechnology

As discussed before, tetraspanins are connected to many pathological processes, which makes them interesting for diagnostic purposes. A topic that has recently emerged is the connection between tetraspanins and EVs. The EVs extracted from peripheral blood provide a powerful tool in diagnostics. This approach is still in its infancy and standardized protocols for EV isolation from bodily fluids are needed, which is a key factor of EV research[112]. Most EV isolations base on different centrifugation steps, which sometimes are accompanied by size filtration[113]. Currently, flow cytometry strategies are used for EV analysis, because they manage without EV isolation or concentration[114] and allow for smaller sample volumes and decreased overall processing time. The detection of EVs in many assays is based on tetraspanins because of their abundancy on EV surfaces. The range of EV-indicated diseases is broad and includes placental dysfunction[116], cancer [117], infections[118] and many more. Even drug abuse (like methamphetamine) can be diagnosed by EV analysis[106].

These small vesicles are not only used in diagnostics but are also potential candidates for therapy strategies. For example, exosomes could be used as novel drug delivery system in cancer therapy[119]. In recent years, multiple synthetic drug delivery systems were developed; however, their applications are limited because of inefficiency, cytotoxicity and/or immunogenicity[120]. As a counter-draft to the synthetic systems, the EVs possess favorable characteristics as drug delivery systems, such as their ability to cross physical barriers, specific targeting, natural intracellular trafficking pathways and their biocompatibility. There are even concepts of EV drug loading via an endogenous cellular EV packing machinery. The targeting towards specific cells could be arranged by modifying the surface protein composition of EVs, which involves heat shock proteins, lysosomal-associated membrane proteins, cytoskeletal proteins and tetraspanins[121]. Besides targeting drugs towards their destination, EVs can act as drugs themselves. This new approach was developed as an anti-SARS-CoV-2 therapy, in which EVs were loaded with a modified CD63 attached to an anti-SARS-CoV-2 nanobody, which binds at the receptor-binding domain of the Spike protein[122]. These modified EVs were able to neutralize the functional SARS-CoV-2 viruses.

The connection between tetraspanins and disease makes them a promising starting point for new therapeutics with a special focus on their large extracellular domain. In cancer therapy, there is the idea of using the LEL to label cancerous cells in

immunotherapy[123], which might also be used to counteract tumor angiogenesis[124]. Moreover, anti-tetraspanin antibodies can be used in the fight against pathogens by blocking the pathogen docking sites. Additionally, there is an approach in which tetraspanin-mimicking peptides are used to counteract infections on the pathogen's site[125,126].

3.7 Aim of the work

The work at hand addresses novel aspects regarding structural variability in membrane protein families. This analysis was done by the example of the tetraspanin membrane protein family with a focus on two main aspects:

(1) The impact of alternative splicing and the result in structural variability. There are examples of the alternative splice analysis for single membrane proteins, but the analysis of a whole protein family is a new approach, to the best of my knowledge. This family-wide strategy was chosen for its potential in detecting underlying patterns and systematic classification. The family of tetraspanins was picked as subject for this analysis, because their 33 members show strong homology. Furthermore, the considerable number of alternatively spliced RNAs on the NCBI database represent a solid foundation for this analysis.

(2) Analysis of small inter-helix connecting loops regarding their structure and interactions by the example of the small intracellular loop in tetraspanins. This particular sequence was selected because it was shown for some family members to be conserved between species[21]. Moreover, it is rather unnoticed in prior tetraspanin research for there is only one previous publication[22] that focuses directly on this domain.

Chapter 4: Alternative splicing within the tetraspanin family

Alternative splicing is a common process within eukaryotic cells, which increases the diversity of the gene products that are encoded in the genome. This phenomenon was first described in 1977 for the mRNA of adenovirus 2[127] and since then an enormous number of other examples was discovered. Today, we know that up to 95% of human multi-exonic genes[9] are affected by alternative splicing leading to approximately 80,000 different gene products based on the 20,000 human genes[2].

While it is well established that most human gene products are affected by alternative splicing, the precise impact on RNA and protein structure has not been studied in most cases. Especially the effects on small membrane proteins are less described and the implications for a membrane protein family as a whole have never been analyzed, to the best of my knowledge. The family of tetraspanins pose an ideal subject for such an endeavor, because there is a great number of alternatively spliced tetraspanin mRNAs listed on the NCBI database in addition to the 33 conventional ones. Only one isoform was previously described, which is the alternative splicing product of CD82[128]. Furthermore, alternative splicing of tetraspanins is likely to affect domain composition. This assumption is based on their rather small size, which results in tightly-packed protein domains. Therefore, a great impact on protein structure and function can be envisioned for alternatively spliced tetraspanins.

4.1 Abstract

The basis for understanding the effects of alternative splicing on the tetraspanin family was the analysis of the available mRNA sequences and the categorization of the different splicing patterns. The 33 human tetraspanin genes yielded a list of 89 mRNAs including 31 mRNAs that code for non-conventional tetraspanins, which are defined by an alternative amino acid sequence. In addition, there were multiple mRNAs that did not differ in their open reading frames, but in their 5'- or 3'-untranslated regions (UTRs). The comparison of the conventional tetraspanins with the alternatively spliced ones allowed for a first prediction of the structural changes (Fig. 8).

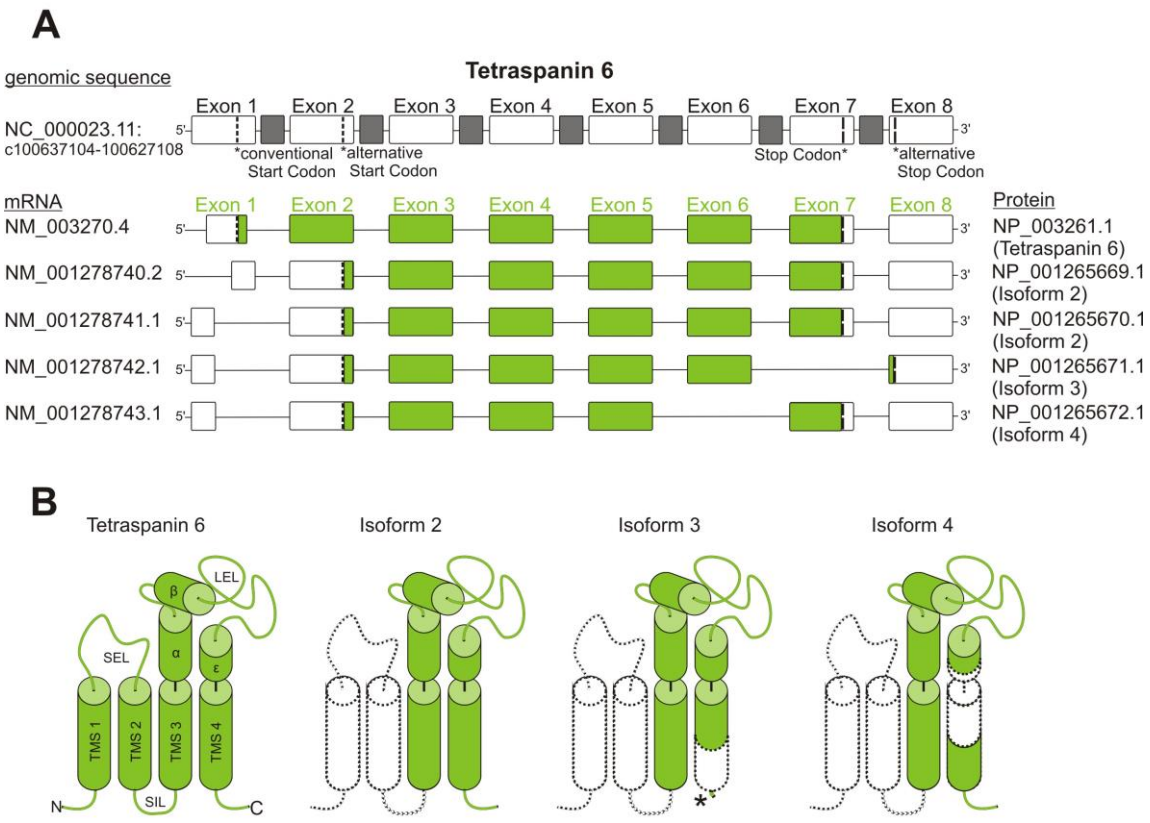


Figure 8: Illustration of different isoforms by the example of Tspan6

(A) Genomic sequence of Tspan6 depicted as a cartoon with exons (white boxes) and introns (grey boxes) as well as the five different mRNAs derived from this genomic sequence. Here, the open reading frame (green) is depicted in reference to the mRNA variant 1 found on the NCBI database. The NCBI database reference numbers are shown for the mRNAs and protein sequences as well. (B) Comparison of the conventional Tspan 6 sequence with the three isoforms. The sequence parts missing in the isoforms (dashed lines and white filling) as well as alternative sequence parts (asterisk) are shown. For a more detailed description of this figure, see the publication: Hochheimer et al. (2019)[129] in Appendix B.

A more detailed analysis of the secondary structure elements and transmembrane segments was done via the structural prediction tools Jpred4[130] and TMHMM 2.0[131,132]. This approach revealed the structural mono-, di-, and trispanins, which were then used to classify the different tetraspanin isoforms. For most tetraspanins, alternative splicing generates several mRNAs per gene with a maximum of five different isoforms in Tspan17. There were almost the same number of conventional and non-conventional tetraspanins found in this analysis. Nevertheless, this might be an underestimation, because only validated/reviewed sequences were included in this study and the NCBI database lacks mRNAs, as the example of the CD53 isoform 4 showed.

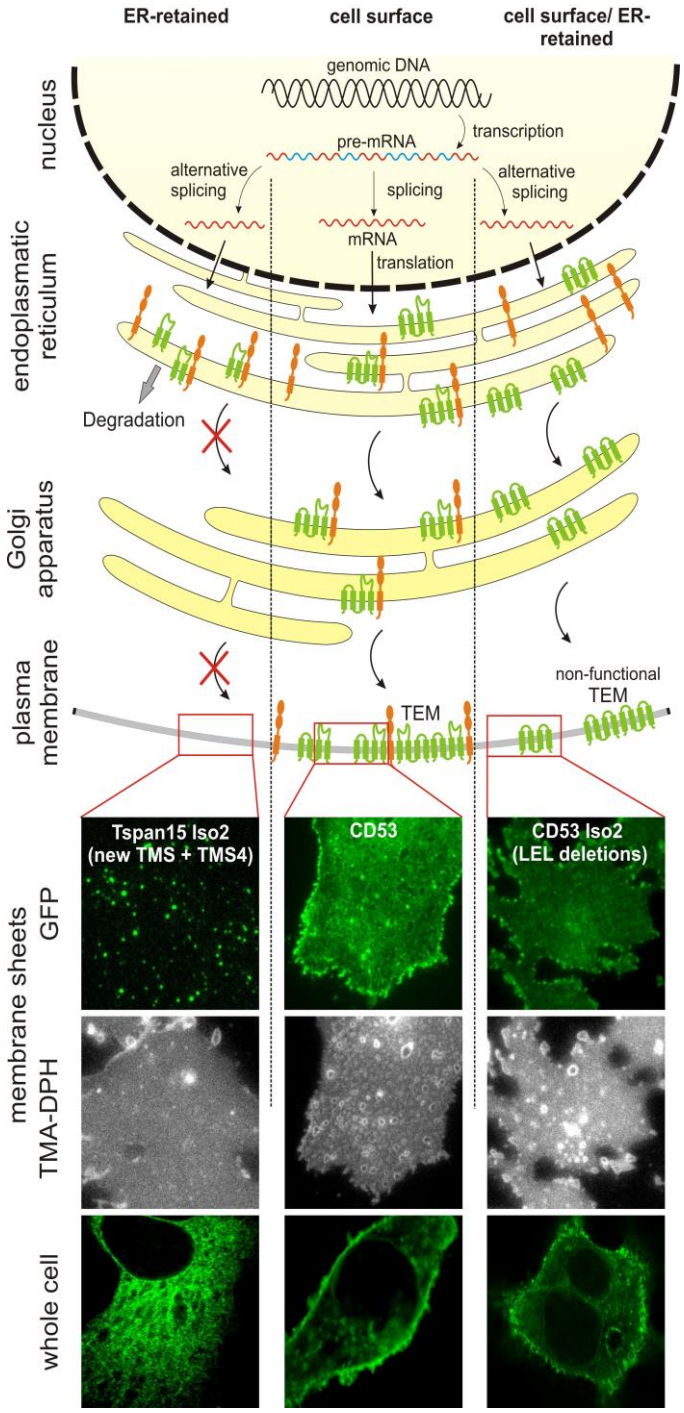


Figure 9: Intracellular trafficking and hypothetical functions of non-conventional tetraspanins.

Illustration of the transcription and splicing process leading to different mRNAs (top). In the following, these different mRNAs are translated and inserted into the ER membrane. Next, the tetraspanin isoforms (green) travel through the ER and Golgi apparatus towards the cell membrane. There are three possible pathways for the isoforms. The conventional tetraspanin (middle lane) interacts with a binding partner (orange) and is co-transported to the plasma membrane, where it forms TEMs. Most isoforms lack TMSs (left lane, dispanin is shown as representative of the largest group of isoforms) and cannot exit the ER. They might still be able to interact with their partner proteins and retain them in the ER, which might lead to their degradation. The group of LEL deletion isoforms (right lane) is unable to interact with its binding partner, but can still exit the ER. This might lead to non-functional TEMs in the plasma membrane, because their co-factors are missing.
For each of these three examples, there is an exemplary confocal micrograph showing HepG2 cells expressing GFP tagged isoforms (bottom, from left to right: Tspan15 Iso2, CD53 or CD53 Iso2). The TMS deletion isoform (Tspan15 Iso2) cannot exit the ER, in contrast to the LEL deletion isoform (CD53 Iso2). For a more detailed description of this figure, see the publication: Hochheimer et al. (2019)[129] in Appendix B.

The implied changes in protein structure and the consequences of these changes on protein function were subject to this study. The described mRNAs encode for isoforms with one, two, three or four transmembrane segments and represent structurally mono-, di-, tri- and tetraspanins. Furthermore, these isoforms differ in transmembrane topology, LEL composition and N- or C-terminus sequence. The tetraspanin isoforms were divided into four classes, based on their number of TMSs. After classifying the differently spliced tetraspanins, the next question arose whether these isoforms are

likely to be expressed. There were no retained introns or premature termination codons found in alternatively spliced mRNAs, which would have made them a target for rapid degradation via nonsense-mediated decay (NMD). Next, we considered sequence alterations in the 5' and 3' untranslated region as well as the ORF and categorized the mRNAs in three groups of "very likely expressed", "likely expressed" and "not likely expressed". This left 15 mRNAs that code for non-conventional tetraspanins from 10 different tetraspanin genes as "very likely expressed".

Based on these findings, we speculated on the effects of non-conventional tetraspanins on cellular processes (Fig. 9). There are two scenarios envisioned for the effect of non-conventional tetraspanins. The first one describes isoforms with TMS deletions and intact LEL that are retained within the ER. These non-conventional tetraspanins could still interact with partner proteins, which might lead to retention in the ER and consequent degradation of these complexes. The second scenario describes the isoforms that are solely lacking parts of the LEL. They can still reach the plasma membrane (Fig. 9, bottom), but are unlikely to aid the co-transport of their interaction partners. This impaired function potentially leads to a deficit in interaction partners within the tetraspanin-enriched microdomains (TEMs) and consequently to malfunctioning TEMs.

In summary, alternative splicing in tetraspanins leads to large structural variability of the resulting isoforms. Based on the structure of non-conventional tetraspanins, we speculate on their role in regulating ER exit, co-transport and their impact on TEM composition.

4.2 Publication: Classes of non-conventional tetraspanins defined by alternative splicing

Classes of non-conventional tetraspanins defined by alternative splicing

Nikolas Hochheimer[1], Ricarda Sies[1], Anna C Aschenbrenner[2,3], Dirk Schneider[4], Thorsten Lang[5]

[1] Department of Membrane Biochemistry, Life & Medical Sciences Institute (LIMES), University of Bonn, Carl-Troll-Straße 31, 53115, Bonn, Germany.
[2] Department of Genomics and Immunoregulation, Life & Medical Sciences Institute (LIMES), University of Bonn, Carl-Troll-Straße 31, 53115, Bonn, Germany.
[3] Department of Internal Medicine and Radboud Center for Infectious Diseases (RCI), Radboud University Nijmegen Medical Center, Geert Grooteplein Zuid 8, Nijmegen, 6525 GA, The Netherlands.
[4] Institute of Pharmacy and Biochemistry, Johannes Gutenberg University Mainz, Johann-Joachim-Becher-Weg 30, 55128, Mainz, Germany.
[5] Department of Membrane Biochemistry, Life & Medical Sciences Institute (LIMES), University of Bonn, Carl-Troll-Straße 31, 53115, Bonn, Germany.
thorsten.lang@uni-bonn.de.

Sci. Rep. **9**, 14075 (2019). DOI: https://doi.org/10.1038/s41598-019-50267-0

My own contributions: conceptualisation, data acquisition, formal analysis, validation, figure preparation, and writing the original draft together with Thorsten Lang and Dirk Schneider.

The publication is listed in Appendix B.

Chapter 5: The small intracellular loop of tetraspanins; more than meets the eye

The seemingly unstructured regions in a protein that connect regular structure elements, such as alpha-helices and beta-sheets are called loops. These loops form a large part of the overall protein structure (up to 50%)[133] and they are crucial for many protein functions such as protein-protein interaction, dimerisation, ligand binding and many more[12]. The central role of this protein structure makes it a promising subject for further analysis. Although short loops (also called turns) in between secondary structure elements have been studied and characterised before (s. Chapter 1.3), there might still be structures that are unaccounted for.

Historically, the most recognised domain of tetraspanins was the large extracellular domain, because of its easy accessibility for antibodies and its role in protein interaction. Next, the complete structure of tetraspanins shifted into focus due to the publications of Seigneuret[26] and Zimmerman[27]. The intracellular domains and especially the small intracellular loop (SIL) have been less represented in tetraspanin research. Regarding the SIL, it is known that this domain is conserved[21] and nevertheless there is still only one publication[22] that focuses on this particular tetraspanin domain.

5.1 Abstract

The first step in studying the small intercellular loop (SIL) of tetraspanins was the sequence analysis of this domain in different species; therefore, the SIL was defined as sequence between TMS2 and TMS3. The SILs of four animals (human, mouse, zebrafish and drosophila) and one plant species (arabidopsis) were subjected to this approach and showed a conserved five amino acid sequence. This SIL core sequence showed conservation between the animal tetraspanins but not the plant one. The conserved sequence for humans was identified as [R/K] E [N/S] [R/K/Q] C with similar sequences in the other animal species. The chemical signature of this stretch can be simplified to positive charge – negative charge – polar – positive charge – polar. The secondary structure prediction of the analysed tetraspanins showed a helical continuation of the TMS helices into the cytosol with their turning point at the central SIL position 3.

The three crystal structures of full tetraspanins all show the funnel-shaped conformation of these proteins. Unfortunately, the structure of CD81 does not resolve the exact positions of the position 2 and 3 of the SIL core. However, within the structures of CD9 and CD53, the whole SIL core is visible and it lies parallel to the membrane surface (Fig 10A). Although the SIL core sequences of these proteins are different (QESQC in CD9 versus KENKC in CD53), the core sequence in both structures showed an M-motif with the position 2 (glutamate in CD9 & CD53) and 4 (glutamine in CD9, lysine in CD53) pointing towards TMS1 and TMS4, respectively (Fig. 10B). The position 1 (glutamine in CD9, lysine in CD53) and 5 (conserved cysteine) are directed away from the tetraspanin center and position 3 (serine in CD9, asparagine in CD53) interacts with the helix backbone C-terminal of the SIL core. In addition, a salt-bridge between the glutamate in position 2 of the SIL and a lysine N–terminally of the TMS1 can be found in CD9. For CD53 there is an interaction of the lysine at position 4 with an asparagine at the C-terminus of TMS4.

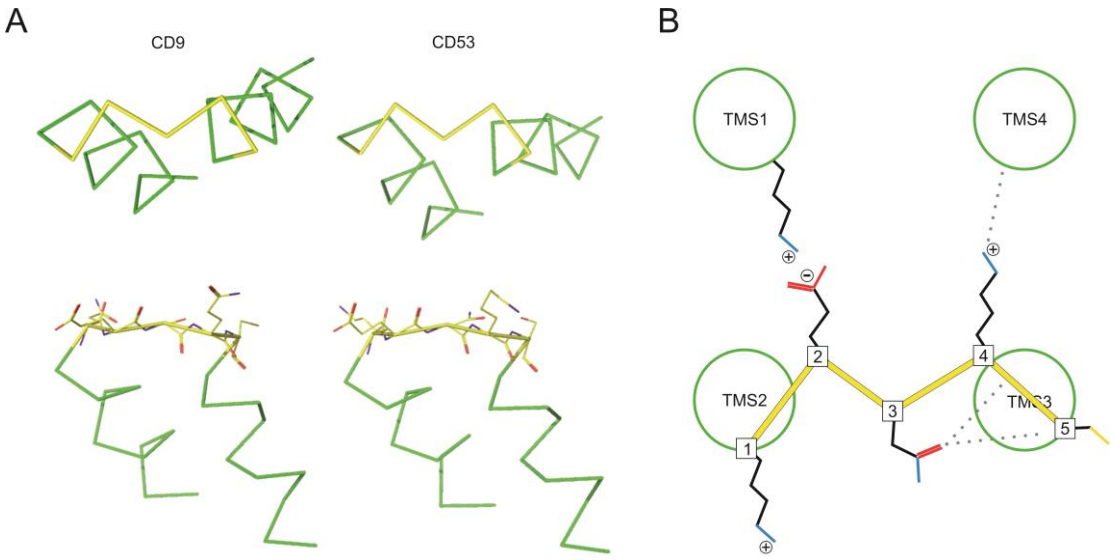


Figure 10: The SIL core resembles an M-motif

(A) Illustration of the SIL core (yellow) with the flanking helical structures (green) of CD9 (PDB 6K4J; left) and CD53 (PDB 6WVG; right) forming an M-motif. The view from the intracellular side (upper panel) as well as the side view (lower panel) with illustrated amino acid side chains of the SIL core is shown. The side view shows the SIL core residues roughly laying in the M-plane. (B) Exemplary interactions between the SIL core sequence and the other tetraspanin domains are shown. An idealized tetraspanin SIL core sequence is depicted. The colours red, blue and yellow of the side chain indicate oxygen, nitrogen and sulphur, respectively. The charge of the amino acids (⊕/⊖) at physiological pH is shown as well. For a more detailed description of this figure, see the publication: Reppert & Lang (2022) in Appendix C.

The question arose, whether this M-motif is a conserved structure element. It does not fit any previously loop or turn classification (s. Chapter 1.3), because the motif's

amino acids are all included in the hydrogen bonding with the flanking helices backbones. There are several structures found in the PDB database for other proteins that incorporate this structural feature, but they do not have a conserved amino acid sequence except for position 3 that was found to be an asparagine in most cases. Consequently, the M-motif resembles a new structural element that is not based on a conserved amino acid sequence.

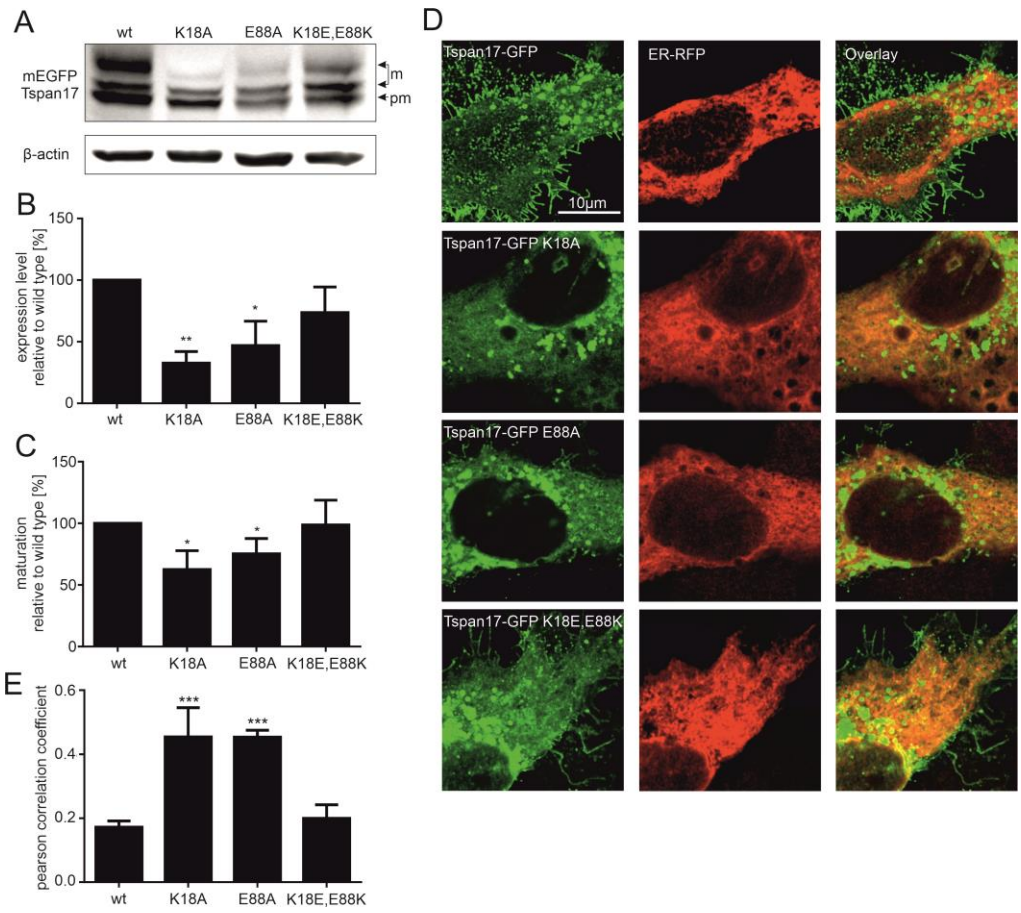


Figure 11: Effect of disrupting the glutamate-lysine interaction in Tspan17

(A) The mEGFP-Tspan17 single mutations (K18A and E88A) and the double mutation (K18E/E88K) were expressed in HepG2 cells and subjected to Western blot analysis. The bands referring to N-glycosylated protein were called mature (m) and none-glycosylated protein was called pre-mature (pm). (B) The expression level of the different constructs was normalised to actin. The single mutants showed a decreased expression level whereas the double mutation showed no difference in comparison to the wild type (set to 100%). (C) The maturation was calculated as ratio between mature and total protein and showed the same pattern as the expression level analysis. (D, E) The mEGFP-Tspan17 constructs were co-localized with an RFP-fused ER marker, which showed a higher overlap for the single mutations compared to the wild type and double mutation. For a more detailed description of this figure, see the publication: Reppert & Lang (2022) in Appendix C.

The amino acid sequence is no prerequisite for the M-motif structure, which makes the conserved sequence in the tetraspanins' M-motif even more interesting. The mutation of single SIL core positions in Tspan17, CD9 and CD53 revealed the most conclusive effect in protein expression levels for the glutamate at position 2. Because

this residue forms a salt-bridge with a lysine at the N-terminal side of the TMS1, the effect of single mutations of both salt-bridge partners to alanine and a double mutation with switched glutamate and lysine was analysed next. For Tspan17, the single mutations resulted in a decreased expression and maturation (Fig. 11A, B, C) and increased ER retention (Fig. 11D,E), whereas the double mutation with switched amino acids showed no significant differences compared to the wild type.

The same analysis was done with CD9, which showed no differences in expression level, maturation or ER retention upon mutation of the SIL core position 2, the N-terminal lysine or the double mutation with switched residues. However, the CD9 mutations were also tested for their interaction capability with EWI-2, which is a known CD9 interaction partner. The single mutations showed a higher affinity to EWI-2 and the double mutation showed no difference compared to the wild type. Regarding CD53, the single mutations had no effect and it was not included in this analysis.

These findings reveal a functional electrostatic interaction in Tspan17 and potentially in CD9, which is formed between the glutamate of the SIL and a conserved lysine N-terminally of the TMS1. Disruption of this interaction led to a loss of expression, maturation and ER exit in Tspan17 and a gain of affinity in CD9. To explain this divergence, we hypothesise that this particular salt-bridge is involved in the stabilisation of the funnel shape conformation of tetraspanins (see Fig. 5). The disruption of this interaction might destabilise the funnel-shaped conformation and force the conformation with a tight helix bundle and an upright LEL. In Tspan17, this might lead to a malfunction in glycosylation causing the above mentioned loss in expression, maturation and ER exit. For CD9, this hypothesis is in good agreement with the cyro-EM structures of tetraspanin complexes (e.g. CD9/EWI-2[36], CD9/EWI-F[37] or CD81/CD19[14]), which show the tetraspanins with the LEL in an upright position and a more compact helix bundle. A mutation that favors the helix bundle conformation might lead to an increased affinity between the tetraspanin and its interaction partner.

5.2 Publication: A conserved sequence in the small intracellular loop of tetraspanins forms an M-shaped inter-helix turn

A conserved sequence in the small intracellular loop of tetraspanins forms an M-shaped inter-helix turn

Nikolas Reppert[1] & Thorsten Lang[1]

[1] Department of Membrane Biochemistry, Life & Medical Sciences Institute (LIMES), University of Bonn, Carl-Troll-Straße 31, 53115, Bonn, Germany.

Sci. Rep. **12**, 4494 (2022). DOI: https://doi.org/10.1038/s41598-022-07243-y

My own contributions: conceptualisation, data acquisition, formal analysis, validation, figure preparation, and writing the original draft together with Thorsten Lang

The publication is listed in Appendix C.

Chapter 6: Summary and Outlook

The work at hand addresses the question of structural variability in membrane protein families. The family of tetraspanins was chosen for this approach, because of the family's large size (33 members in humans), their tight packing of structural domains and their underrepresentation in studies regarding tetraspanins as a protein family. Two different structural aspects were analysed. First, the effect of alternative splicing on the tetraspanin family and the resulting structural implications and secondly, the sequence conservation and structure as well as the interactions and connected functions of the tetraspanins' small intracellular loop.

The alternative splicing analysis for the 33 human tetraspanin genes in Chapter 4 yields a list of 89 gene products in total, including 31 non-conventional isoforms. Based on the amino acid sequences of the isoforms, we predicted their secondary structure elements as well as their transmembrane domains and topology using structural prediction software. These non-conventional isoforms differ in their number of TMSs and resemble structural mono-, di- or trispanins. Furthermore, they are distinguished in their sequence of the LEL, N- and C-terminus. Next, we speculated on the likelihood of expression for each isoform by comparing mRNA features in the 5' and 3' untranslated regions as well as the open reading frame that are known to be important in mRNA degradation and translation. Unfortunately, their actual expression levels remain unclear. The subcellular localisation of exemplary isoforms was tested using transient transfection of GFP tagged isoforms in HepG2 cells. Based on these findings we hypothesised on the function of these non-conventional tetraspanins and found two possible effects. First, isoforms that lack TMSs can be hold up in the ER, which leads to their degradation. Many of these isoforms have an unaltered LEL and might interact with their regular binding partners, which could lead to their ER retention and consequent degradation, too. Second, the isoforms with all four TMSs and a shortened LEL can be localized at the plasma membrane but might be incapable of co-transporting their binding partners. This might lead to a change in TEM composition as well as malfunctioning TEMs.

There is only one example of a tetraspanin splice variant found in literature. This splice variant of CD82[128] (also called Kai1) lacks the majority of the LEL and might be involved in gastric and other cancers. There is no further indicator for tetraspanin isoform expression, which represents a major issue for this analysis and the impact of its results. Therefore, the biggest open question in this chapter is the one of isoform expression. The first step in tackling this question would be a screening for the

isoforms on an mRNA level. After cell lines have been found that express isoform mRNAs, the next step would be their detection on a protein level. This second step would heavily rely on antibodies that recognize the wild type proteins as well as the isoforms. The majority of anti-tetraspanin antibodies target the LEL for its size and accessibility, which possibly constitutes an obstacle for this approach. Many of the isoforms are likely to fold the LEL differently, because they lack one of the LEL-attached TMSs, show an inverted topology or lack the SEL that interacts with the LEL. There are some antibodies against the tetraspanins' intracellular domains, but these domains are also affected by alternative splicing. Consequently, an individual combination of different antibodies targeted against a variety of intra- and extracellular domains would be necessary for this strategy.

The functional analysis of non-conventional tetraspanins could be done via transient expression, potentially linked with the down-regulation of the endogenous background of conventional tetraspanins via gene silencing or knock-out. This approach might make it necessary to down-regulate multiple tetraspanins at the same time, because their functions possibly overlap and mask the effect of a single knock-out as was shown before[134,135]. This strategy would allow testing the two prior discussed functions of non-conventional tetraspanins that are malfunctioning TEMs and altered co-transport. There is a number of tetraspanin/ hetero protein pairs (e.g. CD81/CD19, TspanC8s/ADAM10, CD53/IL-7R alpha) that could be potential test subjects, because they co-locate in TEMs or are co-transported in between cellular organelles.

Beyond being the first systematic analysis of alternative splicing effects in tetraspanins and a membrane protein family in general, this analysis includes the first family-wide analysis of tetraspanin secondary structure. The focus of tetraspanin research rests heavily on family members such as CD81 that have been perceived as the average structure. This has led to an underrepresentation in studies of tetraspanin that stray from this norm. For example, the function of the long C-termini in Tspan12, Tspan17, Tspan22, Tspan23 and Tspan32[129] has only been studied for Tspan22[38], which contains an amphipathic helix involved in membrane interaction. Furthermore, the unusually large N-terminus in Tspan10 (longest N-terminus with 116 aa, second longest in Tspan33 with 24 aa) has not been studied. Additionally, the above average SILs in Tspan22 and Tspan23 harbor an internalisation motif[136], but have not been described further. In general, the function of the tetraspanins' intracellular domains are lesser known.

After the analysis of alternative splicing in tetraspanins, I focused on the smallest and least studied tetraspanin domain: the small intracellular loop. The sequential and structural analysis of the SIL in tetraspanins revealed a conserved SIL core sequence in different animal species, but not in plant tetraspanins. The available crystal structures of tetraspanins show the SIL lying coplanar to the membrane and adopting an M-shape. This M-motif is found in other inter-helix domains as well and only has the asparagine at SIL core position 3 as a conserved residue. The M-motif does not fit any prior turn classification and must be a new kind of inter-helix turn. The amino acids at position 2 and 4 of the SIL core are pointed towards the TMS1 and TMS4, respectively. The position 1 and 5 of the SIL are pointing away from the TMSs, and the position 3 harbors a conserved asparagine or serine, which interacts with the helix backbone of TMS3. In addition, the crystal structure of CD9 shows an interaction of the glutamate at the SIL core position 2 with a conserved lysine at the N-terminus and in CD53 there is an interaction between the lysine at the SIL position 4 and an asparagine at the C-terminus.

In order to analyse the importance of the individual SIL positions and their interactions, a series of point mutations in Tspan17, CD9 and CD53 was analysed. Based on the effects on expression levels of these mutations, we focused on the glutamate/ lysine interaction in Tspan17 and CD9. We used single point mutations of the glutamate at SIL position 2 and the N-terminal lysine in question as well as a double mutations with switched amino acid positions. The single mutations showed effects on protein expression, maturation, ER exit and affinity to binding partners and the double mutations behaved like the wild type proteins in all of these aspects. From these findings, we conclude an electrostatic interaction between the position 2 of the SIL and the N-terminal lysine. We speculated on a role of this interaction in conformation switching in tetraspanins. This hypothesis is based on the position of the interactions at the focal point of the four TMSs in the funnel shaped tetraspanin conformation as well as the gain of function in the co-immunoprecipitation assay of CD9 and EWI-2. Based on these data, a function of this interaction in stabilising the conformation of tetraspanins can be envisioned.

To test this hypothesis, there are several approaches. For example, there are antibodies directed against the tetraspanins' LEL that are supposed to detect certain LEL conformations[137] and would allow to test for an conformational effect of the salt-bridge mutations. By using different antibodies directed against certain parts of the LEL, it would be possible to map different orientations of the LEL, which would allow to conclude on the tetraspanin conformation in general. The analysis of the mutations'

lateral mobility in living cells via FRAP would test the mutations in a more physiological system. Apart from these more indirect analyses of tetraspanin conformation, there are direct methods like molecular docking, which could clarify the conformation impact of those mutations. However, the most direct and straightforward approach would be a structural analysis via crystallography.

The interaction of the conserved glutamate in position 2 and the asparagine/serine in position 3 of the SIL core sequence were discussed in Chapter 5. However, the potential interactions of the other conserved residues are still in question. At position 1, there is a conserved positively charged residue that is directed away from the TMS bundle. This residue might be involved in interactions with negatively charged head groups of surrounding lipids such as phosphatidylinositol phosphates (PiP). Position 4 is occupied by a lysine in CD53, which interacts with an asparagine at the C-terminus of the TMS4 and could be an additionally conserved interaction of the SIL. The cysteine in position 5 is a potentially conserved palmitoylation site, which could influence protein interaction as well[39,138]. Further functional analysis of SIL mutations potentially sheds some light on these open questions.

The SIL and its interaction spectrum could be involved in translating intracellular second messenger signals into different tetraspanin conformations, which could lead to altered tetraspanin interactions. For example, the interaction of the SIL position 2 glutamate with positively charged ions (such as Ca^{2+}) could influence the tetraspanin conformation. Another example could be the above-mentioned interaction with differently phosphorylated PiPs. In conclusion, despite the SIL being a small domain there are many possible functions left to analyse.

The work at hand focuses on alternative splicing and small inter-helix loops in membrane protein families as novel aspects in protein structure variability by the example of tetraspanins. This approach led to a first description and classification of non-conventional family members as well as the detection of unusual family members regarding their intracellular domain length. To my knowledge, these findings resemble the first example of a family-wide alternative splicing analysis of a membrane protein family. The second part of this work addresses the sequential and structural analysis of the small intracellular loop of tetraspanins, which yielded the discovery of the SILs' M-motif as well as the electrostatic interaction with the N-terminus. These findings are based on the initial approach of this study, which was the analysis of tetraspanins as a family and not as individual players in certain cellular processes. I hope that my protein family-based approach to structure variability analysis and its fruitful results

will inspire future endeavors that address similar questions in other protein families as well as studies regarding tetraspanins. In my opinion, there are many open questions left in the tetraspanin field besides the already mentioned ones above. Further fields of research might concern the precise nature of the different tetraspanin conformations, the connection between tetraspanin conformation and membrane curvature and the big question of a potential underlying function of tetraspanins that connects the broad range of tetraspanin functions in different cellular contexts. A better understanding of tetraspanin structure, conformation and function could open up new possibilities in terms of therapeutics, diagnostics and biotechnological applications. We might be just at the beginning of realising their potential applications and understanding the functional and structural spectrum that is navigated by the family of tetraspanins.

References

1. Bianconi, E. *et al.* An estimation of the number of cells in the human body. *Ann. Hum. Biol.* **40**, 463–471 (2013).

2. Klerk, E. de & Hoen, P. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. (2015) doi:10.1016/J.TIG.2015.01.001.

3. Singer, S. J. & Nicolson, G. L. The fluid mosaic model of the structure of cell membranes. *Science* **175**, 720–731 (1972).

4. Kusumi, A. *et al.* Paradigm Shift of the Plasma Membrane Concept from the Two-Dimensional Continuum Fluid to the Partitioned Fluid: High-Speed Single-Molecule Tracking of Membrane Molecules. *Annu. Rev. Biophys. Biomol. Struct.* **34**, 351–378 (2005).

5. Lang, T. & Rizzoli, S. O. Membrane Protein Clusters at Nanoscale Resolution: More Than Pretty Pictures. *Physiology* **25**, 116–124 (2010).

6. Levental, I., Levental, K. R. & Heberle, F. A. Lipid Rafts: Controversies Resolved, Mysteries Remain. *Trends Cell Biol.* **30**, 341–353 (2020).

7. Belfort, M. & Weiner, A. Another Bridge between Kingdoms: tRNA Splicing in Archaea and Eukaryotes. *Cell* **89**, 1003–1006 (1997).

8. Will, C. L. & Lührmann, R. Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.* **3**, a003707 (2011).

9. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).

10. Lewis, B. P., Green, R. E. & Brenner, S. E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 189–192 (2003).

11. Charrin, S., Jouannet, S., Boucheix, C. & Rubinstein, E. Tetraspanins at a glance. *J. Cell Sci.* **127**, 3641–3648 (2014).

12. Espadaler, J., Querol, E., Aviles, F. X. & Oliva, B. Identification of function-associated loop motifs and application to protein function prediction. *Bioinforma. Oxf. Engl.* **22**, 2237–2243 (2006).

13. Homsi, Y. & Lang, T. The specificity of homomeric clustering of CD81 is mediated by its δ-loop. *FEBS Open Bio* **7**, 274–283 (2017).

14. Susa, K. J., Rawson, S., Kruse, A. C. & Blacklow, S. C. Cryo-EM structure of the B cell co-receptor CD19 bound to the tetraspanin CD81. *Science* **371**, 300–305 (2021).

15. Song, Q. *et al.* Predicting Turns in Proteins with a Unified Model. *PLOS ONE* **7**, e48389 (2012).

16. Meissner, M., Koch, O., Klebe, G. & Schneider, G. Prediction of turn types in protein structure by machine-learning classifiers. *Proteins Struct. Funct. Bioinforma.* **74**, 344–352 (2009).

17. Chou, K.-C. Prediction of Tight Turns and Their Types in Proteins. *Anal. Biochem.* **286**, 1–16 (2000).

18. Zheng, C. & Kurgan, L. Prediction of beta-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments. *BMC Bioinformatics* **9**, 430 (2008).

19. Kountouris, P. & Hirst, J. D. Predicting β-turns and their types using predicted backbone dihedral angles and secondary structures. *BMC Bioinformatics* **11**, 1–11 (2010).

20. Duddy, W. J., Nissink, J. W. M., Allen, F. H. & Milner-White, E. J. Mimicry by asx- and ST-turns of the four main types of β-turn in proteins. *Protein Sci. Publ. Protein Soc.* **13**, 3051–3055 (2004).

21. Stipp, C. S., Kolesnikova, T. V. & Hemler, M. E. Functional domains in tetraspanin proteins. *Trends Biochem. Sci.* **28**, 106–112 (2003).

22. Mazurov, D., Heidecker, G. & Derse, D. The Inner Loop of Tetraspanins CD82 and CD81 Mediates Interactions with Human T Cell Lymphotrophic Virus Type 1 Gag Protein*. *J. Biol. Chem.* **282**, 3896–3903 (2007).

23. Oren, R., Takahashi, S., Doss, C., Levy, R. & Levy, S. TAPA-1, the target of an antiproliferative antibody, defines a new family of transmembrane proteins. *Mol. Cell. Biol.* **10**, 4007–4015 (1990).

24. Lazo, P. A. Functional implications of tetraspanin proteins in cancer biology. *Cancer Sci.* **98**, 1666–1677 (2007).

25. Kitadokoro, K. *et al.* CD81 extracellular domain 3D structure: insight into the tetraspanin superfamily structural motifs. *EMBO J.* **20**, 12–18 (2001).

26. Seigneuret, M. Complete Predicted Three-Dimensional Structure of the Facilitator Transmembrane Protein and Hepatitis C Virus Receptor CD81: Conserved and Variable Structural Domains in the Tetraspanin Superfamily. *Biophys. J.* **90**, 212–227 (2006).

27. Zimmerman, B. *et al.* Crystal Structure of a Full-Length Human Tetraspanin Reveals a Cholesterol-Binding Pocket. *Cell* **167**, 1041-1051.e11 (2016).

28. Yang, Y. *et al.* Open conformation of tetraspanins shapes interaction partner networks on cell membranes. *EMBO J.* **39**, e105246 (2020).

29. Palor, M., Stejskal, L., Mandal, P., Shepherd, A. J. & Grove, J. Conformational switching of CD81 controls its function as a receptor for hepatitis C virus. *bioRxiv* 542837 (2019) doi:10.1101/542837.

30. Lang, T. & Hochheimer, N. Tetraspanins. *Curr. Biol. CB* **30**, R204–R206 (2020).

31. Hochheimer, N., Sies, R., Aschenbrenner, A. C., Schneider, D. & Lang, T. Classes of non-conventional tetraspanins defined by alternative splicing. *Sci. Rep.* **9**, 14075 (2019).

32. Stipp, C. S., Kolesnikova, T. V. & Hemler, M. E. Functional domains in tetraspanin proteins. 7 (2003).

33. Scholz, C.-J., Sauer, G. & Deissler, H. Glycosylation of tetraspanin Tspan-1 at four distinct sites promotes its transition through the endoplasmic reticulum. *Protein Pept. Lett.* **16**, 1244–1248 (2009).

34. Kovalenko, O. V., Metcalf, D. G., DeGrado, W. F. & Hemler, M. E. Structural organization and interactions of transmembrane domains in tetraspanin proteins. *BMC Struct. Biol.* **5**, 11 (2005).

35. Seigneuret, M. Complete Predicted Three-Dimensional Structure of the Facilitator Transmembrane Protein and Hepatitis C Virus Receptor CD81: Conserved and Variable Structural Domains in the Tetraspanin Superfamily. *Biophys. J.* **90**, 212–227 (2006).

36. Umeda, R. *et al.* Structural insights into tetraspanin CD9 function. *Nat. Commun.* **11**, 1606 (2020).

37. Oosterheert, W. *et al.* Implications for tetraspanin-enriched microdomain assembly based on structures of CD9 with EWI-F. *Life Sci. Alliance* **3**, (2020).

38. Khattree, N., Ritter, L. M. & Goldberg, A. F. X. Membrane curvature generation by a C-terminal amphipathic helix in peripherin-2/rds, a tetraspanin required for photoreceptor sensory cilium morphogenesis. *J. Cell Sci.* **126**, 4659–4670 (2013).

39. Yang, X. *et al.* Palmitoylation of Tetraspanin Proteins: Modulation of CD151 Lateral Interactions, Subcellular Distribution, and Integrin-dependent Cell Morphology. *Mol. Biol. Cell* **13**, 767–781 (2002).

40. Charrin, S. *et al.* Differential stability of tetraspanin/tetraspanin interactions: role of palmitoylation. *FEBS Lett.* **516**, 139–144 (2002).

41. Yang, X. *et al.* Palmitoylation supports assembly and function of integrin–tetraspanin complexes. *J. Cell Biol.* **167**, 1231–1240 (2004).

42. Matsumoto, A. K. *et al.* Functional dissection of the CD21/CD19/TAPA-1/Leu-13 complex of B lymphocytes. *J. Exp. Med.* **178**, 1407–1417 (1993).

43. Schick, M. R. & Levy, S. The TAPA-1 molecule is associated on the surface of B cells with HLA-DR molecules. *J. Immunol. Baltim. Md 1950* **151**, 4090–4097 (1993).

44. Kummer, D., Steinbacher, T., Schwietzer, M. F., Thölmann, S. & Ebnet, K. Tetraspanins: integrating cell surface receptors to functional microdomains in homeostasis and disease. *Med. Microbiol. Immunol. (Berl.)* **209**, 397–405 (2020).

45. Levy, S. & Shoham, T. The tetraspanin web modulates immune-signalling complexes. *Nat. Rev. Immunol.* **5**, 136–148 (2005).

46. Berditchevski, F. & Odintsova, E. Characterization of Integrin–Tetraspanin Adhesion Complexes: Role of Tetraspanins in Integrin Signaling. *J. Cell Biol.* **146**, 477–492 (1999).

47. Charrin, S. *et al.* Multiple levels of interactions within the tetraspanin web. *Biochem. Biophys. Res. Commun.* **304**, 107–112 (2003).

48. Levy, S. & Shoham, T. Protein-Protein Interactions in the Tetraspanin Web. *Physiology* **20**, 218–224 (2005).

49. Shoham, T., Rajapaksa, R., Kuo, C.-C., Haimovich, J. & Levy, S. Building of the Tetraspanin Web: Distinct Structural Domains of CD81 Function in Different Cellular Compartments. *Mol. Cell. Biol.* **26**, 1373–1385 (2006).

50. Wirth, D. *et al.* Quantitative characterization of Tetraspanin 8 homointeractions in the plasma membrane. *Biochem. J.* (2021) doi:10.1042/BCJ20210459.

51. Berditchevski, F., Odintsova, E., Sawada, S. & Gilbert, E. Expression of the Palmitoylation-deficient CD151 Weakens the Association of α3β1 Integrin with the Tetraspanin-enriched Microdomains and Affects Integrin-dependent Signaling *. *J. Biol. Chem.* **277**, 36991–37000 (2002).

52. Zuidscherwoude, M. *et al.* The tetraspanin web revisited by super-resolution microscopy. *Sci. Rep.* **5**, 12201 (2015).

53. Caparotta, M. & Masone, D. Cholesterol plays a decisive role in tetraspanin assemblies during bilayer deformations. *Biosystems* **209**, 104505 (2021).

54. Matthews, A. L., Szyroka, J., Collier, R., Noy, P. J. & Tomlinson, M. G. Scissor sisters: regulation of ADAM10 by the TspanC8 tetraspanins. *Biochem. Soc. Trans.* **45**, 719–730 (2017).

55. Healy, E. F. How tetraspanin-mediated cell entry of SARS-CoV-2 can dysregulate the shedding of the ACE2 receptor by ADAM17. *Biochem. Biophys. Res. Commun.* **593**, 52–56 (2022).

56. Rajesh, S. *et al.* Structural Basis of Ligand Interactions of the Large Extracellular Domain of Tetraspanin CD81. *J. Virol.* **86**, 9606–9616 (2012).

57. Yu, J. *et al.* The CD9, CD81, and CD151 EC2 domains bind to the classical RGD-binding site of integrin αvβ3. *Biochem. J.* **474**, 589–596 (2017).

58. Seigneuret, M., Delaguillaumie, A., Lagaudrière-Gesbert, C. & Conjeaud, H. Structure of the tetraspanin main extracellular domain. A partially conserved fold with a structurally variable domain insertion. *J. Biol. Chem.* **276**, 40055–40064 (2001).

59. Homsi, Y. *et al.* The extracellular δ-domain is essential for the formation of CD81 tetraspanin webs. *Biophys. J.* **107**, 100–113 (2014).

60. Deventer, S. van, Arp, A. B. & Spriel, A. B. van. Dynamic Plasma Membrane Organization: A Complex Symphony. *Trends Cell Biol.* **31**, 119–129 (2021).

61. Toyo-oka, K. *et al.* Association of a tetraspanin CD9 with CD5 on the T cell surface: role of particular transmembrane domains in the association. *Int. Immunol.* **11**, 2043–2052 (1999).

62. Palor, M. *et al.* Cholesterol sensing by CD81 is important for hepatitis C virus entry. *J. Biol. Chem.* **295**, 16931–16948 (2020).

63. Hosokawa, K., Ishimaru, H., Watanabe, T. & Fujimuro, M. The Lysosome Pathway Degrades CD81 on the Cell Surface by Poly-ubiquitination and Clathrin-Mediated Endocytosis. *Biol. Pharm. Bull.* **43**, 540–545 (2020).

64. Lineberry, N., Su, L., Soares, L. & Fathman, C. G. The Single Subunit Transmembrane E3 Ligase Gene Related to Anergy in Lymphocytes (GRAIL) Captures and Then Ubiquitinates Transmembrane Proteins across the Cell Membrane*. *J. Biol. Chem.* **283**, 28497–28505 (2008).

65. Berditchevski, F. & Odintsova, E. Tetraspanins as Regulators of Protein Trafficking. *Traffic* **8**, 89–96 (2007).

66. Wang, H.-X., Kolesnikova, T. V., Denison, C., Gygi, S. P. & Hemler, M. E. The C-terminal tail of tetraspanin protein CD9 contributes to its function and molecular organization. *J. Cell Sci.* **124**, 2702–2710 (2011).

67. Israels, S. J. & McMillan-Ward, E. M. Palmitoylation supports the association of tetraspanin CD63 with CD9 and integrin alphaIIbbeta3 in activated platelets. *Thromb. Res.* **125**, 152–158 (2010).

68. Charrin, S. *et al.* A physical and functional link between cholesterol and tetraspanins. *Eur. J. Immunol.* **33**, 2479–2489 (2003).

69. Silvie, O. *et al.* Cholesterol contributes to the organization of tetraspanin-enriched microdomains and to CD81-dependent infection by malaria sporozoites. *J. Cell Sci.* **119**, 1992–2002 (2006).

70. Fernandez, L. *et al.* CD82 and Gangliosides Tune CD81 Membrane Behavior. *Int. J. Mol. Sci.* **22**, 8459 (2021).

71. Marsay, K., Roehl, H., Monk, P., Partridge, L. & Carney, T. Tetraspanins in zebrafish development. 2.

72. Fradkin, L. G., Kamphorst, J. T., DiAntonio, A., Goodman, C. S. & Noordermeer, J. N. Genomewide analysis of the Drosophila tetraspanins reveals a subset with similar function in the formation of the embryonic synapse. *Proc. Natl. Acad. Sci.* **99**, 13663–13668 (2002).

73. Boavida, L. C., Qin, P., Broz, M., Becker, J. D. & McCormick, S. Arabidopsis Tetraspanins Are Confined to Discrete Expression Domains and Cell Types in Reproductive Tissues and Form Homo- and Heterodimers When Expressed in Yeast. *Plant Physiol.* **163**, 696–712 (2013).

74. Huang, S. *et al.* The phylogenetic analysis of tetraspanins projects the evolution of cell–cell interactions from unicellular to multicellular organisms. *Genomics* **86**, 674–684 (2005).

75. Zhang, F., Kotha, J., Jennings, L. K. & Zhang, X. A. Tetraspanins and vascular functions. *Cardiovasc. Res.* **83**, 7–15 (2009).

76. Balise, V. D., Saito-Reis, C. A. & Gillette, J. M. Tetraspanin Scaffold Proteins Function as Key Regulators of Hematopoietic Stem Cells. *Front. Cell Dev. Biol.* **8**, 598 (2020).

77. Mahmudi-Azer, S., Downey, G. P. & Moqbel, R. Translocation of the tetraspanin CD63 in association with human eosinophil mediator release. *Blood* **99**, 4039–4047 (2002).

78. Kobayashi, T. *et al.* The Tetraspanin CD63/lamp3 Cycles between Endocytic and Secretory Compartments in Human Endothelial Cells. *Mol. Biol. Cell* **11**, 1829–1843 (2000).

79. Rous, B. A. *et al.* Role of adaptor complex AP-3 in targeting wild-type and mutated CD63 to lysosomes. *Mol. Biol. Cell* **13**, 1071–1082 (2002).

80. Duffield, A., Kamsteeg, E.-J., Brown, A. N., Pagel, P. & Caplan, M. J. The tetraspanin CD63 enhances the internalization of the H,K-ATPase beta-subunit. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 15560–15565 (2003).

81. Liu, L. *et al.* Tetraspanin CD151 Promotes Cell Migration by Regulating Integrin Trafficking *. *J. Biol. Chem.* **282**, 31631–31642 (2007).

82. Bebelman, M. P., Smit, M. J., Pegtel, D. M. & Baglio, S. R. Biogenesis and function of extracellular vesicles in cancer. *Pharmacol. Ther.* **188**, 1–11 (2018).

83. Andreu, Z. & Yáñez-Mó, M. Tetraspanins in Extracellular Vesicle Formation and Function. *Front. Immunol.* **5**, (2014).

84. Doyle, L. M. & Wang, M. Z. Overview of Extracellular Vesicles, Their Origin, Composition, Purpose, and Methods for Exosome Isolation and Analysis. *Cells* **8**, 727 (2019).

85. Borges, F. T., Reis, L. A. & Schor, N. Extracellular vesicles: structure, function, and potential clinical uses in renal diseases. *Braz. J. Med. Biol. Res. Rev. Bras. Pesqui. Medicas E Biol.* **46**, 824–830 (2013).

86. Möbius, W. *et al.* Immunoelectron Microscopic Localization of Cholesterol Using Biotinylated and Non-cytolytic Perfringolysin O. *J. Histochem. Cytochem.* **50**, 43–55 (2002).

87. Babst, M., Katzmann, D. J., Estepa-Sabal, E. J., Meerloo, T. & Emr, S. D. Escrt-III: an endosome-associated heterooligomeric protein complex required for mvb sorting. *Dev. Cell* **3**, 271–282 (2002).

88. Wei, H. *et al.* Regulation of exosome production and cargo sorting. *Int. J. Biol. Sci.* **17**, 163–177 (2021).

89. Rana, S. & Zöller, M. Exosome target cell selection and the importance of exosomal tetraspanins: a hypothesis. *Biochem. Soc. Trans.* **39**, 559–562 (2011).

90. Perez-Hernandez, D. *et al.* The Intracellular Interactome of Tetraspanin-enriched Microdomains Reveals Their Function as Sorting Machineries toward Exosomes *. *J. Biol. Chem.* **288**, 11649–11661 (2013).

91. Barreiro, O. *et al.* Endothelial tetraspanin microdomains regulate leukocyte firm adhesion during extravasation. *Blood* **105**, 2852–2861 (2005).

92. McNamara, R. P. *et al.* Imaging of surface microdomains on individual extracellular vesicles in 3-D. *J. Extracell. Vesicles* **11**, e12191 (2022).

93. Hemler, M. E. Tetraspanin proteins promote multiple cancer stages. *Nat. Rev. Cancer* **14**, 49–60 (2014).

94. Vences-Catalán, F. & Levy, S. Immune Targeting of Tetraspanins Involved in Cell Invasion and Metastasis. *Front. Immunol.* **9**, 1277 (2018).

95. Sadej, R., Grudowska, A., Turczyk, L., Kordek, R. & Romanska, H. M. CD151 in cancer progression and metastasis: a complex scenario. *Lab. Invest.* **94**, 41–51 (2014).

96. Tominaga, N. *et al.* RPN2-mediated glycosylation of tetraspanin CD63 regulates breast cancer cell malignancy. *Mol. Cancer* **13**, 134 (2014).

97. Seipold, L. & Saftig, P. The Emerging Role of Tetraspanins in the Proteolytic Processing of the Amyloid Precursor Protein. *Front. Mol. Neurosci.* **9**, 149 (2016).

98. Guix, F. X. *et al.* Tetraspanin 6: a pivotal protein of the multiple vesicular body determining exosome release and lysosomal degradation of amyloid precursor protein fragments. *Mol. Neurodegener.* **12**, 25 (2017).

99. Miyoshi, E. *et al.* Exosomal tau with seeding activity is released from Alzheimer's disease synapses, and seeding potential is associated with amyloid beta. *Lab. Invest.* 1–13 (2021) doi:10.1038/s41374-021-00644-z.

100. Walther, D. *et al.* Tetraspanin 7 autoantibodies in type 1 diabetes. *Diabetologia* **59**, 1973–1976 (2016).

101. Takeda, Y. *et al.* Double Deficiency of Tetraspanins CD9 and CD81 Alters Cell Motility and Protease Production of Macrophages and Causes Chronic Obstructive Pulmonary Disease-like Phenotype in Mice*. *J. Biol. Chem.* **283**, 26089–26097 (2008).

102. Floren, M. *et al.* Tetraspanin CD82 drives acute myeloid leukemia chemoresistance by modulating protein kinase C alpha and β1 integrin activation. *Oncogene* **39**, 3910–3925 (2020).

103. Iwasaki, T. *et al.* Deletion of Tetraspanin CD9 Diminishes Lymphangiogenesis in Vivo and in Vitro*. *J. Biol. Chem.* **288**, 2118–2131 (2013).

104. Monk, P. N. & Partridge, L. J. Tetraspanins: gateways for infection. *Infect. Disord. Drug Targets* **12**, 4–17 (2012).

105. York, S. B. *et al.* Zika Virus Hijacks Extracellular Vesicle Tetraspanin Pathways for Cell-to-Cell Transmission. *mSphere* e0019221 (2021) doi:10.1128/mSphere.00192-21.

106. Odegaard, K. E. *et al.* Role of Extracellular Vesicles in Substance Abuse and HIV-Related Neurological Pathologies. *Int. J. Mol. Sci.* **21**, 6765 (2020).

107. Thali, M. The Roles of Tetraspanins in HIV-1 Replication. *Curr. Top. Microbiol. Immunol.* **339**, 85–102 (2009).

108. Finke, J., Hitschler, L., Boller, K., Florin, L. & Lang, T. HPV caught in the tetraspanin web? *Med. Microbiol. Immunol. (Berl.)* **209**, 447–459 (2020).

109. Earnest, J. T. *et al.* The tetraspanin CD9 facilitates MERS-coronavirus entry by scaffolding host cell receptors and proteases. *PLOS Pathog.* **13**, e1006546 (2017).

110. Earnest, J. T., Hantak, M. P., Park, J.-E. & Gallagher, T. Coronavirus and Influenza Virus Proteolytic Priming Takes Place in Tetraspanin-Enriched Membrane Microdomains. *J. Virol.* **89**, 6093–6104.

111. Elgawidi, A. *et al.* A role for tetraspanin proteins in regulating fusion induced by Burkholderia thailandensis. *Med. Microbiol. Immunol. (Berl.)* **209**, 473–487 (2020).

112. Witwer, K. W. *et al.* Standardization of sample collection, isolation and analysis methods in extracellular vesicle research. *J. Extracell. Vesicles* **2**, 20360 (2013).

113. Pol, E. van der, Böing, A. N., Harrison, P., Sturk, A. & Nieuwland, R. Classification, Functions, and Clinical Relevance of Extracellular Vesicles. *Pharmacol. Rev.* **64**, 676–705 (2012).

114. Maia, J. *et al.* Employing Flow Cytometry to Extracellular Vesicles Sample Microvolume Analysis and Quality Control. *Front. Cell Dev. Biol.* **8**, 1165 (2020).

115. Jørgensen, M. *et al.* Extracellular Vesicle (EV) Array: microarray capturing of exosomes and other extracellular vesicles for multiplexed phenotyping. *J. Extracell. Vesicles* **2**, 20920 (2013).

116. Salomon, C. *et al.* A gestational profile of placental exosomes in maternal plasma and their effects on endothelial cell migration. *PloS One* **9**, e98667 (2014).

117. Mizenko, R. R. *et al. Tetraspanin immunocapture phenotypes extracellular vesicles according to biofluid source but may limit identification of multiplexed cancer biomarkers*. 2021.03.02.433595 https://www.biorxiv.org/content/10.1101/2021.03.02.433595v1 (2021) doi:10.1101/2021.03.02.433595.

118. Crenshaw, B. J., Sims, B. & Matthews, Q. L. *Biological Function of Exosomes as Diagnostic Markers and Therapeutic Delivery Vehicles in Carcinogenesis and Infectious Diseases*. *Nanomedicines* (IntechOpen, 2018). doi:10.5772/intechopen.80225.

119. Liang, Y., Duan, L., Lu, J. & Xia, J. Engineering exosomes for targeted drug delivery. *Theranostics* **11**, 3183–3195 (2021).

120. Elsharkasy, O. M. *et al.* Extracellular vesicles as drug delivery systems: Why and how? *Adv. Drug Deliv. Rev.* **159**, 332–343 (2020).

121. Vogt, S., Stadlmayr, G., Grillari, J., Rüker, F. & Wozniak-Knopp, G. *Engineering of Surface Proteins in Extracellular Vesicles for Tissue-Specific Targeting*. *Current Topics in Biochemical Engineering* (IntechOpen, 2019). doi:10.5772/intechopen.83537.

122. Scott, T. A. *et al.* Engineered extracellular vesicles directed to the spike protein inhibit SARS-CoV-2. *Mol. Ther. Methods Clin. Dev.* (2022) doi:10.1016/j.omtm.2022.01.015.

123. Bonnet, M. *et al.* Targeting the Tetraspanins with Monoclonal Antibodies in Oncology: Focus on Tspan8/Co-029. *Cancers* **11**, 179 (2019).

124. Vasse, M. *et al.* [Tetraspanins: a new target for antiangiogenic therapy?]. *Ann. Pharm. Fr.* **73**, 100–107 (2015).

125. Karam, J., Méresse, S., Kremer, L. & Daher, W. The roles of tetraspanins in bacterial infections. *Cell. Microbiol.* **22**, e13260 (2020).

126. Zona, L. *et al.* CD81-Receptor Associations — Impact for Hepatitis C Virus Entry and Antiviral Therapies. *Viruses* **6**, 875–892 (2014).

127. Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5’ terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci.* **74**, 3171–3175 (1977).

128. Lee, J. H., Seo, Y.-W., Park, S. R., Kim, Y. J. & Kim, K. K. Expression of a splice variant of KAI1, a tumor metastasis suppressor gene, influences tumor invasion and progression. *Cancer Res.* **63**, 7247–7255 (2003).

129. Hochheimer, N., Sies, R., Aschenbrenner, A. C., Schneider, D. & Lang, T. Classes of non-conventional tetraspanins defined by alternative splicing. *Sci. Rep.* **9**, 14075 (2019).

130. Drozdetskiy, A., Cole, C., Procter, J. & Barton, G. J. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* **43**, W389–W394 (2015).

131. Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175–182 (1998).

132. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).

133. Regad, L., Martin, J., Nuel, G. & Camproux, A.-C. Mining protein loops using a structural alphabet and statistical exceptionality. *BMC Bioinformatics* **11**, 75 (2010).

134. Yeung, L., Hickey, M. J. & Wright, M. D. The Many and Varied Roles of Tetraspanins in Immune Cell Recruitment and Migration. *Front. Immunol.* **9**, 1644 (2018).

135. Haining, E. J. *et al.* The TspanC8 Subgroup of Tetraspanins Interacts with A Disintegrin and Metalloprotease 10 (ADAM10) and Regulates Its Maturation and Cell Surface Expression*. *J. Biol. Chem.* **287**, 39753–39765 (2012).

136. Reppert, N. & Lang, T. A conserved sequence in the small intracellular loop of tetraspanins forms an M-shaped inter-helix turn. *Sci. Rep.* **12**, 4494 (2022).

137. Gutierrez-Lopez, M. D. *et al.* A functionally relevant conformational epitope on the CD9 tetraspanin depends on the association with activated beta1 integrin. *J. Biol. Chem.* **278**, 208–218 (2003).

138. Levy, S. & Shoham, T. Protein-Protein Interactions in the Tetraspanin Web. *Physiology* **20**, 218–224 (2005).

Appendix A: Quick Guide – Tetraspanins

CellPress

Current Biology
Magazine

Quick guide

Tetraspanins

Thorsten Lang*
and Nikolas Hochheimer

What are tetraspanins? Tetraspanins are small transmembrane proteins. The family name refers to their four membrane-spanning segments, which flank a highly conserved topology of two extracellular loops (see inset in Figure 1). These proteins are known as 'master organizers' of the plasma membrane and are involved in an astonishingly large variety of cellular processes.

What do they look like? Apart from the four transmembrane segments (TMSs) and their small and large extracellular loops (SEL and LEL, respectively), tetraspanins contain mainly very short intracellular domains. So far we have only one complete structural model available for tetraspanins and this is for the human tetraspanin CD81. This protein adopts a compact mushroom shape, with the mushroom head and stalk being composed of the extracellular loops and a four-TMS bundle, respectively. Two TMS helices extend from the bundle, reaching out straight into the mushroom head. Consequently, with the stalk embedded into the lipid bilayer, the head sits upright on the membrane. This original model based on a crystal structure of the LEL alone was recently challenged by a crystal structure of the complete CD81 protein revealing two differences. First, the LEL folds down to the membrane because the above-mentioned extending helices are kinked, and second, a cholesterol molecule is enclosed by the four transmembrane helices. This cholesterol precludes formation of a tight four-helix bundle; however, upon release of cholesterol, the gap closes and the LEL adopts a more upright position similar to the initial model.

The LEL mediates the specificity of interactions between a particular tetraspanin and its so-called primary interaction partner. This loop contains a conserved region within which is embedded a variable region. The conserved region contains three α-helices. The variable region is found between the second and third helices and, depending on the tetraspanin, consists of α-helices, β-sheets and/or unstructured loops. The variable region also contains disulfide bridges formed between four, six or eight cysteine residues. In addition, tetraspanins are subject to post-translational modifications, such as palmitoylation of cysteine residues in the intracellular domains and/or glycosylation of the extracellular loop (see inset of Figure 1).

Why do they have weird names? The naming of some tetraspanins is based on the 'cluster of differentiation' (CD) classification aimed at classifying a specific cell-surface molecule identified by several different

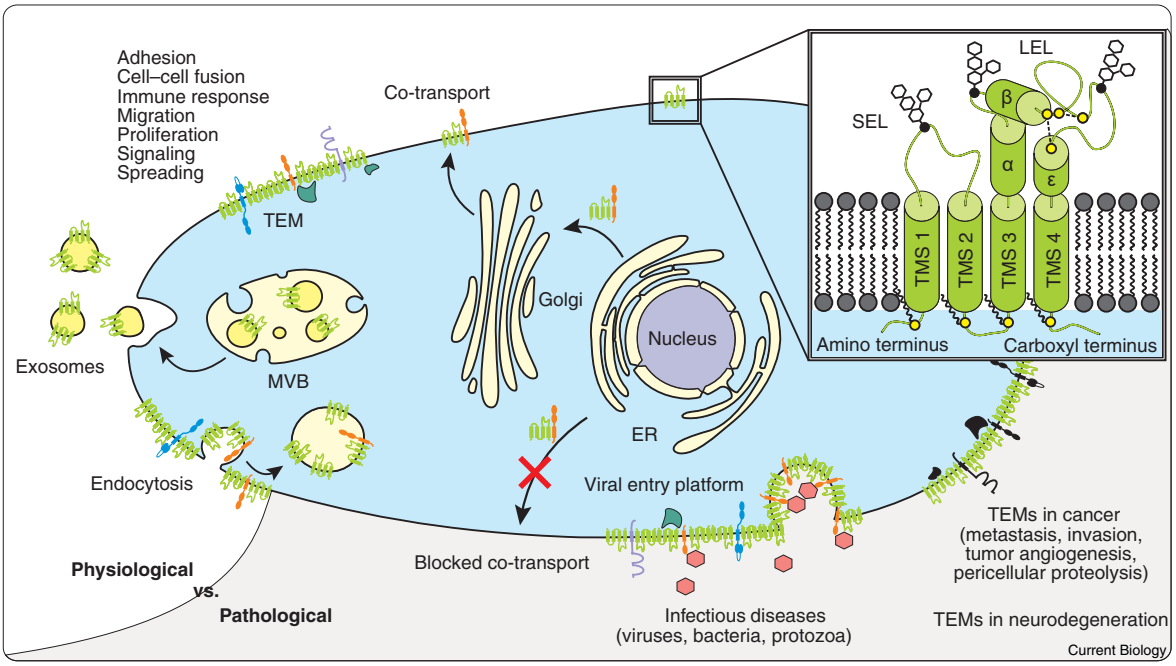


Figure 1. Overview on the involvement of tetraspanins.
The diagram differentiates between physiological roles of tetraspanins and their pathological roles in diseases. MVB, multivesicular body; ER, endoplasmic reticulum; pink hexagons, viral particles. The inset shows the basic tetraspanin structure (adapted from Hochheimer *et al*. (2019); for details see text): the yellow dots indicate cysteine residues that form disulfide bridges (top) or that are palmitoylated (bottom), and the white hexagons depict sugar chains.

Current Biology
Magazine

monoclonal antibodies by using a unique CD number. For example, the target of an antiproliferative antibody (TAPA-1) became CD81. Names of the most intensively studied members originated during the 1980s and are now known as tetraspanins CD9, CD37, CD53, CD63, CD81, CD82 and CD151. Unfortunately, the designation 'CD' and the associated number does not give any information regarding the function of the protein. Other names indicate where the tetraspanins were originally observed, such as the uroplakins, which were first found in the urothelium. More appealing names were given to some *Drosophila* tetraspanins, such as late bloomer and Sunglasses. Apart from the historic names, systematic names exist, e.g. Tspan1–33 for the human tetraspanins, which are used for those that do not possess a catchier name.

Where do we find them?
Tetraspanins are expressed in all multicellular organisms. They have been studied in amoebae, flatworms, fungi, plants, fish, insects and mammals. The number of family members varies between species. For instance, 17 members are encoded in the genome of *Arabidopsis thaliana*, 20 in *Caenorhabditis elegans*, 33 in *Homo sapiens*, and 37 in *Drosophila melanogaster*. At the cellular level, they localize to the cell membrane with their most important functional domain exposed to the extracellular environment. There are also a few exceptions of tetraspanins that localize intracellularly, for example the lysosomal marker CD63. Some are expressed almost ubiquitously, as shown for human CD9 and CD81; others are exclusive to particular tissues or even cell types, such as late bloomer, which in *Drosophila* is transiently expressed during development in motoneurons, or TET16 and TET17, which in *Arabidopsis* are expressed almost solely in closed buds.

What do we (believe we) know about their cellular function?
Tetraspanins are known as master organizers of the cell membrane because they form so-called tetraspanin-enriched microdomains (see next question), which often contain several different tetraspanins along with integrins, immunoglobulin superfamily proteins, proteases and/or other cell-specific receptors. Depending on the cell type and repertoire of molecules that are crowded together, tetraspanins are involved in the following cellular processes (in alphabetical order; see also Figure 1): adhesion, cell–cell fusion, endocytosis, exosome formation, immune response, migration, neurite navigation, pericellular proteolysis, proliferation, signaling, spreading, trafficking, vascular morphogenesis and remodeling, thrombosis, tumor progression and metastasis, viral and other pathogen entry, and viral release. This plurality is fascinating and simultaneously confusing. Is it possible that a family of proteins can have that many functions, just by sorting the right components to one place? Is this the end of the tetraspanin story, or did we miss something? Apart from organizing cell membranes, tetraspanins are also required for the exit from the endoplasmic reticulum by co-transport of their interaction partners.

How do tetraspanin-enriched microdomains form? Tetraspanin-enriched microdomains (TEMs) compartmentalize a cellular process by gathering together a set of components. The formation of TEMs follows a hierarchy of interactions. The first level involves primary interactions between a tetraspanin and its specific binding partner. These primary complexes are very stable and even resistant to harsh detergent treatment. Secondary interactions are weaker and form, for example, between two different tetraspanins. These interactions likely crosslink the primary complexes to larger clusters. Finally, weak interactions mediated via palmitate chains attached to tetraspanins further stabilize the network. At the microscopic level, different tetraspanins form separate nanoclusters of around 100 nm. These nanoclusters apparently gather together into larger entities that could be identical to the TEMs defined by biochemical approaches, and TEMs could further connect together to form a large tetraspanin network.

Can tetraspanins be used in practical applications? There is a high potential for the use of tetraspanins in therapeutic applications because they are relevant to cancer, immune regulation and infections. Most therapeutic approaches aim at targeting the LEL with an antibody or some other binding molecule. The idea is that this interaction with the LEL results in the tetraspanin becoming either stimulated or non-functional. Given the extracellular localization of the LEL, the therapeutic agent does not need to enter the cell. Currently, an anti-CD37 antibody is undergoing testing in clinical trials as a treatment for relapsed/refractory chronic lymphocytic leukemia. Another practical application relates to the emerging field of extracellular vesicles (EVs), where tetraspanins are used as EV biomarkers.

Have we solved the tetraspanin puzzle? The many unrelated processes that are dependent on tetraspanins could in fact mean that we have not yet discovered their universal principle of function. Is there an even more sophisticated, complex, overarching network of mechanisms, or just a simple basic function that, when disturbed or activated, could lead to different outcomes? The Occam's razor principle would suggest that we should go for the latter. One speculation is that the true function of tetraspanins is to mediate membrane curvature. In this case, TEMs would be comparable to clathrin lattices, apart from the fact that tetraspanins have transmembrane segments and predominantly act extracellularly, though one should not completely ignore intracellular domains. We also can speculate on a completely novel function cell biologists have not even thought of yet.

What are the important outstanding issues? In the past, interests in a specific area like immune biology, cancer, infectious diseases or protease regulation have often driven tetraspanin research. Therefore, we know little or nothing about many family members that have no link to these fields. Expanding research

CellPress

Current Biology
Magazine

on such neglected tetraspanins will likely help to identify novel types of interaction partner, which may help us to understand a possible universal tetraspanin function. Another area that requires more study is structural research because a complete structural model is available for only one tetraspanin so far. We should generate more examples and additionally study tetraspanin complexes at atomic resolution. Finally, what is the role of the large intracellular domains that are present in a handful of tetraspanins, or of alternatively spliced forms that lack transmembrane segments? In fact, it appears as if the tetraspanin field is just at its beginning and many exciting discoveries are yet to be made.

Where can I find out more?

Andreu, Z., and Yáñez-Mó, M. (2014). Tetraspanins in extracellular vesicle formation and function. Front. Immunol. *5*, 442.

Berditchevski, F., and Odintsova, E. (2007). Tetraspanins as regulators of protein trafficking. Traffic *8*, 89–96.

Charrin, S., Jouannet, S., Boucheix, C., and Rubinstein, E. (2014). Tetraspanins at a glance. J. Cell Sci. *127*, 3641–3648.

Florin, L., and Lang, T. (2018). Tetraspanin assemblies in virus infection. Front. Immunol. *9*, 1140.

Hemler, M. E. (2014). Tetraspanin proteins promote multiple cancer stages. Nat. Rev. Cancer *14*, 49–60.

Hochheimer, N., Sies, R., Aschenbrenner, A.C., Schneider, D., and Lang, T. (2019). Classes of non-conventional tetraspanins defined by alternative splicing. Sci. Rep. *9*, 14075.

Levy, S., and Shoham, T. (2005). The tetraspanin web modulates immune-signalling complexes. Nat. Rev. Immunol. *5*, 136–148.

Seigneuret, M., Delaguillaumie, A., Lagaudrière-Gesbert, C., and Conjeaud, H. (2001). Structure of the tetraspanin main extracellular domain. A partially conserved fold with a structurally variable domain insertion. J. Biol. Chem. *276*, 40055–40064.

Seipold, L., and Saftig, P. (2016). The emerging role of tetraspanins in the proteolytic processing of the amyloid precursor protein. Front. Mol. Neurosci. *9*, 149.

van Deventer, S.J., Dunlock, V.E., and van Spriel, A.B. (2017). Molecular interactions shaping the tetraspanin web. Biochem. Soc. Trans. *45*, 741–750.

Yeung, L., Hickey, M.J., and Wright, M.D. (2018). The many and varied roles of tetraspanins in immune cell recruitment and migration. Front. Immunol. *9*, 1644.

Zimmerman, B., Kelly, B., McMillan, B.J., Seegar, T.C.M., Dror, R.O., Kruse, A.C., and Blacklow, S.C. (2016). Crystal structure of a full-length human tetraspanin reveals a cholesterol-binding pocket. Cell *167*, 1041–1051.

Department of Membrane Biochemistry, Life & Medical Sciences (LIMES) Institute, University of Bonn, Carl-Troll-Straße 31, 53115 Bonn, Germany.
*E-mail: thorsten.lang@uni-bonn.de

Primer

Bird flocks

Steven J. Portugal

Birds often fly in aggregations and formations, sometimes gathering in vast and impressive numbers. The precise function of these flocks has been the topic of much debate. The many theories proposed to explain aerial flocking are probably not mutually exclusive. Such proposed functions include protection against predators, the sharing of orientation or information, and for certain species, energy conservation.

Flapping flight is the most energetically demanding mode of locomotion in vertebrates. How birds assemble and the dynamics within the group have significant implications for individual energy expenditure within the flock. Bird flocks are a particularly exciting study system due to their dynamic nature, the speed at which events and decision making need to occur, and the potential for collisions and injury. For example, herds of ungulates seemingly make group movement decisions through facing the same direction, or through slow directional movements aimed to illicit a response from other herd members. However, such signals pertaining to group movement and direction cannot be as subtle or slow for animal groupings moving in air, due to the risk of collision between flock members, or falling out of the sky. Birds can only fly so slow before they are no longer creating enough lift to keep themselves airborne. While this speed will vary greatly depending on the size of the bird, all birds will eventually stall below a certain minimum speed. Therefore, for birds in a flock, decisions need to be made literally on the wing, and at a certain speed. Likewise, for all animals travelling in groups collisions or indecisiveness may lead to individuals being more easily picked out by predators. Many studies have demonstrated that individuals who are less well aligned to the group, slower to respond or choose the wrong direction are more likely to be preyed upon. When birds flock in the air they form either clusters or V-formations. These two flock types are thought to provide different benefits and disadvantages.

V-formation flight

For centuries, people have been fascinated by V-formation flight. Pliny the Elder noted that flocks of geese flew ‘like fast galleys, cleaving the air more easily than if they drove at it with a straight front’. Since then, numerous ideas have been proposed to explain the function of these V-shaped flocks that are a such a familiar sight. The notion that these distinctive formations provide energetic savings for those individuals not leading the formation was based on applying fixed-wing aerodynamic theory — typically applied to aircraft — to bird flight, even though many birds flap their wings between four and seven times a second. This notion was based on the potential positive aerodynamic interactions that may be taking place between members of the flock. When a bird flies, the lift required for flight is achieved through a pressure difference between the top and bottom of the bird’s wing. This pressure difference cannot be maintained beyond the wingtips, as there is no longer a surface to create the pressure gradient. As a result, the high-pressure under the wing flows around the tip, and inwards across the dorsal surface of the wing. This in turn forms a stream of air trailing from the wingtip and behind into the birds wake, which is commonly referred to as upwash (Figure 2). This upwash flows outboard of the wing, while there is a region of downwash more centrally behind the main body of the bird. Theoretical studies have demonstrated that, if birds position themselves optimally in a V-formation, they can take advantage of the upwash from the preceding bird to contribute to lift and reduce power requirements. Simultaneously, such positioning also benefits the birds by avoiding the region of downwash, flying into which significantly increases the cost of flight.

The key to capturing upwash and avoiding downwash was believed to be due to appropriate wingtip spacing (Figure 2). Wingtip spacing is defined as the distance between the centers of two birds minus their mean maximum wingspan. In theory, as wingtip spacing decreases, the induced power required for flight also decreases as the bird is flying in stronger upwash. Hence, there is an optimal wingtip spacing that maximizes the reduction in induced power requirements through optimizing

Appendix B: Classes of non-conventional tetraspanins defined by alternative splicing

www.nature.com/scientificreports

SCIENTIFIC REPORTS
nature research

OPEN

Classes of non-conventional tetraspanins defined by alternative splicing

Received: 24 June 2019
Accepted: 10 September 2019
Published online: 01 October 2019

Nikolas Hochheimer[1], Ricarda Sies[1], Anna C. Aschenbrenner[2,3], Dirk Schneider[4] & Thorsten Lang[1]

Tetraspanins emerge as a family of membrane proteins mediating an exceptional broad diversity of functions. The naming refers to their four transmembrane segments, which define the tetraspanins' typical membrane topology. In this study, we analyzed alternative splicing of tetraspanins. Besides isoforms with four transmembrane segments, most mRNA sequences are coding for isoforms with one, two or three transmembrane segments, representing structurally mono-, di- and trispanins. Moreover, alternative splicing may alter transmembrane topology, delete parts of the large extracellular loop, or generate alternative N- or C-termini. As a result, we define structure-based classes of non-conventional tetraspanins. The increase in gene products by alternative splicing is associated with an unexpected high structural variability of tetraspanins. We speculate that non-conventional tetraspanins have roles in regulating ER exit and modulating tetraspanin-enriched microdomain function.

Tetraspanins are small membrane proteins expressed in all multicellular eukaryotes. With a few exceptions, they are localized at the plasma membrane[1]. Tetraspanins are plasma membrane (PM) master organizers or scaffolding proteins[2,3]. They interact with one another and integrins, immunoglobulin superfamily proteins, proteases and receptors. By these interactions they laterally associate a set of components into so-called tetraspanin-enriched microdomains (TEMs)[4]. Additionally, tetraspanin interactions in the ER are required for co-transport of proteins from the endoplasmatic reticulum (ER) to the PM, a mechanism depending on TEMs already assembling in the ER[5]. Depending on the cell type and the group of associated proteins, TEMs mediate many different functions[6]. As a result, tetraspanins regulate trafficking, signaling, cell proliferation, adhesion, spreading, migration, cell-cell fusion, pathogen entry, cancer and other diseases[4,7].

The human genome encodes 33 tetraspanins (Tspans) systematically named Tspan1 – Tspan33[8]. However, the systematic nomenclature is rarely applied to frequently studied tetraspanins as CD9 (Tspan29), CD63 (Tspan30), CD81 (Tspan28), CD82 (Tspan27), and CD151 (Tspan24). These historic names refer to their identification by the "cluster of differentiation (CD)" protocol. Other historic names are for example uroplakin 1A and 1B (Tspan21 and Tspan20), peripherin-2 (Tspan22) or rod outer segment membrane protein (Tspan23) (for a complete list of non-systematic names see Table S1).

Structurally, all tetraspanins share the same topology (Fig. 1). Thus, in a typical tetraspanin about a third of the protein sequence (see also Table S1) orders into the four transmembrane segments (TMSs), which form two transmembrane helical hairpins. On the extracellular site, a small extracellular loop (SEL) connects the first and the second TMS, and a large extracellular loop (LEL) the third and the fourth TMS[1]. For the SEL, no structural data is available, even in the crystal structure of a full-length Tetraspanin[9]. In contrast, LEL crystal structures[9,10] reveal five largely α-helical segments (α - ε)[11] forming a compact structure. The LEL subdivides into a conserved (α, β and ε) and a variable domain (γ and δ). The differences in the variable domains explain the specificity of interactions between tetraspanins and their primary binding partners[12,13]. However, interactions are not restricted to the LEL, but also involve the transmembrane segments and the C-terminus[14]. The N- and C-terminus and

[1]Department of Membrane Biochemistry, Life & Medical Sciences Institute (LIMES), University of Bonn, Carl-Troll-Straße 31, 53115, Bonn, Germany. [2]Department of Genomics and Immunoregulation, Life & Medical Sciences Institute (LIMES), University of Bonn, Carl-Troll-Straße 31, 53115, Bonn, Germany. [3]Department of Internal Medicine and Radboud Center for Infectious Diseases (RCI), Radboud University Nijmegen Medical Center, Geert Grooteplein Zuid 8, Nijmegen, 6525 GA, The Netherlands. [4]Institute of Pharmacy and Biochemistry, Johannes Gutenberg University Mainz, Johann-Joachim-Becher-Weg 30, 55128, Mainz, Germany. Correspondence and requests for materials should be addressed to T.L. (email: thorsten.lang@uni-bonn.de)

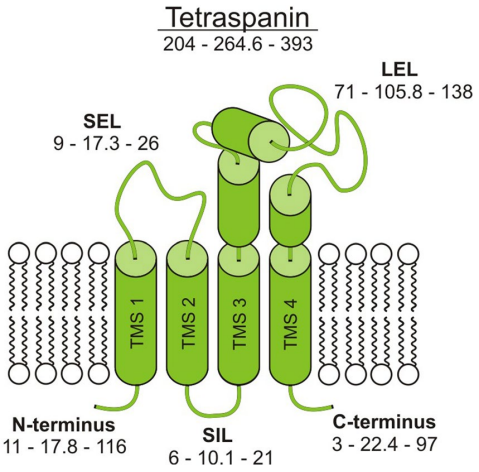


Figure 1. Conventional tetraspanin topology. Depicted is the typical topology of a tetraspanin. Intracellular domains include the N-terminus, the small intracellular loop (SIL), and the C-terminus, which are all short (for exceptions see Table S1). At the extracellular site, a small extracellular loop (SEL) connects transmembrane segment 1 (TMS1) and TMS2 and a large extracellular loop (LEL) TMS3 and TMS4. For the complete tetraspanin and its different segments, the three numbers (xx-yy-zz) indicate the sequence lengths of the shortest sequence (xx), the average sequence (yy) and the longest sequence (zz) (for details see Table S1).

the TMS2-TMS3-linker, called small intracellular loop (SIL), are located at the intracellular site. They are short segments containing putative sorting motifs[5] and palmitoylation sites which stabilize tetraspanin protein interactions[15]. Only five tetraspanins have large C-terminal domains (by definition that their C-termini are two-fold larger than the average; Table S1), from which one has in addition a large N-terminal domain. To date, little is known about their role. For instance, in PRPH-2 an amphipathic helix within the C-terminus partitions into the cytosolic membrane leaflet mediating curvature[16].

As typical in eukaryotes, *tspan* genes have alternating sections of exons (coding) and introns (non-coding)[17]. From a precursor mRNA, the introns are "spliced out", yielding the mature mRNA for translation into the protein[18]. Splicing is catalyzed by the spliceosome either co-transcriptionally during transcription, or immediately afterwards. Moreover, self-splicing introns do exist[19]. Yet, the splice product is not necessarily well defined and a pre-mRNA may undergo several splicing pathways, called alternative splicing (AS)[20]. In fact more than 90% of all human genes are subject to AS[21]. Therefore, AS increases the number of gene products[22]. In human, AS leads to more than 80,000 transcripts encoded in the 20,000 human genes[23]. This equals a 4-fold increase in transcriptome diversity, with possible effects in the regulation of protein function.

In contrast to soluble proteins, only little is known about AS of mRNAs coding for membrane integral proteins[24]. Yet, at least ¼ of all open reading frames in any organism code for membrane proteins[25,26]. Compared to soluble proteins, AS could have a stronger impact on the function of membrane proteins. For instance, in eukaryotes, most membrane proteins integrate co-translationally into the membrane at the Sec translocon[27]. Here, the nascent polypeptide chain enters or crosses the membrane. Statistics show that cytosolic segments near the TMS are more positively charged when compared to extracellular segments (positive-inside rule; negative inside depletion/outside enrichment rule[28,29]). This opens the possibility that AS changes the charge distribution and consequently membrane topology. Moreover, certain features of membrane proteins are required for subsequent trafficking from the ER via the Golgi network to the plasma membrane[30]. In the case of tetraspanins, ER retention is caused by truncation of a TMS in CD82[31], malfunctions in post translational modifications of Tspan1[32], or by Tspan5 mutants folding improperly[33]. However, proper folding may not be sufficient for ER exit, because deletion of the first CD82 TMS precludes ER exit although the LEL has folded properly. Transport is only restored when the TMS is co-expressed as a separate peptide[31]. Altogether, it is likely that AS eliminates sections required for efficient ER exit.

In the following, we have analyzed the variety of AS of tetraspanin pre-mRNA from the human genome that may enrich the tetraspanin gene products.

Results and Discussion

We screened the National Center for Biotechnology Information (NCBI) data bank for human tetraspanin gene products. Taking into account only validated and reviewed sequences, we identified 86 mRNAs originating from the 33 human tetraspanin genes. In addition, we found via PCR the sequence of two novel mRNAs, one from a human whole brain and one from a natural killer cells cDNA library (Fig. S1).

Finally, we included a splice variant of CD82 described in the literature[34]. In total, the 89 gene products include the known 33 conventional tetraspanin proteins and 31 different, non-conventional isoforms. The non-conventional isoforms originate from 18 conventional tetraspanins. For Tspan17, we found the highest number of five isoforms (Table 1).

Compared to the structure of a conventional tetraspanin (Fig. 1), non-conventional tetraspanins display broad structural variability. As examples, we explain the isoforms of Tspan6 (for illustration of isoforms for Tspan2, Tspan3, Tspan16, Tspan17, CD53, CD82, CD63, and Tspan31 see Figs S2–S9, respectively). Figure 2A shows the genomic sequence together with five mRNAs, from which four are derived by AS. In Fig. 2B, we depict the proteins deriving from the mRNA splice variants. Illustrated are remaining and deleted protein segments with reference to the conventional tetraspanin topology (Fig. 1), not yet predicting how the deletion may affect protein topology and/or the numbers of TMSs. Apart from the deletion of protein segments, in all Tspan6 splice variants AS produces additional changes in the 5′-UTR (untranslated region). These changes eventually cause diminished expression (see below).

The first mRNA codes for the conventional Tspan6 (isoform 1). In all other isoforms, the first two TMSs are missing (Fig. 2B). The second and third mRNA differs in their 5′-end but have the same alternative start codon. Therefore, both yield isoform 2 with large part of the N-terminus deleted, including TMS1, the SEL and TMS2. Also in case of isoform 3 this alternative start codon is used, resulting in TMS1/SEL/TMS2-deletion. Moreover, splicing eliminates exon 7 by which an alternative stop codon located in exon 8 is used. This causes deletion of the C-terminal half of TMS4 and an alternative C-terminus. Finally, isoform 4 again uses the alternative start codon, resulting in the N-terminal truncation. Moreover, exon 6 is eliminated, and thus the C-terminal end of the LEL ε-helix and the N-terminal half of TMS4 are not encoded.

We wondered whether such deletions also occur in other species and analyzed tetraspanin isoforms in mouse. Here, the database has lesser entries, as only 31 tetraspanins are described, four of them with provisional status only, and in general there are not that many mRNA variants available. Still, we identify eight non-conventional isoforms, including isoforms with only three predicted TMSs, LEL deletions, and changes in the N-terminus (Table S2). Between the two species, there is no direct correlation on the level of specific tetraspanins, but there is overlap in the type of structural change caused by AS. That not all structural variations occurring in human are also found in mouse maybe explained by the smaller data base and/or that the two species share only about a quarter of alternatively used exons[35,36].

Structural variability defines classes of non-conventional tetraspanins. We next analyzed the topologies of the human non-conventional isoforms. Based on computational analyses of the proteins' transmembrane helices (TMHMM Server, 2.0) we predict protein isoforms with overall one, two, three or four TMSs (Fig. 3). Thus, the isoforms categorize into four major classes, which are tetraspanins that structurally are mono-, di-, tri- and tetraspanins (Fig. 4). The monospan-tetraspanins maintain either TMS 3 or 4, and the dispan-tetraspanins TMSs 1 & 2, 3 & 4, or 4 and form a novel TMS. In the trispan-tetraspanins, any one of the TMSs is deleted, with the exception of TMS2. In one case in which TMS2 is remaining, TMS2 forms an extended TMS together with a half-deleted TMS1 (CD63 Iso2). Please note that for simplicity in the following we refer to e.g. trispan-tetraspanins just as trispanins.

In about half of the cases, AS results in a partially or completely inverted topology (indicated by an asterisk in Fig. 4). Surprisingly, an inverted topology is also predicted for the conventional Tspan15. However, experimental evidence indirectly indicates that murine Tspan15, which is also predicted to have an inverted topology, inserts with the correct topology[37]. Therefore, topology predictions should be treated with caution.

Most classes include representatives with a modified C-terminus. Moreover, in several cases AS affects the LEL, causing almost its complete elimination (CD53 Iso2), or shortening (CD82 Iso4 and Tspan17 Iso2, 3, 4 and 5). Based on the structure of the CD81 LEL[9] and the prediction of secondary structural elements (Jpred 4.0), the short deletions would largely affect the variable domain of the LEL, which is interesting, as this part is supposed to encode the information for specific interactions. Finally, for Tspan10, the only tetraspanin with a large N-terminus, we find an isoform with a truncation in the large N-terminal domain (Tspan10 Iso2).

In summary, for most tetraspanins AS generates several mRNAs, yielding up to five isoforms per gene (see Tspan17). The number of non-conventional tetraspanin isoforms roughly equals the number of conventional tetraspanins. However, it is very likely that this is greatly underestimated as we included only validated/reviewed sequences. Moreover, the discovery of many yet undocumented sequences is expected.

Expression of non-conventional tetraspanins. The question arises as to how likely the protein isoforms express at levels that would affect cellular function. Several factors would play a role, such as mRNA copy number (about which the data bank makes no statement), the stability of the mRNA, and the stability of the expressed protein.

In the following, we evaluate the stability of the 53 listed alternatively spliced mRNAs by analyzing features making mRNA prone to degradation or influencing its expression level (Table 1). All mRNAs lack retained introns and premature termination codons (PTCs), which would promote nonsense-mediated decay (NMD)[38]. This argues against enhanced degradation by such elements. For clarity, we have not included this information in Table 1.

We find 31 spliced mRNA variants with a sole alteration in the 5′ UTR (21 mRNAs coding for 8 conventional tetraspanins) or an alteration in the 5′ UTR and the ORF (10 mRNAs coding for 9 non-conventional tetraspanins). The 5′ UTR contains regulatory elements of translation. Effects on expression level upon alteration of this region are unpredictable[39–41]. To be on the safe side, we make a conservative estimate and assume that expression rather would be diminished. Therefore, the proteins for which these 31 mRNAs code for do not rate as being "very likely expressed", but "likely expressed" (Fig. 5). The expression of two mRNAs coding for conventional and five mRNAs coding for non-conventional tetraspanins is not likely (Fig. 5). In these cases, we find alterations in the 3′ UTR that may cause NMD, retention in the nucleus and miRNA binding sites and/or an uORF (upstream open reading frame), which reduces expression levels 30–80% and/or makes the mRNA more likely subject to NMD[39,42]. However, this may be an overcautious rating as four not alternatively spliced mRNAs

Historic Name	Systematic Name	mRNA (NM_)	Protein (NP_)	Isoform	5′UTR	ORF Exons	3′UTR	uORF	Alternative stop codon	Frame shift
	Tspan1	005727.4	005718.2	1	1–3	3–9	9	X		
	Tspan2	005725.6	005716.2	1	1	1–8	8			
		001308315.1	001295244.1	2		Δ4				
		001308316.1	001295245.1	3		Δ7				
	Tspan3	005724.6	005715.1	1	1	1–7	7			
		198902.3	944492.1	2		Δ3				
		001168412.2	001161884.1	3		Δ2				
	Tspan4	001025237.1	001020408.1	1	1–3	3–9	9	X		
		001025234.1	001020405.1	1	↔1					
		001025235.1	001020406.1	1	↔1					
		001025236.1	001020407.1	1	↔1					
		003271.5	003262.1	1	↔1					
		001025238.2	001020409.1	1	↔1, Δ2					
		001025239.1	001020410.1	2	Δ1, Δ3	Δ3				
	Tspan5	005723.4	005714.2	1	1	1–8	8			
	Tspan6	003270.4	003261.1	1	1	1–7	7–8			
		001278740.2	001265669.1	2	1^{ASS}	1^{ASS}		X		
		001278741.1	001265670.1	2	1^{ASS}	1^{ASS}		X		
		001278742.1	001265671.1	3	1^{ASS}	1^{ASS}, Δ7	Δ7	X	Δ7	
		001278743.1	001265672.1	4	1^{ASS}	1^{ASS}, Δ6		X		
	Tspan7	004615.3	004606.2	1	1	1–7	7–8			
	Tspan8	004616.3	004607.1	1	1–2	2–9	9			
		001369760.1	001356689.1	1	Δ1, 2^{ASS}					
	Tspan9	006675.5	006666.1	1	1–3	3–9	9			
		001168320.1	001161792.1	1	Δ2					
	Tspan10	001290212.1	001277141.1	1	1	1–4	4			
		031945.4	114151.3	2	Δ1	Δ1				
	Tspan11	001080509.2	001073978.1	1	1–2	2–8	8	X		
		001370301.1	001357230.1	2	Δ2	Δ2				
		001370302.1	001357231.1	1	↔1					
	Tspan12	012338.4	036470.1	1	1–2	2–8	8			
	Tspan13	014399.4	055214.1	1	1	1–6	6			
	Tspan14	030927.3	112189.2	1	1–2	2–9	9			
		001128309.2	001121781.1	2		Δ3–5				
		001351266.1	001338195.1	1	+1					
		001351267.3	001338196.1	1	+1					
		001351268.1	001338197.1	1	+1					
		001351269.1	001338198.1	1	↔1					
		001351270.1	001338199.1	1	↔1			X		
		001351271.1	001338200.1	1	↔1					
		001351272.1	001338201.1	1	↔1					
	Tspan15	012339.5	036471.1	1	1	1–8	8			
		001351263.1	001338192.1	2		Δ2–3				
	Tspan16	001282509.2	001269438.1	1	1	1–7	7			
		012466.4	036598.1	2		↔7			↔7	↔7
		001282510.2	001269439.1	3		Δ3				
	Tspan17	012171.3	036303.1	1	1	1–9	9			
		130465.5	569732.2	2		6^{ASS}				
		001006616.3	001006617.2	3		6^{ASS}, 8^{ASS}			8^{ASS}	8^{ASS}
		001366491.2	001353420.1	4	1^{ASS}	6^{ASS}, 7^{ASS}				
		001366492.2	001353421.1	5	1^{ASS}	Δ2–3, 6^{ASS}				
	Tspan18	130783.5	570139.3	1	1–3	3–9	9			
	Tspan19	001100917.2	001094387.1	1	1–2	2–9	9			
UPK1b	Tspan20	006952.4	008883.2	1	1–2	2–8	8			
UPK1a	Tspan21	007000.3	008931.1	1	1–2	2–9	9			
		001281443.1	001268372.1	2		+6			+6	+6

Continued

Historic Name	Systematic Name	mRNA (NM_)	Protein (NP_)	Isoform	5′UTR					
PRPH2	Tspan22	000322.5	000313.2	1	1	1–3	3			
ROM1	Tspan23	000327.3	000318.1	1	1	1–3	3			
CD151	Tspan24	004357.5	004348.2	1	1–3	3–9	9			
		139030.3	620599.1	1	Δ2					
		139029.1	620598.1	1	1^{ASS}					
		001039490.1	001034579.1	1	1^{ASS} Δ2					
CD53	Tspan25	001040033.1	001035122.1	1	1–3	3–9	9			
		000560.4	000551.1	1	Δ1, 2^{ASS}					
		001320638.1	001307567.1	2	Δ1, 2^{ASS}	Δ6–7				
		from Natural Killer Cells cDNA		3	N/A	Δ5–8	N/A	N/A		
CD37	Tspan26	001774.3	001765.1	1	1	1–8	8			
		001040031.2	001035120.1	2		1^{ASS}				
CD82	Tspan27	002231.4	002222.1	1	1–3	3–10	10			
		001024844.1	001020015.1	2		Δ6				
		from Human Brain cDNA		3	N/A	Δ9	N/A	N/A		
		Lee *et al.* 2003		4	N/A	Δ7	N/A	N/A		
CD81	Tspan28	004356.3	004347.1	1	1	1–8	8			
		001297649.1	001284578.1	2	Δ1	Δ1, 2^{ASS}				
CD9	Tspan29	001769.4	001760.1	1	1	1–8	8			
		001330312.2	001317241.1	2	↔1	↔1		X		
CD63	Tspan30	001780.5	001771.1	1	1–2	2–8	8			
		001257389.1	001244318.1	1	↔1			X		
		001257390.1	001244319.1	1	↔1					
		001257391.1	001244320.1	1	Δ1	2^{ASS}				
		001257392.1	001244321.1	2	Δ1	3^{ASS}				
		001257400.1	001244329.1	3	↔1, Δ2	Δ2				
		001257401.1	001244330.1	3	↔1, Δ2	Δ2				
		001267698.1	001254627.1	1	1^{ASS}					
	Tspan31	005981.5	005972.1	1	1	1–6	6			
		001330168.2	001317097.1	2		Δ2–3				
		001330169.2	001317098.1	3		1^{ASS}				
	Tspan32	139022.2	620591.3	1	1	1–10	10			
	Tspan33	178562.5	848657.1	1	1	1–8	8	X		

Table 1. Tetraspanin splice variants. Left, historic names used with priority over the systematic names in the NCBI database. More historic names with lower priority are listed in Table S1. Second and third columns, mRNA variants are sorted by systematic name, next sorted by the NCBI variant number for the mRNA. For Tpan16 and Tspan21, the first mRNAs variants are trispanins, and the second ones are tetraspanins. In these cases, we moved up the second mRNA variants referring to them as conventional tetraspanins (isoforms 1). Forth column, NCBI reference sequence number for protein (NP). Column 5 lists the isoform number. Column 6 lists the exons forming the 5′ UTR of the respective splice variant 1, being the reference sequence for comparisons with the 5′ UTR of the other alternatively spliced variants. For the alternatively spliced variants, the column lists the eliminated exon(s) (Δ), the number of an exchanged exon (↔) or the number of an exon after which another exon has been introduced (+). In case an alternative splice site (ASS) is used, the number of the exon with the ASS is given. Columns 7 and 8 provide the same information for the ORF and the 3′ UTR, respectively. Column9; a cross indicates an open reading frame upstream of the ORF. Modifications by AS generating a new stop codon are shown in column 10, and modifications generating a frame shift in column 11. N/A, not available.

encoding conventional tetraspanins also contain uORFs, arguing against a complete uORF induced decay of tetraspanin mRNAs.

Finally, 15 mRNAs from 10 tetraspanin genes, all coding for non-conventional isoforms, are very likely expressed because they lack 5′ untranslated region (5′ UTR) alterations, uORF, PTCs or 3′ UTR alterations (Table 1) (Fig. 5). From these 15 mRNAs, nine code for proteins that have the extracellular loops on the extracellular site and therefore a membrane topology similar or identical to the respective conventional tetraspanin (Fig. 5), meaning their domains could in principle interact with binding partners. Moreover, there is another isoform with a shortened LEL (CD82 Iso4), from which the mRNA is unknown, wherefore we cannot evaluate its expression probability. Still, from published data we can safely conclude that this isoform expresses at levels that affect cellular function[43].

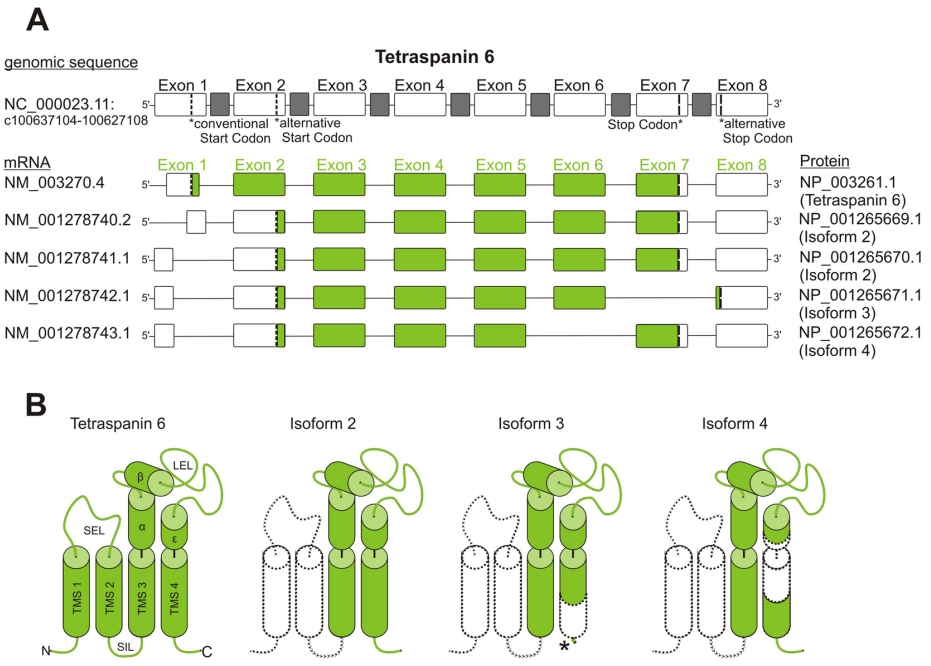


Figure 2. Isoforms of Tspan6. (**A**) Top, cartoon illustrating the genomic sequence of Tspan6 as exons (white boxes) and introns (grey boxes). Exon numbering refers to the genomic sequence. From the genomic sequence, five different mRNAs derive. Here, exon numbering (green) refers to mRNA variant 1. Green boxes mark the open reading frame. Exon-skipping is indicated by leaving out the exon. When compared to splice variant 1, a shortened exon box indicates the use of an AS site. The green exon numbering is used for comparison of the respective mRNA variant 1 to the splice variants in Tab. 1. (**B**) Left, helical structural elements of Tspan6 predicted by Seigneuret *et al.*[12]. The cartoons of the isoforms only illustrate the alterations in the primary structure and are no predictions of the protein topology. TMS, transmembrane segment; SIL, small intracellular loop; SEL, small extracellular loop; LEL, large extracellular loop; α - ε, helices of the LEL. Dashed lines mark missing parts (filled white). The asterisk in isoform 3 marks the alternative C-terminus.

In addition, for those isoforms amplified by PCR, we compared mRNA expression levels of the conventional to the non-conventional isoform(s) by quantitative real-time PCR. In the cDNA library from human brain, the Tspan15 Iso2 mRNA level is about 10% of the conventional form, and this wild-type Tspan15 is expressed at a several fold higher level than RPS9, encoding the 40S ribosomal protein S9, used as a reference (Table S3). Hence, although lower expressed than the conventional, the Tspan15 Iso2 expression level is substantial and in a dominant negative mechanism could be sufficient to alter cellular functions. Moreover, although expressed 10-fold less than Tspan15, Tspan15 Iso2 may dominate in the ER by accumulating there (compare Fig. S13).

CD53 expression in natural killer cells is also dominated by the conventional transcript, which is again found at a several fold higher level than RPS9. In comparison, CD53 Iso2 and Iso3 expression were found to be 5% and 3% of the conventional form. For CD53 Iso2 this is expected, as it has a lower mRNA expression probability when compared to Tspan15 Iso2 (Fig. 5). For CD53 Iso3 the expression probability cannot be evaluated, as only the coding sequence is known. In any case, for the CD53 isoforms it is difficult to predict whether they may influence cell physiology at such low expression levels. However, future analysis of other cDNA libraries may reveal cellular systems with higher expression levels.

Retention in the ER of co-transported factors. What might be the physiological effects of expressing non-conventional tetraspanins? In most cases described here, alternative splicing results in expression of variants with missing TMSs (compare Fig. 4). The role of the individual TMSs for proper folding has been studied to some extent, and especially tight packing of the TMSs1 and 2 appear to be crucial for proper tetraspanin folding[44]. Moreover, all four TMSs of Tspan20 are required for proper protein folding and forward-trafficking from the ER to the plasma membrane[45]. Thus, formation of a proper four-helix bundle structure appears to be crucial for ER exit. In conclusion, it appears to be very unlikely that tetraspanins with missing TMSs will be able to leave the ER. In fact, when studying the distribution of Tspan15 Iso2, which is a dispanin, we find retention in the ER (Fig. 6 and Fig. S13).

Yet, tetraspanin variants retained in the ER could affect cell physiology in two ways: First, complementation of a truncated tetraspanin via interaction with the "missing" helix of its full-length counterpart is possible, eventually resulting in improper folding of the full-length tetraspanin. Via a domino effect, this could result in cross-linked tetraspanins not leaving the ER. Actually, formation of unspecific tetraspanin aggregates has been suggested to be

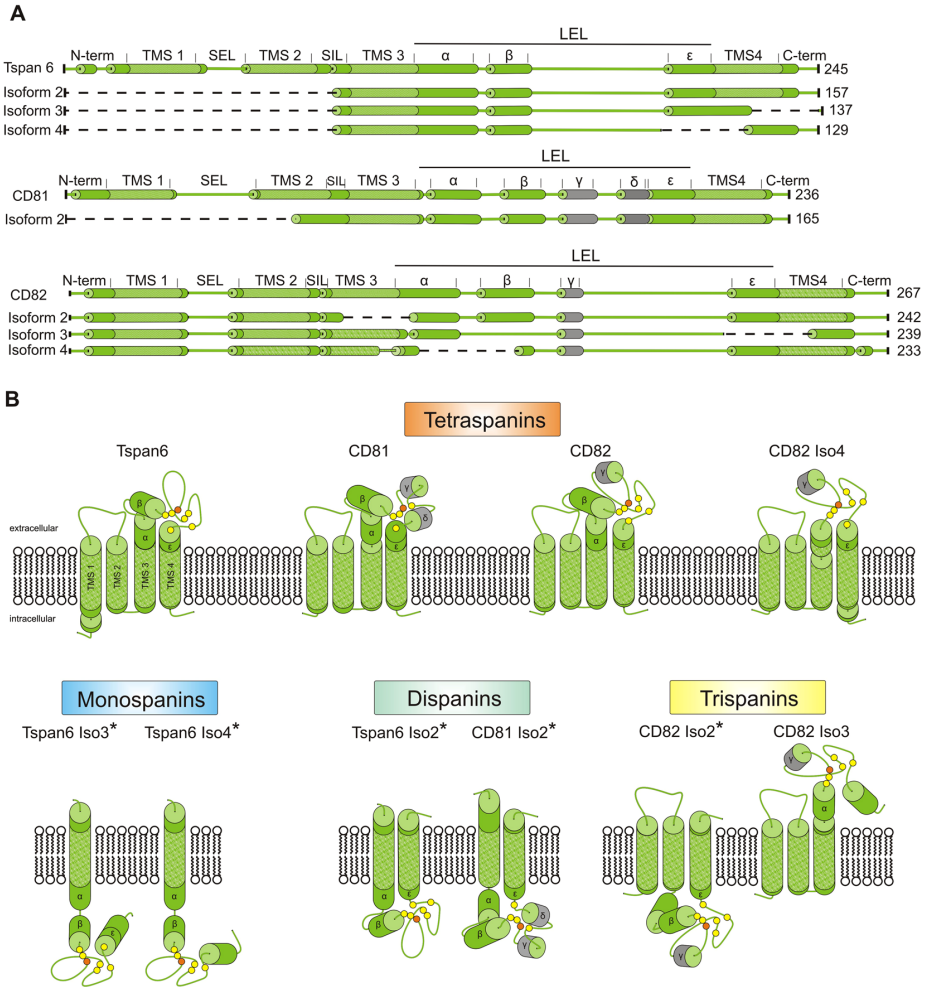


Figure 3. Predicted topology of tetraspanin isoforms. (**A**) Linearized proteins. Dashed lines indicate deleted parts. Green cylinders, α-helical structure predicted by Jpred. Grey cylinders, α-helical structure not predicted by Jpred but by Seigneuret *et al.*[12], and in case of CD81 revealed from crystallographic data (Kitadokoru et al., 2001). Patterned green marks the predicted transmembrane helices (TMHMM Server, 2.0). The length of the sections scales with the number of amino acids. TMS, transmembrane segment; SEL, small extracellular loop; SIL, small intracellular loop; LEL, large extracellular loop; α - ε, α-helices in the LEL[11]. For Tspan6, no alpha helical structure of the variable domain is predicted wherefore no γ- and δ-helix are depicted. The AS of Tspan6 Iso3 leads to an alternative C-terminus. For CD82 in the variable domain only the γ-helix is predicted to be α-helical. (**B**) Topology of the tetraspanin isoforms illustrated in (**A**) with reference to the prediction which parts are intra- and extracellular (TMHMM Server, 2.0). Isoforms with an inverted topology are indicated by an asterisk. Yellow and orange spheres indicate cysteine- and glycine-residues, respectively. Cysteine-residues form disulfide bridges in the LEL; the glycine-residue is part of a conserved CCG-motif.

a mechanism causing ER retention[45]. Likely, these aggregates would be degraded and therefore such a mechanism would decrease the tetraspanin level at the cell-surface. Second, isoforms retained in the ER could still bind to their interaction partners, holding these in the ER and causing their degradation (Fig. 6). Evidence that such a mechanisms could exist comes from a study in which a mutation in the CD81 gene produces an isoform that is lacking TMS4, which is accompanied by a lack of expression of the CD81 interaction partner CD19[46].

Non-functional TEMs. While deletions of TMSs cause ER retention, modifications of the N- or C-terminus, or the LEL may still allow proteins to traffic to the cell membrane. Previously, it has been shown that deletions of segments in the CD81 LEL (deleting the α/β-, γ/δ-, γ- or δ-helical segment(s)) do not result in inefficient plasma membrane targeting[47]. Moreover, deletion of the entire LEL in CD53 Iso2 still allows for trafficking to the cell membrane (Fig. 6). In Jurkat T cells, that express endogenous CD81 at high levels, the additional expression

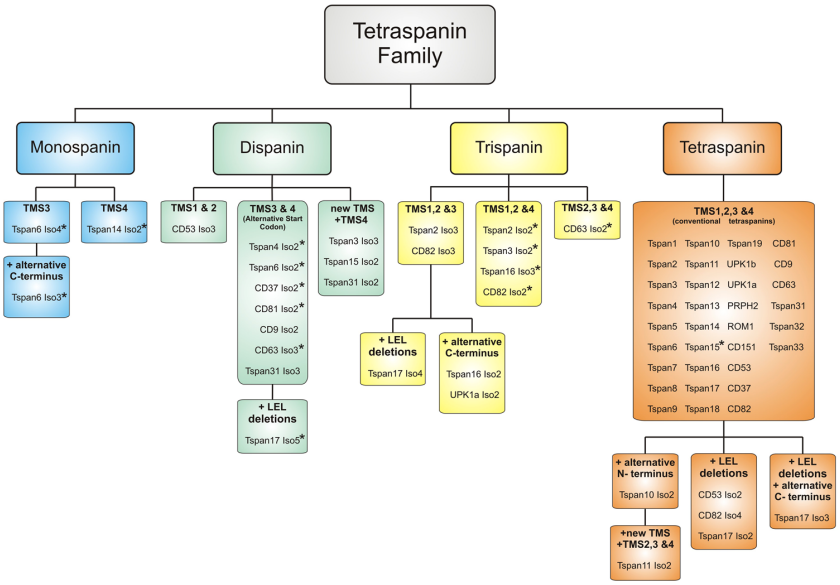


Figure 4. Classes of tetraspanins. Based on the analysis illustrated for the examples Tspan6, CD81 and CD82 (Fig. 3), all tetraspanins and their isoforms were classified as mono-, di-, tri- or tetraspanins. Subclasses result from the type of remaining TMSs, or whether a novel TMS is formed. Alteration of the N- or C- terminus, or the LEL define further subclasses. Isoforms with a partially or completely inverted topology are marked by an asterisk.

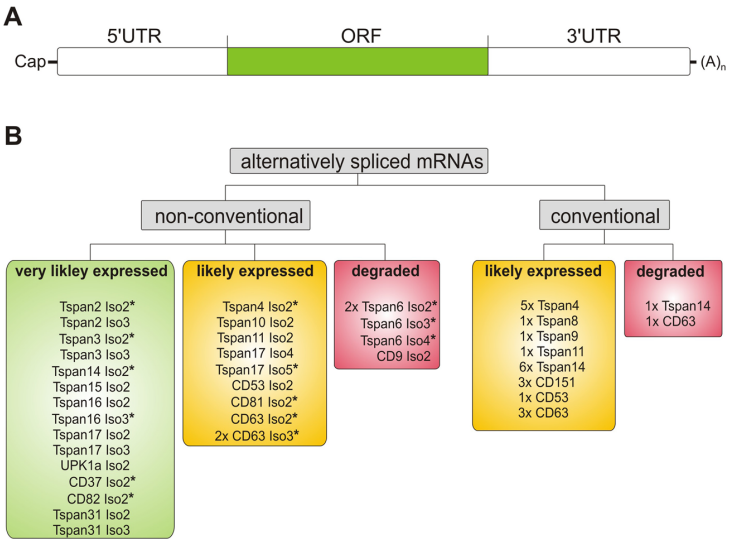


Figure 5. Evaluation of the expression probability of alternatively spliced mRNAs. (**A**) Sections of an mRNA. Cap, 5′-Cap; 5′ untranslated region (5′ UTR); ORF, open reading frame (ORF); 3′ untranslated region (3′ UTR); $(A)_n$, poly(A) tail. (**B**) All alternatively spliced mRNAs lack retained introns and PTCs. In addition, mRNAs were analyzed for an upstream open reading frame (uORF), which induces NMD. They were also tested for alterations in the 3′UTR that could be associated with NMD, retention in the nucleus via nuclear RNA quality control, and miRNA-based gene silencing. Finally, they were analyzed for alteration in the 5′UTR that can alter the expression level of the mRNA. Based on these criteria, the mRNAs were sorted into three groups ranking their expression probability from very likely expressed (green - none of the criteria match), likely expressed (yellow - only alterations in the 5′UTR), or degraded (red - uORF and/or alteration in the 3′UTR).

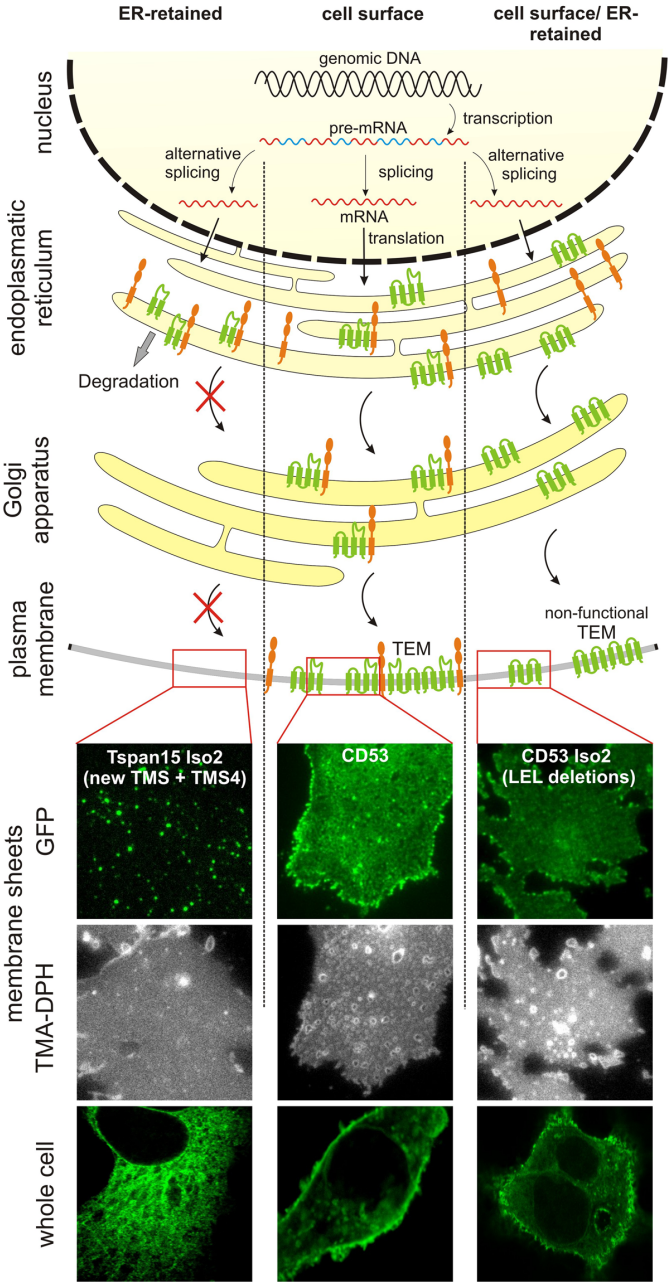


Figure 6. Trafficking and possible functions of non-conventional tetraspanins. Top, illustration of alternative splicing and trafficking from the ER to the plasma membrane. Transcription of the genomic DNA (black) generates pre-mRNA with introns (blue) and exons (red). AS generates two additional different mRNAs. After translation and insertion into the ER membrane, apart from the classical pathway (middle), isoforms may behave differently in two ways. Middle, the conventional tetraspanin (green) interacts with a binding partner (orange) and both are co-transported to the plasma membrane, where the tetraspanin forms a TEM. Left, most isoforms lack TMS. The isoform shown (green) is an example from the largest group of dispanins. They cannot exit the ER, but may still interact with other proteins. Thus, if it is degraded together with the binding partner, the surface expression level of the binding partner is altered. Right, the LEL deleted isoform (green) does not interact with its binding partner (orange) but exits the ER and forms TEMs in the plasma membrane. These TEMs would lack one or more co-factors and would therefore be non-functional or differently acting TEMs. Bottom, the lower panels show confocal micrographs of GFP-labeled Tspan15 Iso2 (the conventional Tspan15

is shown in Fig. S11), CD53 or CD53 Iso2 expressed in HepG2 cells (for non-GFP-expressing control cells see Fig. S10; Western blot analysis documents the correct size of the expressed constructs; see Fig. S12). Tspan15 reaches the plasma membrane (Fig. S11), whereas Tspan15 Iso2 remains in the ER (for co-staining analysis with an ER marker see Fig. S13). Bottom, upper panels, ER retention is confirmed by analysis of cell-free plasma membrane sheets that were visualized by the membrane dye TMA-DPH. In the respective GFP-channel, only a few Tspan15 Iso2 spots are detected, that arise from ER-PM contact sites[50]. In contrast, CD53 and CD53 Iso2 readily reach the plasma membrane, albeit CD53 Iso2 less efficient. CD53 Iso2 has lost its glycosylation sites and therefore appears in Western blot analysis as a single band (Fig. S12).

of the CD81 mutant lacking the δ-helical segment inhibits viral uptake[47], which indicates that the mutant has a dominant negative effect. This suggests that LEL deletion mutants might still be able to integrate into TEMs into which otherwise the conventional tetraspanin locates. However, as the deletion mutant does not properly interact anymore with its interaction partners, the TEM becomes non-functional (Fig. 6). It is also possible that the TEM loses only part of its functionality, resulting e.g. in aberrant cellular signaling.

Conclusion

Little is known about the effect of AS on membrane proteins. Using the tetraspanin family as example, we studied whether AS enriches the gene products, revealing a large structural variability of tetraspanin isoforms. We speculate that non-conventional tetraspanins may regulate ER exit of tetraspanins and their interaction partners, form non-functional TEMs, or TEMs with different roles.

Materials and Methods

Sequence acquisition and cloning. The human tetraspanin sequences are acquired from the National Center for Biotechnology Information (NCBI) database for genes (as of 5th June 2019), listed under 'NCBI Reference Sequence (RefSeq) - mRNA and Protein(s)'. For human tetraspanins, we considered only sequences with the status report 'reviewed' or 'validated'.

Two additional sequences were obtained by PCR from cDNA libraries (kindly provided by the AG Kolanus, LIMES institute, Bonn). The cDNA libraries used as PCR template were from human brain for Tspan15 Iso2 and CD82 Iso3, and from natural killer cells for CD53 Iso2 and CD53 Iso3. We employed Q5 High-Fidelity DNA Polymerase (NEB, # M0491S) and primers aligning with the 5′- and 3′-end of the ORF (without a stop codon) of the corresponding conventional tetraspanin. Primers carried a XhoI restriction site at the 5′- end, before the Kozak sequence and the start codon, and a blunt 3′-end. The following primers were used: Tspan15, 5′-TATTATCTCGAGCATGCCGCGCGGGGACTCGGAGC-3′ (fwd) and 5′-ATTGGGGTAGCACAAGCAGCATCCCG-3′ (rev); CD53, 5′-TATTATCTCGAGCATGGGCATGA GTAGCTTGAAAC-3′ (fwd) and 5′-TAGCCCTATGGTCTGGCTGG-3′ (rev); CD82 5′-TATTATCTCGA GCATGGGCTCAGCCTGTATCAAAGTC-3′ (fwd) and 5′-GTACTTGGGGACCTTGCTGTAGTCTTCGG-3′ (rev). The PCR products were digested with XhoI and the inserts were ligated into a pEGFP-C1 vector (Clontech, #6084-1) containing a monomeric enhanced GFP variant[47]. Also the backbone vector was amplified via PCR using the primers 5′-ATGGTGAGCAAGGGCGAGG-3′ (fwd) and 5′- ATAATACTCGAGGGATCTGACGGTTCACTAAACC- 3′ (rev). The amplified sequences were XhoI (NEB, #R0146S) digested and ligated using T4 Polynucleotide Kinase (NEB, #M0201S) and T4 DNA Ligase (NEB, #M0202S). The constructs were verified by sequencing (Eurofins GATC Biotech GmbH).

Expression analysis by quantitative real-time PCR. Isoform-specific expression was measured in human cDNA samples from brain and natural killer cells by quantitative real-time PCR using the Maxima SYBR Green qPCR Master Mix (Thermo Scientific, #K0221). Primers used for intron-spanning assay: all CD53 isoforms, 5′-TCATGGTAGTTGCCTTCCTGG -3′ (fwd); CD53, 5′-CACATACTCATTCAGCTTCTGTTC-3′ (rev); CD53 Iso2, 5′-CGCATAGCAACCCTTCTGTTC-3′ (rev); CD53 Iso3, 5′-CATCCCCAACACCGACATAAG-3′ (rev); all Tspan15 isoforms, 5′-CTGCAGTCGTGGTACTGATTC-3′ (rev); Tspan15, 5′-TCCGGAACCAGACCATTGAC-3′ (fwd); Tspan15 Iso2, 5′-CCGTGTTCTGGACCATTGAC-3′ (fwd); RSP9, 5′-CTGCTGACGCTTGATGAGAA-3′ (fwd) and 5′-CAGCTTCATCTTGCCCTA-3′ (rev). Reactions were run in a total volume of 10 µl containing 200 pM of each primer with the following program: 40 cycles with incubation at 95 °C for 10 min, followed by 30 s at 60 °C and 30 s at 72 °C on a Roche LightCycler® 480 II. The experiment was performed twice with each reaction pipetted in duplicates. Data were analyzed by advanced relative quantification in the LightCycler® 480 Software using RPS9 as a reference.

Analysis of transmembrane segments. The protein sequences were analyzed employing the program TMHMM Server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/[48]). The program predicts with a certain likelihood the length and the position of transmembrane segments, and the intra- and extracellular localization of the segments connected to the TMSs. These TMSs lengths and the lengths of the interconnecting segments are shown in Table S1. In addition, for Tspan10, 19 and 22, we considered TMS as positively predicted if the segment had a total length of 15–35 residues. From these residues, the central 13–33 residues had a transmembrane probability ≥54%, and were flanked by one intracellular and one extracellular residue with a lower transmembrane probability[48]. In some cases, TMSs were shifted or shortened due to AS by a few amino acids. Here, we classified the TMS as a novel one if the shift was greater than five amino acids or more than 1/3 of the original TMS was replaced. Finally, the analysis indicates which domains or segments are changed by AS.

Structure prediction. The helical structural elements of the isoforms shown in Fig. 3 were predicted combining the results from the TMS prediction by the program TMHMM Server v. 2.0 and a secondary structure

prediction of the complete protein (JPred4; http://www.compbio.dundee.ac.uk/jpred4[49]). The relative lengths of the segments containing predicted helices, their position in the membrane, and the interconnecting segments were illustrated. The approximate position of the cysteine residues within the variable part of the LEL and the glycine and cysteine residues in the ubiquitously conserved cysteine-cysteine-glycine motif (CCG motifs)[12] are also shown. The β-strands were not illustrated. The structural suggestions of the interconnecting domains refer to previous drawings of tetraspanins and claim no accuracy.

Analysis of AS generated changes of mRNA. To determine the alterations by AS, we used the program BioEdit v7.0.5 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). Exons were defined with reference to the NCBI database. The changes generated by AS were identified by sequence alignment of the splice variant with the conventional tetraspanin sequence and are summarized in Table 1. Specifically, we tested for alterations in the 5′UTR and 3′UTR, NMD (nonsense mediated decay) initiating PTCs and uORFs (upstream open reading frames), which all may affect the mRNA expression level. The 5′UTR and 3′UTR alterations were directly extracted from the alignment described above. The PTCs were tested for their NMD initiation potential by measuring the distance between the PTC and the most 3′ exon-exon junction (EEJ). PTCs more than 50 nucleotides upstream of the most 3′-EEJ were categorized as NMD promoting. However, none of the PTCs fulfilled this criterion. The sequences were also tested for uORFs by translating them to their corresponding amino acid sequence (ExPASy Bioinformatics Resource Portal; https://web.expasy.org/translate/) and testing them for an in frame open reading frame 5′ upstream of the tetraspanin start codon.

Expression and imaging of GFP-labelled tetraspanins. HepG2 cells were transfected essentially as described[47] with a vector for expression of GFP (pEGFP-N1, clonetech, #6085-1) or the above described vectors for expression of GFP fused to the C-terminus of Tspan15, Tspan15 Iso2, CD53 or CD53 Iso2. Cell-free membrane sheets were produced by short ultra-sound pulses[47]. If not stated otherwise, epi-fluorescence microscopy was employed for imaging membrane sheets and whole cells that in this case additionally were visualized with the membrane dye TMA-DPH (Invitrogen, #T204). TMA-DPH and GFP-fluorescence were imaged by epi-fluorescence microscopy essentially as described[47]. For confocal microscopy, cells additionally expressed KDEL-RFP and were stained as described below. They were imaged in the confocal mode of a 4-channel easy3D superresolution STED optics module (Abberior Instruments) coupled to an Olympus IX83 confocal microscope (Olympus, Tokyo, Japan), equipped with an UPlanSApo 100x (1.4 NA) objective (Olympus, Tokyo, Japan). For imaging details see below. Additionally, GFP was excited with a 485 nm laser and recorded with a 525/50 nm filter.

Western blotting of GFP-tagged Tetraspanins. HepG2 cells were lysed 22 h after transfection by addition of buffer A (82.25 mM Tris-HCl, 32.9% (w/v) glycerol, 2.6% SDS, pH 6.8). The lysate was mixed 1:4 with buffer A with additional 5% β-mercaptoethanol and was heated to 95 °C for 10 min. Proteins were separated on a 12% SDS-PAGE and blotted on a Nitrocellulose membrane (Carl Roth, #HP40.1) using a tank blot system (Bio-Rad, #1703930). The membrane was blocked with blocking buffer, a 1:1 mixture of TBS-T (50 mM Tris, 150 mM NaCl, 0.05% Tween 20 (w/v), pH 7.4) and Odyssey Blocking Buffer (Li-Cor, #927-40000). Membranes were incubated with primary antibodies in blocking buffer and incubated over night at 4 °C. Primary antibodies used were rabbit polyclonal anti-GFP (Thermo Fisher Scientific, #A-11122) diluted 1:2,000 and mouse monoclonal anti-beta-Actin (Cell Signaling, #3700) diluted 1:5,000. The secondary antibodies donkey anti-mouse coupled to IRDye 680RD (Li-Cor, #926-68072) and donkey anti-rabbit coupled to IRDye 800CW (Li-Cor, #926-32213) were diluted 1:10,000 in blocking buffer. The membrane was imaged using an Odyssey Classic Imaging System (Li-Cor).

Colocalization of Tetraspanin 15 with the endoplasmic reticulum. HepG2 cells were transfected with GFP-labelled Tspan15 or Tspan15 Iso2 as described above. Six hours after transfection, cells were transduced with a KDEL-RFP fusion construct (BacMam 2.0, Life Technologies, # C10591) specifically targeting the ER according to the manufacturer's instructions with 20 particles per cell for an additional 16 h. Cells were fixed with 4% paraformaldehyde (PFA) in PBS for 30 minutes. Fixation solution was removed and residual PFA was quenched with 50 mM NH_4Cl in PBS for 30 minutes. Cells were then permeabilized with 0.2% Triton X-100 in PBS for 2 minutes and blocked with 3% BSA for 1 hour at room temperature (RT). To enhance the GFP- and RFP-signal, samples were incubated with GFP-Booster Atto647N (Chromotek, # gba647n) and RFP-Booster Atto594 (Chromotek, # rba594) diluted 1:200 in 1% BSA for 1 hour at RT. At last, samples were washed with PBS and mounted onto microscopy slides with ProLong® Gold antifade mounting medium (Invitrogen, #P36930). Coverslips were cured for 24 hours and sealed with nail polish. Cells were imaged in the confocal mode of the superresolution STED microscope described above. Atto594/RFP was excited with a 561 laser and detected with a 580–630 nm filter (red channel). Atto647N was excited with a 640 nm laser and recorded with a 650–720 nm filter (long red channel). For all images, pixel size was set to 50 nm and pinhole size was set to 60 µm.

Colocalization analysis was performed with the program ImageJ. Regions of interest (ROIs) were placed into the red channel to an area that showed the typical ER network structure, and then propagated to the long red channel (illustrated in the figure employing a green lookup table). The Pearson correlation coefficient (PCC) between the two areas marked by the ROIs was calculated with a custom made ImageJ macro.

References

1. Boucheix, C. & Rubinstein, E. Tetraspanins. *Cell. Mol. Life Sci. CMLS* **58**, 1189–1205 (2001).
2. Maecker, H. T., Todd, S. C. & Levy, S. The tetraspanin superfamily: molecular facilitators. *FASEB J.* **11**, 428–442 (1997).
3. Miranti, C., Bergsma, A. & Spriel, A. *Tetraspanins as Master Organizers of the Plasma Membrane. In Cell Membrane Nanodomains: From Biochemistry to Nanoscopy* 59–86 Edited by A. Cambi and D.S. Lidke, CRC Press (2014).
4. Charrin, S., Jouannet, S., Boucheix, C. & Rubinstein, E. Tetraspanins at a glance. *J Cell Sci* **127**, 3641–3648 (2014).

5. Berditchevski, F. & Odintsova, E. Tetraspanins as Regulators of Protein Trafficking. *Traffic* **8**, 89–96 (2007).
6. Yáñez-Mó, M., Barreiro, O., Gordon-Alonso, M., Sala-Valdés, M. & Sánchez-Madrid, F. Tetraspanin-enriched microdomains: a functional unit in cell plasma membranes. *Trends Cell Biol.* **19**, 434–446 (2009).
7. Hemler, M. E. Tetraspanin functions and associated microdomains. *Nat. Rev. Mol. Cell Biol.* **6**, 801–811 (2005).
8. Charrin, S. *et al*. Lateral organization of membrane proteins: tetraspanins spin their web. *Biochem. J.* **420**, 133–154 (2009).
9. Zimmerman, B. *et al*. Crystal Structure of a Full-Length Human Tetraspanin Reveals a Cholesterol-Binding Pocket. *Cell* **167**, 1041–1051.e11 (2016).
10. Kitadokoro, K. *et al*. CD81 extracellular domain 3D structure: insight into the tetraspanin superfamily structural motifs. *EMBO J.* **20**, 12–18 (2001).
11. Schmidt, T. H., Homsi, Y. & Lang, T. Oligomerization of the Tetraspanin CD81 via the Flexibility of Its δ-Loop. *Biophys. J.* **110**, 2463–2474 (2016).
12. Seigneuret, M., Delaguillaumie, A., Lagaudrière-Gesbert, C. & Conjeaud, H. Structure of the tetraspanin main extracellular domain. A partially conserved fold with a structurally variable domain insertion. *J. Biol. Chem.* **276**, 40055–40064 (2001).
13. Seigneuret, M. Complete predicted three-dimensional structure of the facilitator transmembrane protein and hepatitis C virus receptor CD81: conserved and variable structural domains in the tetraspanin superfamily. *Biophys. J.* **90**, 212–227 (2006).
14. Stipp, C. S., Kolesnikova, T. V. & Hemler, M. E. Functional domains in tetraspanin proteins. *Trends Biochem. Sci.* **28**, 106–112 (2003).
15. Charrin, S. *et al*. Differential stability of tetraspanin/tetraspanin interactions: role of palmitoylation. *FEBS Lett.* **516**, 139–144 (2002).
16. Khattree, N., Ritter, L. M. & Goldberg, A. F. X. Membrane curvature generation by a C-terminal amphipathic helix in peripherin-2/rds, a tetraspanin required for photoreceptor sensory cilium morphogenesis. *J. Cell Sci.* **126**, 4659–4670 (2013).
17. Garcia-España, A. & DeSalle, R. Intron sliding in tetraspanins. *Commun. Integr. Biol.* **2**, 394–395 (2009).
18. Sharp, P. A. Split genes and RNA splicing. *Cell* **77**, 805–815 (1994).
19. Kornblihtt, A. R. *et al*. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.* **14**, 153–165 (2013).
20. Wang, Y. *et al*. Mechanism of alternative splicing and its regulation. *Biomed. Rep.* **3**, 152–158 (2015).
21. Wang, E. T. *et al*. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
22. Early, P. *et al*. Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways. *Cell* **20**, 313–319 (1980).
23. Klerk, Ede & Hoen, P. A. C. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends Genet.* **31**, 128–139 (2015).
24. Mittendorf, K. F., Deatherage, C. L., Ohi, M. D. & Sanders, C. R. Tailoring of Membrane Proteins by Alternative Splicing of Pre-mRNA. *Biochemistry (Mosc.)* **51**, 5541–5556 (2012).
25. Almén, M. S., Nordström, K. J., Fredriksson, R. & Schiöth, H. B. Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biol.* **7**, 50 (2009).
26. Fagerberg, L., Jonasson, K., von Heijne, G., Uhlén, M. & Berglund, L. Prediction of the human membrane proteome. *Proteomics* **10**, 1141–1149 (2010).
27. Hegde, R. S. & Keenan, R. J. Tail-anchored membrane protein insertion into the endoplasmic reticulum. *Nat. Rev. Mol. Cell Biol.* **12**, 787–798 (2011).
28. Baker, J. A., Wong, W.-C., Eisenhaber, B., Warwicker, J. & Eisenhaber, F. Charged residues next to transmembrane regions revisited: "Positive-inside rule" is complemented by the "negative inside depletion/outside enrichment rule". *BMC Biol.* **15**, 66 (2017).
29. Heijne, Gvon The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J.* **5**, 3021–3027 (1986).
30. Lee, M. C. S., Miller, E. A., Goldberg, J., Orci, L. & Schekman, R. Bi-Directional Protein Transport Between the Er and Golgi. *Annu. Rev. Cell Dev. Biol.* **20**, 87–123 (2004).
31. Cannon, K. S. & Cresswell, P. Quality control of transmembrane domain assembly in the tetraspanin CD82. *EMBO J.* **20**, 2443–2453 (2001).
32. Scholz, C.-J., Sauer, G. & Deissler, H. Glycosylation of tetraspanin Tspan-1 at four distinct sites promotes its transition through the endoplasmic reticulum. *Protein Pept. Lett.* **16**, 1244–1248 (2009).
33. Saint-Pol, J. *et al*. New insights into the tetraspanin Tspan5 using novel monoclonal antibodies. *J. Biol. Chem.* **292**, 9551–9566 (2017).
34. Lee, J. H., Seo, Y.-W., Park, S. R., Kim, Y. J. & Kim, K. K. Expression of a splice variant of KAI1, a tumor metastasis suppressor gene, influences tumor invasion and progression. *Cancer Res.* **63**, 7247–7255 (2003).
35. Barbosa-Morais, N. L. *et al*. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593 (2012).
36. Modrek, B. & Lee, C. J. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.* **34**, 177–180 (2003).
37. Prox, J. *et al*. Tetraspanin15 regulates cellular trafficking and activity of the ectodomain sheddase ADAM10. *Cell. Mol. Life Sci.* **69**, 2919–2932 (2012).
38. McGlincy, N. J. & Smith, C. W. J. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends Biochem. Sci.* **33**, 385–393 (2008).
39. Mockenhaupt, S. & Makeyev, E. V. Non-coding functions of alternative pre-mRNA splicing in development. *Semin. Cell Dev. Biol.* **47–48**, 32–39 (2015).
40. Resch, A. M., Ogurtsov, A. Y., Rogozin, I. B., Shabalina, S. A. & Koonin, E. V. Evolution of alternative and constitutive regions of mammalian 5′UTRs. *BMC Genomics* **10**, 162 (2009).
41. Sobczak, K. & Krzyzosiak, W. J. Structural Determinants of BRCA1 Translational Regulation. *J. Biol. Chem.* **277**, 17349–17358 (2002).
42. Somers, J., Pöyry, T. & Willis, A. E. A perspective on mammalian upstream open reading frame function. *Int. J. Biochem. Cell Biol.* **45**, 1690–1700 (2013).
43. Miller, J. *et al*. Differential tumor biological role of the tumor suppressor KAI1 and its splice variant in human breast cancer cells. *Oncotarget* **9**, 6369–6390 (2018).
44. Kovalenko, O. V., Metcalf, D. G., DeGrado, W. F. & Hemler, M. E. Structural organization and interactions of transmembrane domains in tetraspanin proteins. *BMC Struct. Biol.* **5**, 11 (2005).
45. Tu, L., Kong, X.-P., Sun, T.-T. & Kreibich, G. Integrity of all four transmembrane domains of the tetraspanin uroplakin Ib is required for its exit from the ER. *J. Cell Sci.* **119**, 5077–5086 (2006).
46. van Zelm, M. C. *et al*. CD81 gene defect in humans disrupts CD19 complex formation and leads to antibody deficiency. *J. Clin. Invest.* **120**, 1265–1274 (2010).
47. Homsi, Y. *et al*. The Extracellular δ-Domain is Essential for the Formation of CD81 Tetraspanin Webs. *Biophys. J.* **107**, 100–113 (2014).
48. Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175–182 (1998).
49. Drozdetskiy, A., Cole, C., Procter, J. & Barton, G. J. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* **43**, W389–W394 (2015).
50. Merklinger, E., Schloetel, J.-G., Spitta, L., Thiele, C. & Lang, T. No Evidence for Spontaneous Lipid Transfer at ER-PM Membrane Contact Sites. *J. Membr. Biol.* **249**, 41–56 (2016).

Acknowledgements
We would like to thank Prof. Dr. Dietmar Schmucker (University of Bonn) for comments on the manuscript and Prof. Dr. Waldemar Kolanus (University of Bonn) for providing cDNA libraries. This work was supported by grants from DynaMem (state of Rhineland-Palatinate) to D.S.

Author Contributions
N.H., conceptualization, data acquisition, formal analysis, validation, figure preparation, and writing-original draft, R.S. and A.C.A. data acquisition, formal analysis, validation, figure or table preparation, D.S., writing original draft, conceptualization, T.L., conceptualization, validation, supervision and writing original draft.

Additional Information
Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-50267-0.

Competing Interests: The authors declare no competing interests.

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

Supplementary Data

Classes of non-conventional tetraspanins defined by alternative splicing

Nikolas Hochheimer[1], Ricarda Sies[1], Anna C. Aschenbrenner[2,3], Dirk Schneider[4] and Thorsten Lang[1,*]

[1]Department of Membrane Biochemistry and [2]Department of Genomics and Immunoregulation, Life & Medical Sciences Institute (LIMES), University of Bonn, Carl-Troll-Straße 31, 53115 Bonn, Germany, [3]Department of Internal Medicine and Radboud Center for Infectious Diseases (RCI), Radboud University Nijmegen Medical Center, Geert Grooteplein Zuid 8, 6525GA Nijmegen, The Netherlands, [4]Institute of Pharmacy and Biochemistry, Johannes Gutenberg University Mainz, Johann-Joachim-Becher-Weg 30, 55128 Mainz, Germany

*correspondence should be addressed to thorsten.lang@uni-bonn.de

Supplementary Tables S1 – S3 and Figures S1 – S13

historic names	systematic	Length of Tetraspanin Domains: N-term	TMS1	SEL	TMS2	SIL	TMS3	LEL	TMS4	C-term	full length	LEL-Cys
NET-1, TM4C	Tspan1	11	23	19	23	12	23	103	23	4	241	6
NET-3, TSN2	**Tspan2**	12	23	19	23	11	23	76	23	11	221	4
TM4-A, TM4SF8	**Tspan3**	12	23	14	23	12	23	104	23	19	253	6
NAG-2, TM4SF7	**Tspan4**	11	23	19	23	6	23	96	23	14	238	6
NET-4, TM4SF9	Tspan5	19	23	19	23	8	23	119	23	11	268	8
T245, TM4SF6	**Tspan6**	19	23	14	23	12	23	96	23	12	245	6
A15, CCG-B7	Tspan7	16	23	19	20	11	23	102	23	12	249	6
CO-029, TM4SF3	Tspan8	12	23	14	23	12	23	99	23	8	237	6
NET-5, PP1057	Tspan9	11	23	19	23	6	23	99	23	12	239	6
OCSP	**Tspan10**	116	23	19	23	12	23	116	20	41	393	8
VSSW1971	**Tspan11**	20	23	14	23	12	23	106	23	9	253	6
EVR5, NET-2	Tspan12	12	23	19	23	12	23	115	23	55	305	6
NET-6, TM4SF13	Tspan13	12	23	9	23	6	20	71	23	17	204	6 (+2)
DC-TM4F2, TM4SF14	**Tspan14**	19	23	19	23	6	23	119	23	15	270	8
2700063A19Rik, NET-7	**Tspan15**	21	23	12	23	14	23	118	23	37	294	8
TM-8, TM4-B	**Tspan16**	12	23	19	23	11	23	107	23	3	244	6
FBX23, TM4SF17	**Tspan17**	19	23	21	23	8	23	120	23	72	332	8
TSPAN	Tspan18	12	23	14	23	12	23	114	23	4	248	6
	Tspan19	11	23	21	23	6	23	106	17	18	248	6
UPK1b, UPIB, UPK1	Tspan20	12	23	19	23	11	23	121	23	5	260	6
UPK1a, UP1A, UPIA	**Tspan21**	12	23	24	23	11	23	116	23	3	258	6
PRPH2, AOFMD, RDS	Tspan22	20	23	18	17	21	23	130	23	71	346	6 (+1)
ROM1, ROSP1, RP7	Tspan23	20	23	17	23	19	23	138	23	65	351	6 (+1)
CD151, GP27, PETA-3	Tspan24	19	23	14	23	12	23	107	23	9	253	6
CD53, MOX44	**Tspan25**	12	23	14	23	11	23	75	23	15	219	4
CD37, GP52-40	**Tspan26**	12	26	19	23	6	23	134	23	15	281	6
CD82, 4F9, KAI1	**Tspan27**	12	23	19	23	6	23	123	23	15	267	6
CD81, CVID6, TAPA1	**Tspan28**	12	23	26	23	6	23	90	23	10	236	4
CD9, BTCC-1, DRAP-27	**Tspan29**	12	23	24	23	6	23	83	23	11	228	4
CD63, LAMP-3, ME491	**Tspan30**	12	23	14	23	12	23	98	23	10	238	6
SAS	**Tspan31**	11	23	9	23	6	20	81	20	17	210	6 (+2)
ART1, TSSC6	Tspan32	19	23	17	23	6	23	89	23	97	320	4
PEN	Tspan33	24	23	14	23	12	23	119	23	22	283	8
	Mean	17.8	23.1	17.3	22.7	10.1	22.8	105.8	22.6	22.4	264.6	6.1
	Min	11	23	9	17	6	20	71	17	3	204	4
	Max	116	26	26	23	21	23	138	23	97	393	8

Table S1: Segment lengths of conventional tetraspanins. Listed are historic and systematic names. The historic names include at least two examples from the NCBI database listed under “Aliases”. In case historic names are used with priority over the systematic names, up to three historic names are shown. The tetraspanins with isoforms are written bold. The length of the protein domains are based on the TMHMM2.0 TMS predictions using the NCBI tetraspanin sequences. For the LEL, the number of disulfide bond forming cysteines is indicated together with the cysteines from the CCG motif (written in brackets). Bottom; mean, minimal and maximal values for each column.

historic Name	systematic Name	mRNA (NM_)	Protein (NP_)	Isoform	5'UTR	ORF Exon	3'UTR	Class	Changes
	Tspan1*	133681.4	598442.1	1	1 - 2	2 - 8	8		
	Tspan2	027533.3	081809.2	1	1	1 - 8	8		
		001243132.1	001230061.1	2	↔1	↔1		Tetraspanin	alt. N-term, new TMS1, SEL shortened
	Tspan3	019793.3	062767.3	1	1	1 - 7	7		
	Tspan4	053082.3	444312.1	1	1 - 2	2 - 8	8		
		001252588.1	001239517.1	1	↔1				
	Tspan5	019571.5	062517.1	1	1	1 - 8	8		
	Tspan6	019656.3	062630.2	1	1	1 - 7	7 - 8		
	Tspan7	019634.2	062608.2	1	1	1 - 7	7 - 8		
	Tspan8	146010.2	666122.1	1	1	1 - 8	8		
		001168679.1	001162150.1	1	1^{ASS}				
		001168680.1	001162151.1	1	1^{ASS}				
	Tspan9	175414.5	780623.1	1	1 - 3	3 - 9	9		
		001356301.1	001343230.1	1	Δ2				
	Tspan10**	145363.2	663338.2	1	1	1 - 3	3		
	Tspan11	026743.3	081019.1	1	1 - 2	2 - 8	8		
	Tspan12	173007.4	766595.1	1	1 - 3	3 - 9	9		
		001363814.1	001350743.1	2	Δ1	Δ8		Tetraspanin	LEL deletions
	Tspan13*	025359.3	079635.1	1	1	1 - 6	6		
	Tspan14	145928.2	666040.1	1	1 - 2	2 - 9	9		
		001316748.1	001303677.1	1	↔1				
	Tspan15	197996.2	932113.2	1	1	1 - 8	8		
	Tspan16	N/A	N/A	N/A					
	Tspan17	028841.3	083117.2	1	1	1 - 8	8		
	Tspan18	183180.2	899003.1	1	1 - 5	5 - 10	10		
	Tspan19	N/A	N/A	N/A					
UPK1b	Tspan20	178924.4	849255.2	1	1 - 2	2 - 8	8		
UPK1a	Tspan21*	026815.2	081091.1	1	1 - 2	2 - 8	8		
PRPH2	Tspan22	008938.2	032964.1	1	1	1 - 3	3		
ROM1	Tspan23	009073.4	033099.3	1	1	1 - 3	3		
CD151	Tspan24	009842.3	033972.2	1	1 - 3	3 - 8	8		
		001111049.1	001104519.1	1	1^{ASS}				
		001111050.1	001104520.1	1	↔1, Δ2				
CD53	Tspan25	007651.3	031677.1	1	1 - 2	2 - 8	8		
CD37	Tspan26	001290802.1	001277731.1	1	1	1 - 9	9		
		001290804.1	001277733.1	2		Δ2		Tetraspanin	alt. N-term -TMS1- SEL-TMS2
		007645.4	031671.1	3	Δ1, 2^{ASS}	Δ1, 2^{ASS}		Tetraspanin	alt. N-term
CD82	Tspan27	007656.5	031682.1	1	1 - 3	3 - 10	10		
		001136055.2	001129527.1	1	↔1				
		001271432.1	001258361.1	1	+1				
		001271430.1	001258359.1	1	Δ2				
		001271462.1	001258391.1	1	↔1				

		001271461.1	001258390.1	1	↔1				
		001271431.1	001258360.1	1	↔1, Δ2				
CD81	Tspan28	133655.2	598416.1	1	1	1 - 8	8		
CD9	Tspan29	007657.4	031683.1	1	1	1 - 8	8		
CD63	Tspan30	001042580.1	001036045.1	1	1 - 2	2 - 8	8		
		007653.3	031679.1	1	↔1				
		001282966.1	001269895.1	1	↔1				
	Tspan31*	025982.4	080258.1	1	1	1 - 6	6		
	Tspan32	001128080.2	001121552.1	1	1	1 - 9	9		
		020286.3	064682.1	2	Δ1	Δ1, 2^{ASS}		Trispanin	Δ N-term - TMS1
		001128081.1	001121553.1	3	Δ1	Δ1, 2^{ASS}, Δ7		Trispanin	Δ N-term - TMS1, LEL deletion, new TMS4
		001128082.1	001121554.1	4	Δ1	Δ1, 2^{ASS}, Δ5, Δ7		Trispanin	Δ N-term - TMS1, LEL deletion, new TMS4
	Tspan33	146173.3	666285.1	1	1	1 - 8	8		
		001301407.1	001288336.1	2		3^{ASS}		Tetraspanin	Deletion in SEL

Table S2: Tetraspanin splice variants in mouse.
Mouse tetraspanin sequences are from the National Center for Biotechnology Information (NCBI) database for genes (as of 28^{h} August 2019). Left, historic names, second column systematic names (asterisk indicates that the status in the database is provisional and not reviewed or validated). Third column, mRNA variants are sorted by the NCBI variant number. Forth column, NCBI reference sequence number for protein (NP). Column 5 lists the isoform number. Column 6 lists the exons forming the 5' UTR of the respective splice variant 1 (grey), being the reference sequence for comparisons with the 5' UTR of the other alternatively spliced variants. For the alternatively spliced variants (white), the column gives the number of an eliminated exon (Δ), the number of an exchanged exon (↔) or the number of an exon after which another exon has been introduced (+). In case an alternative splice site (ASS) is used, the number of the exon with the ASS is given. Columns 7 and 8 provide the same information for the ORF and the 3' UTR, respectively. For the eight non-conventional isoforms, columns 9 indicates the class and column 10 lists the specific changes. **Conventional tetraspanin is predicted to have only three TMSs.

Target	Target Mean Cp	Reference	Reference Mean Cp	Mean Target/Reference	SD	Mean Ratio over conventional	SD
Tspan15	19.89			4.96	1.86	1.00	
		RPS9	22.08				
Tspan15 Iso2	23.43			0.41	0.11	0.09	0.02
CD53	18.34			4.79	1.30	1.00	
		RPS9	20.54				
CD53 Iso2	22.55			0.25	0.05	0.05	0.01
CD53 Iso3	23.66			0.12	0.04	0.03	0.00

Table S3: Isoform-specific expression of CD53 and Tspan15.
Listed are mean Cp (crossing point) values for expression of Tspan15 and CD53 isoforms in human brain and natural killer cells, respectively, as well as the mean target/reference ratios. "Ratio over conventional" compares isoform transcript levels to the transcript of each gene. Values are given as means ± SD (n = 4).

CD53 Iso3 sequence from natural killer cell cDNA

5'- ATGGGCATGAGTAGCTTGAAACTGCTGAAGTATGTCCTGTTTTTCTTCAACTTGCTCT
TTTGGATCTGTGGCTGCTGCATTTTGGGCTTTGGGATCTACCTGCTGATCCACAACAAC
TTCGGAGTGCTCTTCCATAACCTCCCCTCCCTCACGCTGGGCAATGTGTTTGTCATCGT
GGGCTCTATTATCATGGTAGTTGCCTTCCTGGGCTGCATGGGCTCTATCAAGGAAAACA
AGTGTCTGCTTATGTCGGTGTTGGGGATGTCCTTTGCACTGACCCTGAACTGCCAGATT
GACAAAACCAGCCAGACCATAGGGCTA -3'

CD82 Iso3 sequence from human brain cDNA

5'- ATGGGCTCAGCCTGTATCAAAGTCACCAAATACTTTCTCTTCCTCTTCAACTTGATCT
TCTTTATCCTGGGCGCAGTGATCCTGGGCTTCGGGGTGTGGATCCTGGCCGACAAGAG
CAGTTTCATCTCTGTCCTGCAAACCTCCTCCAGCTCGCTTAGGATGGGGGCCTATGTCT
TCATCGGCGTGGGGGCAGTCACTATGCTCATGGGCTTCCTGGGCTGCATCGGCGCCG
TCAACGAGGTCCGCTGCCTGCTGGGGCTGTACTTTGCTTTCCTGCTCCTGATCCTCATT
GCCCAGGTGACGGCCGGGGCCCTCTTCTACTTCAACATGGGCAAGCTGAAGCAGGAG
ATGGGTGGCATCGTGACTGAGCTCATTCGAGACTACAACAGCAGTCGCGAGGACAGCC
TGCAGGATGCCTGGGACTACGTGCAGGCTCAGGTGAAGTGCTGCGGCTGGGTCAGCT
TCTACAACTGGACAGACAACGCTGAGCTCATGAATCGCCCTGAGGTCACCTACCCCTG
TTCCTGCGAAGTCAAGGGGGAAGAGGACAACAGCCTTTCTGTGAGGAAGGGCTTCTGC
GAGGCCCCCGGCAACAGGACCCAGAGTGGCAACCACCCTGAGGACTGGCCTGTGTAC
CAGGAGCTCCTGGGGATGGTCCTGTCCATCTGCTTGTGCCGGCACGTCCATTCCGAAG
ACTACAGCAAGGTCCCCAAGTAC -3'

Figure S1: Nucleotide sequences isolated from cDNA libraries for CD53 Iso3 and CD82 Iso3.

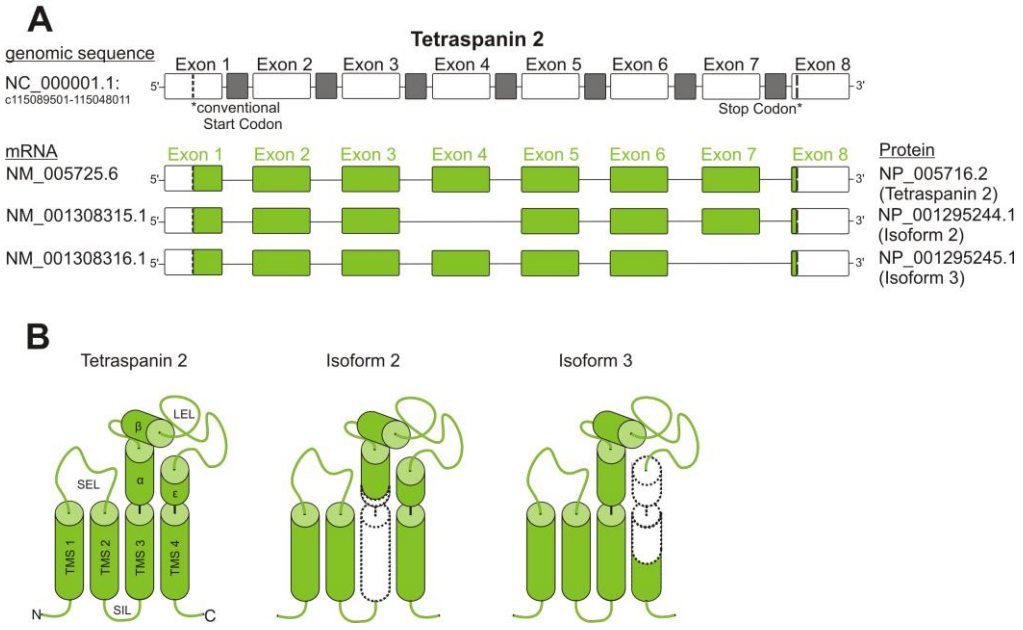


Figure S2: Tspan2 isoforms.

For scheme explanation, please see legend to figure 2.

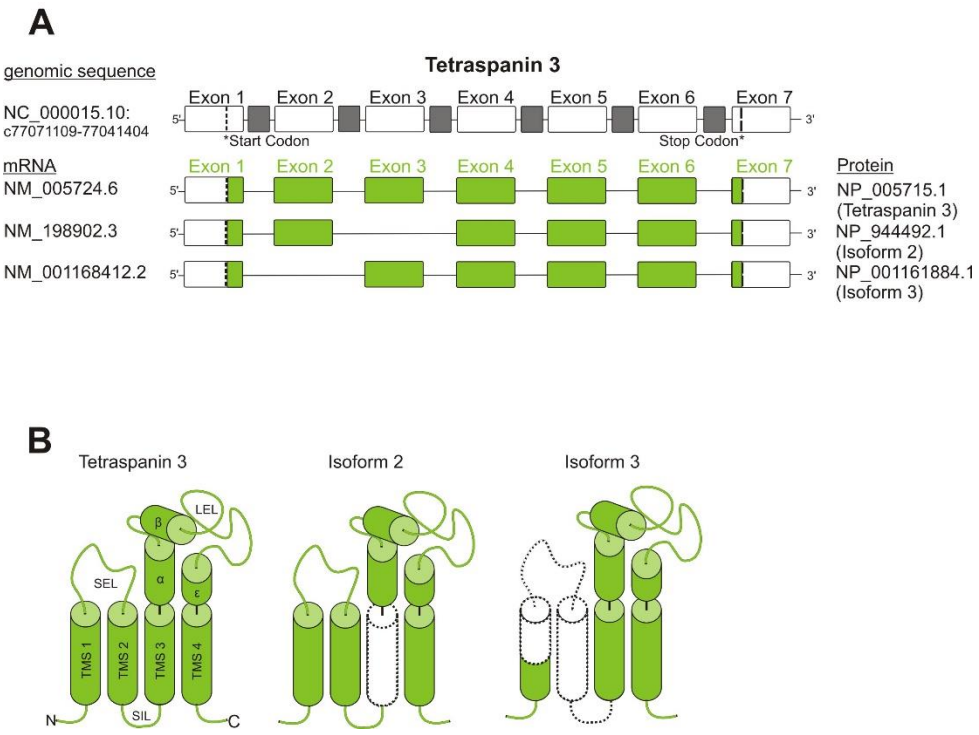


Figure S3: Tspan3 isoforms.

For scheme explanation, please see legend to figure 2.

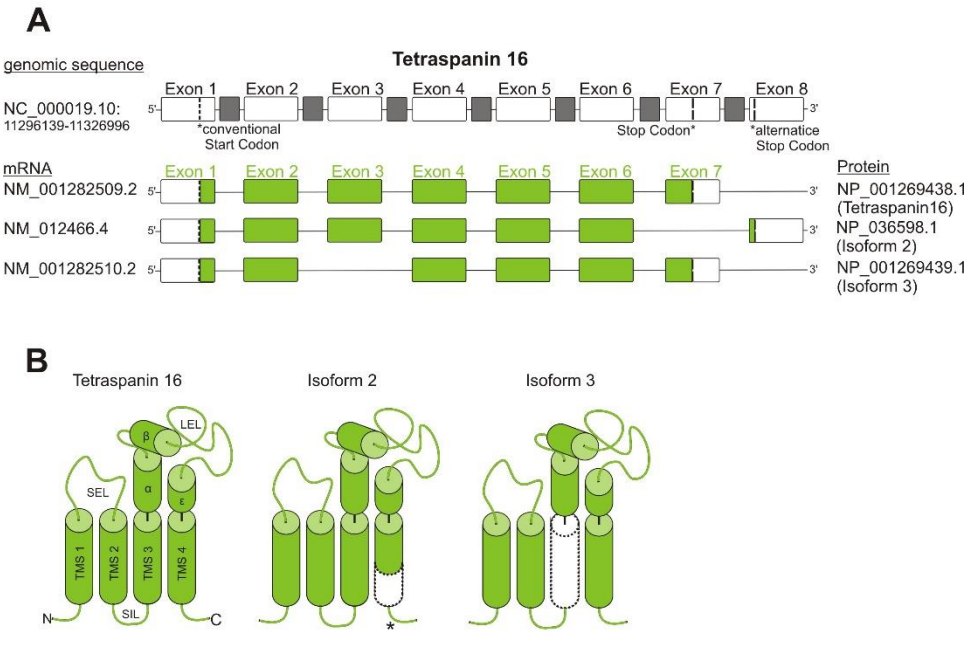


Figure S4: Tspan16 isoforms.

For scheme explanation, please see legend to figure 2. Asterisk indicates alternative C-terminus.

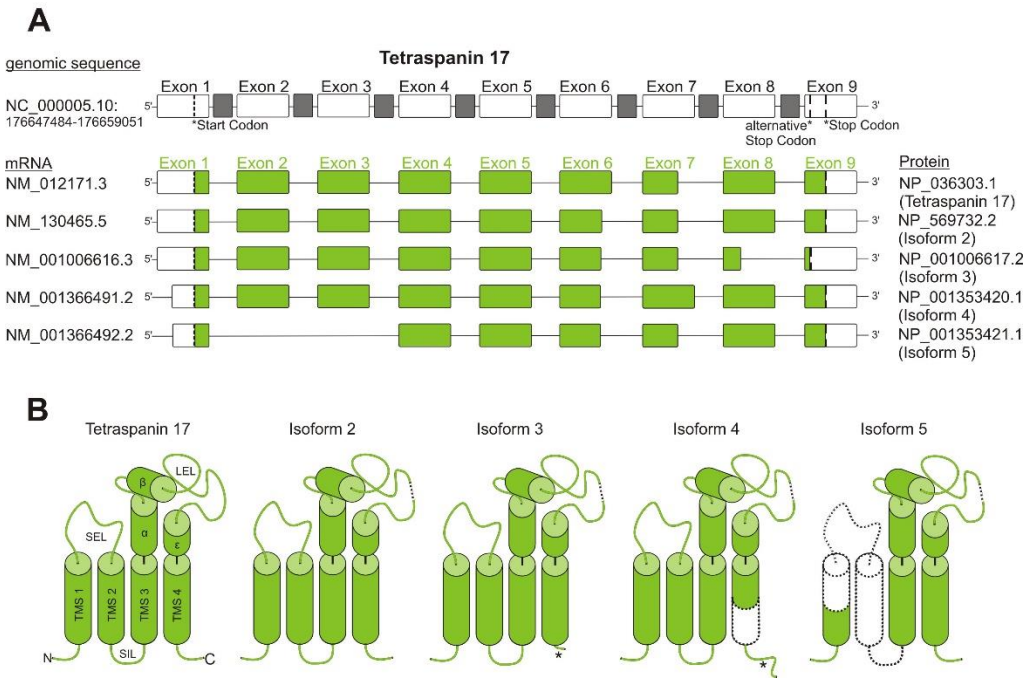


Figure S5: Tspan17 Isoforms.

For scheme explanation, please see legend to figure 2. Asterisk indicates alternative C-terminus.

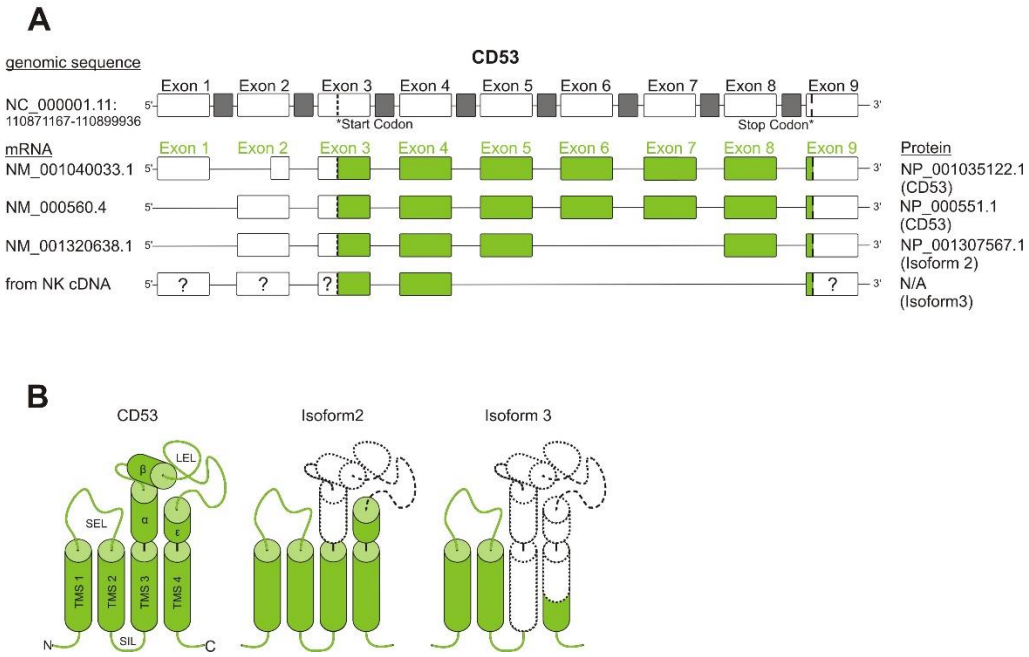


Figure S6: CD53 isoforms.

For scheme explanation, please see legend to figure 2. Please note that for isoform 3 the sequence of the untranslated regions is unknown.

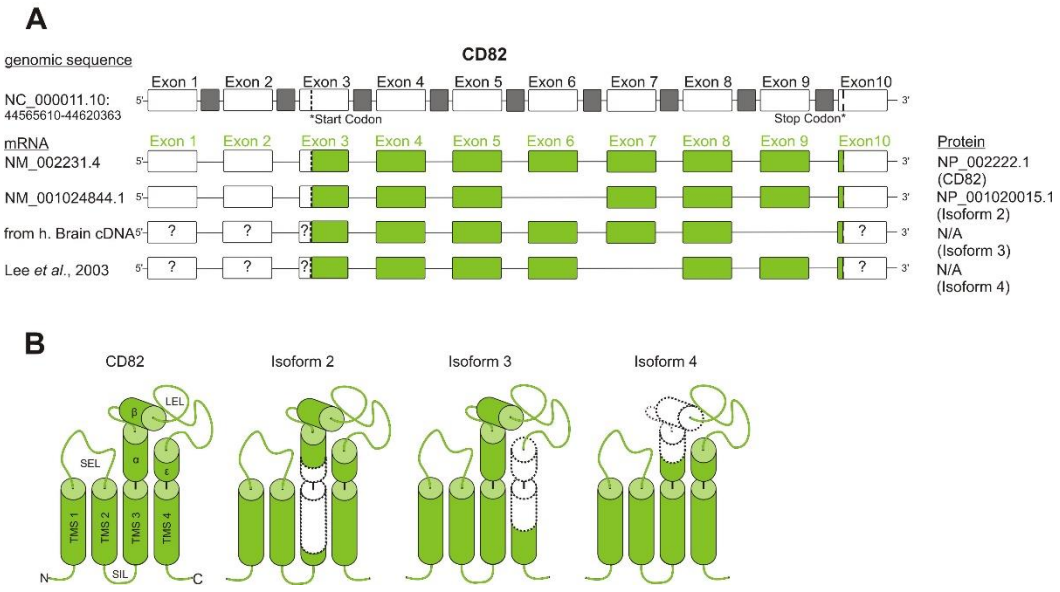


Figure S7: CD82 Isoforms.

For scheme explanation, please see legend to figure 2. Please note that for isoform 3 and 4 the sequence of the untranslated regions is unknown.

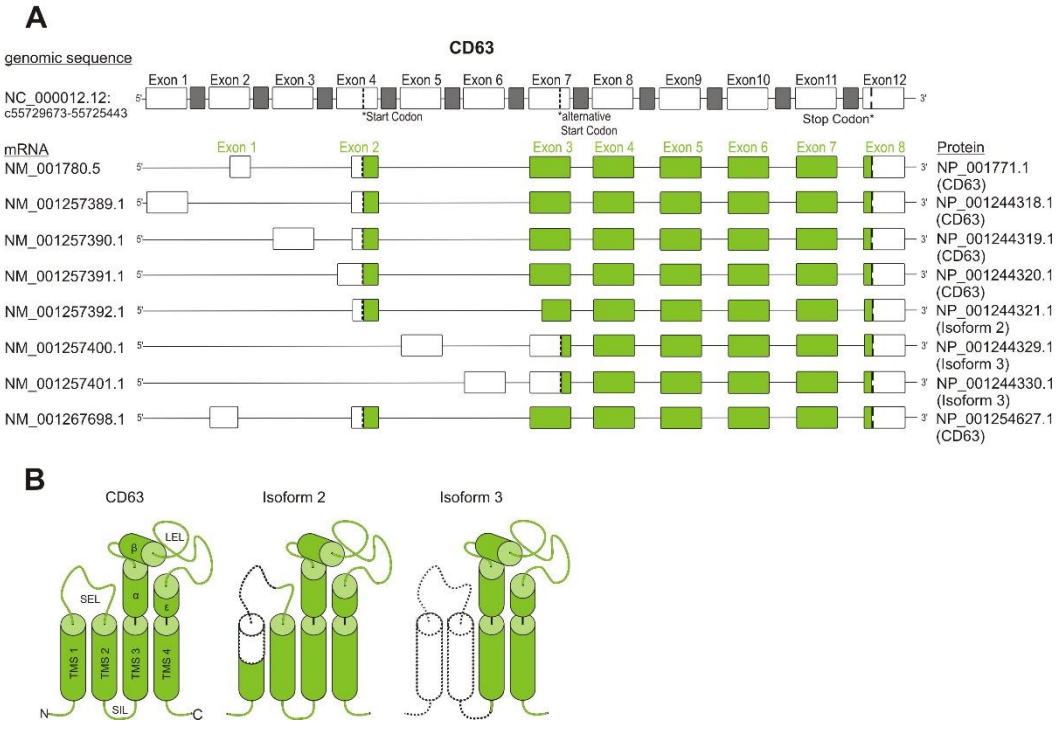


Figure S8: CD63 Isoforms.

For scheme explanation, please see legend to figure 2.

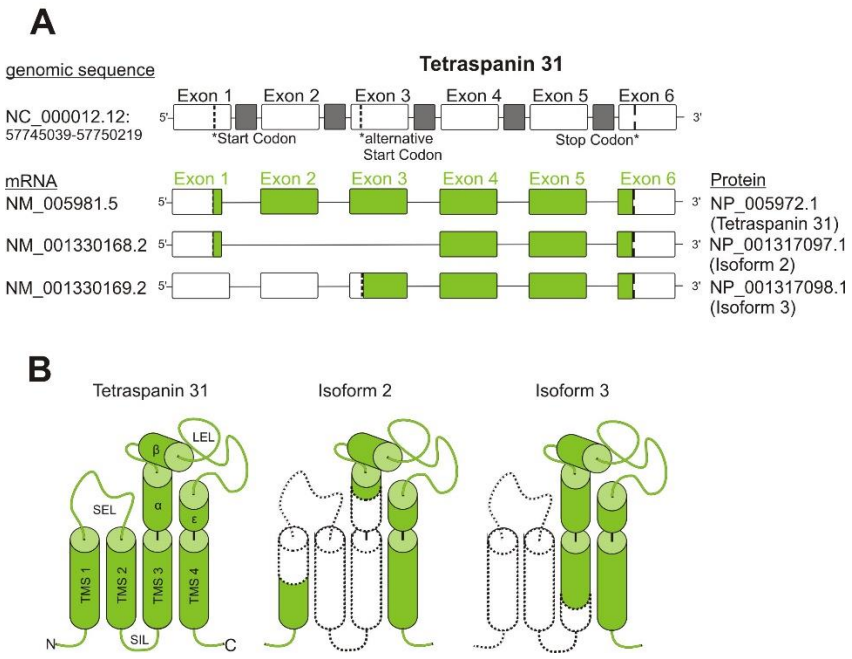


Figure S9: Tspan31 Isoforms.

For scheme explanation, please see legend to figure 2.

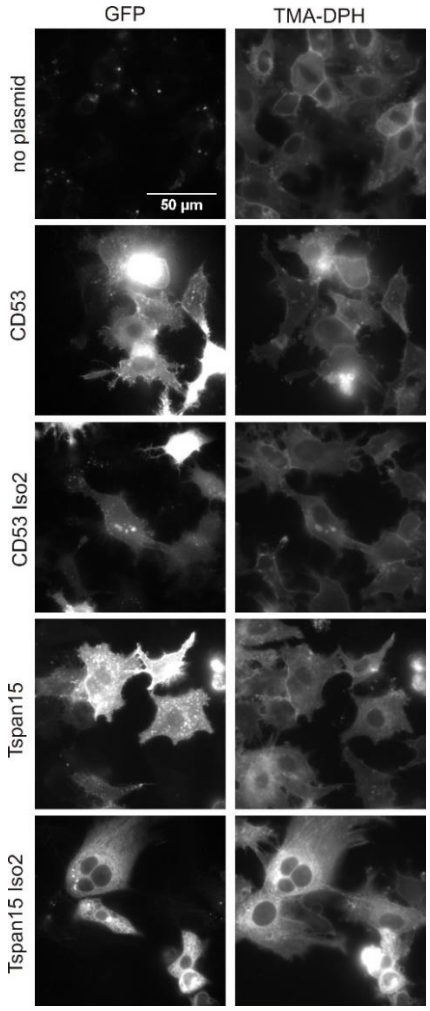


Figure S10: Intensity of GFP signal in comparison to untransfected cells.

HepG2 cells were transfected as described in the methods section without plasmid or plasmids coding for the indicated GFP-labelled tetraspanins. After 22 h cells were visualized by the membrane dye TMA-DPH and imaged in the green (left, GFP signal) and the blue (right, TMA-DPH fluorescence) channel. For evaluation of the GFP intensity, images in the green channel are scaled identically. The blue channel indicates the presence of cells. Comparison to the green channel indicates that not all cells are expressing the respective GFP-construct.

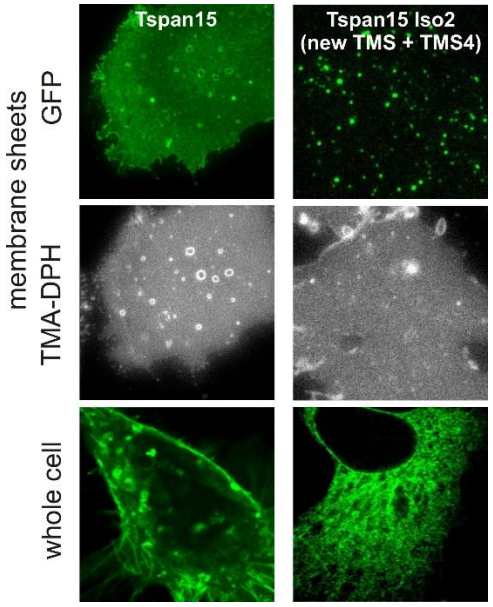


Figure S11: Cellular distribution of Tspan15.
Extension of Figure 6, illustrating a HepG2 cell expressing GFP-labelled conventional Tspan15 and a membrane sheet generated from a HepG2 cell expressing Tspan15-GFP. For comparison to Tspan15 Iso2, again the respective panels from Figure 6 are shown. For details, please see legend of Figure 6.

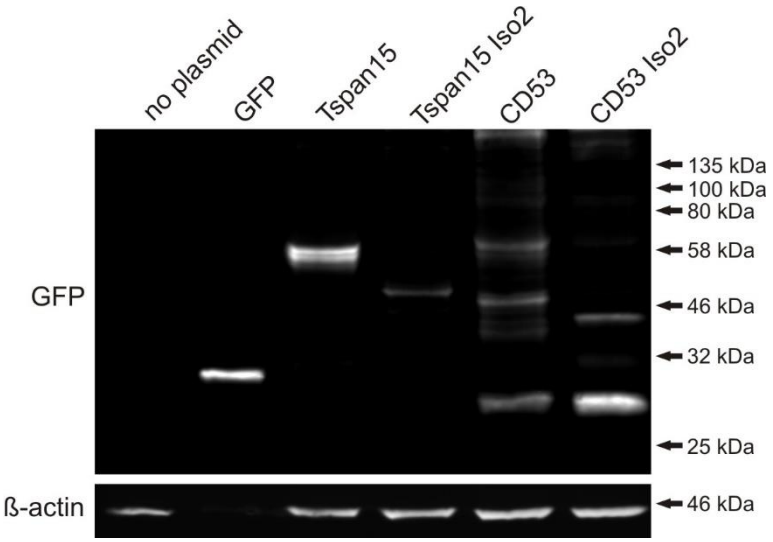


Figure S12. Expression of GFP-labeled tetraspanins
Top, HepG2 cells were transfected with no plasmid, a plasmid for GFP expression (≈ 27 kDa), or for expression of GFP-labelled conventional Tspan15 (≈ 60 kDa), Tspan15 Iso2 (≈ 51 kDa), conventional CD53 (≈ 51 kDa) or CD53 Iso2 (≈ 45 kDa). In case of GFP expression, the amount of loaded material was only 1/8 as otherwise the GFP signal in the Western Blot would be too strong. Please note that the conventional forms appear as double bands, which is due to glycosylation. In case of CD53, glycosylation causes a strong shift > 10 kDa toward higher apparent molecular mass, already described in the literature[1,2]. In contrast, CD53 Iso2 appears as a single band, as the glycosylation sites are deleted. CD53/CD53 Iso2 degradation products are of lower mass than GFP, indicating that they are not fluorescent and consequently not visualized in fluorescence microscopy. See Figure S14 for the full-length blot. Bottom, on the same membrane we immunoblotted for actin as a loading standard, visualized in a different channel. See Figure S15 for the full-length blots.

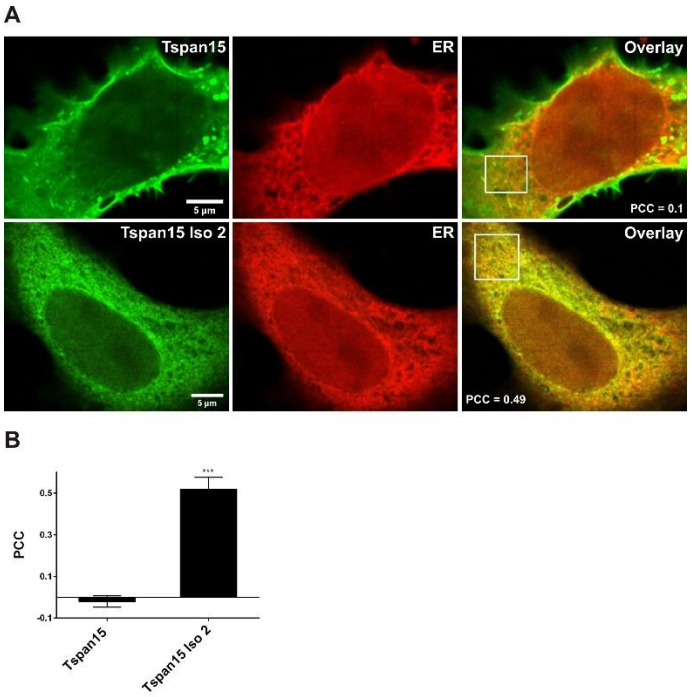


Figure S13. Colocalization of Tspan15 and Tspan 15 Iso2 with the ER.
(A) Confocal micrographs showing equatorial sections of HepG2 cells expressing GFP-labelled conventional Tspan15 or Tspan15 Iso2 (shown in the green channel) and the ER marker KDEL-RFP (red channel). White boxes mark regions of interest (ROIs), in which colocalization between the two channels was measured by calculation of the Pearson correlation coefficient (PCC). For the shown cells, PCC values of 0.1 and 0.49 were obtained for Tspan15/ER and Tspan15 Iso2/ER, respectively (B) Average Pearson correlation coefficient. Values are given as means ± SEM (n = 4 biological replicates, for each replicate the values from 5 – 20 cells were averaged). Unpaired student's t-test (***, $p < 0.001$).

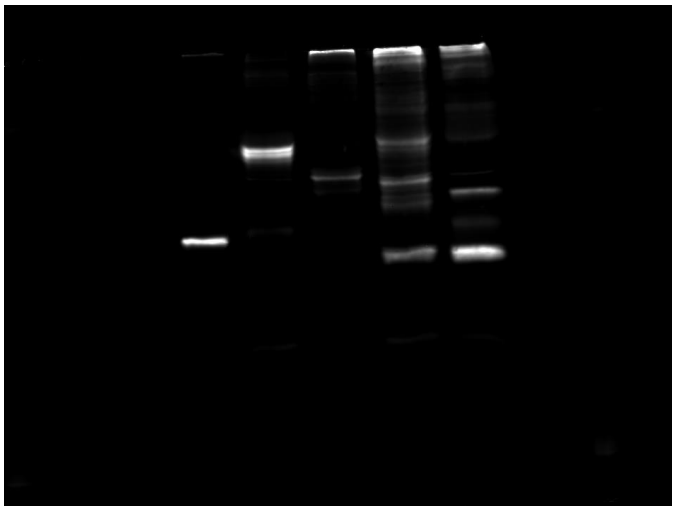


Figure S14. Full-length blot stained for GFP (used for Figure S12).

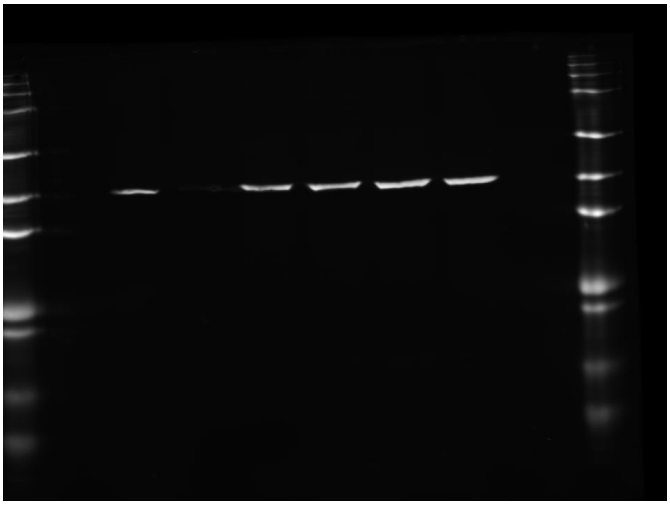


Figure S15. Full-length blot stained for actin (used for Figure S12).

References

1. Mollinedo, F., Martín-Martín, B., Gajate, C. & Lazo, P. A. Physiological activation of human neutrophils down regulates CD53 cell surface antigen. *J. Leukoc. Biol.* **63**, 699–706 (1998).

2. Angelisová, P., Vlček, Č., Štefanová, I., Lipoldová, M. & Hořejší, V. The human leucocyte surface antigen CD53 is a protein structurally similar to the CD37 and MRC OX-44 antigens. *Immunogenetics* **32**, 281–285 (1990).

Appendix C: A conserved sequence in the small intracellular loop of tetraspanins forms an M-shaped inter-helix turn

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

A conserved sequence in the small intracellular loop of tetraspanins forms an M-shaped inter-helix turn

Nikolas Reppert✉ & Thorsten Lang✉

Tetraspanins are a family of small proteins with four transmembrane segments (TMSs) playing multiple roles in human physiology. Nevertheless, we know little about the factors determining their structure. In the study at hand, we focus on the small intracellular loop (SIL) between TMS2 and TMS3. There we have identified a conserved five amino acid core region with three charged residues forming an M-shaped backbone, which we call M-motif. The M´s plane runs parallel to the membrane surface and the central amino acid constitutes the inter-helix turning point. At the second position of the M-motif, in tetraspanin crystal structures we identified a glutamate oriented towards a lysine in the juxtamembrane region of TMS1. Using Tspan17 as example, we find that by mutating either the glutamate or juxtamembrane-lysine, but not upon glutamate/lysine swapping, expression level, maturation and ER-exit are reduced. We conclude that the SIL is more than a short linking segment but propose it is involved in shaping the tertiary structure of tetraspanins.

Tetraspanins comprise a family of small membrane proteins expressed in all multicellular organisms. The human genome encodes 33 family members[1]. They are involved in many cellular processes of either physiological or pathological nature, including adhesion, cell–cell fusion, endocytosis, exosome formation, immune response, migration, neurite navigation, pericellular proteolysis, proliferation, signalling, spreading, trafficking, vascular morphogenesis and remodelling, thrombosis, tumor progression and metastasis, viral and other pathogen entry, and viral release[1]. The basis for this broad range of functions is their capability to form so-called tetraspanin-enriched microdomains (TEMs)[2], sometimes also referred to as tetraspanin web[3]. The underlying mechanisms include weaker secondary interactions among themselves and stronger primary interactions[4] with a variety of non-tetraspanins, for instance integrins, members of the immunoglobulin superfamily, and signalling receptors[3,5,6].

Until recently, studies have concentrated on members that locate to the plasma membrane, a characteristic that has led to their nick name 'master organizers of the plasma membrane'[7]. However, lately, their intracellular roles have become more and more obvious. Especially their function in extracellular vesicle formation and targeting[8] is shifting into focus.

Tetraspanins are membrane anchored via a bundle of four transmembrane segments (TMSs). They share a conserved topology (Fig. 1A) comprised of a small and a large extracellular loop (LEL) connecting extracellularly the first and last TMS pairs, respectively. The LEL, which is often glycosylated, contains up to five helical segments (A–E) and up to four disulphide bridges[9]. With few exceptions, the intracellular N- and C-termini, as well as the small intracellular loop (SIL) connecting TMS2 and TMS3, are short segments[1,10].

For their tertiary structure, two different models are envisioned. On the one hand, using the crystal structure of the LEL of CD81 as starting point, an early model predicts a tight four transmembrane helix bundle (similar to Fig. 1B, left). The alpha helical structure of TMS3 and TMS4 protrudes and merges with the alpha-helical segments A and E of the LEL, respectively. As a result, the more bulky extracellular domain sits enthroned on top of the bundle[13]; the entire structure resembles a mushroom. In cryo-EM, a closely related open conformation is found for CD81 (Fig. 1B, left) and CD9 in complex with their primary binding partner CD19[14] and EWI-F[15], respectively. On the other hand, all known crystal structures of complete tetraspanins (CD81, CD9 and CD53)[12,16,17] reveal a funnel shaped arrangement of the TMSs opening towards the extracellular site (Fig. 1B, right), with a cholesterol bound inside the cavity of CD81[12]. It is important to note that here the alpha-helical connections between TMS3/4 and the LEL helices are disrupted by a kink, causing the LEL to fold-back onto the membrane, thereby closing the cholesterol cavity. Removal of the cavity-bound cholesterol in molecular dynamics simulation, the connecting segments become more alpha-helical and the LEL unbends, similar to the early

Department of Membrane Biochemistry, Life and Medical Sciences (LIMES) Institute, University of Bonn, Carl-Troll-Straße 31, 53115 Bonn, Germany. ✉email: nikolas.reppert@uni-bonn.de; thorsten.lang@uni-bonn.de

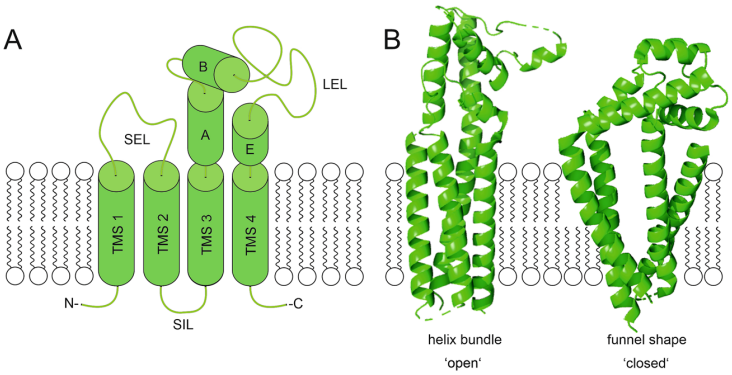


Figure 1. Tetraspanin topology and open-/closed-conformation. Tetraspanin embedded into a membrane bilayer; top and bottom correspond to the extracellular and intracellular side, respectively. (**A**) Schematic depiction of the tetraspanin's topology including four transmembrane segments (TMS1–4), a small extracellular loop (SEL) and a large extracellular loop (LEL, shown with the conserved helical domains A,B,E), a small intracellular loop (SIL) and a N and C terminus. This illustration was modified from Hochheimer et al.[11]. (**B**) Different conformations proposed for CD81 referred to as helix bundle/'open' (left, based on PDB: 7JIC) and funnel shape/'closed' (right, basd on PDB: 5TCX) that differ in the arrangement of the TMSs and the orientation of the LEL; the right structure represents the closed conformation as defined by Zimmerman et al.[12]. The scheme was created using CorelDRAW 2019 (www.corel.com) and the cryo-EM and crystal structure images were created using PyMOL 2.5 (https://pymol.org/2/).

model, but the funnel-shaped TMSs arrangement persists[12]. Altogether, the combined structural data result in a model of a conformational switch between an open- and closed-state (Fig. 1B), which modulates the potential interactions and functions of tetraspanins[12,18]. The impact of the switch lies in the reorientation of the LEL that plays a pivotal role in primary interactions[14,19–21], but the TMSs and intracellular parts are also crucial[16,22–25]. Additionally, a hypothesis claims that the funnel shaped TMS bundle influences membrane curvature generation and/or could locate the protein towards high curvature membrane regions[16].

The SIL is the smallest segment in human tetraspanins[11] and may only be a linking stretch between TMS2 and TMS3, with any arbitrary sequence. However, a study found a conservation of 33% between four analysed human tetraspanins and a high conservation of 78% between these human tetraspanins with their homologs in zebrafish[26]. Moreover, Mazurov et al. showed the SIL to be decisive for the interaction of human T cell lymphotrophic virus type 1 gag protein with CD81 and CD82[27]. Apart from that little is known, which makes the SIL a poorly understood tetraspanin section. In this study, we examine whether the SIL is more than just a connecting segment, and identify a conserved SIL core sequence between TMS2 and TMS3. The sequence adopts an M-shape, harbouring at its second position a glutamate. In Tspan17, the glutamate seems to interact electrostatically with a lysine in the juxtamembrane region of TMS1, and by this regulates Tspan17 glycosylation and ER-exit.

Material and methods

Cloning of constructs. The tetraspanin sequences of Tspan15 (NM_012339.5), Tspan17 (NM_130465.5), CD53 (NM_001040033.1), CD37 (NM_001774.3), CD82 (NM_002231.4), CD81 (NM_004356.3), CD9 (NM_001769.4) and CD63 (NM_001780.5) were GFP-tagged by cloning the sequences into the pEGFP-C1 vector (Clontech, #6084-1) as described before[11]. EWI-2 (NM_001206665.2) was cloned using the pEGFP-C1 vector without the GFP sequence employing the NEbuilder HiFi DNA Assembly Cloning Kit (NEB, #E5520S). The Myc-tag was fused to EWI-2 at the C-terminus without a linker region. The tag was attached employing two forward primers harbouring the myc tag and a reverse primer complementary to the EWI-2 C-terminus. The used primers are listed below. Tetraspanins were C-terminally fused to monomeric enhanced GFP (mEGFP[11]) with the exception of Tspan17 that is better expressed when N-terminally fused to GFP. For the generation of mutants, we employed Q5 High-Fidelity DNA Polymerase (NEB, # M0491S) and back-to-back primers carrying the desired mutations. The following primers were used (the mutated nucleotides are lowercase): Tspan17-K18A: 5′-gcATACTTCCTGTTTGGCTTCAAC-3′ (fwd) and 5′-CCCGCAGCAGCCGACCTCAG-3′ (rev); Tspan17-K18E, E88K: 5′-gAATACTTCCTGTTTGGCTTCAAC-3′ (fwd) and 5′-CCCGCAGCAGCCGACCTCAG-3′ (rev) then 5′-aaGAACACCTTCCTGCTCAAG-3′ (fwd) and 5′-CCGGAGGGCCCCAATGCAGC-3′ (rev); Tspan17-R87A: 5′-gcGGAGAACACCTTCCTGCTCAAGTTTTTCTCCG-3′ (fwd) and 5′-GAGGGCCCCAATGCAGCCAGC-3′ (rev); Tspan17-E88A: 5′-GcGAACACCTTCCTGCTCAAG-3′ (fwd) and 5′-CCGGAGGGCCCCAATGCAGC-3′ (rev) Tspan17-N89A: 5′-CGGGAGgcCACCTTCCTGCTCAAGTTTTTCTCCG-3′ (fwd) and 5′-GAGGGCCCCAATGCAGCCAGC-3′ (rev); Tspan17-T90A: 5′-CGGGAGAACgcCTTCCTGCTCAAGTTTTTCTCCG-3′ (fwd) and 5′-GAGGGCCCCAATGCAGCCAGC-3′ (rev); Tspan17-F91A: 5′-CGGGAGAACACCgcCCTGCTCAAGTTTTTCTCCG-3′ (fwd) and 5′-GAGGGCCCCAATGCAGCCAGC-3′ (rev); CD9-K11A: 5′-gcATACCTGCTGTTCGGATTTAAC-3′ (fwd) and 5′-GATGCACTTGGTGCC

TCCTTTG-3′ (rev); CD9-K11E, E84K: 5′-gAgTACCTGCTGTTCGGATTTAAC-3′ (fwd) and 5′-GATGCACTT GGTGCCTCCTTTG-3′ (rev) then 5′-aAGTCCCAGTGCATGCTGGGAC-3′ (fwd) and 5′-CTGCACAGCCCC GCAGCAGC-3′ (rev); CD9-Q83A: 5′-GGGGCTGTGgcGGAGTCCCAGTGCATGC-3′ (fwd) and 5′-GCAGCA GCCCAGGAAGCCCACCAGCATC-3′ (rev); CD9-E84A: 5′-GcGTCCCAGTGCATGCTGGGAC-3′ (fwd) and 5′-CTGCACAGCCCCGCAGCAGC-3′ (rev); CD9-S85A: 5′-ATGCTGGGACTGTTCTTCGGCTTCCTCTTG G-3′ (fwd) and 5′-GCACTGGgcCTCCTGCACAGCCCCGCAGC-3′ (rev); CD9-Q86A: 5′-GAGTCCgcGTG CATGCTGGGACTGTTCTTCGGC-3′ (fwd) and 5′-CTGCACAGCCCCGCAGCAGC-3′ (rev); CD9-C87A: 5′-GAGTCCCAGgcCATGCTGGGACTGTTCTTCGGC-3′ (fwd) and 5′-CTGCACAGCCCCGCAGCAGC-3′ (rev); CD53-K7A: 5′-gcACTGCTGAAGTATGTCCTG-3′ (fwd) and 5′-CAAGCTACTCATGCCCATGC-3′ (rev); CD53-K7E, E77K: 5′-gAACTGCTGAAGTATGTCCTG-3′ (fwd) and 5′-CAAGCTACTCATGCCCAT GC-3′ (rev) than 5′-AAGaAAACAAGTGTCTGCTTATGTCGTTCTTC-3′ (fwd) and 5′-GATAGAGCCCAT GCAGCCCAGGAAGGC-3′ (rev); CD53-K10A: 5′-gcGTATGTCCTGTTTTTCTTCAACTTGC-3′ (fwd) and 5′-CAGCAGTTTCAAGCTACTCATGCC-3′ (rev); CD53-K10E, E77K: 5′-gAGTATGTCCTGTTTTTCTTCA ACTTGC-3′ (fwd) and 5′-CAGCAGTTTCAAGCTACTCATGCC-3′ (rev) then 5′-AAGaAAACAAGTGTCT GCTTATGTCGTTCTTC-3′ (fwd) and 5′-GATAGAGCCCATGCAGCCCAGGAAGGC-3′ (rev); CD53-K76A: 5′-gcGGAAAACAAGTGTCTGCTTATGTCGTTCTTC-3′ (fwd) and 5′-GATAGAGCCCATGCAGCCCAG GAAGGC-3′ (rev); CD53-E77A: 5′-AAGGcAAACAAGTGTCTGCTTATGTCGTTCTTC-3′ (fwd) and 5′-GAT AGAGCCCATGCAGCCCAGGAAGGC-3′ (rev); CD53-N78A: 5′-gcCAAGTGTCTGCTTATGTCG-3′ (fwd) and 5′-TTCCTTGATAGAGCCCATGC-3′ (rev); CD53-K79A: 5′-AAGGAAAACgcGTGTCTGCTTATGTC GTTCTTC-3′ (fwd) and 5′-GATAGAGCCCATGCAGCCCAGGAAGGC-3′ (rev); CD53-C80A: 5′-AAGGAA AACAAGgcTCTGCTTATGTCGTTCTTC-3′ (fwd) and 5′-GATAGAGCCCATGCAGCCCAGGAAGGC-3′ (rev); Tspan15-D87A: 5′-GcCAACCTGTACCTTCTCCAAGC-3′ (fwd) and 5′-ACGGAGGGACGCCAGCAC ACCAATG-3′ (rev); CD37-E82A: 5′-GcGCTCCGCTGCCTCCTGGGCC-3′ (fwd) and 5′-CTTGAGGGCCCC CACACAACCC-3′ (rev); CD82-E80A: 5′-GcGGTCCGCTGCCTGCTGGGGC-3′ (fwd) and 5′-GTTGACGGC GCCGATGCAGCCC-3′ (rev); CD81-E86A: 5′-cATCCCAGTGCCTGCTGGGGACGTTC-3′ (fwd) and 5′-CCT GGATGGCCCCGTAGCAGCCCAGG-3′ (rev); CD63-E78A: 5′-GcGAACTATTGTCTTATGATCAC-3′ (fwd) and 5′-CTTGCAGGCCCCGCAGCAGCC-3′ (rev); EWI-2 C-terminal Myc-tag: 5′-TCTGAAGAAGATCTG-taaagcggccgcgactctag-3′ (fwd), 5′-GAACAAAAACTTATTTCTGAAGAAGATCTGtaaag-3′ (fwd) and 5′-CCG TTTTCGAAGCCTCTTCATGAAGCAGCAAGTG-3′ (rev). The PCR products were run on a 1% agarose (Carl Roth, #2267.4) gel in TAE (40 mM Tris, 1 mM Na_2EDTA, 0.1% acetic acid, pH 8.3). The wanted bands were cut out and cleaned up using the Monarch DNA Gel Extraction Kit (NEB, #T1020L). The linear plasmid was phosphorylated using T4 Polynucleotide kinase (NEB, #M0201S) and ligated by T4 DNA Ligase (NEB, #M0202S). The protein coding sequence was verified by sequencing (Eurofins GATC Biotech GmbH).

Alignment of amino acid sequences and obtaining the consensus sequences. All alignments were done with BioEdit[28] v7.0.5 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). The tetraspanins' SIL sequences were aligned with the most N-terminally glutamate or aspartate and if there was none the sequences were aligned at the most N-terminally lysine or arginine. The sequences without a glutamate, aspartate or lysine and arginine were aligned due to overall similarity to the other sequences. The tetraspanins' N-terminal sequences with the first five amino acids of the TMS1 were aligned at their most C-terminal lysine or arginine.

The consensus sequences of SIL and N-terminus were obtained by counting the frequency of each amino acid at the given position. An amino acid was counted as consensus, if its frequency (f) was equal to or higher than the mean frequency (f_{mean}) plus its standard deviation ($f \geq f_{mean} + SD$).

For the Claudin SIL sequence alignment we used the ClustelW multi alignment tool of BioEdit.

Structural predictions and depiction. The TMSs of each tetraspanin[11] were defined using TMHMM Server v. 2.0[29,30] (http://www.cbs.dtu.dk/services/TMHMM/) and their domains such as N-terminus and SIL were defined as N-terminally of TMS1 and the linker between TMS2 and TMS3, respectively. The secondary structures were analysed by Jpred4[31] (http://www.compbio.dundee.ac.uk/jpred/). The helices containing TMS1 of each human Tspan were analysed by HeliQuest[32] (https://heliquest.ipmc.cnrs.fr/cgi-bin/ComputParams.py). The images of crystal and cryo-EM structures as well as the dihedral angles were obtained using PyMOL 2.5 (https://pymol.org/2/). The orientation of the SIL towards the membrane was adopted from the Orientations of Proteins in Membranes (OPM) database (https://opm.phar.umich.edu/). The illustration of amino acid frequency was done via Weblogo[33] 3.7.4 (https://weblogo.berkeley.edu/logo.cgi). The data composition for the figures was done employing CorelDRAW 2019 (www.corel.com).

Confocal microscopy. HepG2 cells (ATCC, #HB-8065) were grown in MEM Eagle (Pan Biotech, # P04-08509) with 10% FBS (Pan Biotech, # P303031), 1% penicillin–streptomycin (Pan Biotech, #P06-07050) and 1% stabile glutamine (Pan Biotech, # P04-82100) up to 80% confluence. The cells were detached using trypsin (Pan Biotech, #P10-0231SP) and diluted in DPBS (Gibco, #14190-094) to 14.4 million cells per ml. The cells were transiently transfected with the above described vectors using the Neon transfection system (Thermo Fisher Scientific, #MPK10096) with settings 1200 V, 50 ms, 1 puls. The cells were stained, imaged and analysed essentially as described before[11]. Briefly, the cells ER was visualized by a KDEL-RFP fusion construct (BacMam 2.0, Life Technologies, #C10591) and the signal was enhanced using a RFP-Booster Atto594 (Chromotek, #rba594). Cells were imaged in the confocal mode using a 4-channel easy3D superresolution STED optics module (Abberior Instruments) coupled to an Olympus IX83 confocal microscope (Olympus, Tokyo, Japan), equipped with an UPlanSApo 100x (1.4 NA) objective (Olympus, Tokyo, Japan). GFP was excited with a 485 nm laser and recorded with a 500–520 nm filter. Atto594 was excited with a 561 nm laser and recorded with a 580–630 nm

filter. The pixel size was set to 25 nm. The Pearson correlation coefficient (PCC) between the protein of interest and the ER was calculated with a custom made macro using Fiji-ImageJ[34] (www. https://imagej.net/).

Tunicamycin assay. HepG2 cells were treated 4 h after transient transfection with tunicamycin (Sigma-Aldrich, #T7765) up to a final concentration of 5 µg/ml in MEM Eagle with 10% FBS. After 18 h of incubation, cells were scraped off and washed with ice cold DPBS and lysed in 200 µl RIPA buffer (Santa Cruz, #sc-24948). After the cell debris was removed by centrifugation for 20 min at 16,000 g the lysate was mixed with 4 × Lämmli buffer with beta-mercaptoethanol (250 mM Tris HCl, 8% (w/v) SDS, 40% glycerol, 20% beta-mercaptoethanol, bromophenol blue, pH 6.8). The samples were heated for 30 min at 37 °C and subsequently analysed via western blot.

Co-immunoprecipitation assay. HepG2 cells were transiently transfected with equal amounts of vector DNA (6 µg of each plasmid per transfection) of the CD9-mEGFP variants and the EWI-2-myc construct. For each condition, the cells of two transfections (equals to 3.6 million cells) were pooled and seeded. After 22 h of incubation, cells were scraped off in ice cold DPBS, lysed in lysis buffer (150 mM NaCl, 5 mM $MgCl_2$, 25 mM HEPES containing 1% CHAPS (Sigma-Aldrich, #C5070-5G)) and subjected to co-immunoprecipitation with GFP-Trap beads (Chromotek, #gta-20) following the manufacturers protocol (1 h at 4 °C incubation with beads, protein was dissociated from the beads by boiling in 2 × Laemmli + 20% β-mercaptoethanol). The Co-IP samples were analysed by Western blot.

Western blot. The cells were scraped off in ice cold DPBS and lysed 30 min in 200 µl RIPA buffer (Santa Cruz, #sc-24948). The SIL core and lysine mutations of Tspan17 and CD9 were lysed directly in 250 µl 2.5 × Lämmli buffer (156 mM Tris HCl, 5% (w/v) SDS, 25% glycerol, pH 6.8). After a 20 min spin down of the insoluble cell debris, the lysate was mixed with 4 × Lämmli buffer with or without 20% β-mercaptoethanol and incubated for 30 min at 37 °C (CD53) or 5 min at 95 °C (all others). The proteins were separated by a 10% SDS PAGE and blotted onto a nitrocellulose membrane (BIO-RAD, #1620112) in ice cold towbin buffer (25 mM Tris HCl, 192 mM glycine, 20% methanol, pH 8.3). The membrane was blocked using Intercept blocking buffer (Li-Cor, #210218) and incubated with primary antibodies in blocking buffer with added 0.05% Tween-20 (Carl Roth, # 9127.1) over night at 4 °C. The used primary antibodies were rabbit polyclonal anti-GFP (Thermo Fisher Scientific, #A-11122) diluted 1:2000, mouse monoclonal anti-beta-actin (Cell Signalling, #3700) diluted 1:10,000, goat anti-EWI-2 (R&D systems, #A3117) diluted 1:500 and mouse anti-CD9 (Merck, #CBL162) diluted 1:1000. The secondary antibodies donkey anti-mouse coupled to IRDye 680RD (Li-Cor, #926-68072) and donkey anti-rabbit coupled to IRDye 800CW (Li-Cor, #926-32213) were diluted 1:10,000 in blocking buffer with added 0.05% Tween 20. For the Co-IP the secondary antibodies donkey anti-goat coupled to IRDye 800CW (Li-Cor, #926-32214) and donkey anti-rabbit coupled to IRDye 680RD (Li-Cor, #926-68073) or donkey anti-rabbit coupled to IRDye 800CW (Li-Cor, #926-32213) were used. The blots were imaged using an Odyssey Classic Imaging System (Li-Cor) and the images were analysed using Fiji-ImageJ[34].

Statistical analysis. All experiments were performed at least three times independently. Microscopy data were averaged per day. Data were analysed using a two tailed and paired t-test or a repeated measures ANOVA. The analysis and illustration of data was performed using GraphPad Prism version 6.04 for Windows (www.graphpad.com). Results are expressed as mean ± standard deviation. Statistical significance was determined (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$).

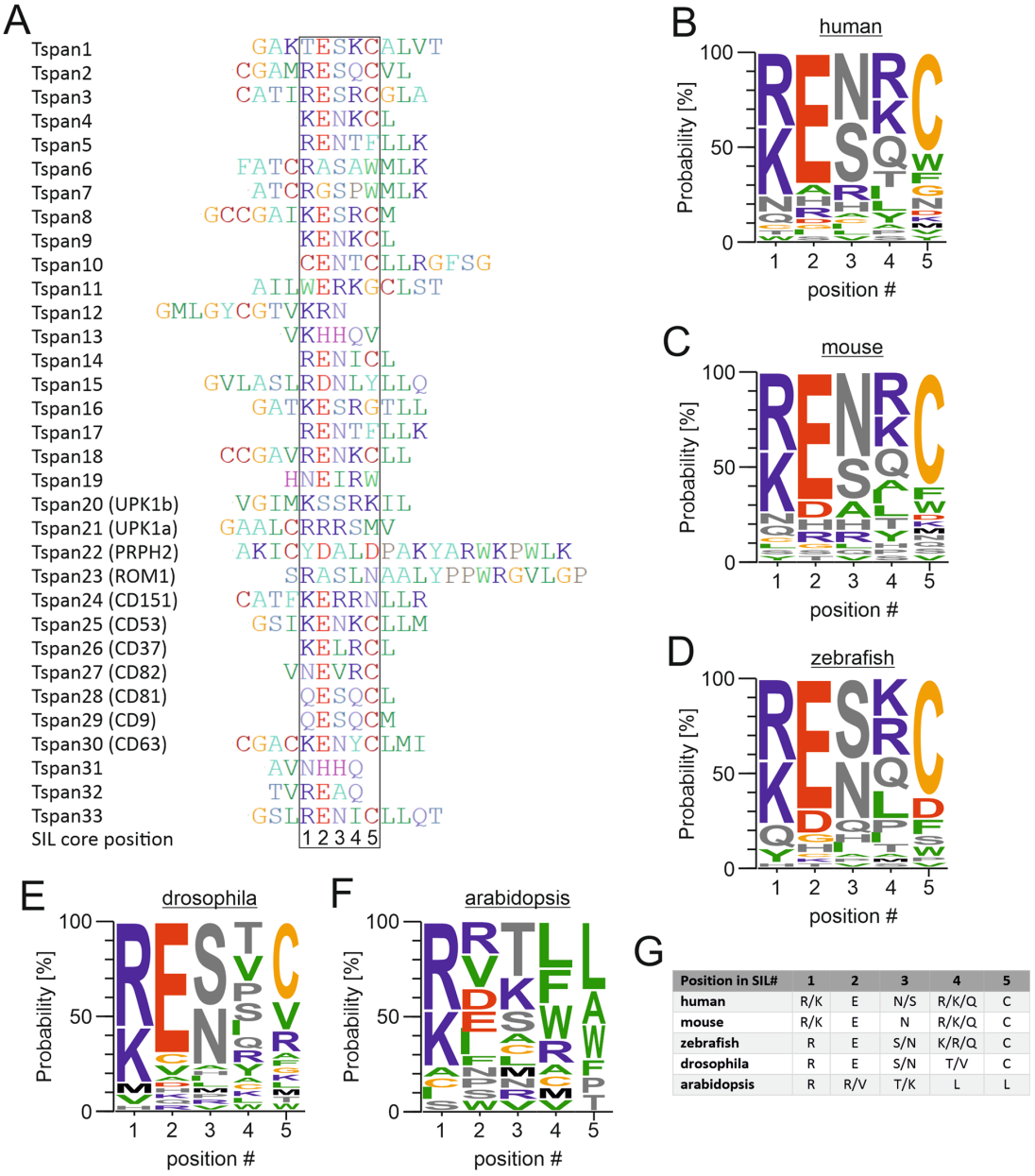
Results

Definition of the SIL core sequence. In human Tetraspanins, the shortest and longest SIL sequences comprise six and 21 amino acids, respectively[11]. Comparing these sequences, we frequently find a positively charged amino acid directly followed by a negatively charged one, to which we assigned 'position 1' and 'position 2' of the core sequence (Fig. 2A). We started the alignment of the 33 tetraspanins with position 2, assigning to it the most upstream glutamate (in 23 SILs) or aspartate (2 SILs). In eight SILs there is neither a glutamate nor an aspartate present, which is why we used position 1 for further alignment, finding altogether 24 matches (13 arginines and 11 lysines). Finally, we searched for frequently occurring amino acids in positions 3–5 (Fig. 2A; for details see figure legend). As a result, we identified a [R/K] E [N/S] [R/K/Q] C core sequence (Fig. 2B). With the exception of glutamine at positon 4 that is a polar amino acid, the chemical signature of the sequence is positive charge—negative charge—polar—positive charge—polar.

To compare the human SIL with other species we determined the SIL consensus sequences, as it was done for human, from Uniprot database sequences in mouse (31 family members in *Mus musculus*[11]), zebrafish (50 in *Danio rerio*[35], but only 39 available in the Uniprot database), fruit fly (34 in *Drosophila melanogaster*[36]) and arabidopsis (17 in *Arabidopsis thaliana*[37]). We obtained the following sequences: [R/K] E N [R/K/Q] C (mouse), R E [S/N] [K/R/Q] C (zebrafish), R E [S/N] [T/V] C (fruit fly) and R [R/V] [T/K] L L (Arabidopsis) (Fig. 2C–F).

Hence, with the exception of plant tetraspanins (Fig. 2F), the degree of conservation between the SILs of all animal species is high (see also Fig. 2G). The question arose whether the same SIL core region is present in similarly structured proteins. Members of the claudin family have four TMSs and two extracellular loops[38] as well. However, they exhibit a longer and differently structured SIL between TMS2 and TMS3, harbouring a short beta-strand (Fig. S2).

Following, the secondary structure of the SIL core sequences was predicted employing Jpred4. In animal SILs, alpha-helicity gradually decreases from the edges to the central position (Fig. S3). Moreover, it should be noted that the extraordinarily long human SILs in Tspan22 (21 aa) and Tspan23 (19 aa) contain a non-helical stretch



G

Position in SIL#	1	2	3	4	5
human	R/K	E	N/S	R/K/Q	C
mouse	R/K	E	N	R/K/Q	C
zebrafish	R	E	S/N	K/R/Q	C
drosophila	R	E	S/N	T/V	C
arabidopsis	R	R/V	T/K	L	L

Figure 2. The SIL core sequence. (**A**) Alignment of the SILs from human tetraspanins (SILs segments are defined as in Hochheimer et al.[11]), as described in text, yields an amino acid core sequence with five positions (box). Tspan31 did not feature any of the alignment criteria and was aligned with reference to its similarity to Tspan13. In Tspan12, Tspan31 and Tspan32 some of the positions 4 and 5 are empty, as these amino acids are supposed to be integral part of TMS3. For the SIL alignments of other species see Fig. S1. (**B**–**F**) Residue probability in the SIL core of human (*Homo sapiens*), mouse (*Mus musculus*), zebrafish (*Danio rerio*) drosophila (*Drosophila melanogaster*) and Arabidopsis (*Arabidopsis thaliana*) tetraspanins. (**G**) SIL core consensus sequences. Part of the consensus sequence are amino acids occurring with a frequency f equal or higher than the mean frequency plus one time the frequencies standard deviation ($f \geq f_{mean} + SD$). In cases of more than one, they are listed in order of abundancy from high to low. The alignment was created using BioEdit[28] v7.0.5 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html) and the illustration of residue frequency was done using Weblogo[33] 3.7.4 (https://weblogo.berkeley.edu/logo.cgi).

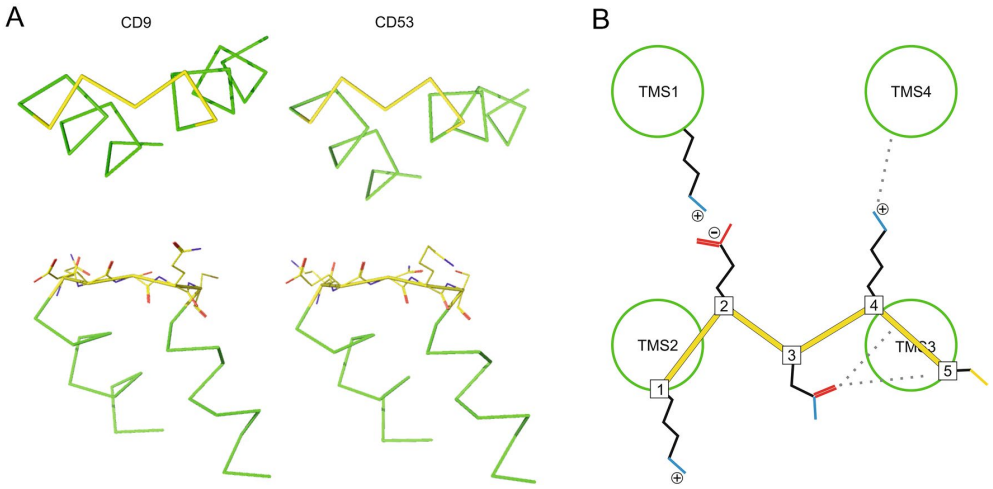


Figure 3. The SIL core forms an M-motif. (**A**) Illustration of the M-motif (yellow) and attached helical structure (green) based on the published crystal structures of CD9 (PDB 6K4J; left) and CD53 (PDB 6WVG; right). Upper panels, top view from the intracellular side. Lower panels, side views. TMS2 (left) and TMS3 (right) are connected to the M-motif and form with the other two TMSs a membrane-embedded funnel (for the complete structures see Fig. S4). The symmetric embedding of the funnel (see OPM database) aligns the M-plane parallel to the membrane. The side view illustrates that the M-motif backbone bonds are co-planar and that, with the exception of position 4, all residues lay roughly in the M-plane. (**B**) Summary of interactions between consensus sequence amino acids that were identified in the crystal structures of CD9 and CD53. Position 2: salt-bridge between glutamate and lysine of the N-terminus in CD9 (for sequence alignment of the N-termini see Fig. S6); in CD53 similar arrangement but distance is too long for a salt-bridge (for details see Fig. S4). Position 3: interactions between the carbonyl-group of an asparagine with the backbone of TMS3 in CD53 (Fig. S4). Position 4: interaction between the lysine side chain and the backbone of TMS4 in CD53 (Fig. S4B). In the stick representation of the side chains, red, blue and yellow indicate oxygen, nitrogen and sulphur, respectively. ⊕/⊖ indicate amino acid residue charges at physiological pH. The scheme was created using CorelDRAW 2019 (www.corel.com) and the crystal structure images were created using PyMOL 2.5 (https://pymol.org/2/).

with a YXXΦ internalization motif outside the core region, pointing towards possibly additional specialized roles of these two SILs.

In the crystal structures of CD9 and CD53, albeit different on the level of amino acids (QESQC versus KENKC; note that for crystallographic reasons in the CD53 structure cysteine is exchanged by a serine), the SIL core structures are essentially identical. The bonds defining the SIL backbone are coplanar, forming an "M", while the two bottom M-endings mark the transition to the helical structure (Fig. 3A). The non-helical central polar residue (position 3) not only marks the turning point of the protein backbone, but also interacts with the backbone and residues of TMS3 (Fig. 3B). The glutamate (position 2) and glutamine/lysine (position 4) residues are oriented towards TMS1 and TMS4, respectively, whereas the amino acid residues at positions 1 and 5 are not oriented towards any of the TMSs. With the exception of position 4, all residues are roughly lying in the "M" plane (Fig. 3A, bottom) that runs parallel to the membrane surface. In CD9, the glutamate at position 2 forms a salt-bridge with a lysine (K11) of the N-terminus (Fig. S4A), and in CD53, the lysine at position 4 interacts with an asparagine (N207) within the C-terminus (Fig. S4B).

This raised the question if the M-motif is a structural element for which the SIL consensus sequence is a prerequisite. For verification, we screened the PDB database for other α-helix rich proteins exhibiting M-motifs. We readily found more examples (Fig. 4A) from which the detailed structure of a selection is shown in Fig. S5, that however have a sequence very different from the SIL consensus sequence (Fig. 4B). The only overlap is the asparagine at position 3, which is involved in interactions between the side chain carbonyl group and the backbone of the C-terminally attached helix in tetraspanins and non-tetraspanins. On the other hand, when measuring the φ and ψ angles in the M-motif, there was clear position-dependent segregation into φ/ψ angle categories. Amino acids in positions 1, 4 and 5 adopt angles typical for alpha helices, in position 2 we find angles associated with a left-handed helix, and for position 3 the angles are typical of a beta-strand (Fig. 4C)[39]. These specific dihedral angles are the basis of the M-motifs shape (Fig. 4D).

Crosstalk between the SIL and the N terminus. As outlined above, the SIL could be simply linking TMS2 and TMS3. On the other hand, it might influence the tertiary structure. Of particular significance for stabilizing the protein could be a salt bridge between the SIL glutamate and the N-terminal lysine, such as seen in the CD9 crystal structure (Fig. S4A).

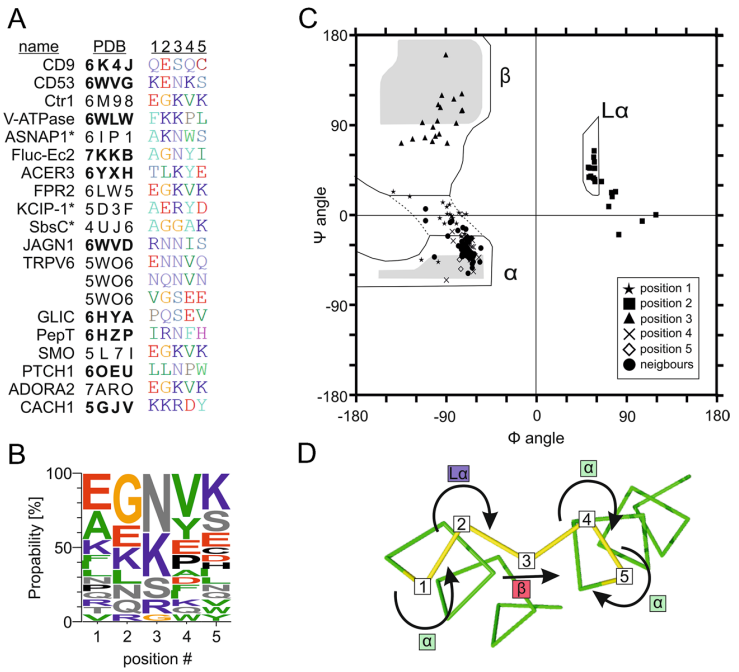


Figure 4. The M-motif is a common structural element formed independent of the SIL consensus sequence. (**A**) The two known tetraspanin structures with resolved M-motif and additional 16 examples of crystal structures containing in total 20 inter-helix turn M-motifs. Proteins are listed using short names, followed by the PDB entry number and the sequence of the M-motif. Mostly, the M-motif was found in membrane proteins but also in three cytosolic ones (marked with an asterisk). Examples in which the motif connects two TMSs were highlighted (PDB number in bold). For structural illustration of the M-motifs in the high-affinity copper transporter Ctrl1, V-ATPase and alpha-soluble NSF attachment protein see Fig. S5. (**B**) Amino acid frequencies in the M-motifs of the proteins listed in A. With the exception of the central asparagine, there is hardly overlap with the animal SIL core consensus sequence. (**C**) Analysis of the dihedral angles employing a Ramachandran plot reveals alpha-helical properties in positions 1, 4 and 5, left-handed helix in position 2, and beta-strand features in position 3. As expected, neighbouring residues are essentially alpha-helical. (**D**) The 180° turn of the direction of the protein backbone involves three structural elements. The turn is initiated by the left-handed character adopted by position 2. Between position 2 and 4 the gap is spanned by the beta-strand character of position 3. Finally, at position 4 already alpha-helicity in the opposite direction is realized. Please note that arrows shown at position 2 and 4 look identical but indicate different angle combinations, which is due to simplification in structure presentation. The alignment was created using BioEdit[28] v7.0.5 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html) and the illustration of residue frequency was done using Weblogo[33] 3.7.4 (https://weblogo.berkeley.edu/logo.cgi). The presentation of the dihedral angles was done using GraphPad Prism version 6.04 for Windows (www.graphpad.com) and PyMOL 2.5 (https://pymol.org/2/).

We had some preliminary data pointing towards the functional importance of the SIL in Tspan17. As for Tspan17 no crystal structure is available, we used another type of analysis to predict whether crosstalk between the SIL glutamate and the N-terminal lysine is possible. The N terminus constitutes of 19 amino acids with two lysines at positions 4 and 18. An analysis of the TMS1 N-terminal sequence shows that the positive charge, mostly provided by a lysine, is highly conserved across all animal species (Fig. S6). Based on this conserved position in the alpha helix, relative to a previously described conserved asparagine[12,13], the lysine of Tspan17 and most other tetraspanins is likely oriented towards the middle of the four helix bundle, and consequently towards the SIL (Fig. S7). This suggest that some crosstalk between the SIL and the N-terminus is possible.

To test experimentally the hypothesis that the SIL interacts with the N-terminus, we analysed the expression levels of Tspan17 after mutating single SIL core amino acids to alanine (Fig. S8), including for comparison CD9 and CD53 as well. With the exception of CD53, we observed strongest diminishment of expression after mutation of the SIL glutamate at position 2.

If the SIL glutamate interacts electrostatically with the N-terminal lysine, exchanging in Tspan17 the lysine to alanine should affect expression just like the glutamate mutation. Because reduced expression levels can have many explanations, in the following we included as well the analysis of glycosylation by western blot and ER-exit by microscopy. In the latter assay, retention in the ER is revealed by an increase in overlap between Tspan17 and an ER-marker. As shown in Fig. 5, mutation of either the SIL glutamate or the N-terminal lysine reduces

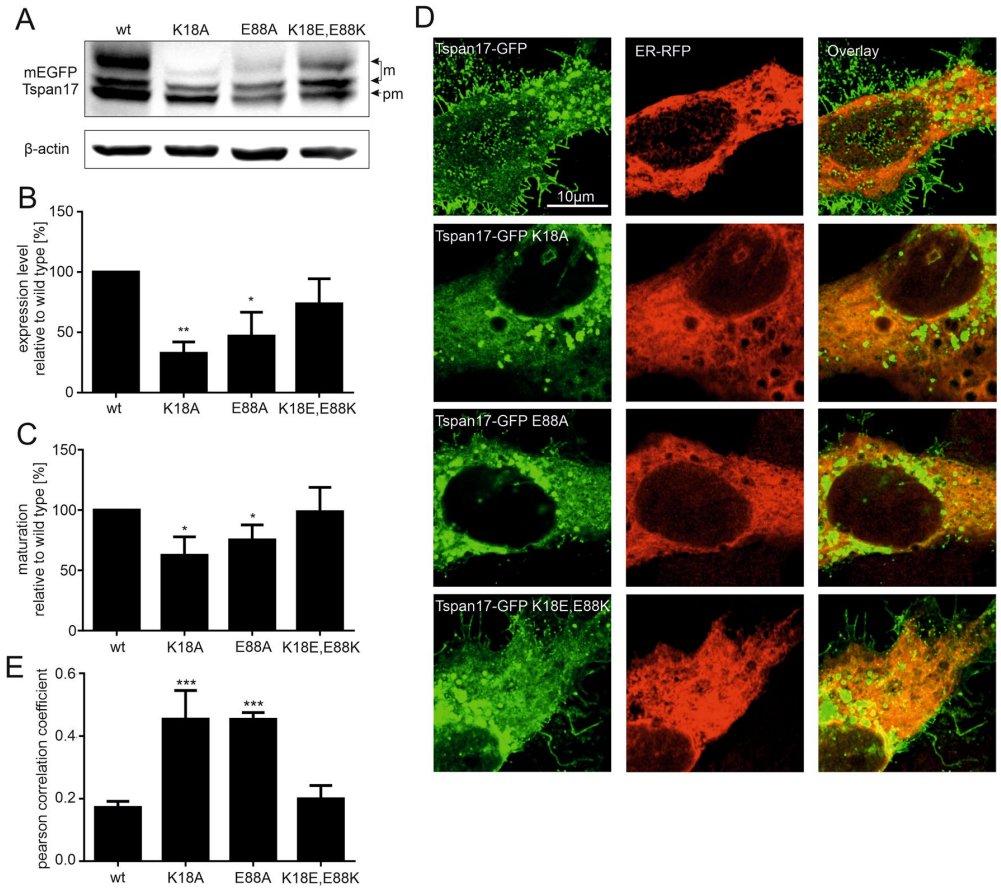


Figure 5. Analysis of the glutamate-lysine interaction. (**A**) In mEGFP-Tspan17, glutamate or lysine were exchanged to alanine (mutants K18A and E88A) or interchanged (K18E/E88K). Constructs are expressed in HepG2 cells, and expression levels are analysed by Western Blot analysis. Maturation of Tspan17 involves N-glycosylation and the different bands representing pre-mature (pm) and matured (m) protein are marked with an arrow (for identification of the bands see Fig. S9). (**B**) The actin normalized GFP signal shows a drop in expression level upon glutamate or lysine mutation (wild type values are set to 100%). The mutation with interchanged charged residues restores the expression level. The drop in Tspan17 expression and maturation is detectable shortly after expression starts (Fig. S10). (**C**) The maturation was calculated as ratio between mature to total protein and was compared to the wild type (set to 100%). The glutamate-/lysine-mutation leads to a drop in protein maturation, whereas the double mutation has no effect. (**D**) The co-localization of the mEGFP-Tspan17 constructs with the ER were analyzed by confocal microscopy, comparing the distribution of the GFP signal to an ER marker fused to RFP. For quantification, the Pearson correlation coefficient between the two channels was calculated. Exemplary images of Tspan17 wt and the mutants are shown. (**E**) The Pearson correlation coefficient showed an increase of ER co-localization for the glutamate-/lysine-mutants and no effect of the double mutation. Values are given as means ± SD (n = 4 for Western blot analysis and n = 3 for microscopy; for each biological replicate 10 cells were imaged). The statistical analysis was done employing a repeated measures ANOVA comparing each mutation with the wild type. The full blots are shown in the supplementary data (Fig. S20). The data analysis and illustration was performed using Fiji-ImageJ[34] (https://imagej.net/) and GraphPad Prism version 6.04 for Windows (www.graphpad.com), respectively.

expression and glycosylation and causes ER retention. Next, we exchanged the positions of glutamate and lysine, which may neutralize the effect of the single mutations as the putative electrostatic interaction may work as well with exchanged positions of the charges. As shown in Fig. 5, all effects are back to normal in the double mutant.

The same mutations in other tetraspanins, e.g. CD9, yield the same expression pattern as for Tspan17 (Fig. S11A) but no effect on ER-exit (Fig. S12). Because in our assay we detect no change in the CD9 band pattern after tunicamycin treatment (Fig. S9), we did not employ maturation analysis of CD9 via probing its glycosylation status. Instead, we examined whether the interaction with the primary binding partner EWI-2 is affected. As shown in Fig. S13, single mutations precipitated more EWI-2 than the wild-type or the double

mutant. Finally, in CD53, single mutations had no effect on expression and maturation (Fig. S11B), although the double mutation drastically diminished expression.

Altogether, the data across different tetraspanins is not consistent, in particular not between CD9 and CD53. Trying to understand better the different roles of the glutamates in CD9 and CD53, we took a closer look at the crystal structures. The short N-termini (12 residues) of CD9 and CD53 contain three and two positively charged amino acids, respectively. In CD9, as already mentioned above, the SIL core E84 forms a salt bridge (distance 2.8 Å) with the second last amino acid of the N-terminal peptide (K11) (for illustration see Fig. S4A). In CD53, although the SIL core E77 is oriented towards K7 and K10 of the N terminus, the distances are about one angstrom too long to establish a salt bridge (Fig. S4B; K7-E77, 5.1 Å and K10-E77, 5.2 Å instead of 4 Å required to form a salt bridge[40]). This suggests a weaker or no electrostatic interaction between the CD53 SIL and the N-terminus and could explain the lacking effect in the CD53 mutants.

Discussion

The SIL core sequence. The tetraspanin sequence analysis of the SIL between TMS2 and TMS3 reveals a conserved [R/K] E [N/S] [R/K/Q] C core sequence in human, similar to mouse and zebrafish. In fruit fly, the positive charge in position 4 is lacking, and in arabidopsis the sequence is very different. In proteins with the same topology, as in the family of claudins, the SIL is longer (Fig. S2), has a longer unstructured stretch with a predicted beta-strand, is more diverse, and exhibits no similarity to the tetraspanin core sequence. This may point to a specific function of the SIL in mammalian tetraspanins.

It is known that positively charged residues close to the cytosolic site of a TMS are beneficial for its membrane insertion, whereas negatively charged or polar residues decrease the TMS insertion[41], known as the "positive-inside rule" and the "negative inside depletion/outside enrichment rule"[42]. Therefore, the presence of positively charged amino acids in the SIL is not surprising, as it aids the correct insertion of the nascent protein into the ER membrane[43]. The negatively charged glutamate neutralizes one positive charge. Apparently, this is not relevant for expression, as in CD9 and Tspan17 mutants with swapped glutamate/lysine express equally well as wild-type (Fig. 5, Fig. S11A).

Apart from that, tetraspanins are known to be palmitoylated at several intracellular cysteine-residues, among them cysteines in the SIL of CD9 and CD81[44], explaining the abundancy of cysteines at the end of the SIL core region (Fig. 2G).

The SIL forms an M-motif. In animal tetraspanins, the three central residues of the SIL core region are predicted to be less helical, which is consistent with the crystal structures of CD53 and CD9. Please note that a crystal structure of CD81 is also available, but could not be used for detailed SIL analysis as the 2nd and 3rd amino acids of the SIL core are unresolved.

In all animal tetraspanins, the non-helical part of the SIL is on average 2.1 amino acids in length (see also Fig. S3), which is close to the shortest possible linker between two TMSs, that is two amino acids[45]. Roughly speaking, the five core residues form a U-turn with helical arms continued by the TMS helices.

The amino acids at positions 2 and 4 constitute the upper two tips of the M-motif and their residues point towards TMS1 and TMS4, respectively. The position 1, 3, and 5 define the lower three tips of the M-motif, all pointing away from the centre of the TMS-bundle. The M-motif shape is not exclusive to tetraspanins but found in many other soluble and membrane proteins (Fig. 4, Fig. S5), although the amino acid sequence is different from the SIL core sequence, with the exception of the central asparagine.

The left-handed character of position 2 and the beta-strand character of position 3 define the starting point and the bridge of the U-turn (Fig. 4D). Moreover, they are involved in stabilizing interactions as shown by frequent examples for the residue at position 3 that interacts with the backbone of the C-terminal helix (e.g. CD9, PDB: 6K4J and POT family transporter, PDB: 6HZP) or forms a salt-bridge with position 1 (e.g. adenosine A2A receptor, PDB: 7ARO or Smoothened, PDB: 5L7I). Additionally, the residue of position 2 can form salt-bridges with adjacent structures (e.g. CD9, PDB: 6K4J or voltage-gated calcium channel Cav1.1, PDB: 5GJV). In conclusion, the M-motif is less defined by a specific amino acid sequence (Fig. 4B) but rather by its secondary structure and interactions.

There are two known groups of short loops/turns connecting secondary structure elements, which are both not defined by a characteristic secondary structure. The first is classified by its length and that the loops' residues are not incorporated into the hydrogen bonding of the neighbouring secondary structure elements[46]. The other group is defined by the side chain (typically Asp, Asn, Ser or Thr) that interacts with the backbone but only moderately changes the backbone orientation and does not result in a pair of antiparallel helices[47]. The residues of the M-motif are all forming backbone hydrogen bonds with the neighbouring alpha-helices, which excludes the M-motif from the first group of turns. Frequently, there is an asparagine/serine in the M-motif that interacts with the C-terminal helix backbone, but in the M-motif, a complete turn is formed. Therefore, the M-motif does not strictly fit into any of the two known groups and defines its own category of inter-helix turns.

Role of the SIL glutamate in Tspan17. For Tspan17, we find that mutating the SIL glutamate or the TMS1 N-terminal lysine reduces glycosylation and ER-exit (Fig. 5). In addition, the expression level is reduced, which could be a secondary effect of disturbed trafficking through the ER. Altogether, the three assays yield a consistent picture.

Importantly, glycosylation, ER-exit and expression are back to normal levels when the SIL glutamate and N-terminal lysine are swapped (Fig. 5). This points towards a functionally important glutamate-lysine interaction between the SIL and the TMS1 N-terminal lysine. Because the positions of the two oppositely charged amino acids can be interchanged, we speculate that the interaction is most likely of electrostatic nature.

In other tetraspanins, the SIL glutamate seems to be of relevance as well, although the overall picture is unclear. For instance, in CD9, glutamate/lysine mutations have no effect on ER exit (Fig. S12), but increase CD9 association with EWI-2 (Fig. S13). This is very interesting as it implies two things. First, without salt-bridge, CD9 still adopts a functional conformation, or in other words, lack of the salt-bridge does not lead to complete misfolding. Second, its higher affinity to EWI-2 may be explained by a switch towards an open conformation, as shown for CD81 that interacts in the open conformation stronger with its primary binding partner CD19[12]. In cryo-EM, CD9 interacts with EWI-2 not in a complete but partial open conformation[16]. The four helices still arrange in a funnel shape but the LEL is folded more upright. In complex with EWI-F, that is a EWI-2 homolog, cryo-EM reveals a CD9 conformation resembling the open conformation shown in Fig. 1B (see also reference[15]). Hence, elimination of the salt-bridge could trigger partial CD9 opening and enhance binding to EWI-2. Moreover, from the 33 human tetraspanins, we have performed mutational analysis of eight family members. In three cases each, mutation of the SIL glutamate either significantly decreases or increases expression (Figs. S8, S14).

Altogether, the picture is neither consistent nor complete and we are just at the beginning of understanding the mechanism by which the SIL modulates tetraspanin structure. In fact, we find it is not surprising that equivalent mutations produce different effects in different tetraspanins, as they have different binding partners and functions.

Conclusion

In this study, we show that the SIL of tetraspanins contains a conserved five amino acid core sequence forming a structural motif that resembles the letter M. Using Tspan17 as example, we find that mutation of the SIL glutamate or the N-terminal lysine adjacent to TMS1 reduces glycosylation, ER-exit, and expression. All effects are back to normal levels upon position swapping of the two oppositely charged amino acids. We speculate that glutamate and lysine interact electrostatically, which might impact the tertiary structure and as a result modulate the interaction network of Tspan17.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 24 September 2021; Accepted: 15 February 2022
Published online: 16 March 2022

References

1. Lang, T. & Hochheimer, N. Tetraspanins. *Curr. Biol.* **30**, R204–R206 (2020).
2. Yáñez-Mó, M., Barreiro, O., Gordon-Alonso, M., Sala-Valdés, M. & Sánchez-Madrid, F. Tetraspanin-enriched microdomains: A functional unit in cell plasma membranes. *Trends Cell Biol.* **19**, 434–446 (2009).
3. Charrin, S. *et al.* Multiple levels of interactions within the tetraspanin web. *Biochem. Biophys. Res. Commun.* **304**, 107–112 (2003).
4. Hemler, M. E. Specific tetraspanin functions. *J. Cell Biol.* **155**, 1103–1108 (2001).
5. van Deventer, S. J., Dunlock, V.-M.E. & van Spriel, A. B. Molecular interactions shaping the tetraspanin web. *Biochem. Soc. Trans.* **45**, 741–750 (2017).
6. Termini, C. M. & Gillette, J. M. Tetraspanins function as regulators of cellular signaling. *Front. Cell Dev. Biol.* **5**, 34 (2017).
7. Miranti, C., Bergsma, A. & Spriel, A. Tetraspanins as Master Organizers of the Plasma Membrane. In 59–86 (2014). https://doi.org/10.1201/b17634-6.
8. Andreu, Z. & Yáñez-Mó, M. Tetraspanins in extracellular vesicle formation and function. *Front. Immunol.* **5**, 25 (2014).
9. Seigneuret, M., Delaguillaumie, A., Lagaudrière-Gesbert, C. & Conjeaud, H. Structure of the tetraspanin main extracellular domain: A partially conserved fold with a structurally variable domain insertion*. *J. Biol. Chem.* **276**, 40055–40064 (2001).
10. Charrin, S., Jouannet, S., Boucheix, C. & Rubinstein, E. Tetraspanins at a glance. *J. Cell Sci.* **127**, 3641–3648 (2014).
11. Hochheimer, N., Sies, R., Aschenbrenner, A. C., Schneider, D. & Lang, T. Classes of non-conventional tetraspanins defined by alternative splicing. *Sci. Rep.* **9**, 14075 (2019).
12. Zimmerman, B. *et al.* Crystal structure of a full-length human tetraspanin reveals a cholesterol-binding pocket. *Cell* **167**, 1041-1051.e11 (2016).
13. Seigneuret, M. Complete predicted three-dimensional structure of the facilitator transmembrane protein and hepatitis C virus receptor CD81: Conserved and variable structural domains in the tetraspanin superfamily. *Biophys. J.* **90**, 212–227 (2006).
14. Susa, K. J., Rawson, S., Kruse, A. C. & Blacklow, S. C. Cryo-EM structure of the B cell co-receptor CD19 bound to the tetraspanin CD81. *Science* **371**, 300–305 (2021).
15. Oosterheert, W. *et al.* Implications for tetraspanin-enriched microdomain assembly based on structures of CD9 with EWI-F. *Life Sci. Alliance* **3**, 25 (2020).
16. Umeda, R. *et al.* Structural insights into tetraspanin CD9 function. *Nat. Commun.* **11**, 1606 (2020).
17. Yang, Y. *et al.* Open conformation of tetraspanins shapes interaction partner networks on cell membranes. *EMBO J.* **39**, e105246 (2020).
18. Palor, M., Stejskal, L., Mandal, P., Shepherd, A. J. & Grove, J. Conformational switching of CD81 controls its function as a receptor for hepatitis C virus. *BioRxiv* https://doi.org/10.1101/542837 (2019).
19. Levy, S. & Shoham, T. Protein-protein interactions in the Tetraspanin web. *Physiology* **20**, 218–224 (2005).
20. Homsi, Y. *et al.* The extracellular δ-domain is essential for the formation of CD81 tetraspanin webs. *Biophys. J.* **107**, 100–113 (2014).
21. Martin, F. *et al.* Tetraspanins in viral infections: A fundamental role in viral biology?. *J. Virol.* **79**, 10839–10851 (2005).
22. Kovalenko, O. V., Metcalf, D. G., DeGrado, W. F. & Hemler, M. E. Structural organization and interactions of transmembrane domains in tetraspanin proteins. *BMC Struct. Biol.* **5**, 11 (2005).
23. Kummer, D., Steinbacher, T., Schwietzer, M. F., Thölmann, S. & Ebnet, K. Tetraspanins: Integrating cell surface receptors to functional microdomains in homeostasis and disease. *Med. Microbiol. Immunol. (Berl.)* **209**, 397–405 (2020).
24. Khattree, N., Ritter, L. M. & Goldberg, A. F. X. Membrane curvature generation by a C-terminal amphipathic helix in peripherin-2/rds, a tetraspanin required for photoreceptor sensory cilium morphogenesis. *J. Cell Sci.* **126**, 4659–4670 (2013).
25. Berditchevski, F. & Odintsova, E. Tetraspanins as regulators of protein trafficking. *Traffic* **8**, 89–96 (2007).
26. Stipp, C. S., Kolesnikova, T. V. & Hemler, M. E. Functional domains in tetraspanin proteins. *Trends Biochem. Sci.* **28**, 106–112 (2003).

27. Mazurov, D., Heidecker, G. & Derse, D. The inner loop of tetraspanins CD82 and CD81 mediates interactions with human T cell lymphotrophic virus type 1 gag protein*. *J. Biol. Chem.* **282**, 3896–3903 (2007).
28. Dagona, A. G. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **20**, 20 (1999).
29. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
30. Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175–182 (1998).
31. Drozdetskiy, A., Cole, C., Procter, J. & Barton, G. J. JPred4: A protein secondary structure prediction server. *Nucleic Acids Res.* **43**, W389–W394 (2015).
32. Gautier, R., Douguet, D., Antonny, B. & Drin, G. HELIQUEST: A web server to screen sequences with specific alpha-helical properties. *Bioinform. Oxf. Engl.* **24**, 2101–2102 (2008).
33. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: A Sequence Logo Generator. 3.
34. Schindelin, J. *et al.* Fiji: An open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
35. Marsay, K., Roehl, H., Monk, P., Partridge, L. & Carney, T. Tetraspanins in zebrafish development. (2017).
36. Fradkin, L. G., Kamphorst, J. T., DiAntonio, A., Goodman, C. S. & Noordermeer, J. N. Genomewide analysis of the *Drosophila tetraspanins* reveals a subset with similar function in the formation of the embryonic synapse. *Proc. Natl. Acad. Sci.* **99**, 13663–13668 (2002).
37. Boavida, L. C., Qin, P., Broz, M., Becker, J. D. & McCormick, S. *Arabidopsis tetraspanins* are confined to discrete expression domains and cell types in reproductive tissues and form homo- and heterodimers when expressed in yeast. *Plant Physiol.* **163**, 696–712 (2013).
38. Tsukita, S., Tanaka, H. & Tamura, A. The Claudins: From tight junctions to biological systems. *Trends Biochem. Sci.* **44**, 141–152 (2019).
39. Hollingsworth, S. A. & Karplus, P. A. A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *Biomol. Concepts* **1**, 271–283 (2010).
40. Donald, J. E., Kulp, D. W. & DeGrado, W. F. Salt bridges: Geometrically specific, designable interactions. *Proteins* **79**, 898–915 (2011).
41. Lerch-Bader, M., Lundin, C., Kim, H., Nilsson, I. & von Heijne, G. Contribution of positively charged flanking residues to the insertion of transmembrane helices into the endoplasmic reticulum. *Proc. Natl. Acad. Sci.* **105**, 4127–4132 (2008).
42. Baker, J. A., Wong, W.-C., Eisenhaber, B., Warwicker, J. & Eisenhaber, F. Charged residues next to transmembrane regions revisited: "Positive-inside rule" is complemented by the "negative inside depletion/outside enrichment rule". *BMC Biol.* **15**, 66 (2017).
43. Shao, S. & Hegde, R. S. Membrane protein insertion at the endoplasmic reticulum. *Annu. Rev. Cell Dev. Biol.* **27**, 25–56 (2011).
44. Charrin, S. *et al.* Differential stability of tetraspanin/tetraspanin interactions: Role of palmitoylation. *FEBS Lett.* **516**, 139–144 (2002).
45. Liò, P. & Goldman, N. Using protein structural information in evolutionary inference: Transmembrane proteins. *Mol. Biol. Evol.* **16**, 1696–1710 (1999).
46. Hoang, N. H. *et al.* Twists or turns: Stabilising alpha vs beta turns in tetrapeptides. *Chem. Sci.* **10**, 10595–10600 (2019).
47. Dhar, J. & Chakrabarti, P. Defining the loop structures in proteins based on composite β-turn mimics. *Protein Eng. Des. Sel.* **28**, 153–161 (2015).

Acknowledgements
We would like to thank Dr. Jerome Finke (University of Bonn) for his help with imaging, Dominik Sons (University of Bonn) for providing the EWI-2-myc construct and Sara C. Schmidt (University of Bonn) for technical assistance. Furthermore, we would like to thank Elke Reppert and Eileen C. Reppert for proof reading the manuscript.

Author contributions
N.R.: conceptualization, data acquisition, formal analysis, validation, figure preparation, and writing-original draft; T.L.: guidance, mentoring and writing original draft.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-07243-y.

Correspondence and requests for materials should be addressed to N.R. or T.L.

Reprints and permissions information is available at www.nature.com/reprints.

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

Supplementary Data

A conserved sequence in the small intracellular loop of tetraspanins forms an M-shaped inter-helix turn

Nikolas Reppert[1,*] and Thorsten Lang[1,*]

[1]Department of Membrane Biochemistry, Life & Medical Sciences (LIMES) Institute, University of Bonn, Carl-Troll-Straße 31, 53115 Bonn, Germany

[*]Correspondence should be addressed to Nikolas Reppert (Nikolas.Reppert@uni-bonn.de) and Thorsten Lang (thorsten.lang@uni-bonn.de)

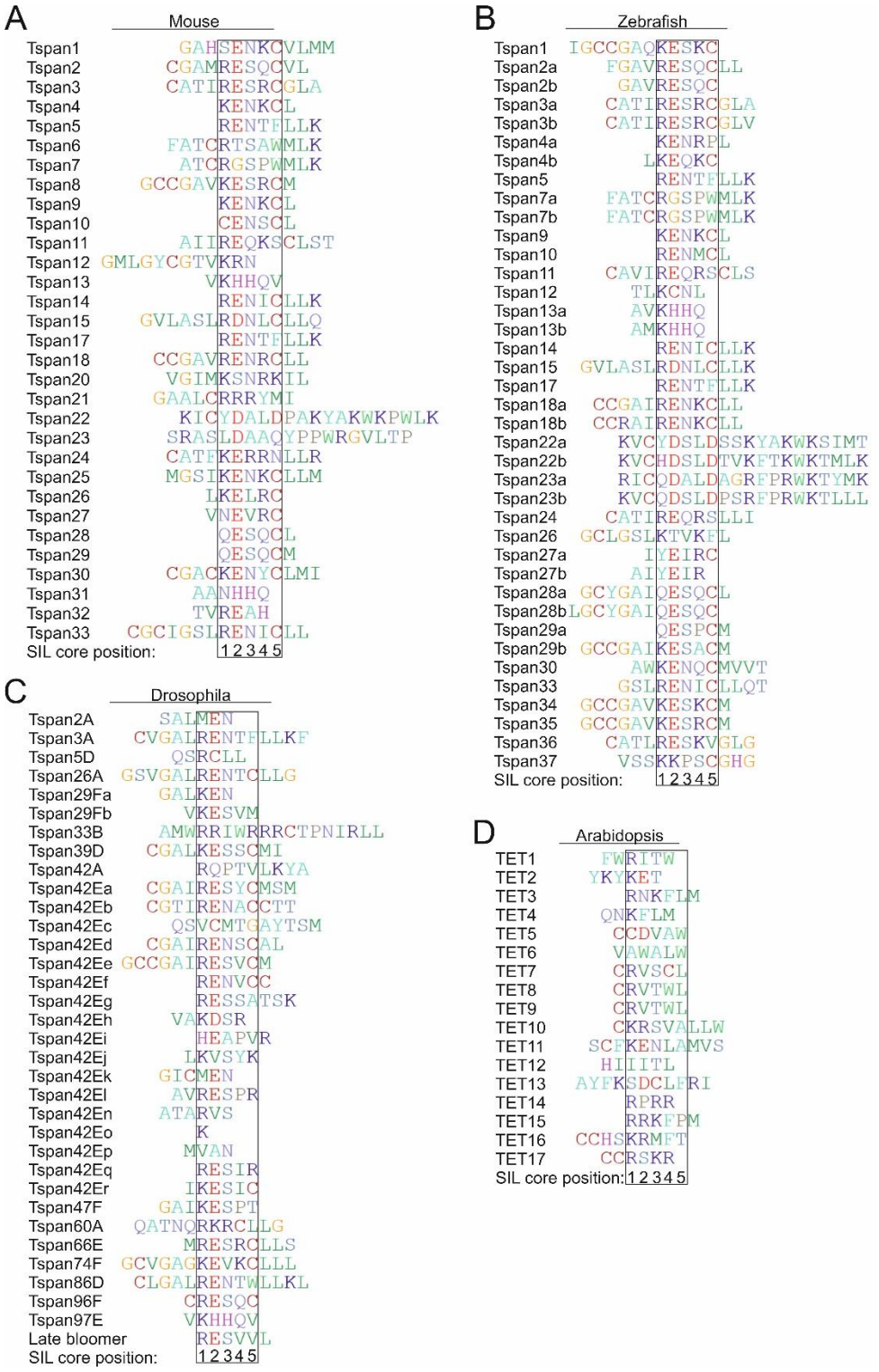


Figure S1: SIL core sequence alignments of other species.
Employing the same criteria as for human tetraspanins, SIL core sequences (boxes) are aligned from tetraspanins from (A) mouse (*Mus musculus*), (B) zebrafish (*Danio rerio*), (C) fruit fly (*Drosophila melanogaster*) and (D) *arabidopsis thaliana*. The alignment was created using BioEdit[1] v7.0.5 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html).

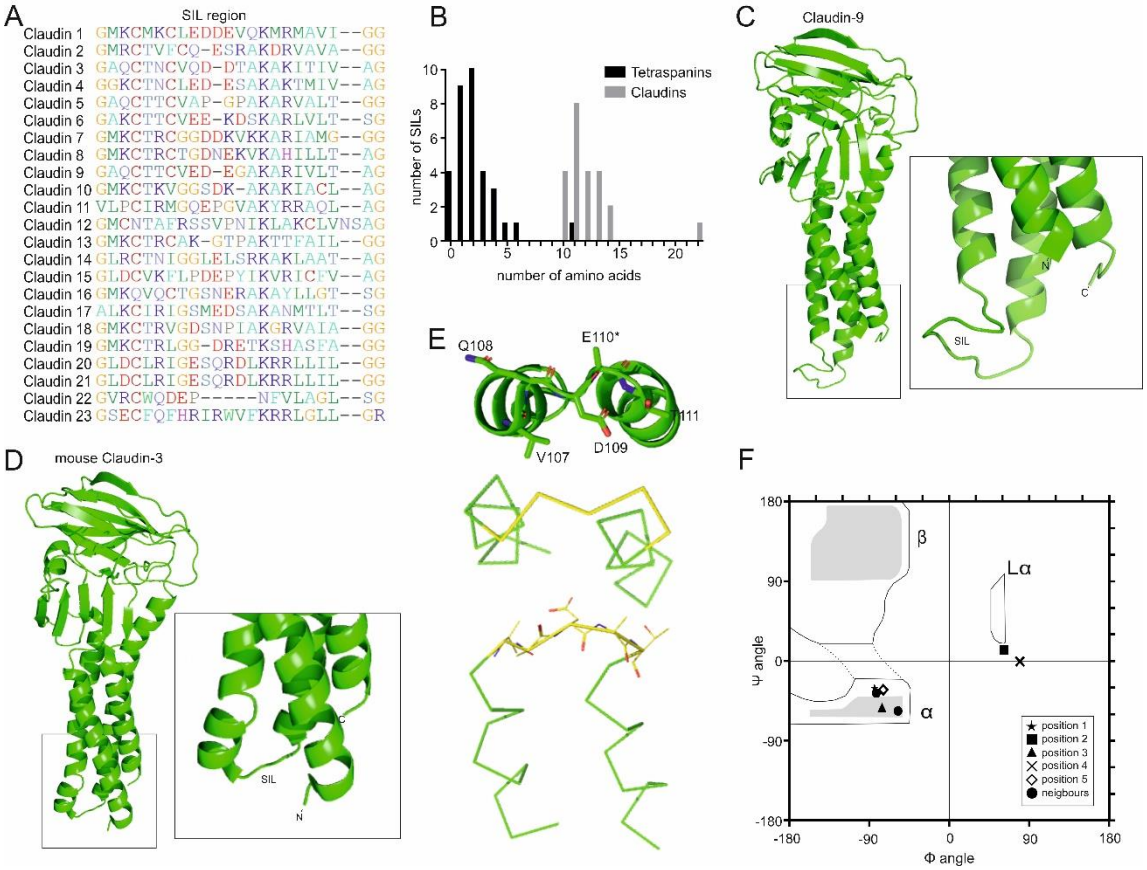


Figure S2: SIL in Claudins, a similarly structured protein family.
(A) ClustelW alignment of 23 human Claudins identified in the Uniprot database. SIL region between TMS2 and TMS3. The alignment was created using BioEdit[1] v7.0.5 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). (B) Distribution of the number of non-helical amino acids present between TMS2 and TMS3 (predicted by Jpred4) in human tetraspanins (black) and Claudins (grey). Please note that in some tetraspanins helical continuity between the two transmembrane segments is predicted, yielding a value of zero. In 19 Claudins, in the longer connecting segment a short beta-strand is predicted using Jpred4[2] (http://www.compbio.dundee.ac.uk/jpred/).(C) Crystal structure of human Claudin-9 (pdb: 6OV3) showing its SIL loop points away from the helix bundle. (D) Crystal structure of mouse Claudin-3 (pdb: 6AKF) with only a short SIL loop resembling an M-motif-like structure as seen for tetraspanins. From the boxed region, a magnified view is shown. (E) Different views from the Claudin-3 M-motif. Asterisk at E110 indicates that part of the side chain is unresolved in the crystal structure. (F) Ramachandran Plot of the Claudin-3 SIL. It does not fulfil the tetraspanin M-motif secondary structure criteria because position 3 is alpha-helical and position 4 is left handed helical. Therefore, we refer to this motif and more such examples (e.g. pdb: 4UJ6, AGGAK; pdb: 6I9D, DGSPD and pdb: 5WO6, YDNNE) as M-like-motifs. The data illustration was performed using GraphPad Prism version 6.04 for Windows (www.graphpad.com) and PyMOL 2.5 (https://pymol.org/2/).

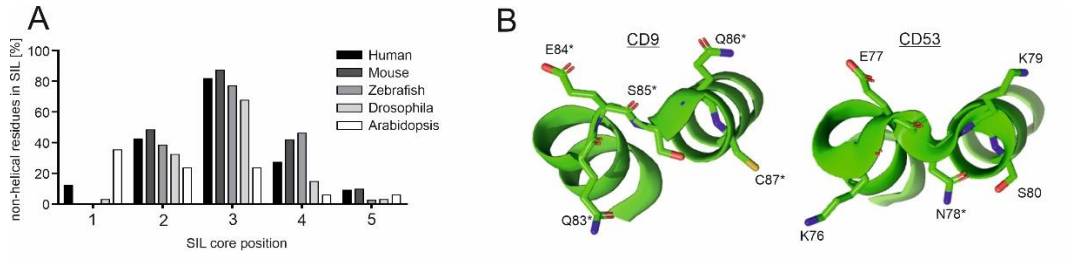


Figure S3: Non-helical structural elements in the SIL core.
(A) In different species (see legend), employing Jpred4 the percent of non-helical amino acids per SIL core position is predicted. In a minority of tetraspanins (Tspan2, 13, 28 and 32) the program suggests unrealistic continuous alpha-helicity between TMS2 and TMS3, that however is unrealistic due to the topology of the protein. For human Tspan22 the predicted helix gap was C-terminal of the SIL core sequence. The data illustration was performed using GraphPad Prism version 6.04 for Windows (www.graphpad.com). (B) Comparison of the theoretical average structure in (A) with crystallographic data of CD9 (pdb: 6K4J) and CD53 (pdb: 6WVG). CD9-S85 and CD53-N78 correspond to the third position of the SIL core. In line with the prediction, the amino acids at the SIL periphery are increasingly alpha helical. An asterisk indicates the amino acids that were predicted to be non-helical. The images were created using PyMOL 2.5 (https://pymol.org/2/).

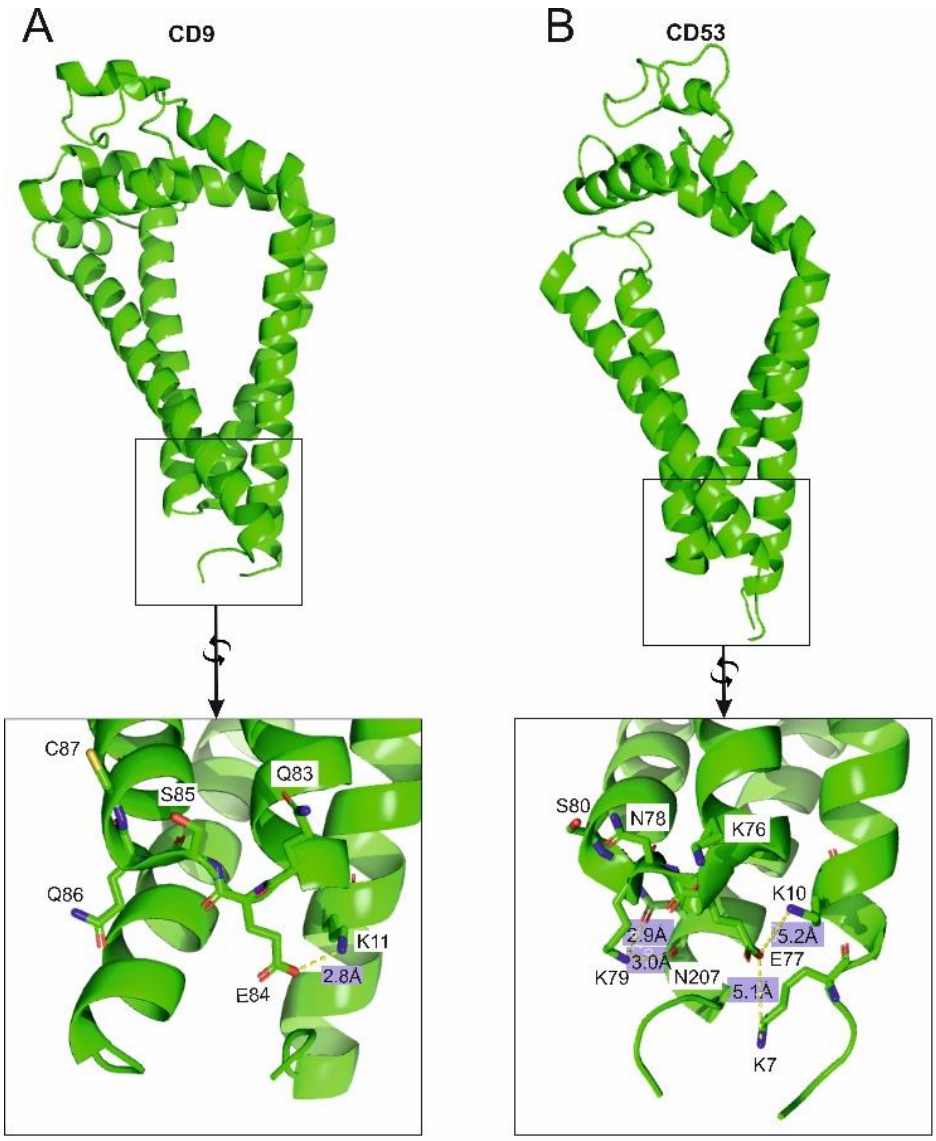


Figure S4: SIL-TMS interactions in CD9 and CD53.
Crystal structures of CD9 (pdb 6K4J; left) and CD53 (pdb 6WVG; right), illustrating the alpha-helical structure of the LELs (top) and the funnel-shaped arrangements of the TMSs (lower part). Large boxes; magnified views from the small boxed regions from a different perspective, such that top views of the M-motifs are visible, together with the N- and C-terminal segments. Amino acids of the SIL core regions (CD9, aa 83 – 87; CD53, 76 - 80), lysines of the N-terminal peptides (CD9, K11; CD53, K7, K11), and the asparagine (N207) within the C-terminal helix of CD53 are indicated. In CD9, a salt-bridge between E84 and K11 is illustrated by the dotted yellow line. The distance between the charged atoms is 2.8 Å (highlighted in purple). In CD53, distances between E77 and K7/K10 (yellow dotted lines) are too long (5.1Å/ 5.2Å, highlighted in purple) for a salt-bridge to form. The distances between the nitrogen in K79 and the carbonyl groups in N207 (side chain, 2.9Å; backbone, 3.0Å) allow for polar interactions (maximal interaction distance between C=O – H-N is 3.0Å[3]). The images were created using PyMOL 2.5 (https://pymol.org/2/).

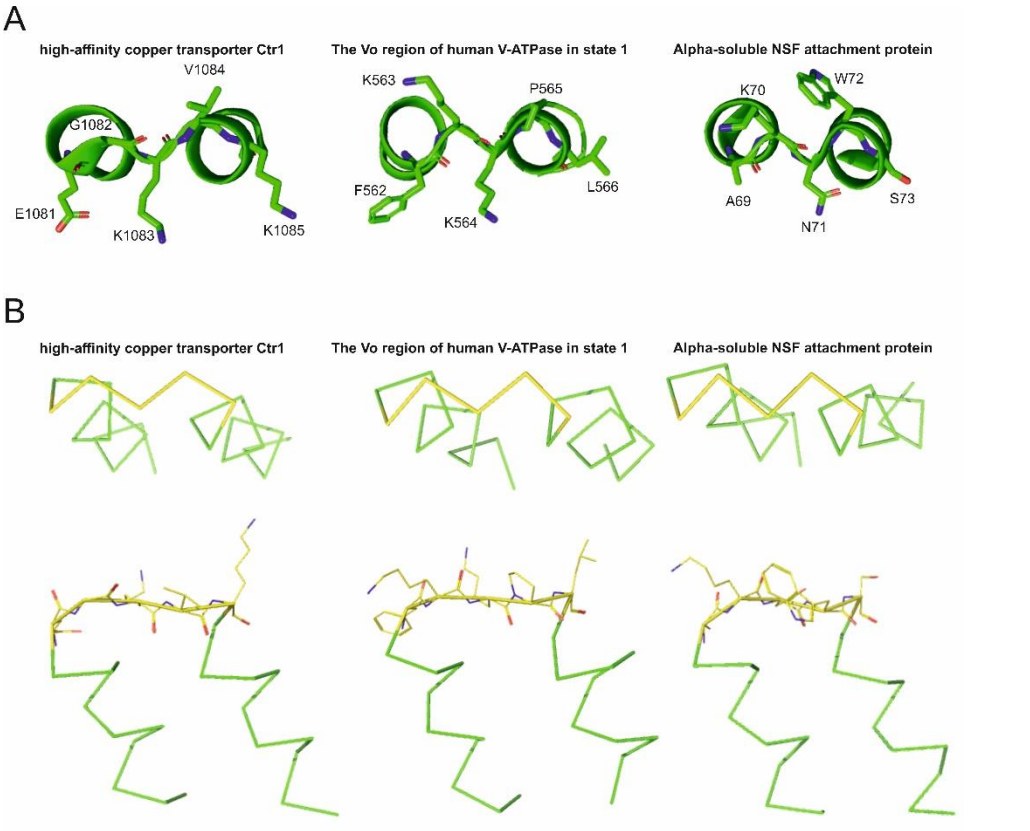


Figure S5: The M-motif in more non-tetraspanin proteins.
Different views from the M-motifs of the high-affinity copper transporter Ctr1 (pdb: 6M98), V-ATPase (pdb: 6WLW) and alpha-soluble NSF attachment protein (pdb: 6IP1). (A) Cartoon and (B) ribbon diagrams with the amino acids of the SIL core shown in detail as yellow sticks (the naming of the modes of illustration refer to PyMOL). Side chain colour code is red for oxygen and blue for nitrogen. Residue numbering in (A) refers to the protein sequence and may differ from the sequence number in the crystal structures. For M-motifs in Claudin-3 see Fig. S2E. Images were created using PyMOL 2.5 (https://pymol.org/2/).

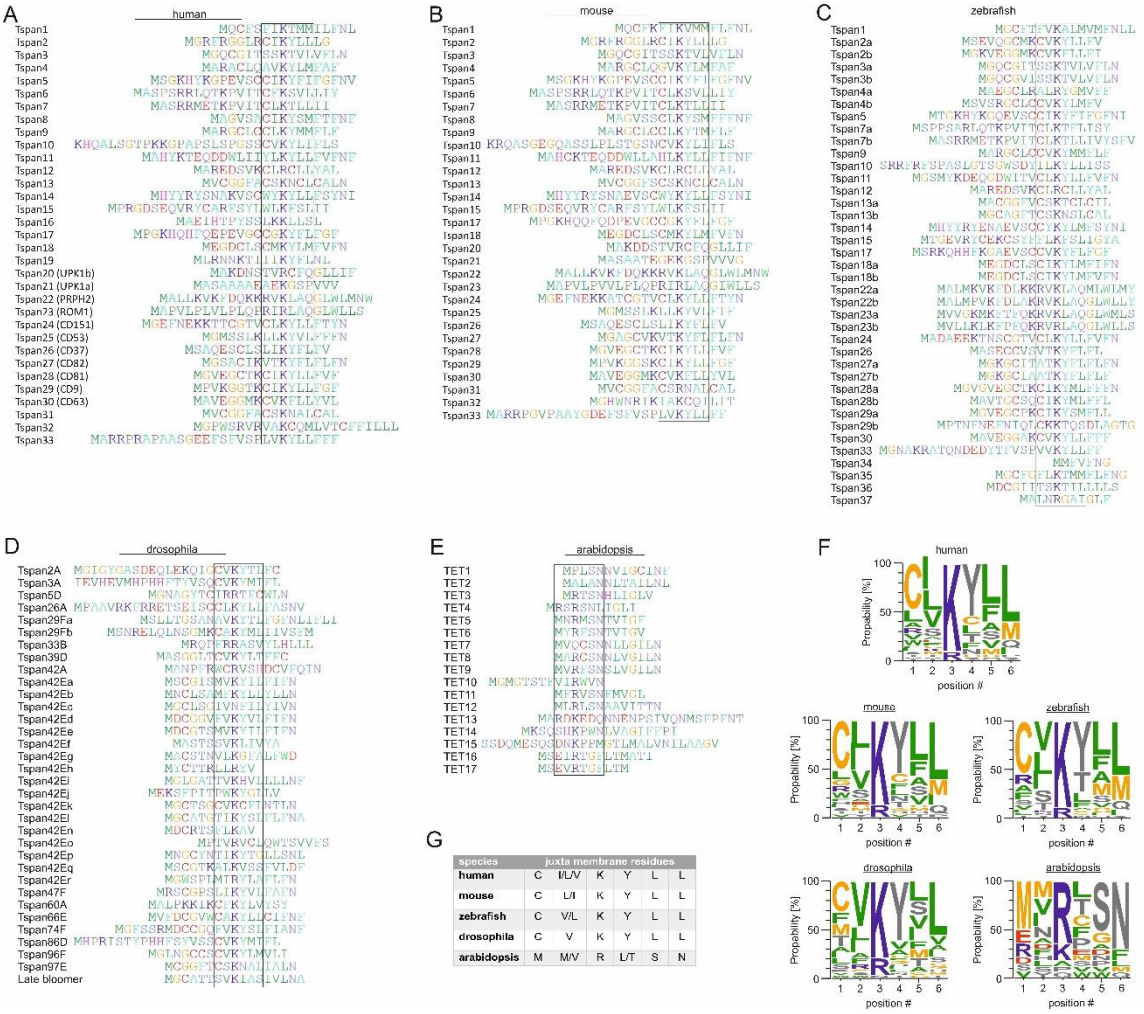


Figure S6: Alignment of the N-terminal tetraspanin sequences from different species.
(A - E) N-terminal sequences plus the first five amino acids of TMS1 are aligned using tetraspanin sequences of (A) human (*Homo sapiens*), (B) mouse (*Mus musculus*), (C) zebrafish (*Danio rerio*), (D) fruit fly (*Drosophila melanogaster*) and (E) arabidopsis (*Arabidopsis thaliana*). The sequences were aligned with reference to the most C-terminal positively charged amino acid of the N-terminus, yielding in the juxta membrane region a conserved sequence of six amino acids including the lysine/arginine used for alignment (see boxes). (F) Residue probability of the six amino acids based on (A - E). The sequence is similarly conserved in animal tetraspanins but different in arabidopsis. (G) Consensus sequences in the juxtamembrane regions defined as described in the legend of Figure 2. The alignment was created using BioEdit[1] v7.0.5 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html) and the illustration of residue frequency was done using Weblogo[4] 3.7.4 (https://weblogo.berkeley.edu/logo.cgi).

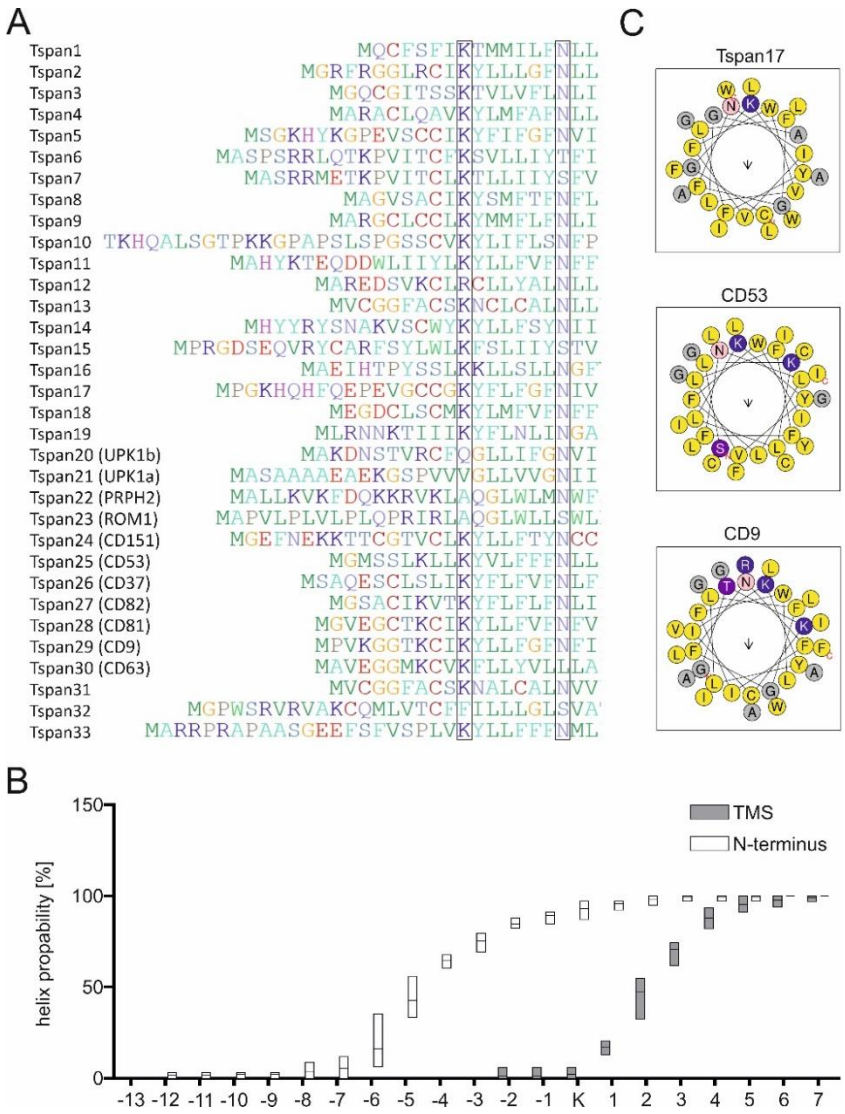


Figure S7: Orientation of the conserved juxtamembrane lysine in relation to the conserved asparagine in TMS1.
(A) Alignment of the human tetraspanin juxtamembrane/TMS1 starting sequences with reference to a conserved asparagine (right box). The left box marks the amino acid seven positions upstream, which frequently is a lysine (see also Fig S6). The long N-terminus of Tspan10 is deleted for this illustration. Seven positions distance roughly equal two turns of an alpha helix (one alpha helical turn corresponds to 3.6 residues[5]) and as a result a similar orientation of the residues. (B) To test whether the conserved lysine and asparagine are located to the same alpha-helix the structure of each tetraspanins' juxtamembrane region (s. Fig S6) was predicted employing Jpred4[2] (http://www.compbio.dundee.ac.uk/jpred/) or TMHMM[6,7] Server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/). The results whether an amino acid of a certain tetraspanin is within an alpha-helix and/or within a TMS were averaged per species and illustrated in a box plot (highest and lowest values framing the box, line in the middle is the mean value for all species; conserved lysine was named K, residues N-terminal of the lysine are named -13 to -1 and C-terminal 1 to 7). This illustration shows that this particular conserved lysine is located to an alpha-helix (in 93% of analysed tetraspanins). The predicted TMS1 probability increases directly C-terminal of the lysine. The lysine is not located to the TMS1 but to the N-terminal helix extension of it. Therefore, it is located to the same helix as the conserved asparagine. (C) The wheel plots of the juxtamembrane/TMS1 sequences of Tspan17, CD53 and CD9 show again the similar orientation of the lysine and asparagine. The wheel plots analyse the amphipathic character of the helix, yielding its hydrophobic moment, which is indicated by the orientation and length of the arrow in the center of the helix. The wheel plots were created using HeliQuest[8] (https://heliquest.ipmc.cnrs.fr/cgi-bin/ComputParams.py).

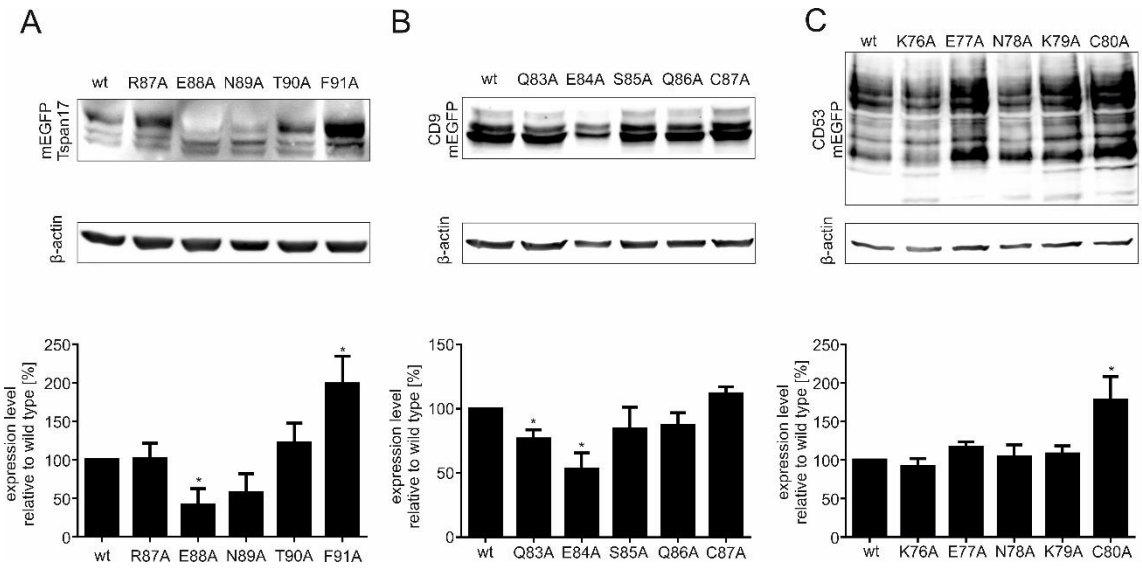


Figure S8: Mutations in the SIL consensus region occasionally change the expression level.
In GFP-tagged (A) Tspan17, (B) CD9 and (C) CD53, single SIL core amino acids are exchanged to alanine. Constructs are expressed in HepG2 cells, and expression levels are analysed by Western Blot. Upper panels, representative Western Blot membranes. Please note that upon mutation of the second and third position in Tspan17 the higher running GFP-Tspan17 band vanishes (A), which is due to a lack in glycosylation (see also Fig. S9). Lower panels, quantification of the Western blot bands. GFP signal is related to actin used as loading control, and normalized to wild-type (set to 100%). Values are given as means ± SD (n = 4). The statistical analysis was done employing a repeated measures ANOVA comparing each mutation to the wild-type (*P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001). The full blots are shown in the supplementary data (Fig. S17-S19). The data analysis and illustration was performed using Fiji-ImageJ[9] (https://imagej.net/) and GraphPad Prism version 6.04 for Windows (www.graphpad.com), respectively.

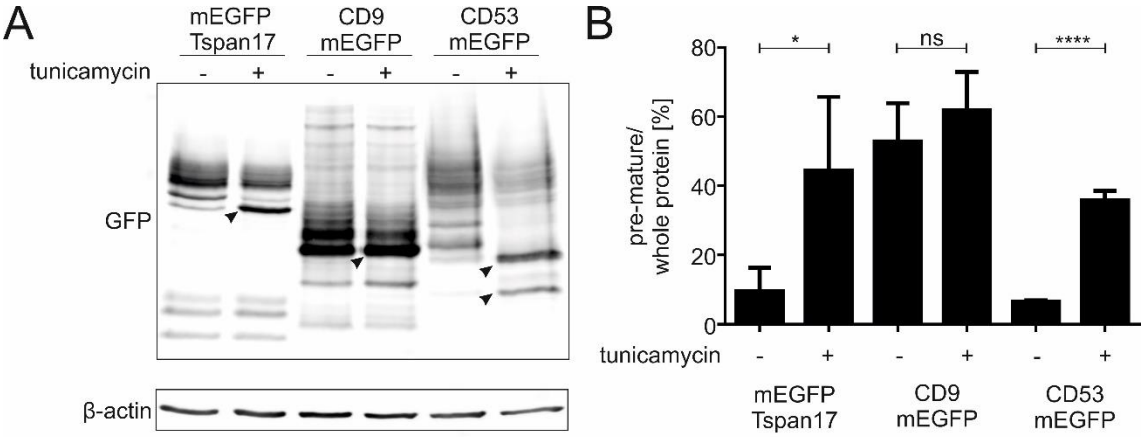


Figure S9: Band pattern of Tspan17, CD9 and CD53 without and with inhibition of glycosylation by tunicamycin.
(A) HepG2 cells are transfected with mEGFP-Tspan17, CD9-mEGFP or CD53-mEGFP and 5 µg/ml tunicamycin is added to the medium. Tunicamycin inhibits N-glycosylation; there are three potential N-glycosylation sites in Tspan17, two in CD53 and two in CD9 (predicted by NetNGlyc[10] 1.0; http://www.cbs.dtu.dk/services/NetNGlyc/). After 18h of incubation, cells are lysed and analysed via Western blot. In the presence of tunicamycin, some bands increase in intensity. The ones with the highest increase were highlighted (see arrowheads) and assumed to represent the pre-mature protein. (B) The diminishment of protein maturation by tunicamycin is quantified by relating the amount of pre-mature to whole protein. There is a significant increase in pre-mature protein in Tspan17 and CD53 but not in CD9. Still, we observe at least two prominent CD9 bands that could reflect a palmitoylated- and non-palmitoylated state, as also non-glycosylated CD81[11] shows up in the palmitoylated- and non-palmitoylated state as two clearly separated bands[12]. CD9 maybe not glycosylated because both overlapping glycosylation sites are located in the small extracellular loop, that could be shielded by the large extracellular loop[13]. The result is used in Fig. 5 to assign the western blot bands to the mature and pre-mature forms of Tspan17. Values are given as means ± SD (n = 4). For statistics a two-tailed paired t-test is used (*P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001). Blots used in this analysis are shown in Fig. S25. The data analysis and illustration was performed using Fiji-ImageJ[9] (www.https://imagej.net/) and GraphPad Prism version 6.04 for Windows (www.graphpad.com), respectively.

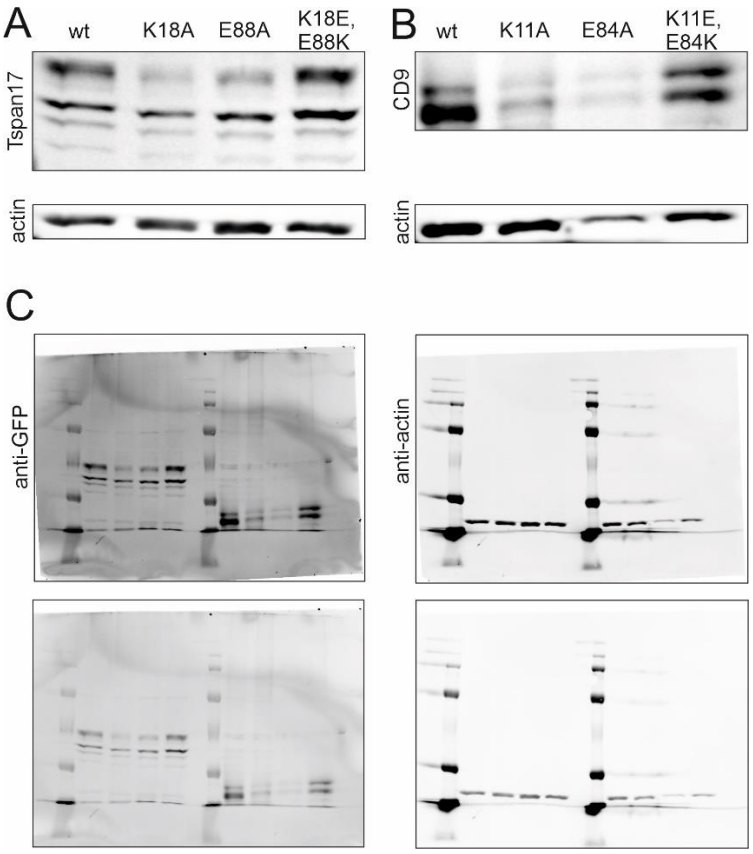


Figure S10: Expression of Tspan17 and CD9 glutamate-/lysine-mutations shortly after transient transfection

The expression of (A) mEGFP-Tspan17 constructs and (B) CD9-mEGFP constructs as indicated 5h after transient transfection (which we identified as the onset of expression) visualized by Western blot analysis. The upper panel shows the GFP signal and the lower one the actin signal. Compared to wild-type tetraspanins, though the respective actin bands are similar to control, the Tspan17-K18 and CD9-K11A bands are weaker, indicating lower expression levels already 5 h after transfection. In case of CD9, this is perhaps due to toxicity of the constructs (see Fig. S15) and not due to a maturation or trafficking defect. (C) Whole blots are shown in two different scalings. The data illustration was performed using Fiji-ImageJ[4] (https://imagej.net/).

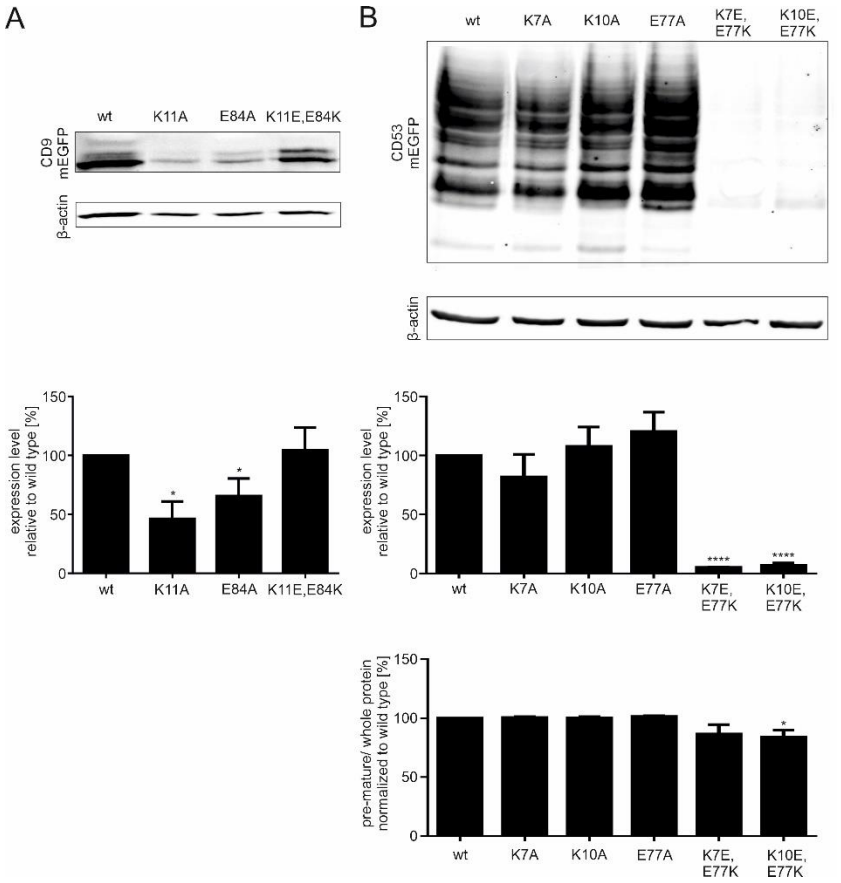


Figure S11: Expression level after disrupting/rescuing the glutamate-lysine interaction in CD9 and CD53. In GFP-tagged (A) CD9 and (B) CD53, glutamate or lysine are exchanged to alanine, or their positions are exchanged (respective mutants in Tspan17 are analysed in Fig. 5). Constructs are expressed in HepG2 cells, and expression levels are quantified by Western blot. Upper panels show representative Western blot membranes. Below, quantification of the expression levels. To this end, the GFP band intensities are related to the actin band intensity (to correct for different loading). Values are then further related to wild-type (set to 100%). The strong reduction in expression of the CD53 double mutants is unexpected but could be explained by the change in charge distribution caused by the glutamate/lysine swap. The negative effect of altered charges in the juxtamembrane region could be strong in a tetraspanin without a stabilizing salt-bridge (please note that in the crystal structure of CD53 the distance between the two amino acids actually is too long for a salt-bridge to form). Lower right panel; for CD53, additionally the pre-mature protein is related to whole protein to test for any defects in glycosylation (see Fig. S9). Values are given as means ± SD (n = 4). The statistical analysis was done employing a repeated measures ANOVA comparing each mutation with the wild type (*P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001). The full blots are shown in Fig. S21 and S22. The data analysis and illustration was performed using Fiji-ImageJ[9] (www. https://imagej.net/) and GraphPad Prism version 6.04 for Windows (www.graphpad.com), respectively. For a diminishment in Tspan17 and CD9 expression shortly after transfection see Fig. S10. CD9 mutants seems to have a cytotoxic effect (Fig. S15).

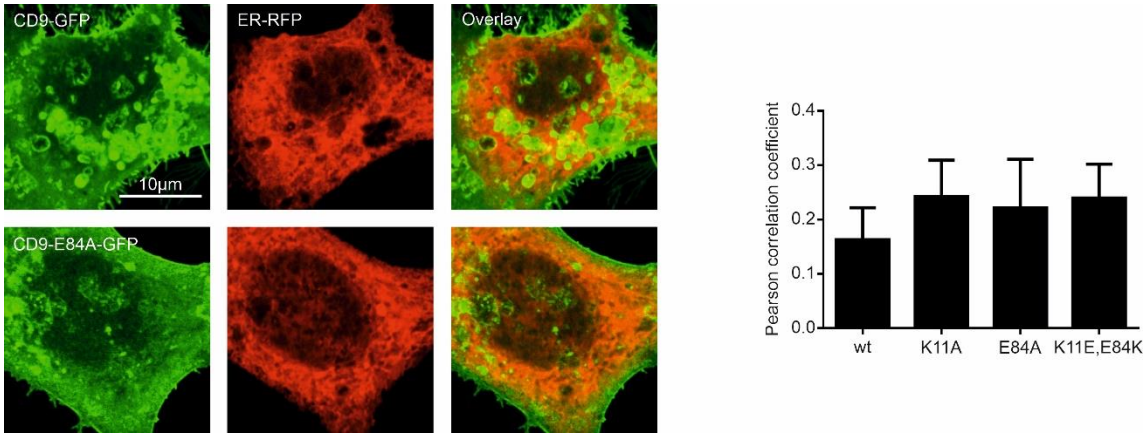


Figure S12: Co-localization of the ER with CD9 or CD9 constructs in which the salt-bridge is disrupted, or positions of the charged amino acids are interchanged

The co-localization of the ER with CD9-mEGFP and salt-bridge mutants was analyzed recording confocal images of the GFP and the RFP channel, the latter showing RFP fused to an ER marker. For quantification, the Pearson correlation coefficient was calculated. Values are given as means ± SD (n = 3; for each biological replicate 10 cells were imaged). The statistical analysis was done employing a repeated measures ANOVA comparing each mutation with the wild type (*P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001). The data analysis and illustration was performed using Fiji-ImageJ[4] (www. https://imagej.net/) and GraphPad Prism version 6.04 for Windows (www.graphpad.com), respectively.

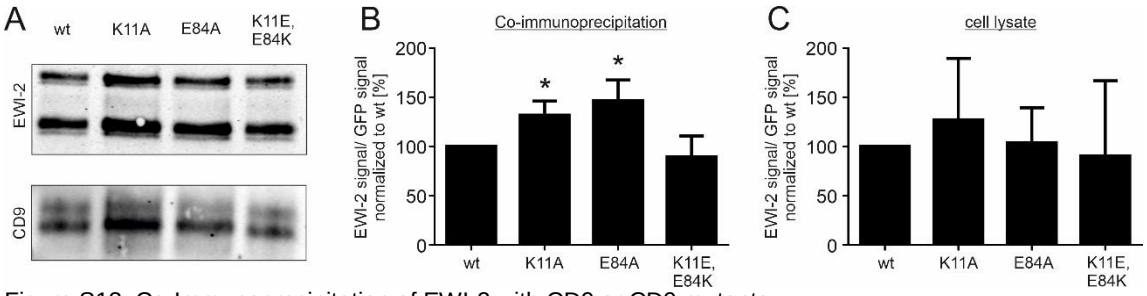


Figure S13: Co-Immunoprecipitation of EWI-2 with CD9 or CD9 mutants.

HepG2 cells express myc-tagged EWI-2 together with GFP labelled CD9 wild-type, CD9 mutants in which the salt-bridge is disrupted (K11A and E84A), or a mutant in which charges are swapped (K11E, E84K). Using an antibody that recognizes the LEL only if disulfide bridges are formed, we tested whether the disulfide-bridge forms in the mutants[14]. Albeit less efficient, all mutants were readily recognized by the antibody (see Fig. S16). For maintaining the tetraspanin-interaction network, cells were lysed with 1% CHAPS, followed by immunoprecipitation of GFP using GFP-trap beads. (A) Samples are analysed by Western blot, staining membranes for EWI-2 (top; please note that two bands are visible, one for EWI-2 and EWI-2-Wint each, both included in the quantification) and GFP (bottom; visualizing the GFP-labelled CD9 constructs as indicated). (B) Quantification of the precipitate. The ratio between the EWI-2 and GFP band intensities, normalized to wild-type (set to 100%), is used as a measure for EWI-2 pull down efficiency. Values are given as means ± SD (n = 4). (C) Relative expression levels of EWI-2 and GFP-labelled CD9 and CD9 mutants in the cell lysate. Although expression levels are highly variable, the ratio between EWI-2 and the GFP-construct is similar in the different immunoprecipitation experiments. Please note that blot #1 shown in Fig. S.26 could not be included in (C) due to low signal in the lysate. Values are given as means ± SD (n = 3). For statistical analysis, we employed a repeated measures ANOVA test comparing the mutations to the wild type (*P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001). Full blots are shown in Fig. S26. The data analysis and illustration was performed using Fiji-ImageJ[9] (www. https://imagej.net/) and GraphPad Prism version 6.04 for Windows (www.graphpad.com).

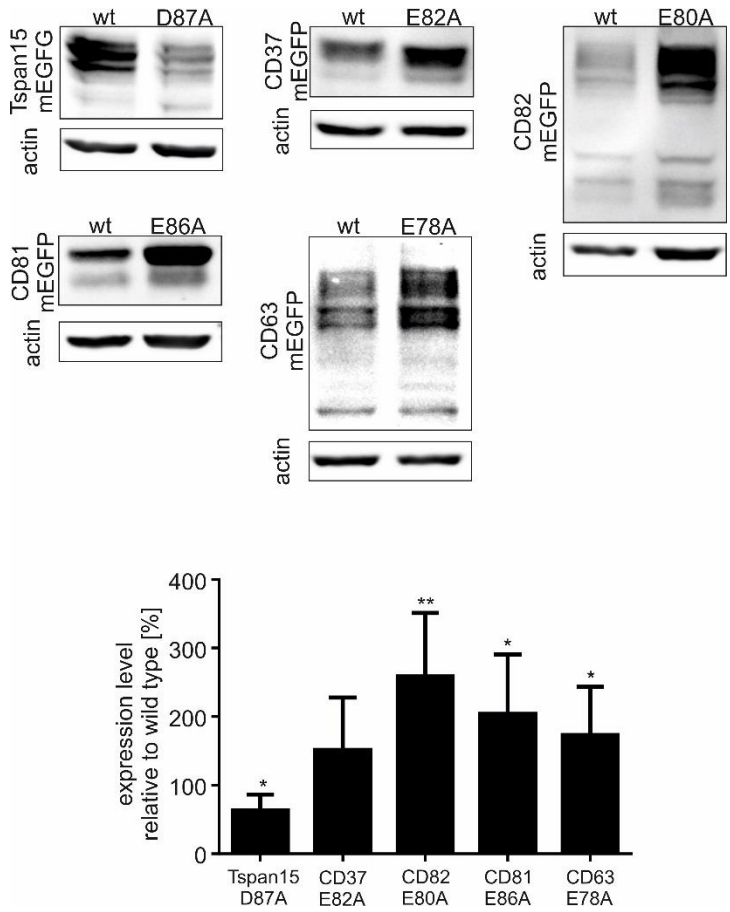


Figure S14: Mutation of SIL position 2 in Tspan15, CD37, CD82, CD81 and CD63
Same experiment as in Fig. S8, mutating only glutamate/aspartate at position 2 to alanine in Tspan15 (D87A), CD37 (E82A), CD82 (E80A), CD81 (E86A) and CD63 (E78A). Expression levels are normalized to actin and related to the respective wild-type protein (set to 100%). Values are given as means ± SD (n = 6 - 7). For statistical analysis a two-tailed paired t-test was used (*P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001). Full blots are shown in Fig. S23 and S24. The data analysis and illustration was performed using Fiji-ImageJ[9] (https://imagej.net/) and GraphPad Prism version 6.04 for Windows (www.graphpad.com), respectively.

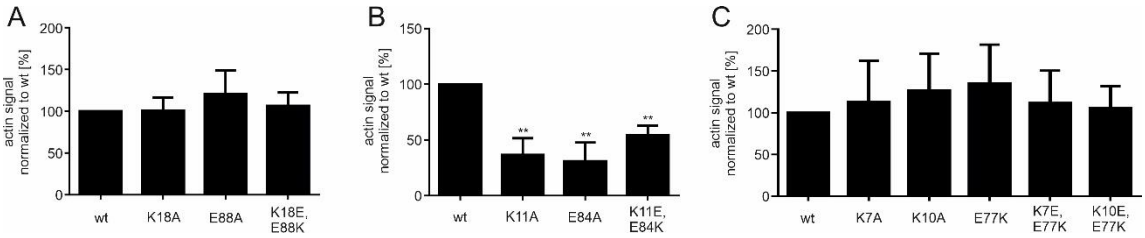


Figure S15. Cell toxicity of some mutants
Expression of (A) Tspan17, (B) CD9 or (C) CD53 or the mutants as indicated. Same experiments as in Figs. 5 and S11, showing quantification of the actin band intensities. In these experiments, a defined amount of HepG2 cells (1.8 Mio) is transfected with the GFP-labelled constructs. Under the assumption that actin levels are proportional to cell survival, we observe toxicity of the CD9 mutants. Should cytotoxicity preferentially affect strongly expressing cells, we would overestimate the reduction in expression in Fig. S11A. The statistical analysis was done by a repeated measures ANOVA test comparing the mutations to the respective wild-type protein (*P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001). Values are given as means ± SD (n = 4). The data analysis and illustration is performed using Fiji-ImageJ[4] (www. https://imagej.net/) and GraphPad Prism version 6.04 for Windows (www.graphpad.com), respectively.

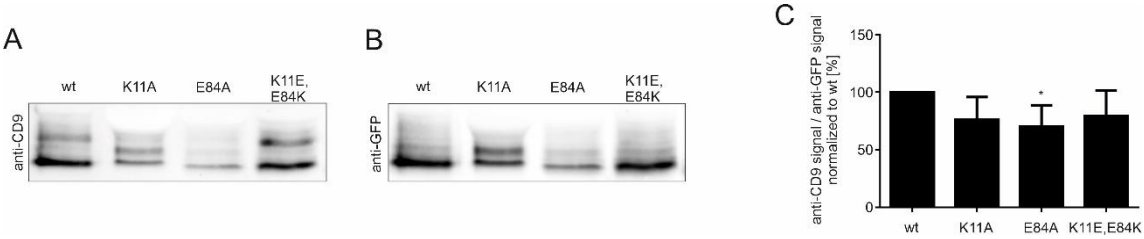


Figure S16: LEL disulfide-bond formation in CD9 mutants
HepG2 cells expressing GFP-labelled CD9 wild type or mutants are subjected to SDS-PAGE Western blot analysis under non-reducing conditions. Membranes are double stained with (A) a CD9 antibody recognizing the LEL only if a disulfide bond has formed[14] and (B) a GFP antibody. A decrease in the CD9-LEL signal relative to the GFP signal indicates absence/diminishment of the disulfide bond. The complete GFP and CD9-LEL signal (as seen in (A) and (B)) was used for this analysis. The CD9 wild type:GFP ratio is used as reference and set to 100%. Values are given as means ± SD (n = 6). Please note that a fraction of the K11A mutation runs at higher molecular weight as wild type, which however, is not reproducible (Fig. S27). Full blots are shown in Fig. S27. The data analysis and illustration was performed using Fiji-ImageJ[4] (www. https://imagej.net/) and GraphPad Prism version 6.04 for Windows (www.graphpad.com), respectively.

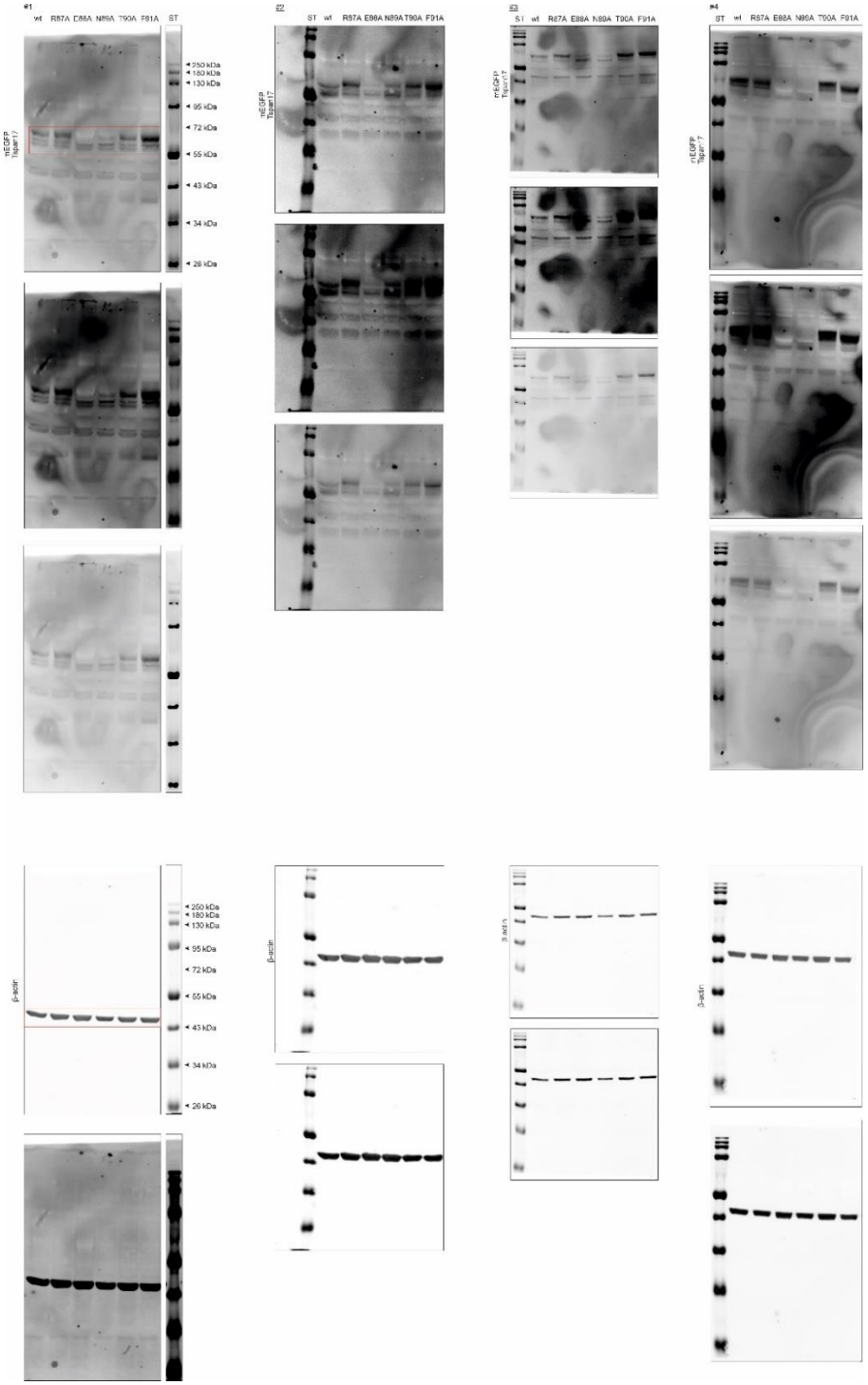


Figure S17: Full-length blots used for Figure S8.
The four blots (#1-4) for mEGFP-Tspan17 SIL core sequence mutations are depicted stained for GFP (upper panel) and for actin (lower panel). They represent the four blots used in the analysis in figure S8 (n=4). They are shown with protein standard (ST, 10-250 kDa range; NEB, #P7719S) and with three (GFP) or two (actin) different scalings to ensure clear visibility of all bands. The cutouts shown in figure S8 are highlighted by a red box in blot #1. During blot imaging the focus lay on the area framed by the protein standard, therefore all blots used for this analysis (Fig. S8) are shown. The data illustration was performed using Fiji-ImageJ[9] (https://imagej.net/).

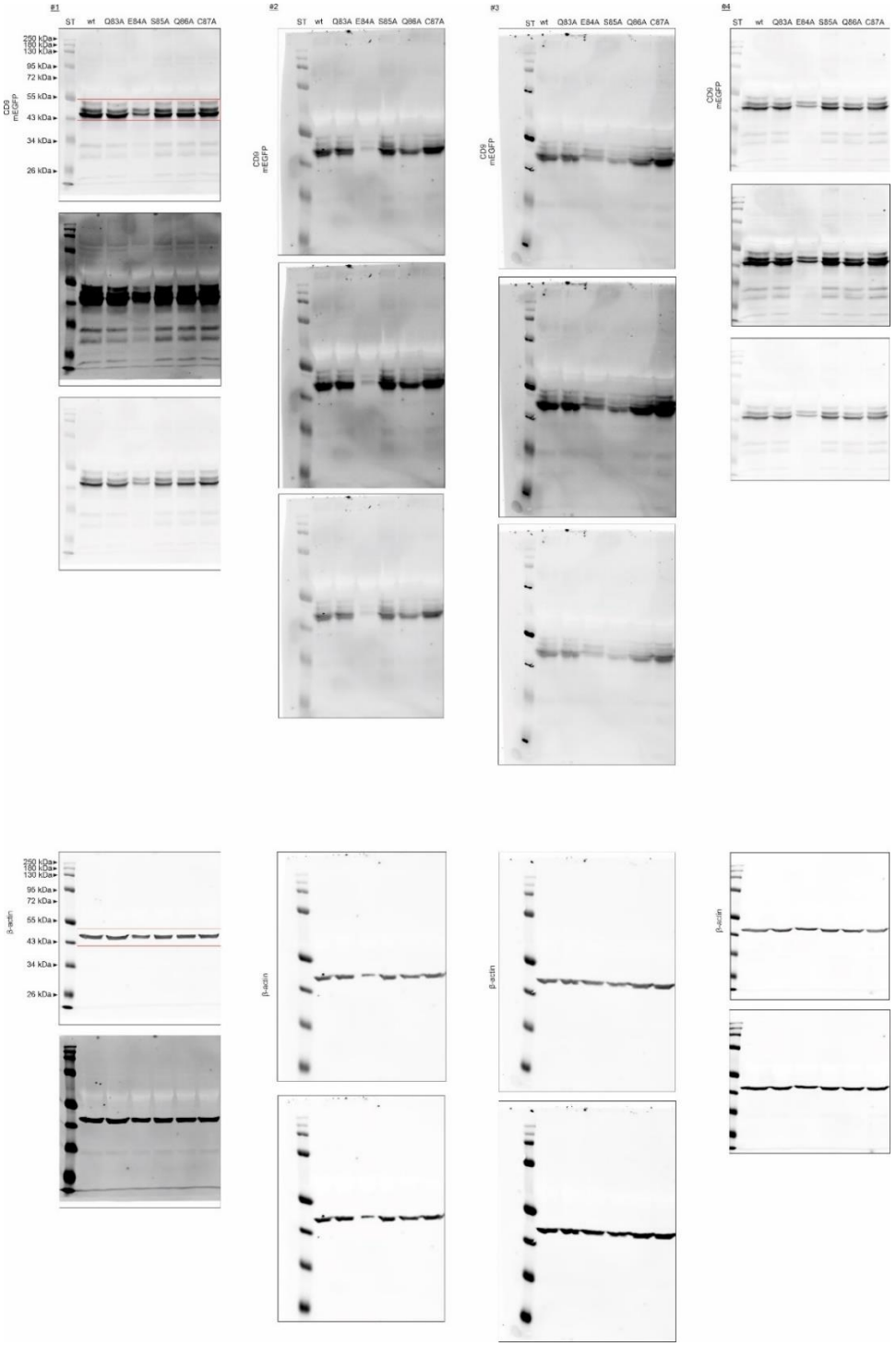


Figure S18: Full-length blots used for Figure S8.
The four blots (#1-4) for CD9-mEGFP SIL core sequence mutations are depicted stained for GFP (upper panel) and for actin (lower panel). They represent the four blots used in the analysis in figure S8 (n=4). They are shown with protein standard (ST, 10-250 kDa range; NEB, #P7719S) and with three (GFP) or two (actin) different scalings to ensure clear visibility of all bands. The cutouts shown in figure S8 are highlighted by a red box in blot #1. During blot imaging the focus lay on the area framed by the protein standard, therefore all blots used for this analysis (Fig. S8) are shown. The data illustration was performed using Fiji-ImageJ[9] (https://imagej.net/).

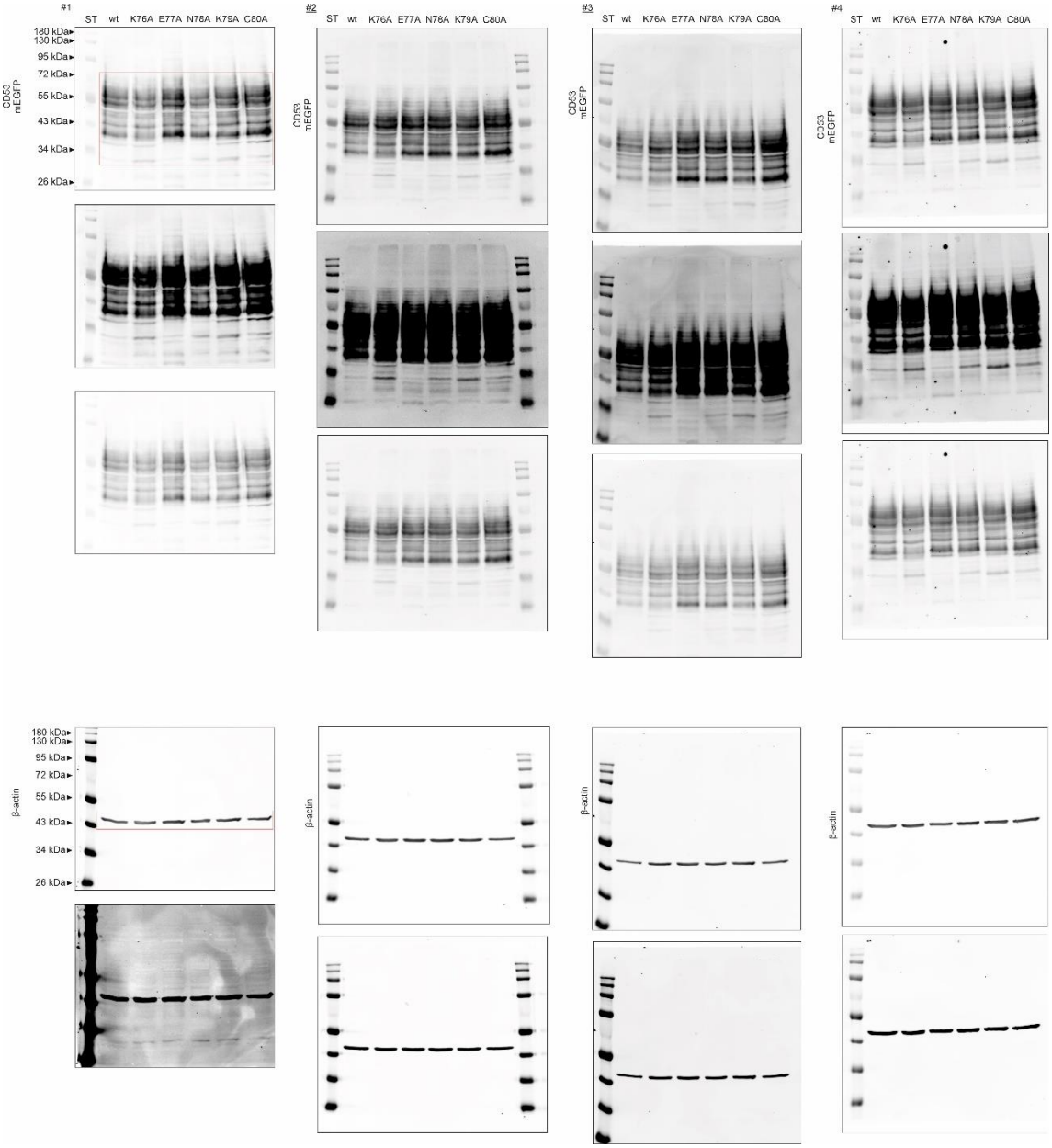


Figure S19: Full-length blots used for Figure S8.
The four blots (#1-4) for CD53-mEGFP SIL core sequence mutations are depicted stained for GFP (upper panel) and for actin (lower panel). They represent the four blots used in the analysis in figure S8 (n=4). They are shown with protein standard (ST, 10-250 kDa range; NEB, #P7719S) and with three (GFP) or two (actin) different scalings to ensure clear visibility of all bands. The cutouts shown in figure S8 are highlighted by a red box in blot #1. During blot imaging the focus lay on the area framed by the protein standard, therefore all blots used for this analysis (Fig. S8) are shown. The data illustration was performed using Fiji-ImageJ[9] (https://imagej.net/).

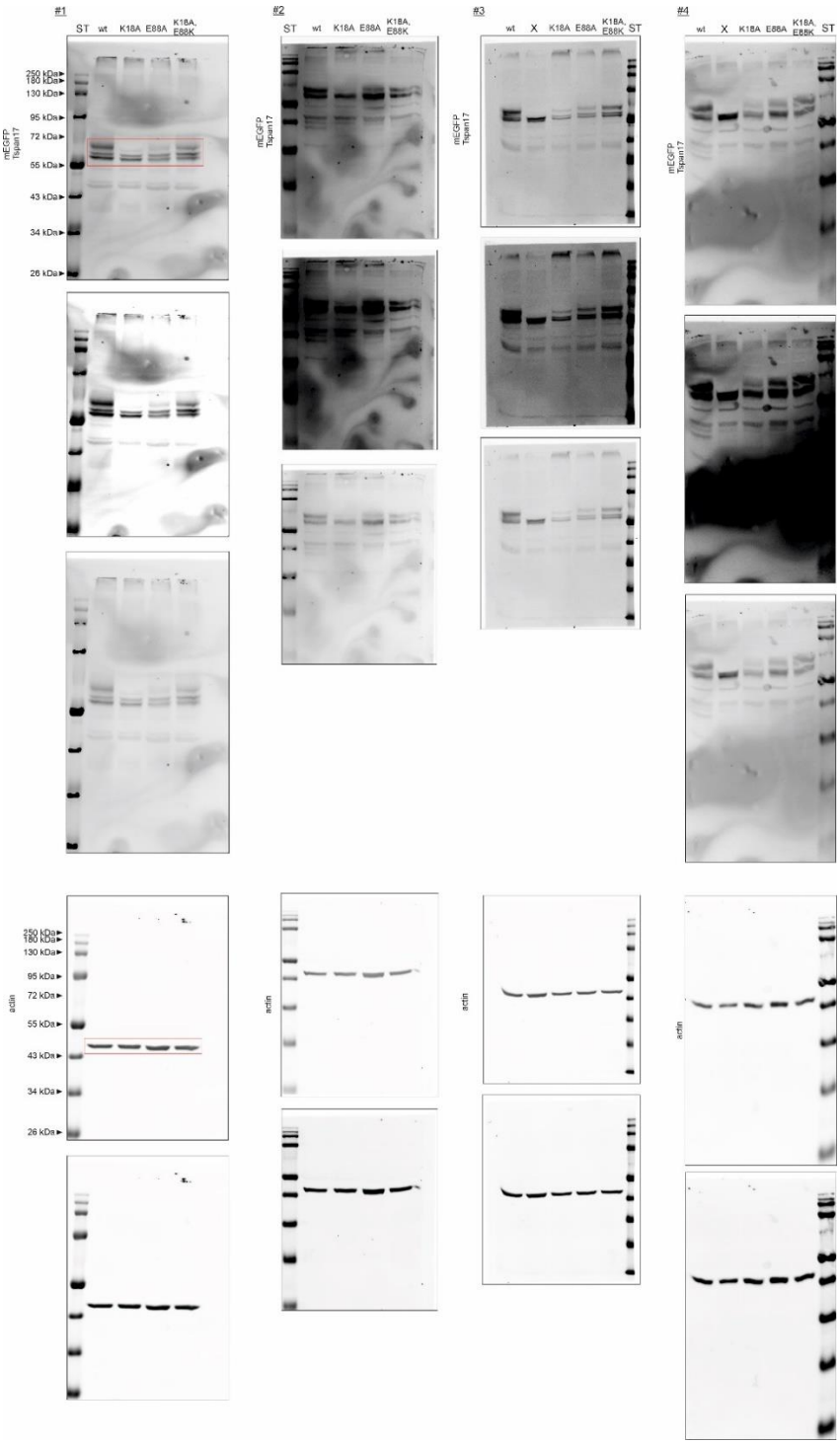


Figure S20: Full-length blots used in Figure 5
The four blots (#1-4) for mEGFP-Tspan17 salt-bridge mutations and rescue mutation are depicted stained for GFP (upper panel) and for actin (lower panel). They represent the four blots used in the analysis in figure 5 (n=4). They are shown with protein standard (ST, 10-250 kDa range; NEB, #P7719S) and with three (GFP) or two (actin) different scalings to ensure clear visibility of all bands. The cutouts shown in figure 5 are highlighted by a red box in blot #1. During blot imaging the focus lay on the area framed by the protein standard, therefore all blots used for this analysis (Fig. 5) are shown. The data illustration was performed using Fiji-ImageJ[9] (https://imagej.net/).

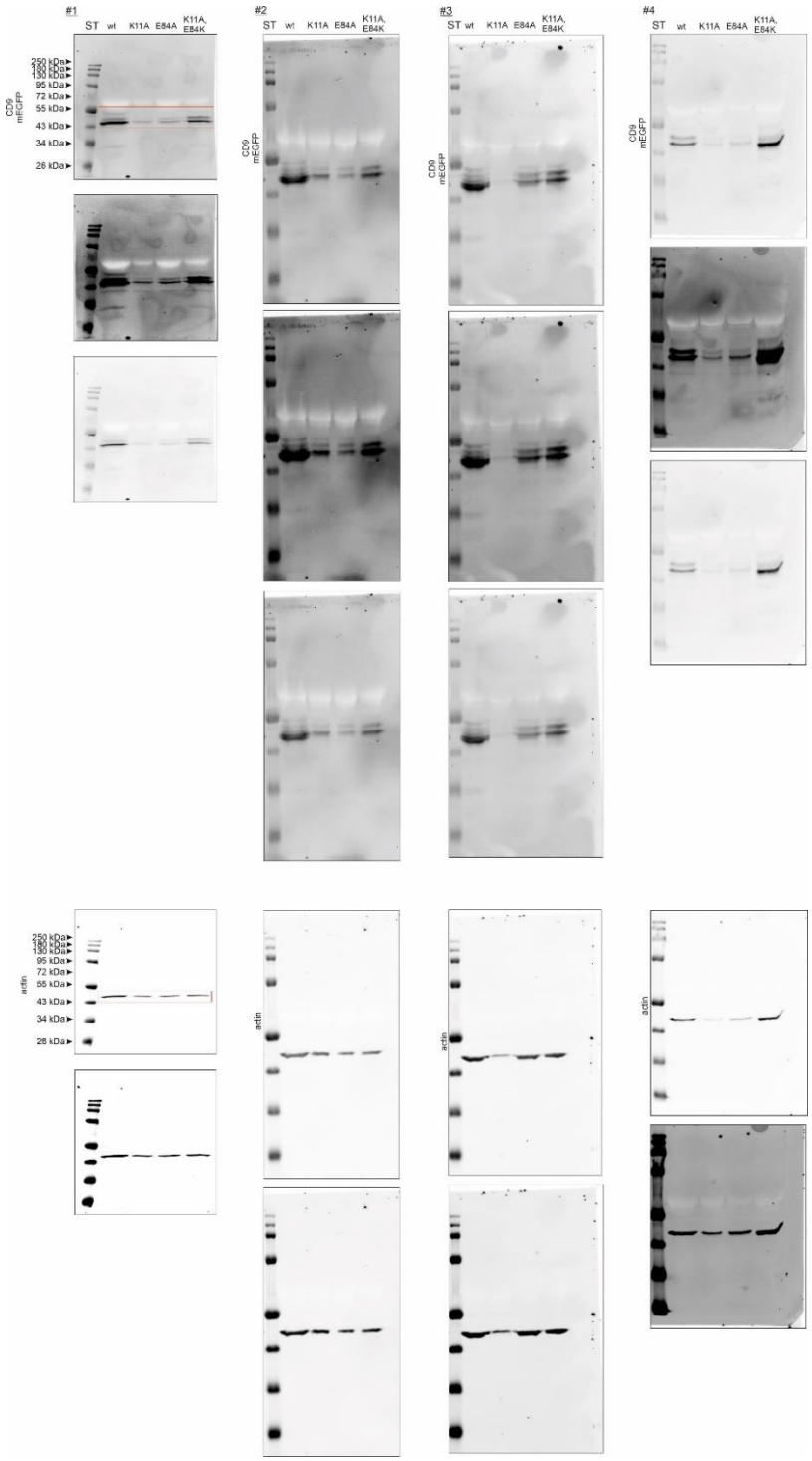


Figure S21: Full-length blots used in Figure S11
The four blots (#1-4) for CD9-mEGFP salt-bridge mutations and rescue mutation are depicted stained for GFP (upper panel) and for actin (lower panel). They represent the four blots used in the analysis in figure S11 (n=4). They are shown with protein standard (ST, 10-250 kDa range; NEB, #P7719S) and with three (GFP) or two (actin) different scalings to ensure clear visibility of all bands. The cutouts shown in figure S11A are highlighted by a red box in blot #1. During blot imaging the focus lay on the area framed by the protein standard, therefore all blots used for this analysis (Fig. S11) are shown. The data illustration was performed using Fiji-ImageJ[4] (https://imagej.net/).

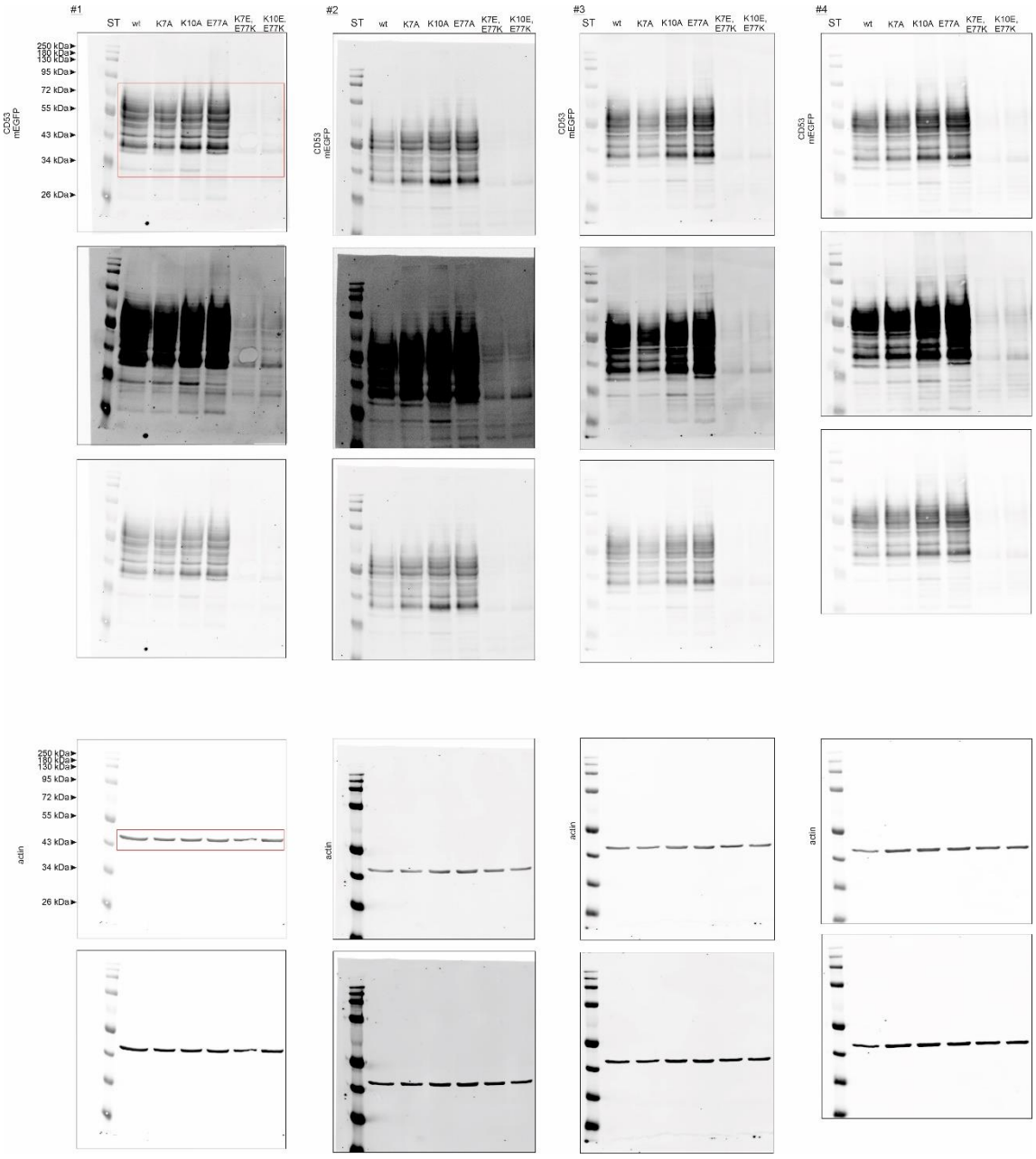


Figure S22: Full-length blots used in Figure S11
The four blots (#1-4) for CD53-mEGFP salt-bridge mutations and rescue mutation are depicted stained for GFP (upper panel) and for actin (lower panel). They represent the four blots used in the analysis in figure S11 (n=4). They are shown with protein standard (ST, 10-250 kDa range; NEB, #P7719S) and with three (GFP) or two (actin) different scalings to ensure clear visibility of all bands. The cutouts shown in figure S11B are highlighted by a red box in blot #1. During blot imaging the focus lay on the area framed by the protein standard, therefore all blots used for this analysis (Fig. S11) are shown. The data illustration was performed using Fiji-ImageJ[4] (https://imagej.net/).

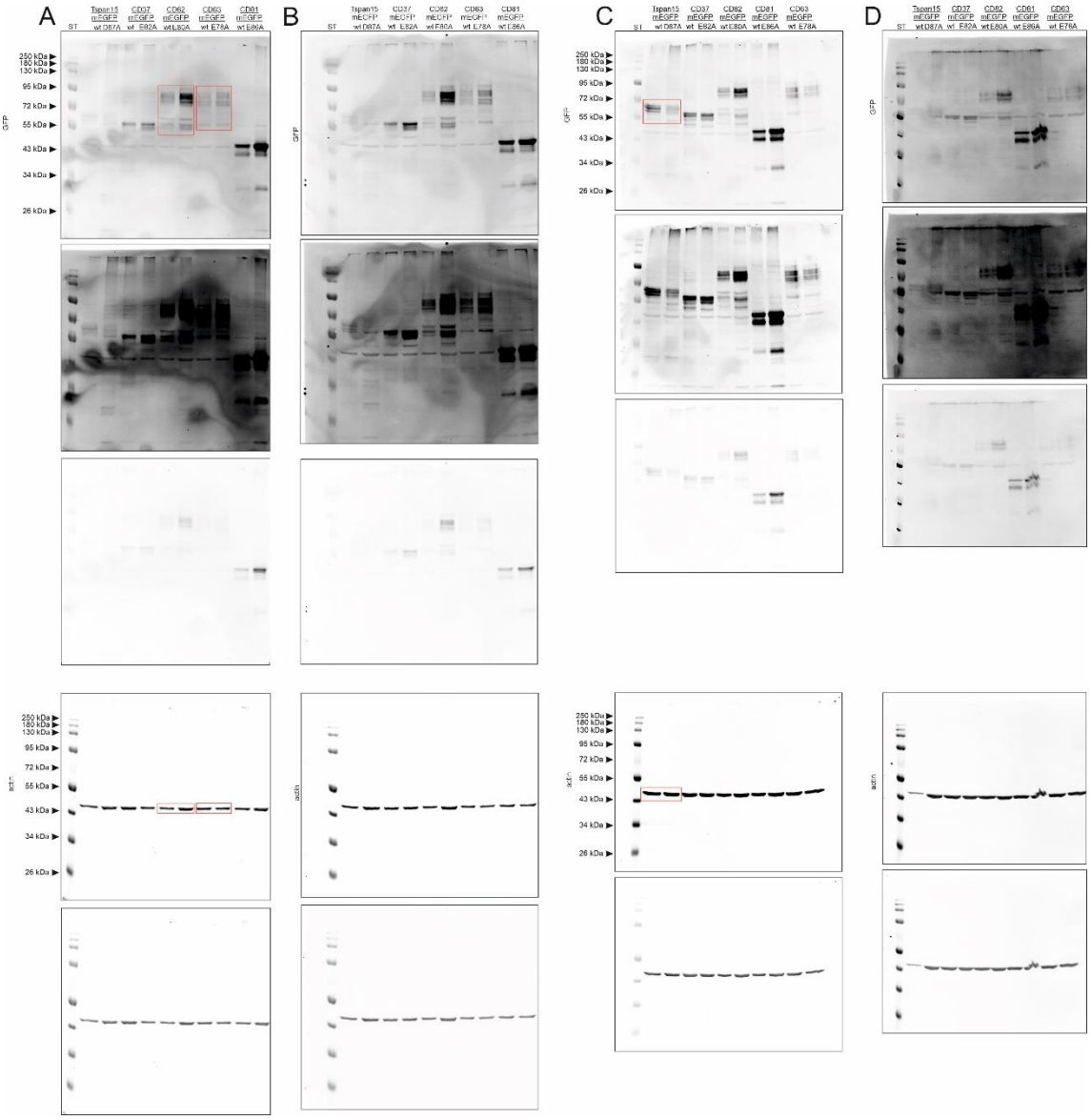


Figure S23: Full-length blots used in Figure S14.
The blots for figure S14 were divided onto two figures (Fig. S23 and S24). The seven blots (A-G, for E-G see Fig S24) for Tspan15, CD37, CD82, CD81 and CD63 are depicted stained for GFP (upper panel) and for actin (lower panel). They are shown with protein standard (ST, 10-250 kDa range; NEB, #P7719S) and with three (GFP) or two (actin) different scalings to ensure clear visibility of all bands. The cutouts used in figure S14 are highlighted by a red box. During blot imaging the focus lay on the area framed by the protein standard, therefore all blots used for this analysis (Fig. S14) are shown. The data illustration was performed using Fiji-ImageJ[9] (https://imagej.net/).

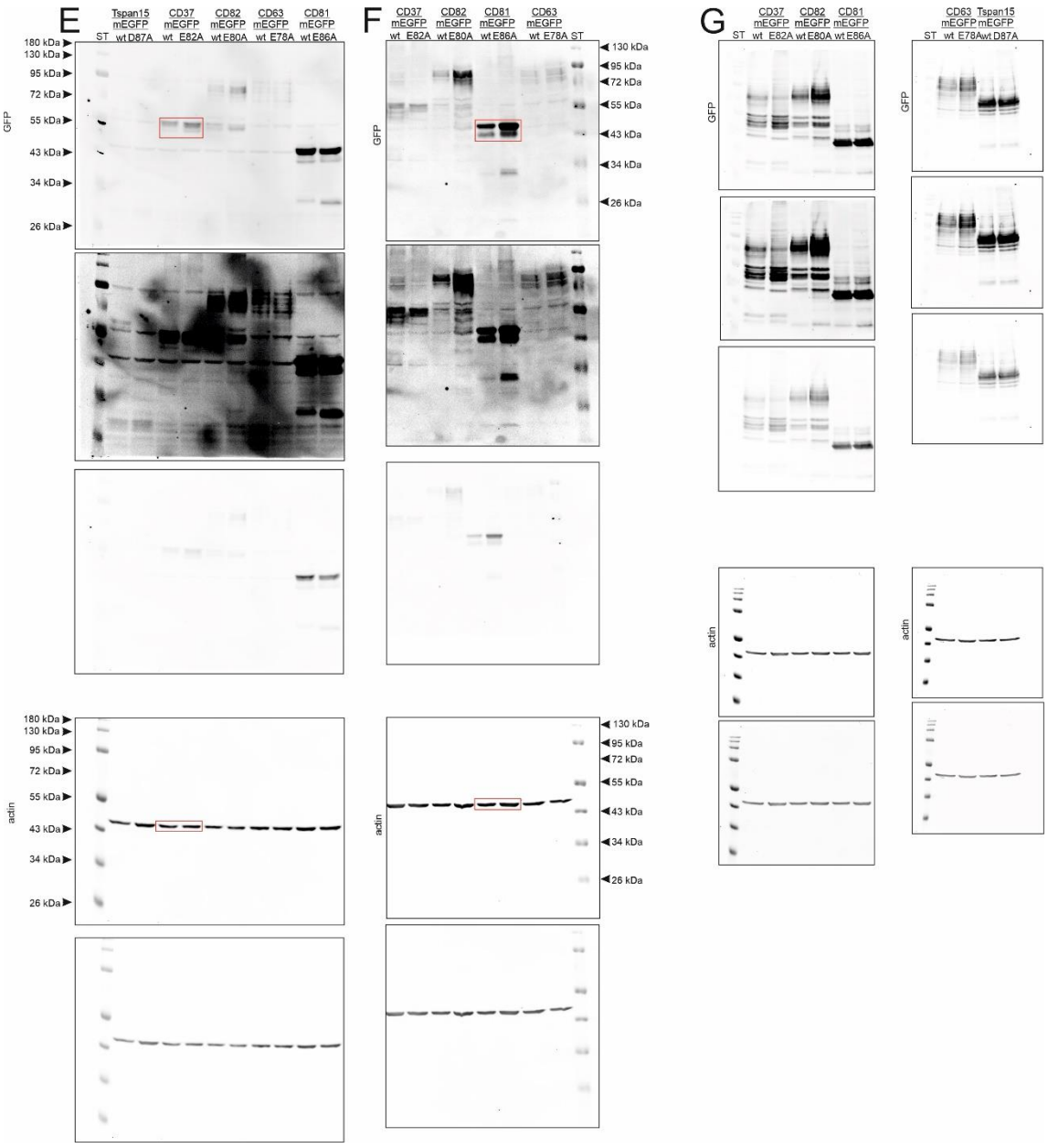


Figure S24: Full-length blots used in Figure S14.
The blots for figure S14 were divided onto two figures (Fig. S23 and S24). The seven blots (A-G, For A-D see Fig S23) for Tspan15, CD37, CD82, CD81 and CD63 are depicted stained for GFP (upper panel) and for actin (lower panel). They are shown with protein standard (ST, 10-250 kDa range; NEB, #P7719S) and with three (GFP) or two (actin) different scalings to ensure clear visibility of all bands. For Tspan15 there were only 6 samples analysed (n=6) which is the reason Tspan15 is missing on blot (F). The cutouts used in figure S14 are highlighted by a red box. During blot imaging the focus lay on the area framed by the protein standard, therefore all blots used for this analysis (Fig. S14) are shown. The data illustration was performed using Fiji-ImageJ[9] (https://imagej.net/).

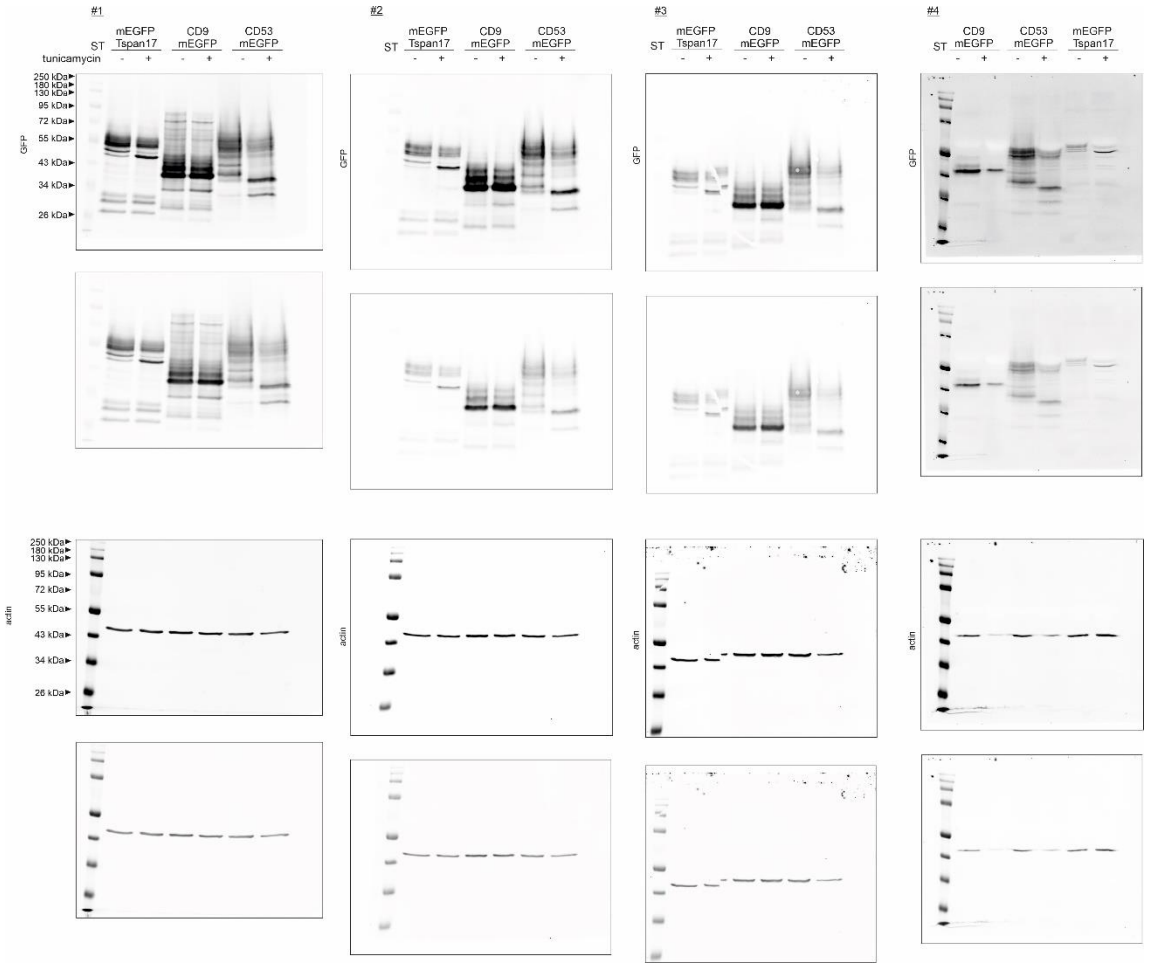


Figure S25: All blots used for the analysis in Figure S9
The four blots (#1-4) were stained for GFP (upper panel) and for actin (lower panel) and show the effect of tunicamycin treatment on the expression of mEGFP-Tspan17, CD9-mEGFP and CD53-mEGFP. They are shown with a protein standard (ST, 10-250 kDa range; NEB, #P7719S) and with two different scalings to ensure clear visibility of all bands. During blot imaging the focus lay on the area framed by the protein standard, therefore all blots used for this analysis (Fig. S9) are shown. The data illustration was performed using Fiji-ImageJ[9] (https://imagej.net/).

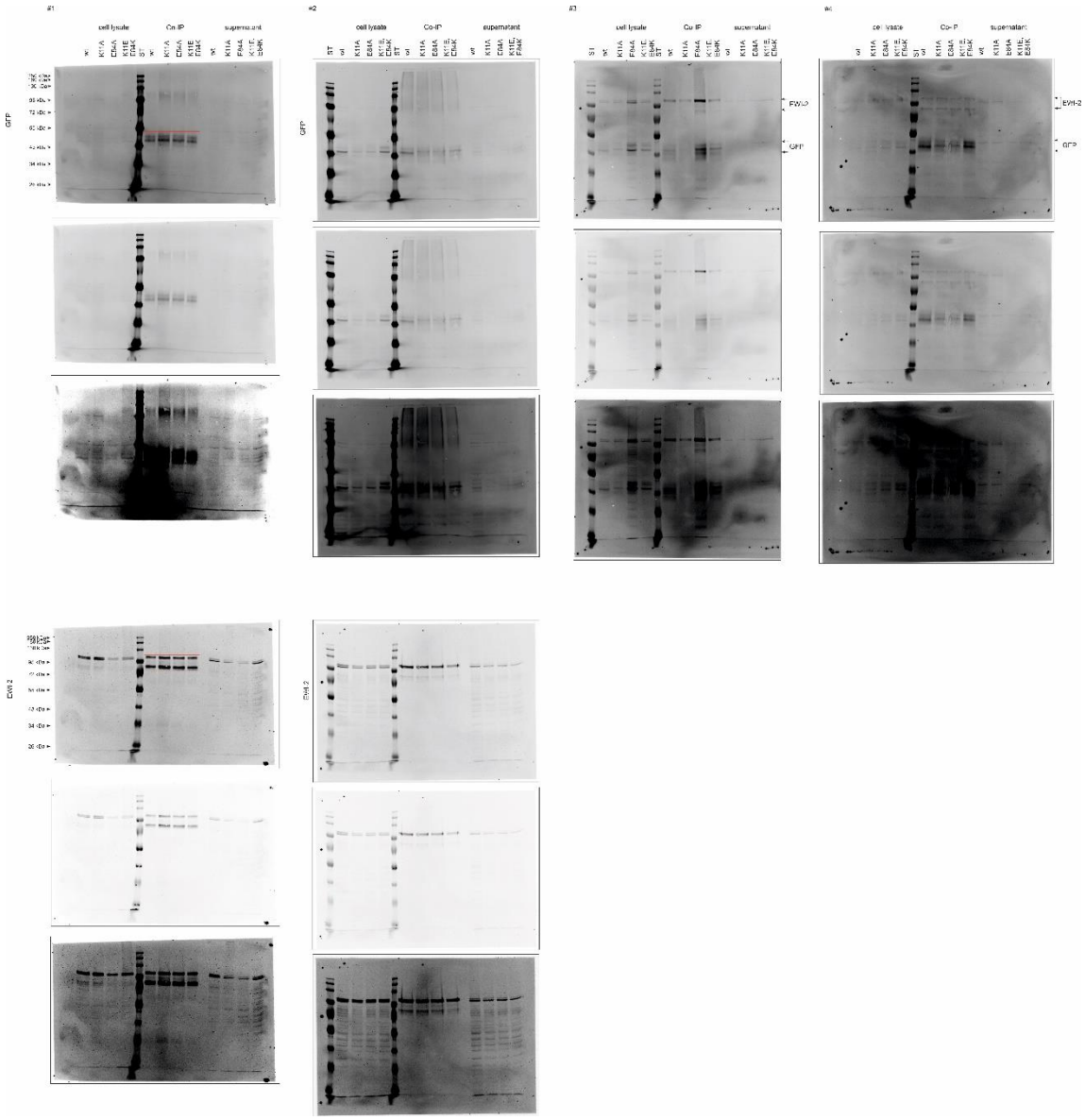


Figure S26: Full blot images used in Figure S13.
The four blots (#1-4) show the cell lysate, Co-Immunoprecipitation (Co-IP) and supernatant of the CD9-mEGFP salt-bridge mutations. The immunoprecipitation was stained for GFP (upper panel) and EWI-2 (lower panel). The EWI-2 molecule (upper band) is processed to EWI-2-Wint (lower band). The blots #3 and #4 show the GFP and EWI-2 staining in the same image, because they were stained using a secondary antibody attached to the same fluorescence dye (CW800). Therefore, the bands corresponding to GFP and EWI-2 staining are highlighted with arrows at the side of the blot. EWI-2-myc was transiently transfected and was stained via a EWI-2 antibody, which also detected the endogen levels and explains the double band. The blots are shown with a protein standard (ST, 10-250 kDa range; NEB, #P7719S) and with different scalings to ensure clear visibility of all bands. During blot imaging the focus lay on the area framed by the protein standard, therefore all blots used for this analysis (Fig. S13) are shown. The data illustration was performed using Fiji-ImageJ[9] (https://imagej.net/).

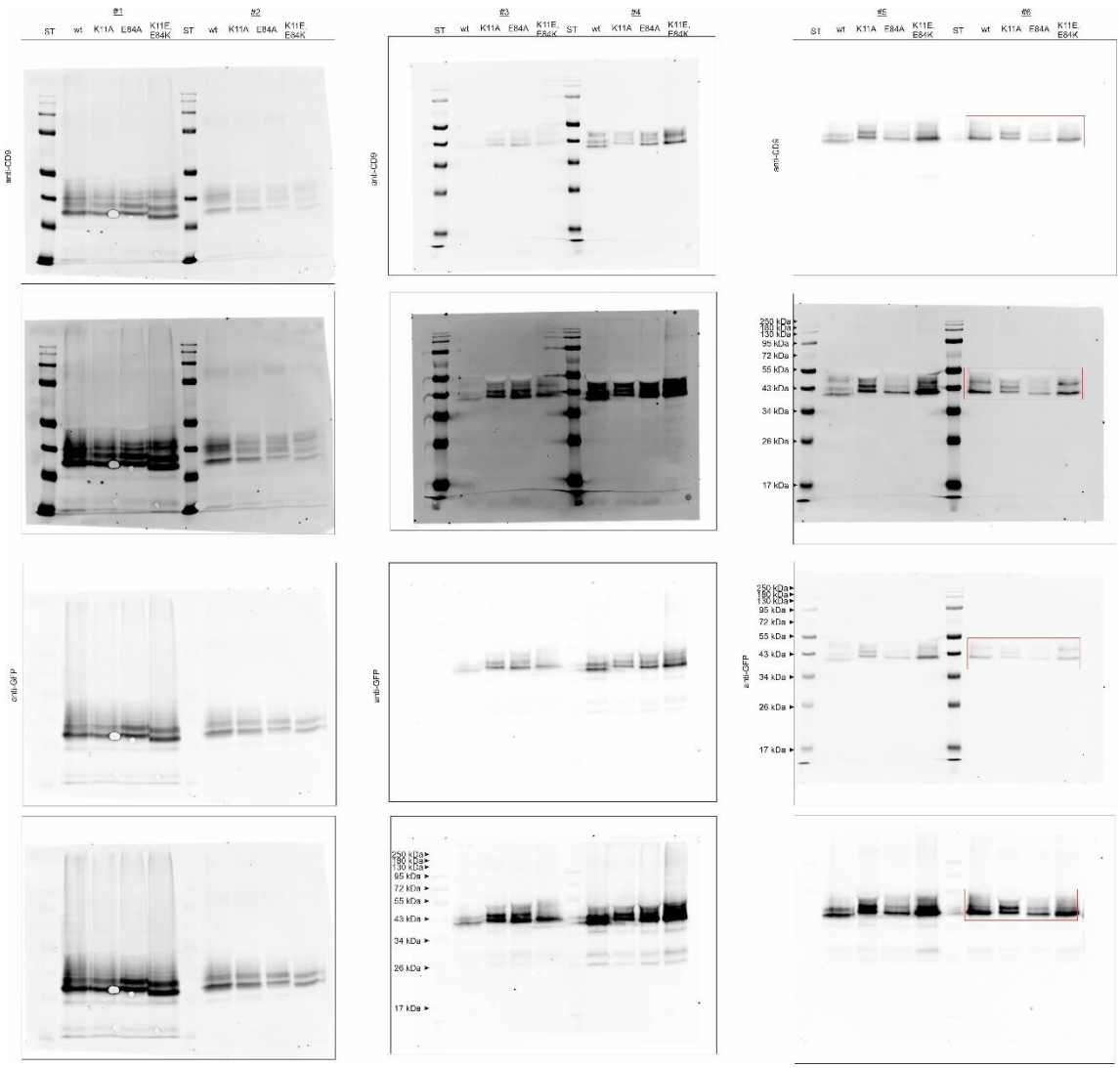


Figure S27: Full blot images used in Figure S16
The whole blots (#1-6 are the n=6) used for this analysis are shown in two different scalings and the cutout shown under (A) is highlighted by a red box in sample #6. The blots are shown with a protein standard (ST, 10-250 kDa range; NEB, #P7719S). The statistical analysis was done employing a repeated measures ANOVA comparing each mutation to the wild type (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$). The data analysis and illustration was performed using GraphPad Prism version 6.04 for Windows (www.graphpad.com) and Fiji-ImageJ[4] (https://imagej.net/).

References

1. Dagona, A. G. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* (1999).

2. Drozdetskiy, A., Cole, C., Procter, J. & Barton, G. J. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* **43**, W389–W394 (2015).

3. Kyte, J. *Structure in Protein Chemistry*. (Garland Science, 2006).

4. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: A Sequence Logo Generator. 3.

5. Eisenberg, D. The discovery of the α-helix and β-sheet, the principal structural features of proteins. *Proc. Natl. Acad. Sci.* **100**, 11207–11210 (2003).

6. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).

7. Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175–182 (1998).

8. Gautier, R., Douguet, D., Antonny, B. & Drin, G. HELIQUEST: a web server to screen sequences with specific alpha-helical properties. *Bioinforma. Oxf. Engl.* **24**, 2101–2102 (2008).

9. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).

10. Gupta, R. & Brunak, S. Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* 310–322 (2002).

11. Takayama, H., Chelikani, P., Reeves, P. J., Zhang, S. & Khorana, H. G. High-Level Expression, Single-Step Immunoaffinity Purification and Characterization of Human Tetraspanin Membrane Protein CD81. *PLoS ONE* **3**, e2314 (2008).

12. Zhu, Y.-Z. *et al.* Significance of palmitoylation of CD81 on its association with tetraspanin-enriched microdomains and mediating hepatitis C virus cell entry. *Virology* **429**, 112–123 (2012).

13. Seigneuret, M. Complete Predicted Three-Dimensional Structure of the Facilitator Transmembrane Protein and Hepatitis C Virus Receptor CD81: Conserved and Variable Structural Domains in the Tetraspanin Superfamily. *Biophys. J.* **90**, 212–227 (2006).

14. Cannon, K. S. & Cresswell, P. Quality control of transmembrane domain assembly in the tetraspanin CD82. *EMBO J.* **20**, 2443–2453 (2001).