A unified evolutionary origin for SecY and YidC

Aaron J. O. Lewis and Ramanujan S. Hegde

MRC Laboratory of Molecular Biology
Francis Crick Avenue, Cambridge, CB2 0QH, UK

Correspondence: aaron.ohare.lewis@gmail.com or rhegde@mrc-lmb.cam.ac.uk

Abstract

Cells use transporters to move protein across membranes, but the most ancient transporters' origins are unknown. Here, we analyse the protein-conducting channel SecY and deduce a plausible path to its evolution. We find that each of its pseudosymmetric halves consists of a three-helix bundle interrupted by a two-helix hairpin. Unexpectedly, we identify this same motif in the YidC family of membrane protein biogenesis factors, which is similarly ancient as SecY. In YidC, the two-helix hairpin faces the cytosol and facilitates substrate delivery, whereas in SecY it forms the substrate-binding transmembrane helices of the lateral gate. We propose that SecY originated as a YidC homolog which formed a channel by juxtaposing two hydrophilic grooves in an antiparallel homodimer. Archaeal and eukaryotic YidC family members have repurposed this interface to heterodimerise with conserved partners. Unification of the two ancient membrane protein biogenesis factors reconstructs a key step in the evolution of cells.

Introduction

By the time of the last universal common ancestor (cenancestor), cells had already evolved a hydrophobic membrane and integral membrane proteins (IMPs) which carried out core metabolic functions (Lombard et al., 2012). Among those IMPs was SecY, a protein-conducting channel (Park and Rapoport, 2012). As is typical for channels, SecY (termed Sec61 in eukaryotes) catalyses the translocation of hydrophilic substrates across the hydrophobic membrane by creating a conducive hydrophilic environment inside the membrane. The substrates which it translocates are secretory proteins and the extracytoplasmic segments of IMPs.

SecY requires that its hydrophilic translocation substrates be connected to a hydrophobic α-helix (von Heijne, 1985; Krogh et al., 2001; Petersen et al., 2011). These hydrophobic helices are essential because they serve as signals which open the SecY channel (Jungnickel and Rapoport, 1995; Li et al., 2016; Voorhees and Hegde, 2016). SecY is comprised of two rigid halves (Van den Berg et al., 2004), which open like a clamshell when a helix binds to the lipid interface between them (Figure 1a). Spreading apart the clamshell destabilises a plug which sits between the halves, opening a hydrophilic pore that spans the membrane. Binding at this site also threads one of the signal's hydrophilic flanking regions through the hydrophilic pore, thereby initiating its translocation.

The site between SecY's halves where signals bind is called the lateral gate. After binding and initiating translocation, the hydrophobic signal can diffuse away from the lateral gate into the surrounding hydrophobic membrane. Many signals, particularly those at the N-terminus of secretory proteins, are then cleaved off by signal peptidase, a membrane-anchored protease whose active site resides on the extracytoplasmic side of the membrane (Paetzel et al., 2002). Longer and more hydrophobic signals that are not cleaved serve as the transmembrane helices (TMHs) of IMPs (White and von Heijne, 2005).

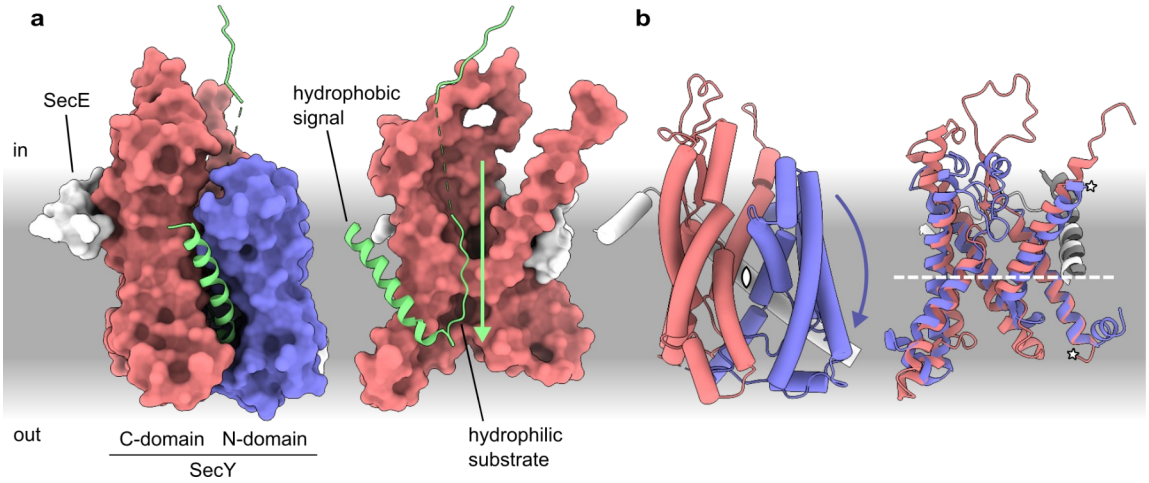

Figure 1. Structure and pseudosymmetry of the SecY protein channel. **a** Left: Structure of the channel engaged by a secretory substrate, represented by solvent-excluded surfaces computed for each domain (*Geobacillus thermodenitrificans* SecYE/proOmpA, Protein Data Bank ID [PDB] 6itc, Ma et al., 2019. The SecA translocase, which is present in the PDB model bound to the substrate and drives unidirectionality in the prokaryotic post-translational pathway, is not shown). Right: Rotated SecYE/substrate model with the SecY N-domain and SecE C-terminus cut away. **b** Pseudosymmetry of the N- and C-domains. Left: Tube representation of the SecYE model shown in **a,** with the pseudo-C_2 symmetry axis denoted by a pointed oval. A curved blue arrow indicates that the N-domain was rotated around the pseudosymmetry axis to obtain the alignment shown at right. Right: Rotated SecY model in ribbon representation, with the N-domain aligned to the C-domain, and SecE divided where it intersects the symmetry axis into N-terminal (white) and C-terminal (grey) segments. A dashed white line indicates the pseudo-C_2 symmetry axis. Stars indicate the point where N- and C-domains were split.

SecY is the only universally conserved transporter for protein secretion. There is however a second universally conserved protein transporter that is specialised for IMP integration: YidC (Hennon et al., 2015). Unlike SecY, YidC is asymmetric (Kumazaki et al., 2014a), and so widely divergent across species that its universality was only recently appreciated (Anghel et al., 2017). Nonetheless a conserved core of three TMHs is evident in all the YidC homologs of known structure: YidC of the bacterial plasma membrane (Kumazaki et al., 2014b, 2014a), Ylp1 of the archaeal plasma membrane (Borowska et al., 2015), and TMCO1, EMC3, and GET1 of the eukaryotic endoplasmic reticulum (ER) (Anghel et al., 2017). The chloroplast and mitochondrial inner membranes also contain YidC homologs, Alb3 and Oxa1, respectively (Bauer et al., 1994; Bonnefoy et al., 1994; Sundberg et al., 1997). Thus, each membrane that is topologically and functionally equivalent to the plasma membrane of the cenancestor contains at least one YidC homolog.

Like SecY, YidC facilitates the integration of IMPs by translocating their extracytoplasmic segments across the membrane (Hell et al., 1998; Samuelson et al., 2000). Unlike SecY substrates however, YidC substrates are limited in the length of translocated polypeptide, typically to less than 30 amino acids (Shanmugam et al., 2019). This limitation may be due to its lack of a membrane-spanning hydrophilic pore; instead, structures of YidC show a membrane-exposed hydrophilic groove that only penetrates partway into the membrane (Kumazaki et al., 2014a).

The halves of the SecY clamshell are structurally similar and related by a two-fold rotational (C_2) pseudosymmetry axis which bisects the membrane plane (Van den Berg et al. 2004; Figure 1b). Such pseudosymmetry is common among membrane proteins, and is believed to arise when the gene encoding an asymmetric progenitor undergoes duplication and fusion (Forrest, 2015). Channels are particularly likely to have a membrane-bisecting C_2 axis of structural symmetry because they have the same axis of functional symmetry: they passively facilitate bi-directional diffusion across the membrane. Although transport through SecY is usually unidirectional, it is no

exception to this symmetry rule; polypeptides can indeed slide through it bidirectionally (Ooi and Weiss, 1992), with unidirectionality arising from other factors (Erlandson et al., 2008; Matlack et al., 1999).

The ubiquity and essentiality of the SecY channel motivated us to investigate how it might have evolved. By focusing on structural elements which were conserved between domains and across species, we identified a core motif of three TMHs which bury a hydrophilic patch inside the membrane. Speculating that this three-TMH motif may resemble an evolutionary ancestor, we compared it to cenancestor membrane proteins and detected a previously unrecognised similarity with YidC.

Drawing on the extensive functional literature about SecY and YidC, we analyse their structural similarities and differences in terms of functional consequences. Based on this analysis, we propose that in a parsimonious model, SecY evolved from a dimeric YidC homologue by gene duplication and fusion. One prediction of this model is that YidC should conserve a tendency to form dimers via the same interface as the SecY progenitor, and indeed we discover novel heterodimers formed via this interface by archaeal and eukaryotic YidC. We discuss the implications of this model for the evolution of YidC itself, and other components of the general secretory pathway.

Results

Conserved ancestral features in the SecY structure

Although the N- and C-domains of SecY have significantly diverged, the architecture of their five TMHs is conserved (Figure 1b; Figure 2a; Van den Berg et al., 2004). The conservation of these five TMHs between the N- and C-domains indicates that the same five TMHs were also present in their last common ancestor. This last pre-duplication ancestor of the SecY domains we term proto-SecY. To facilitate comparisons, we label these five consensus helices H1-H5 (Figure 2a). A prefix of N or C is used when referring to a specific instance of a consensus element in the N- or C-domain of SecY. For example, TM6 of SecY is labelled C.H1 because it is located in the C-domain and corresponds to H1 of proto-SecY, as does TM1 (N.H1) in the N-domain. Flanking and intervening segments are labelled using lower-case references to the nearest consensus elements. For example, the ribosome-binding loop between C.H1 and C.H2 is C.h1h2. The N-terminal peripheral helix of each domain, which we argue later was probably also conserved in proto-SecY, is named H0.

The nearest non-duplicated ancestor of a channel is rarely detectable (Hennerdal et al., 2010), presumably because it is made redundant by the duplicated, fused form. Indeed, no obvious candidates for proto-SecY are evident in sequence or structure databases. We speculated that an even earlier ancestor not redundant to SecY might persist in a more divergent form that nonetheless retains recognisable similarity with SecY. To facilitate a search for this more distant ancestor, we considered sub-domains within the five-TMH core of proto-SecY that might represent a more widely conserved precursor structure. Our analysis was guided by the fact that the folding pathway of a protein imposes constraints on how it is subsequently elaborated during evolution.

In this context, we noted that H1, which is synthesized and inserted into the membrane first, makes negligible contacts with the tightly packed H2/3 hairpin (Figure 2b). Instead, H1 packs against H4 and H5, both of which are located between H1 and the H2/3 hairpin. This arrangement, where sequential TMHs are separated, is highly unusual in IMPs (Bowie, 1997), suggesting that it represents a divergence from the ancestral fold. Because H4/5 passes between H1 and H2/3 from

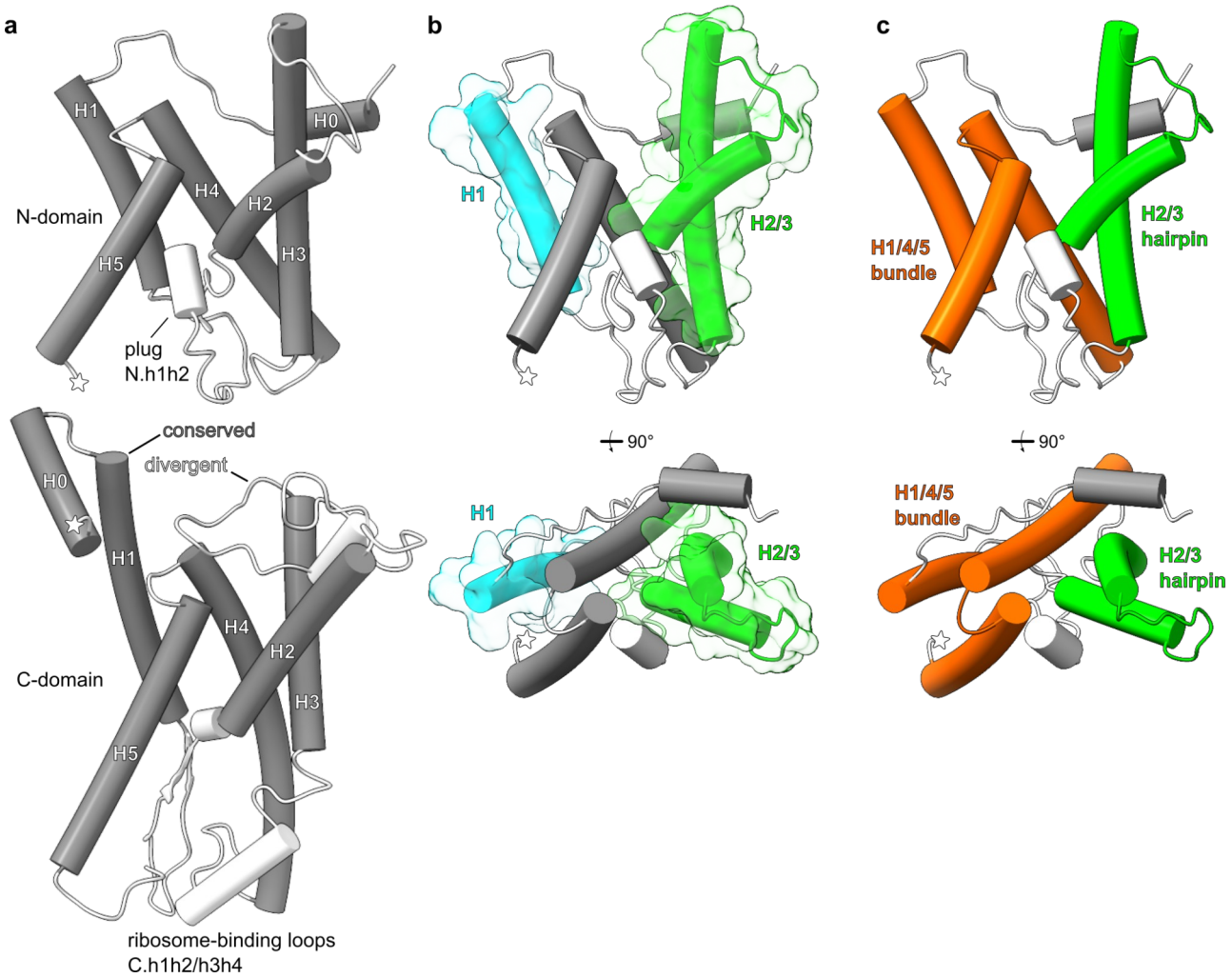

Figure 2. Identification and analysis of the conserved proto-SecY core. All models show *Methanocaldococcus jannaschii* SecY (PDB 1rhz). **a** Consensus elements (grey) in the N-domain (top) and C-domain (bottom). Stars indicate the point where N- and C-domains were split. **b** The N-domain as in **a**, except with H1 (cyan) and H2/3 (green) coloured and overlaid with a semitransparent representation of their solvent-excluded surfaces. Lateral (top) and axial (bottom) views are shown. **c** As in **b**, except with the H1/4/5 three-TMH bundle (orange) recoloured.

one side to the other, it is unlikely that H1 and H2/3 were separated by gradual adjustments to the helices' tilts or positions, as would seem likely if instead H4/5 penetrated the space between them from a single side. Instead the most conservative explanation for the separation of H1 and H2/3 is that one or both of them were not transmembrane in the ancestral fold.

Omission of the H2/3 hairpin preserves H1/4/5 as a three-TMH bundle (Figure 2c), whereas omission of H1 isolates H5, yielding a much less compact fold. This suggests that an ancestral fold containing the H1/4/5 three-TMH bundle subsequently acquired the H2/3 hairpin, which packed against the bundle's surface. Acquisition of a transmembrane hairpin is highly plausible because it is a common transition in membrane protein evolution. Mutations which increase the hydrophobicity of a structural element tend to promote its membrane insertion. Insertion as a hairpin (generally defined as an antiparallel self-associating motif) is both more physically favourable (Engelman and Steitz, 1981) and less topologically disruptive than insertion as a single TMH, which would by contrast invert the topology of any subsequent TMHs. One of many examples of acquired transmembrane hairpins is provided by SecE, which in some proteobacteria such as *E. coli* has acquired a transmembrane hairpin in its N-terminal peripheral helix (Cao and Saier 2003; Figure 2-Figure supplement 1).

These considerations led us to posit that although the five-TMH precursor of SecY may have been lost, the three-TMH bundle of H1/4/5 might persist in a protein family that diverged at an earlier point. Notably, a significant portion of the hydrophilic pore of SecY is lined by the three-TMH bundle, hinting that even the putative distant ancestor could have had a transport function.

Identification of YidC as a candidate proto-SecY homolog

Any non-duplicated homolog of SecY that persists today would also have been present in the cenancestor, which was the last common ancestor of the two fundamental phylogenetic domains, Archaea and Bacteria. Attribution to the cenancestor is difficult, but several IMPs that are widely conserved display divergences between phylogenetic domains that allow us to exclude trans-domain gene transfer as an explanation for their conservation. These include SecY, SecE, TatC, signal peptidase, the rotor-stator ATPase, the multiple resistance and pH (Mrp) sodium-proton antiporters, and certain redox factors (Lombard et al., 2012). An additional IMP, YidC, recently joined this group; its divergence across domains is so great that its universality was only gradually uncovered.

When YidC was first studied in mitochondria (as Oxa1/2), sequence similarity alone was sufficient for detection of its homologues in bacteria (SpoIIIJ, YidC; Bauer et al. 1994, 199; Bonnefoy et al. 1994), and also subsequently in chloroplasts (Alb3; Sundberg et al. 1997). Archaeal homologues were not detectable by these methods, but homology candidates were subsequently identified using position-specific scoring matrices (Yen et al., 2001; Zhang et al., 2009). These candidates were, however, only sporadically distributed within a single archaeal phylum, Euryarcheota, and their marginal detectability engendered little consensus between reports. Validation and extension to Crenarchaeota came from genomic neighbourhood analysis, which identified a widely conserved cluster of YidC, SecY, and ribosomal proteins (Makarova et al., 2015). YidC thus displays the two-domain phylogeny and wide conservation required for confident attribution to the cenancestor.

Because of YidC's divergence across phylogenetic domains, its conserved structural features were also only recently identified. Crystal structures of bacterial YidC show at least five TMHs (Kumazaki et al., 2014b, 2014a), but a crystal structure of an archaeal YidC-like protein (Ylp1 from *Methanocaldococcus jannaschii*) showed only three TMHs (Borowska et al., 2015). Because the crystal structure of Ylp1 displayed domain swapping between adjacent chains, the three-TMH fold's physiological relevance was initially questioned (Kuhn and Kiefer, 2017). Homology modelling does show, however, that an unswapped three-TMH fold is consistent with co-evolutionary and physical restraints (Figure 3-Figure supplement 1).

The three-TMH fold is now well supported by the identification of similar homologs in the eukaryotic ER (Anghel et al., 2017). Although not detected by less sensitive methods, comparisons of profile hidden Markov models for archaeal YidC and eukaryotic proteins identified three YidC paralogs in the ER: TMCO1, EMC3, and GET1. Recent single-particle electron cryomicroscopy (cryo-EM) studies yielded structural models for all three paralogs (Bai et al., 2020; McDowell et al., 2020; McGilvray et al., 2020; Miller-Vedam et al., 2020; O'Donnell et al., 2020; Pleiner et al., 2020), which together with the prokaryotic crystal structures reveal a conserved core consisting of a three-TMH bundle interrupted after the first TMH by a cytosolic helical hairpin. In the prokaryotic forms, a sixth, N-terminal peripheral helix is also present.

Among the cenancestor IMPs, the hairpin-interrupted three-TMH motif of YidC is strikingly similar to the consensus proto-SecY elements identified above. Each consensus helix from the YidC family can be matched to a consensus helix from proto-SecY, unambiguously and with the same connectivity (Figure 3, Table 1). This surprising structural similarity identifies the YidC family as a

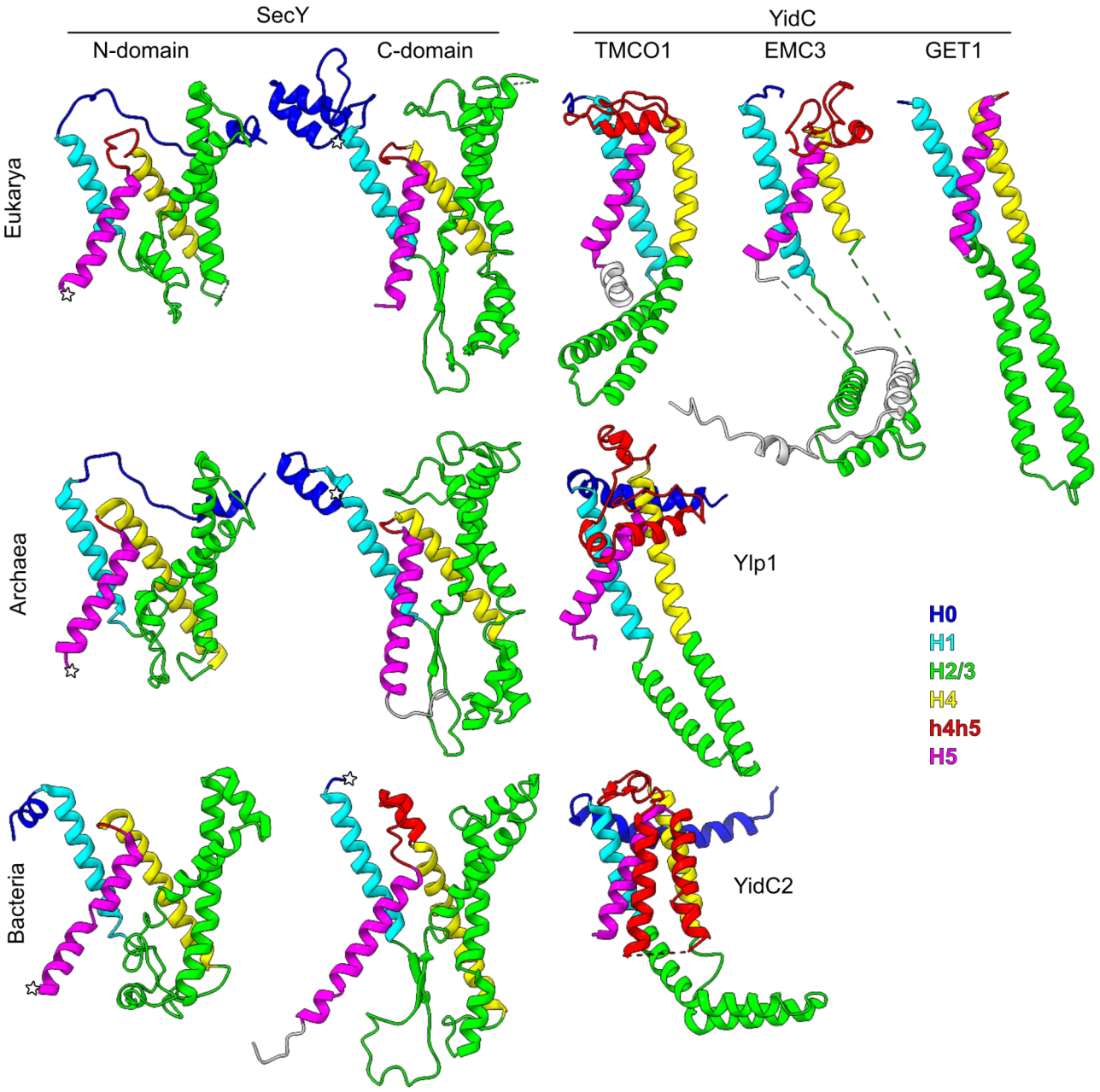

Figure 3. Correspondence between structural elements of the SecY and YidC protein channels. Consensus elements and the intervening element h4h5 are coloured according the colour key shown and the correspondences in Table 1. Other intervening elements are coloured to match a neighbouring consensus element. Flanking elements are coloured white. The models are, from left to right and then top to bottom: *Canis lupus familiaris* Sec61A1 (PDB 6fti, Braunger et al. 2018), *Homo sapiens* TMCO1 (6w6L, McGilvray et al., 2020), *H. sapiens* EMC3 (6ww7, Pleiner et al., 2020), *H. sapiens* GET1 (6so5, McDowell et al., 2020), *M. jannaschii* SecY (1rhz, Van den Berg et al., 2004), *M. jannaschii* Ylp1 (5c8j, Borowska et al., 2015, extended by structure prediction to include the originally unmodelled H2/3; see Figure 3-Figure supplement 1), *G. thermodenitrificans* SecY (6itc, Ma et al., 2019), *Bacillus halodurans* YidC2 (3wo6, Kumazaki et al., 2014a).

uniquely good candidate for the origin of proto-SecY. The functional similarity between SecY and YidC as mediators of IMP integration further supports this idea. The remainder of this paper articulates the evidence for and implications of this possibility. For convenience, we use YidC to refer to the family as a whole, as we have used SecY to refer to both SecY and its Sec61 homologs, and will specify particular clades only when discussing characteristics not shared by the whole.

Table 1. Consensus nomenclature for SecY and YidC.

	SecY		Consensus structure element	YidC			
				Archaea-Eukarya	Bacteria		
		Domain			Gram-positive	Gram-negative	
			h0			TM1, P1	
			H0	EH1	EH1	EH1	
	TM1		H1	TM1	TM1	TM2	
Plug	TM2a		h1h2				
Lateral gate	TM2b	N	H2	CH1	CH1	CH1	C1
	TM3		H3	CH2	CH2	CH2	
	TM4		H4	TM2	TM2	TM3	
			h4h5	EH2	TM3/4	TM4/5	
	TM5		H5	TM3	TM5	TM6	
			h0			TM1, P1	
			H0	EH1	EH1	EH1	
	TM6		H1	TM1	TM1	TM2	
	C4		h1h2				
Lateral gate	TM7	C	H2	CH1	CH1	CH1	C1
	C5		h2h3				
Lateral gate	TM8		H3	CH2	CH2	CH2	
	TM9		H4	TM2	TM2	TM3	
			h4h5	EH2	TM3/4	TM4/5	
	TM10		H5	TM3	TM5	TM6	

CH, cytoplasmic helix; EH, extracytoplasmic helix; P, periplasmic domain; C, cytoplasmic domain

The cores of SecY and YidC are structurally and functionally similar

Like SecY, YidC facilitates the diffusion of hydrophilic protein segments across the hydrophobic membrane by burying hydrophilic groups inside the membrane (Figure 4a). SecY buries a hydrophilic funnel on each side of the membrane and thereby forms a continuous hydrophilic pore across it. By contrast, YidC's hydrophilic groove is only open to the cytosol, and only penetrates part-way into the membrane. Biophysical considerations and molecular dynamics simulations suggest that the groove's exposure of hydrophilic groups to the hydrophobic membrane distorts and thins the membrane in its vicinity (Chen et al., 2017).

YidC's hydrophilic groove is similar to those recently observed in components of the retrotranslocation machinery for ER-associated degradation (ERAD) (Wu et al., 2020). Here, the membrane proteins Hrd1 and Der1 each display hydrophilic grooves, which are open to the cytosol and ER lumen, respectively. The juxtaposition of these two 'half-channels' forms a nearly continuous hydrophilic pore, interrupted by only a thin membrane through which polypeptide translocation is thought to occur. A YidC-derived proto-SecY can similarly be considered a half-channel capable of forming a near-complete channel by antiparallel homodimerisation.

As argued above, the most ancient core of proto-SecY is the three-TMH bundle of H1/4/5 which lines the hydrophilic translocation pore. Strikingly, the corresponding H1/4/5 bundle in YidC forms its hydrophilic translocation groove (Figure 4a,b). The three-TMH bundles in both SecY and YidC have a right-handed twist, with H1 and H4 near parallel and H5 packing crossways against them. Of the three helices, H5 makes the closest contacts with the translocating hydrophilic substrate in SecY (Figure 4b) and in *E. coli* YidC as determined by chemical crosslinking experiments (He et al., 2020). These crosslinking data indicate that YidC's substrates initiate translocation in a looped configuration, analogous to that of SecY's substrates (Mothes et al., 1994; Shaw et al., 1988; Figure 4a).

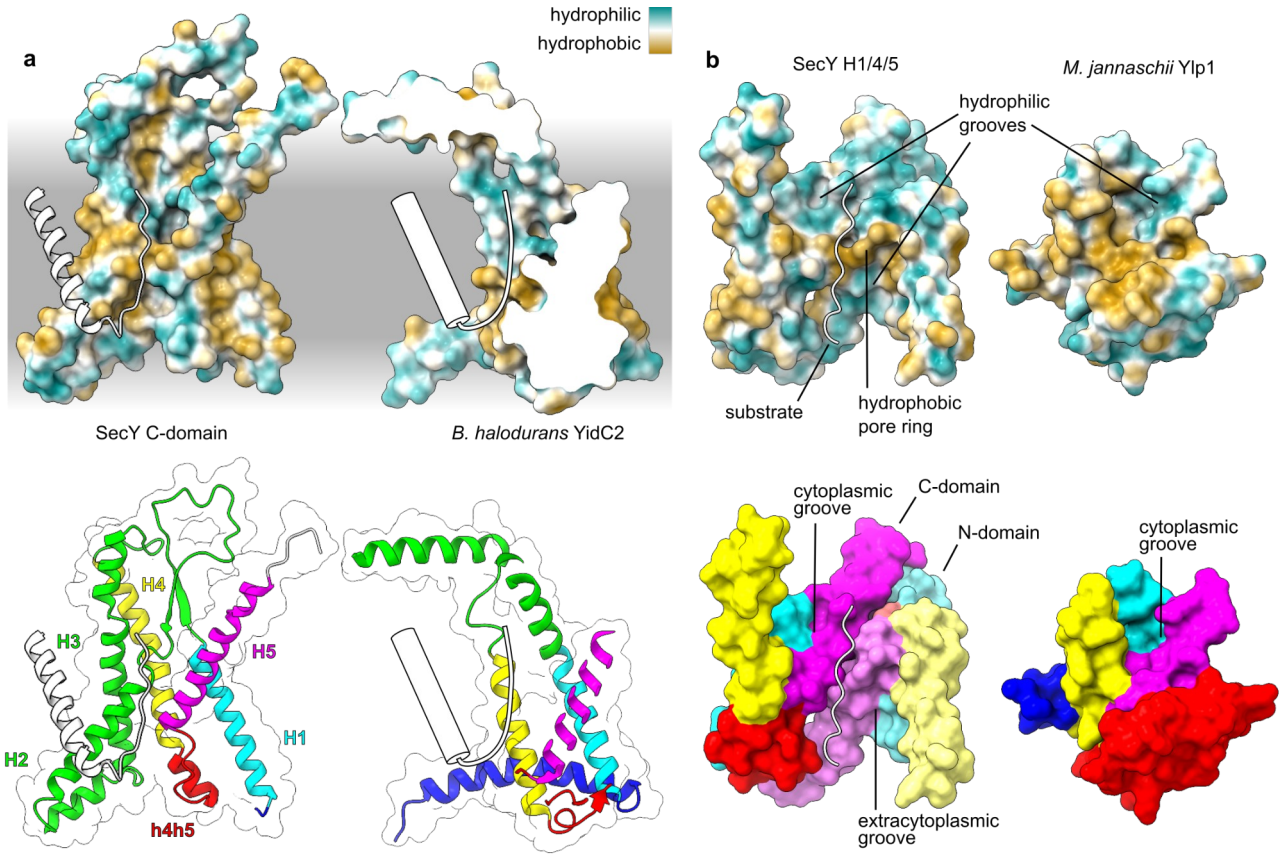

Figure 4. Structural and functional comparison of SecY and YidC. The SecY/substrate model is *G. thermodenitrificans* SecY/proOmpA (PDB 6itc). **a** Comparison of the solvent-excluded surface hydropathy (top) and structure (bottom) of the SecY C-domain (left) and YidC (right; 3wo6). The hydropathy of the lipidic and aqueous phases is represented on a separate scale, ranging from hydrophilic (white) to hydrophobic (grey). The hydropathy of the interfacial layers is approximated by linear gradients, each half the width of the hydrophobic layer. The algorithm used to estimate the membrane thickness and relative position does not account for any anisotropic membrane thinning which lipid-exposed hydrophilic residues may induce (see Methods), and thus none is shown. A schematic representation of a substrate signal and translocating polypeptide is superimposed on YidC, indicating the experimentally determined interface across which substrates translocate. The YidC surface and model are clipped to allow a lateral view of the hydrophilic groove which would otherwise be occluded by the non-conserved h4h5 transmembrane hairpin (*B. halodurans* YidC2 TM3/4). **b** Left: Lateral view of the SecY/substrate complex, with H1/4/5 shown as a solvent-excluded surface and the translocating substrate shown as a cartoon, with its signal helix hidden. The surface is shown colour-coded by hydropathy (top) or by consensus element (bottom). Right: as at left, except for *M. jannaschii* Ylp1 (5c8j).

In YidC, the point where H4 and H5 meet forms the hydrophobic end of the hydrophilic groove. This intersection remains hydrophobic in SecY, but is shifted toward the pseudosymmetry axis. The consequence of this shift is that the corresponding amino acids of the N- and C-terminal halves of SecY are now juxtaposed to form a ring of hydrophobic amino acids known as the pore ring (Figure 4b). The pore ring lies close to the center of the membrane and represents the point where the hydrophilic vestibules from each side of the membrane connect. Thus, key structural features of YidC are not only recognizable in SecY, but also match with similar functional roles. This structure-function correspondence satisfies an important prediction for a putative SecY homolog.

Alongside these elements which are both structurally and functionally similar, there is also an element, H0, which is structurally similar despite not being known to have any direct function in translocation (Figure 5a). It is clearly not essential for function in SecY, having been largely eliminated from some bacterial SecY (Figure 3). Archaeal and eukaryotic SecY N.H0 and YidC H0 are similar in their orientation, length, contact with H4, and the position of that contact site (Figure

5a). SecY C.H0 is similarly peripheral but different in length and orientation, a difference which is attributable to the confining effect of its fusion to N.H5.

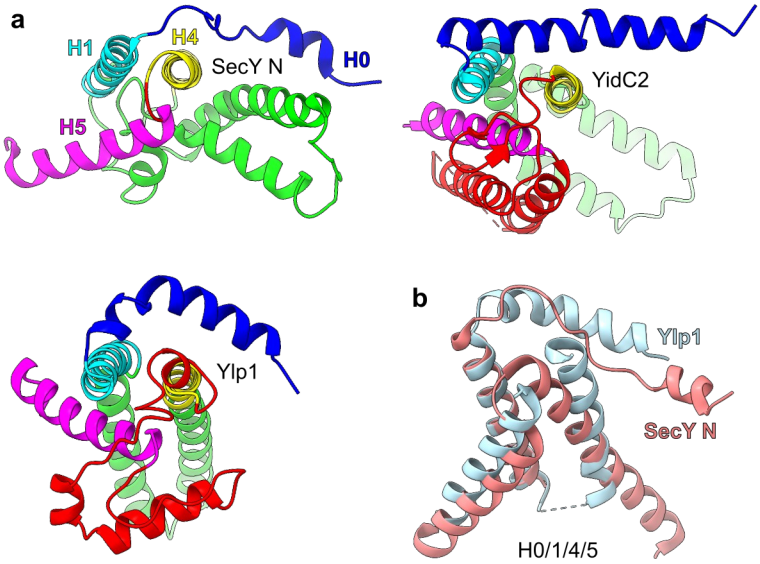

Figure 5. Structural comparison of in H0 of SecY and YidC. **a** Matched views of the SecY N-domain (*M. jannaschii* SecY, PDB 1rhz), bacterial YidC (*B. halodurans* YidC2, 3wo6), and archaeal YidC (*M. jannaschii* Ylp1, see Figure 3-Figure supplement 1). **b** Alignment of Ylp1 and the SecY N-domain from *M. jannaschii*.

The similarity between SecY N.H0 and YidC H0 is particularly good evidence for homology because without a direct functional role in SecY, it is unlikely to be the result of convergent evolution. Instead it indicates a conserved structural role. This independent evidence supports homology as an explanation for the structural and functional similarity of their conserved cores. Considered together, we conclude that SecY and YidC share a structural core composed of a membrane-embedded H1/4/5 bundle and a peripheral H0 brace (Figure 5b).

SecY's structural differences from YidC support its unique secretory function

Whereas the conserved cores of SecY and YidC are similar, their structural differences are concentrated in regions which are hypervariable among YidC homologs: h4h5 and H2/3 (Figure 3). The H2/3 hairpin takes many structural forms among the YidC and SecY families. The relatively compact cytosolic hairpin in bacterial and archaeal YidC is markedly elongated and rigid in GET1, tethered via long flexible loops in EMC3, and retained in a roughly similar architecture in TMCO1 (Figure 3; Figure 3-Figure supplement 2). In contrast to each of these examples, the H2/3 hairpin in SecY is folded back toward the H1/4/5 bundle and is embedded in the membrane. Despite all these differences in topology, length, and linker properties, H2/3 appears to uniformly retain strong coupling between its two helices, and their lengths in all but GET1 remain within a relatively narrow range of ~15-30 amino acids.

These similarities are consistent with H2/3 in YidC and SecY sharing a common evolutionary origin. As already noted above, transmembrane hairpin acquisition is frequently observed during membrane protein evolution. In addition to the SecE example noted above (Figure 2-Figure supplement 1), YidC h4h5 is a peripheral helix in archaea but a transmembrane hairpin in bacteria (Figure 3). In the same way, H2/3 represents a hairpin that has inserted alongside the H1/4/5 three helix bundle. This hairpin is cytosolic or membrane-peripheral in YidC, but became more hydrophobic and membrane-embedded to generate the five-TMH fold of proto-SecY.

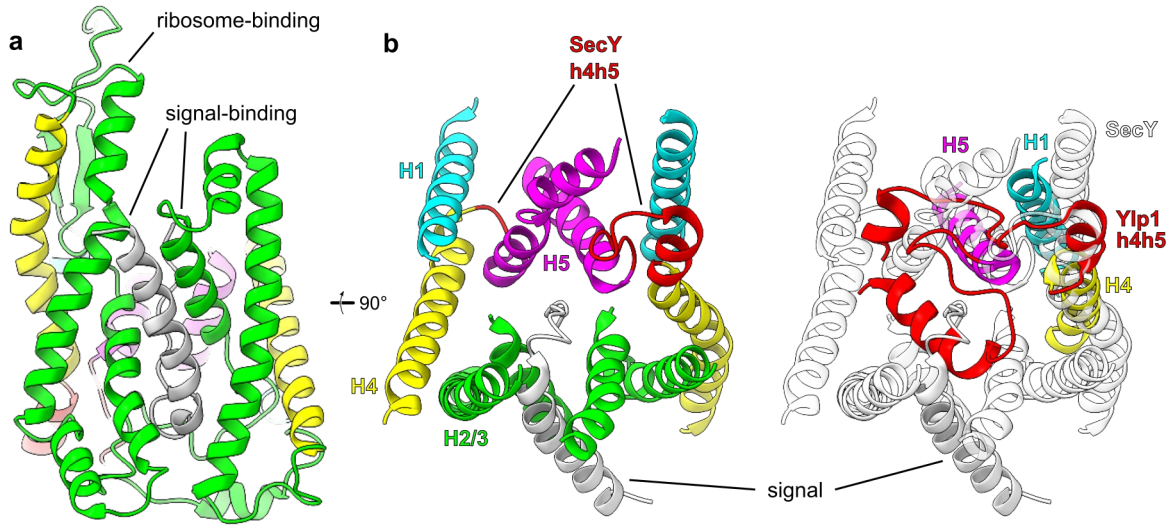

Figure 6. Structural features unique to SecY which enable signal binding and substrate translocation. SecY is *G. thermodenitrificans* SecY/proOmpA (PDB 6itc). **a** Signal-binding and ribosome-binding sites on SecY H2/3, viewed laterally. **b** The substrate translocation channel, viewed from its extracytoplasmic side. Only H1-5 and h4h5 of SecY are shown. SecY is colour-coded by consensus element as in Figure 3 (left), or rendered transparent and superimposed by the corresponding elements of *M. jannaschii* Ylp1 (5c8j), aligned to the SecY C-domain (right).

Consistent with this idea of a common evolutionary origin, H2/3 in SecY and YidC displays not only structural but also functional similarity: it participates in signal recognition in both SecY (Figure 6a) and across diverse YidC homologs. The methionine-rich membrane-facing side of YidC H2/3 is thought to initially engage the TMHs of its substrates, at least in bacteria (Kumazaki et al., 2014a), archaea (Borowska et al., 2015), and eukaryotic TMCO1 and EMC3 (McGilvray et al., 2020; Pleiner et al., 2020). In contrast to direct TMH interaction, the rigid and elongated H2/3 coiled coil of GET1 (McDowell et al., 2020) forms a binding site for the substrate targeting factor GET3 (Figure 3-Figure supplement 2; Mariappan et al. 2011; Stefer et al. 2011; F. Wang et al. 2011). This adaptation may be due to the particularly hydrophobic TMHs inserted by this pathway (Guna et al., 2018), warranting a specialised machinery to shield them in the cytosol.

The migration of H2/3 into the membrane in SecY encloses the translocation channel which in YidC is exposed to the membrane (Figure 6b). This allows SecY to create a significantly more hydrophilic and aqueous environment for its hydrophilic substrates, facilitating their translocation. This is particularly important for SecY's secretory function, which involves translocating much larger hydrophilic domains than those translocated by YidC.

As a secondary consequence, transmembrane insertion of H2/3 makes the site where signals initiate translocation proteinaceous and partially hydrophilic (Figure 4a; Gogala et al., 2014; Park et al., 2014; Plath et al., 1998; Voorhees and Hegde, 2016; Weng et al., 2020). Because of this, translocation via SecY can be initiated via signals which are much less hydrophobic than the TMHs which initiate translocation via YidC. This, too, is important for SecY's secretory function, because the signal peptides of secretory proteins are distinguished from TMHs by their relative hydrophilicity (von Heijne, 1985). This biophysical difference allows signal peptidase to specifically recognise and cleave them (Paetzel et al., 2002). Cleavage frees the translocated domain from the membrane to complete secretion.

After H2/3, the next most conspicuous difference between SecY and YidC is in h4h5, which is nearly eliminated in SecY (Figure 3). Whereas the H2/3 transmembrane insertion differentiates how SecY and YidC receive and recognise hydrophobic domains, the h4h5 elimination helped clear the

channel through which hydrophilic substrates translocate. As mentioned previously, h4h5 is, like H2/3, hypervariable among YidC homologs, forming a peripheral helix in archaea and eukarya and a transmembrane hairpin in bacteria. If h4h5 were not altered in proto-SecY, its dimerisation would place h4h5 inside the hydrophilic funnel of the opposite monomer, instead of in contact with the membrane (Figure 6b). Thus, atrophy of h4h5 in SecY, driven by a change in chemical environment, would have opened a membrane-spanning hydrophilic pore and facilitated translocation.

Both proto-SecY and YidC use the distal face of H5 for dimerisation

SecY's channel is formed from similar hydrophilic grooves buried on each side of the membrane (Figure 4b), suggesting that proto-SecY functioned as an antiparallel homodimer. This is consistent with how more recent antiparallel fusions are inferred to have evolved, via trajectories which consistently feature antiparallel homodimerisation as an intermediate step (Lolkema et al., 2008; Rapp et al., 2006). Subsequent gene duplication and fusion yields a pseudosymmetric protein, in which each domain can specialise for a single orientation.

Antiparallel homodimerisation requires that the protomer possess two characteristics: a tendency to be produced in opposite topologies, and an interface suitable for dimerisation. Although dual-topology is not evident in the YidC proteins which have been assayed, distant ancestors could easily have had this property with relatively few changes. A variety of studies have shown that making only a few changes to charged amino acids flanking the first TMH of an IMP can influence its topology, and that an inverted first TMH can invert an entire IMP containing several TMHs (Beltzer et al., 1991; Brown et al., 2018; Rapp et al., 2006, 2007). Such changes in topology occur naturally in protein evolution (Rapp et al., 2006; Sääf et al., 1999), and YidC is not known to contain any conserved charges in its soluble segments that would impede this evolutionary process.

More important is the second required characteristic: possession of a proto-SecY-like interface suitable for dimerisation. Because SecY N.H2/3 and C.H2/3 separate during gating, the major interaction between the N and C domains occurs on the opposite side of the channel, via N.H5 and C.H5 (Figure 6b). The face of H5 used for this interaction is the side furthest from the rest of the H1/4/5 bundle, and so we will refer to it as the distal face of H5. From this feature of the SecY structure, we infer that proto-SecY formed antiparallel homodimers via the distal face of H5.

The evolutionary relationship between proto-SecY and YidC suggests that some extant YidC proteins may retain a tendency to form protein-protein interactions via the distal face of H5. This surface forms an intramolecular interaction with the h4h5 hairpin in bacterial YidC, but remains exposed in archaeal and eukaryotic YidC (Figure 3). There are no data available about YidC biochemistry in archaeal cells, but at least two eukaryotic YidC proteins, EMC3 and GET1, form functionally important complexes, and structural models show that they use the distal face of H5 to do so (Figure 3-Figure supplement 2; Pleiner et al. 2020; Bai et al. 2020; O'Donnell et al. 2020; Miller-Vedam et al. 2020; McDowell et al. 2020). These interactions via H5 are heterodimeric, rather than homodimeric, but nonetheless demonstrate that EMC3 and GET1 can dimerise (with EMC6 and GET2, respectively) along the same interface as proto-SecY without impeding their translocation activities.

Dimerisation via YidC H5 is ancient and widely conserved

Given that pre-cenancestral proto-SecY and eukaryotic YidC both form dimers via the distal face of H5 (Figure 7a), archaeal YidC should also form dimers via this interface, because eukaryotes descended from the cenancestor via archaea. To test this prediction, we queried nine diverse

archaeal proteomes for homologs of *H. sapiens* EMC6 or GET2 using HHpred (Zimmermann et al. 2018). Although none displayed significant similarity with GET2, every proteome queried contained exactly one protein similar to EMC6 (Figure 7b). Among these archaeal proteins, those most similar to eukaryotic EMC6 tend to come from the species most closely related to Eukarya: the Asgard archaean, then the TACK archaean, and then the euryarchaeans. This phylogenetic

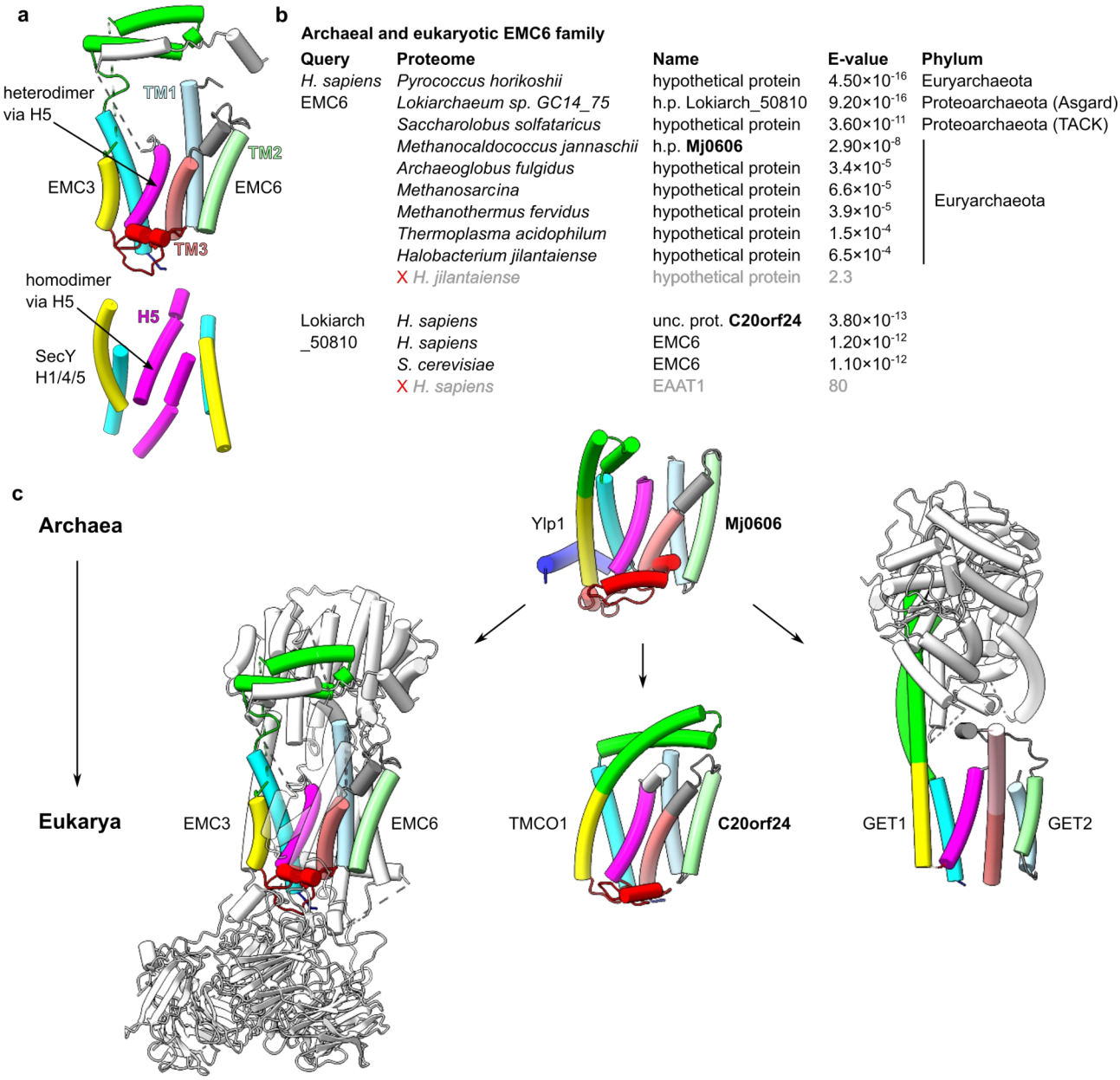

Figure 7. Ubiquitous dimers formed by YidC via the same interface as proto-SecY. **a** Comparison of the EMC3/6 and proto-SecY dimerisation interfaces. EMC3/6 is from *S. cerevisiae* (PDB 6wb9), and proto-SecY is represented by *G. thermodenitrificans* SecY (6itc). **b** Table of the archaeal and eukaryotic proteins most similar to the listed query protein, as measured by comparison of profile hidden Markov models (HHpred). Results for each query are sorted by similarity, and a red cross 'X' and grey text indicates the first rejected result. For sequence accession numbers, see Methods. **c** Structural models of the archaeal and eukaryotic heterodimers of YidC and EMC6 superfamily proteins. Experimentally determined structures for *S. cerevisiae* EMC (6wb9) and *H. sapiens* GET1/2/3 (6so5) are shown; a second GET1/2 dimer is present in the PDB model but omitted here. The acquired segment of GET2 TM3 identified by alignment to EMC6 (Lokiarch_50810) is shown in pink, whereas the rest of TM3 is darker. The *M. jannaschii* Ylp1/Mj0606 dimer is represented by Ylp1 (5c8j, extended by structure prediction to include the originally unmodelled H2/3; see Figure 3-Figure supplement 1) and the likeliest predicted structure for Mj0606 as determined by trRosetta. They are positioned and oriented relative to each other by alignment to the EMC3/6 dimer. The *H. sapiens* TMCO1/C20orf24 dimer is represented similarly, except the TMCO1 model does not differ from that in the PDB (6w6L).

relationship indicates that the archeal proteins are homologs of the eukaryotic protein, and that their ubiquity is due to an ancient origin.

Reciprocal queries of *H. sapiens* and *S. cerevisiae* proteomes with the putative Asgard EMC6 protein (Lokiarch_50810) identified EMC6 in both cases as high-confidence hits. Unexpectedly, the *H. sapiens* search also identified a second equally confident hit, C20orf24 (Figure 7b). These proteins all share a highly similar core structure of three TMHs (Figure 7-Figure supplement 1a; Figure 7-Figure supplement 2), evident in both *de novo* and homology-templated structures predicted by trRosetta (Yang et al., 2020). This structural similarity strongly supports their assignment to the EMC6 family.

The patterns of co-evolution between archaeal YidC and archaeal EMC6 (determined using RaptorX ComplexContact; Zeng et al. 2018) showed that their highest-probability contacts all occur along the same interface used for dimerisation by eukaryotic EMC3/6 (Figure 7-Figure supplement 1b). Thus the distal face of H5 is used for heterodimerisation not only by eukaryotic EMC3 and GET1 but also by archaeal YidC. The third eukaryotic YidC paralog, TMCO1, displays a similar pattern of coevolution with C20orf24, indicating that they form a similar heterodimer (Figure 7-Figure supplement 1b). TMCO1-C20orf24 interaction is consistent with the aforementioned absence of C20orf24 from *S. cerevisiae* (Figure 7b) because *S. cerevisiae* also lacks TMCO1.

Although GET2 lacks significant sequence similarity with these EMC6 family proteins, its structural similarity with EMC6 was immediately recognised (McDowell et al., 2020; Pleiner et al., 2020). Our identification of an ancient EMC6 family reveals a plausible origin for GET2. Consistent with this, although our GET2 query of the lokiarchaean proteome did not identify any high-similarity proteins, the most similar was indeed EMC6 (Lokiarch_50810). This similarity was not detected in the archaeal proteomes more distant from Eukarya. Moreover the aligned columns between GET2 and Lokiarch_50810 correspond exactly to their structurally similar transmembrane domains. The single large gap in this alignment corresponds to the cytosolic extension of GET2 TM3, which brings it into contact with GET3 (Figure 7c). Thus the major difference between GET2 and EMC6 can be explained as a functional adaptation for GET3 recognition, not unlike GET1's elongation of H2/3. We therefore propose that GET2 is a member of the EMC6 superfamily.

The absence of a similar heterodimer in bacteria suggests that EMC6 was acquired in Archaea after divergence from Bacteria, which instead acquired the H5-occluding transmembrane hairpin in h4h5 (Figure 3). An archaeal origin for EMC6 is consistent with its genomic location, which is distant from the widely conserved cluster of cenancestral ribosomal genes, SecY and YidC (Makarova et al., 2015). In the putative period of YidC evolution prior to EMC6, H5 would not have been occupied by heterodimerisation with EMC6, and instead could have mediated homodimerisation, as in proto-SecY. YidC's universal tendency to occlude the distal face of H5 supports this possibility.

Reductive selection in symbionts demonstrates the functional range of YidC

YidC cannot efficiently translocate some SecY substrates (Welte et al., 2011), but if proto-SecY originated in the YidC family, YidC may initially have been the cell's only transporter for the extracytoplasmic parts of IMPs. How severe a constraint would strictly YidC-dependent biogenesis have imposed on the primordial cell's membrane proteome? The looser this constraint, the likelier it is that YidC could have preceded SecY.

Insight into this question of *in vivo* sufficiency can be obtained by inspection of the only clade known to have survived SecY deletion: the mitochondrial symbionts. SecY has been lost from all mitochondrial genomes for which sequences are available, except in one group of eukaryotes, and it

has not been observed to relocate to the nuclear genome (Janouškovec et al., 2017). The exceptional group is the jakobids, only a subset of which retain SecY. The incomplete presence of SecY in this group implies that it was lost multiple times from the jakobids and their sister groups. SecY deletion is therefore a general tendency of mitochondria, rather than a single deleterious accident.

Mitochondria retain two YidC homologues, Oxa1 and Oxa2 (Cox18), the genes for which relocated from the mitochondrial genome to the nuclear genome (Bauer et al., 1994; Bonnefoy et al., 1994). As nuclear-encoded mitochondrial proteins, they are translated by cytosolic ribosomes and then imported into mitochondria via channels in the inner and outer mitochondrial membranes (Wiedemann and Pfanner, 2017). These channels are essential for the import of nuclear-encoded proteins, but they are not known to function in the integration of mitochondrially encoded IMPs (meIMPs), which instead requires export from the matrix, where they are synthesized by mitochondrial ribosomes. This export is generally Oxa1-dependent (Hell et al., 2001).

The meIMPs have diverse properties, including 1 to 19 TMHs and exported parts of various sizes and charges (Figure 8a-c). Oxa1's sufficiency for their biogenesis *in vivo* is consistent with *in vitro* results showing that *E. coli* YidC is sufficient for the biogenesis of certain 6- and 12-TMH model substrates (Serdiuk et al., 2019; Welte et al., 2011). The only apparent constraint on the meIMPs is that they tend to have short (~15 a.a.) soluble segments. This is consistent with observations from *E. coli* that fusing long soluble segments to a YidC-dependent IMP can induce SecY dependence (Andersson and von Heijne, 1993; Kuhn, 1988; Shanmugam et al., 2019). Among the meIMPs, Cox2 is an exception which proves the rule, because Oxa1 cannot efficiently translocate its exceptionally long (~140 a.a.) C-terminal tail; instead it is translocated by Oxa2 in cooperation with two accessory proteins (Saracco and Fox, 2002).

This constraint is less consequential than it may at first appear, because prokaryotic IMPs in general tend to have only short soluble segments (Figure 8c; Wallin and von Heijne, 1998). Thus most prokaryotic IMPs may be amenable to SecY-independent, YidC-dependent biogenesis. Consistent with this, in *E. coli*, the signal recognition particle (SRP) has been found to promiscuously target nascent IMPs to either SecY or YidC (Welte et al., 2011), and YidC is present at a concentration 1-2x that of SecY (Schmidt et al., 2016). By contrast, IMPs with large translocated domains became much more common in eukaryotes (Wallin and von Heijne, 1998) concomitant with YidC's divergence into three niche paralogs, none of which are essential at the single-cell level (Guna et al., 2018; Jonikas et al., 2009; McGilvray et al., 2020).

Even without extrapolating from the meIMPs to other similar IMPs, it is clear that chemiosmotic complexes are amenable to YidC-dependent, SecY-independent biogenesis (Figure 8d). These complexes couple chemical reactions to the transfer of ions across the membrane, and are sufficient for the membrane's core bioenergetic function. Although the complexes shown participate in aerobic metabolism, which presumably post-dates cyanobacteria's photosynthetic oxygenation of Earth's atmosphere, they have homologs which enable chemiosmosis in anaerobes. In particular, chemiosmosis in methanogenic and acetogenic archaea employs the rotor-stator ATPase, Mrp antiporters, and an energy-converting hydrogenase (Ech; Lane and Martin 2012), all of which have homologs of their IMP subunits among the meIMPs (Figure 8d). These archaeans' metabolism has been proposed to resemble primordial anaerobic metabolism at alkaline hydrothermal vents (Weiss et al., 2016).

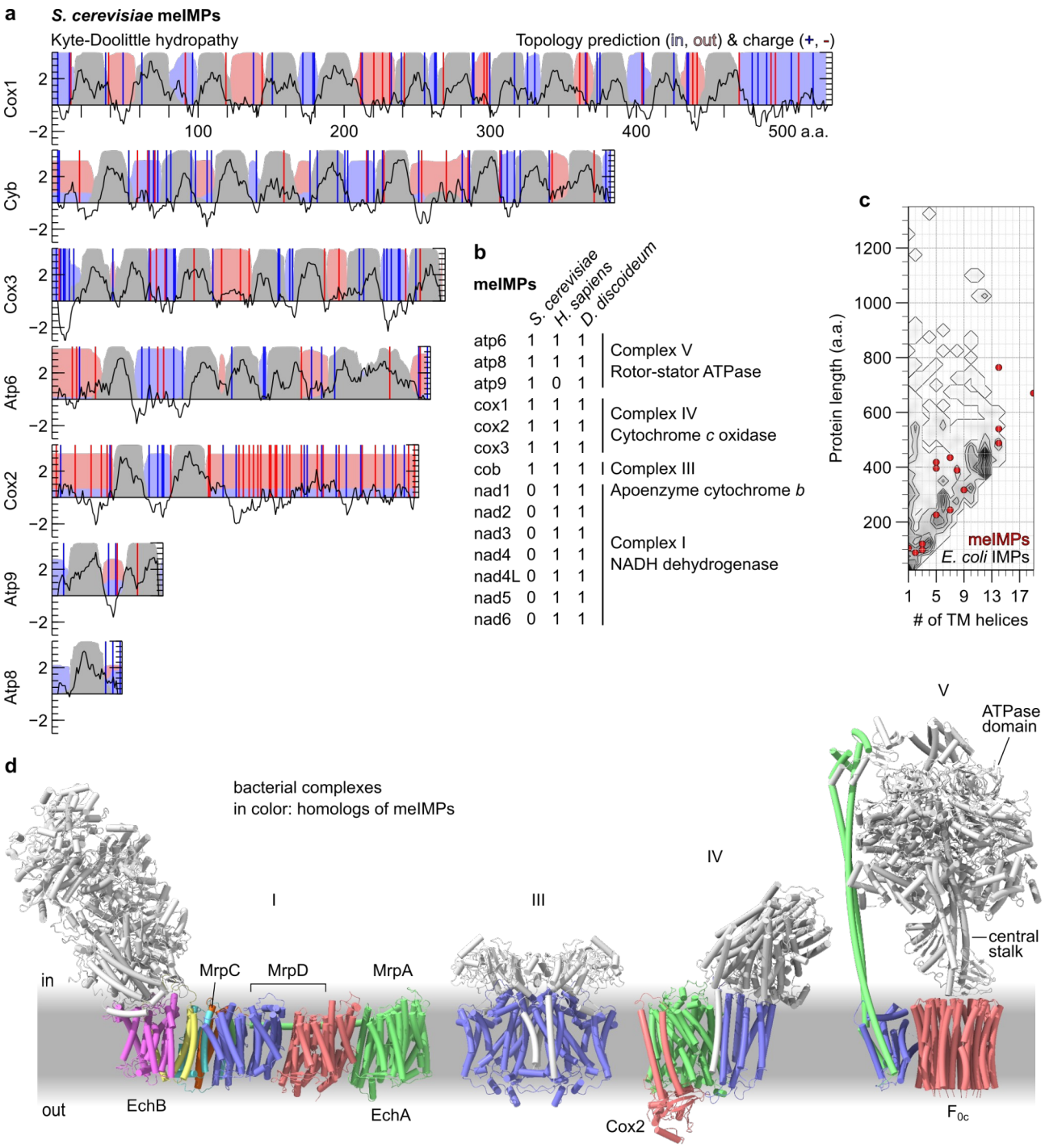

Figure 8. Substrates of the mitochondrial SecY-independent pathway for IMP integration. **a** Sequence characteristics of the mitochondrially-encoded IMPs (meIMPs) from *S. cerevisiae*. Hydropathy on the Kyte-Doolittle scale (left axis) is plotted as an average over a moving window 9 a.a. wide (black line). Topology predictions were computed by TMHMM (right axis) to indicate regions which are retained in the mitochondrial matrix (light blue field), inserted into the membrane (grey field), or exported to the intermembrane space (light red field). Positively (blue) and negatively (red) charged residues are marked with vertical bars. **b** Table of all meIMPs in a fungus (*S. cerevisiae*), a metazoan (*H. sapiens*) and an amoebozoan (*Dictyostelium discoideum*). **c** Scatter plot of the length and number of TM helices for all meIMPs of a single eukaryote (*D. discoideum*), superimposed on a contour plot and heat-map of all 910 IMPs from a proteobacterium (*E. coli;* 170 monotopic, 740 polytopic). Protein lengths were binned in 25 a.a. increments. Each contour represents an increase of 3 proteins per bin. Protein lengths and TMH counts were obtained from UniProt (The UniProt Consortium, 2019). **d** Structures of prokaryotic complexes homologous to meIMPs. Subunits not homologous to the meIMPs listed in **b** are shown in white. Homo-oligomers are represented by a single colour. From left: I, NADH dehydrogenase (*Thermus thermophilus*, 6y11; Gutiérrez-Fernández et al. 2020), III, cytochrome bc_1, (*Rhodobacter sphaeroides*, 6nhh; Esser et al. 2019), IV, cytochrome *c* oxidase (*R. sphaeroides*, PDB 1m57; Svensson-Ek et al. 2002), V, rotor-stator ATPase (*Bacillus* sp. PS3, 6n2y; Guo, Suzuki, and Rubinstein 2019). The labelled subunits of NADH dehydrogenase (I) are homologous to the two IMP subunits of the energy-converting hydrogenase (EchA/B), and/or to subunits of the multiple-resistance and pH (Mrp) antiporters. The labelled subunits of IV and V indicate those referenced in the text.

Thus if YidC had preceded SecY, it would have been sufficient for the biogenesis of diverse and important IMPs, but likely not the translocation of large soluble domains. This is supported by the results of reductive selection in chloroplasts, which retain both SecY (cpSecY) and YidC (Alb3) (Xu et al., 2020). cpSecY imports soluble proteins across the chloroplast's third, innermost membrane, the thylakoid membrane (Peltier et al., 2002). This thylakoid membrane was originally part of the chloroplast inner membrane (equivalent to the bacterial plasma membrane), much like the mitochondrial cristae, but subsequently detached and now forms a separate compartment (Vothknecht and Westhoff, 2001). Because the thylakoid membrane is derived from the plasma membrane, import across the thylakoid membrane is homologous to secretion across the plasma membrane. Thus when symbiosis removed the need for secretion, SecY was eliminated from mitochondria, whereas it was retained in chloroplasts for an internal function homologous to secretion.

A primordial YidC-dependent cell may simply not have secreted protein, or may instead have used a different secretion system. Notably one primordial protein secretion system has been proposed: a protein translocase homologous to the rotor-stator ATPases (Mulkidjanian et al., 2007). This putative primordial protein translocase used its ATPase domain to unfold and feed substrates through the homo-oligomeric channel formed by F_{0c}, now occupied by the central stalk (Figure 8d). The strict YidC-dependence of F_{0c} biogenesis in *E. coli* (Yi et al., 2003) hints that they shared a primordial era of co-evolution, as a laterally closed channel for the secretion of soluble proteins (F_{0c}) and a laterally open channel for the integration of membrane proteins (YidC), including F_{0c} itself. The subsequent advent of a laterally gated channel, SecY, would have enabled the biogenesis of a hybrid class of proteins: IMPs with large translocated domains.

Discussion

Analysis of the SecY structure in the context of the principles of membrane protein folding and evolution led us to re-frame its architecture. In addition to its long-recognised pseudosymmetry, we identified within each half a three-helix bundle abutting a two-helix transmembrane hairpin. The frequent acquisition of transmembrane hairpins during membrane protein evolution argued that the three-helix bundle was its ancestral core. This core element seeded our search for a SecY precursor among cenancestor proteins, leading us to the YidC family. Although the structural similarity of the three-TMH cores of YidC and SecY is striking on its own, the overlaying of key functions onto each structural element strongly reinforced the hypothesis that they are evolutionarily related. Reasoning that YidC may conserve a tendency to dimerise via the same interface as the SecY precursor, we identified a ubiquitous and ancient family of YidC-interacting proteins homologous to EMC6. The surprising conclusion of our study is that an ancestral YidC both preceded and evolved into a proto-SecY whose gene duplication and fusion originated the present-day SecY family.

Evaluation of the homology hypothesis

Inferences about early evolution are necessarily made on the principle of maximum parsimony, i.e. by preferring the hypothesis which explains the most evidence while invoking the fewest *ad hoc* assumptions (Koonin, 2003). By that standard, we weigh the hypothesis that SecY and YidC are homologues against the null hypothesis that they are unrelated. This null hypothesis holds that the similarities shown here all arose by convergent evolution or random chance.

The only positive evidence for the null hypothesis is the presence of certain structural differences between SecY and YidC. We showed, however, that these structural differences are concentrated in

the hypervariable H2/3 and h4h5 regions, and can be explained as adaptations which created a bilayer-spanning pore and lateral gate. Moreover the major difference between SecY and YidC, transmembrane insertion of the H2/3 hairpin, is no greater than the difference between archaeal YidC and bacterial YidC (acquisition and transmembrane insertion of the bacterial h4h5 hairpin). The positive evidence for SecY-YidC homology, by contrast, includes the key similarity in both structure and function of their hydrophilic channels. Homology is further supported by the fact that YidC ubiquitously forms proto-SecY-like interactions via the distal face of H5.

Convergence alone cannot explain the structural similarity between SecY and YidC, because it is not a necessary consequence of their functional similarity. This is evident in the fact that there are other laterally open protein transporters, such as the symbiont import channels and the ERAD channel, which have not converged on the YidC fold (Figure 9). Furthermore there are structural similarities which have no essential function in SecY, namely in the peripheral helix H0, and thus admit no convergent explanation.

The null hypothesis therefore requires that we invoke random chance to explain the similarity between SecY and YidC. But chance appears unlikely to create two protein transporters this similar, because all other laterally open helical protein transporter families of known structure are dissimilar to each other (Figure 9). This inference is not sensitive to the set of transporters considered;

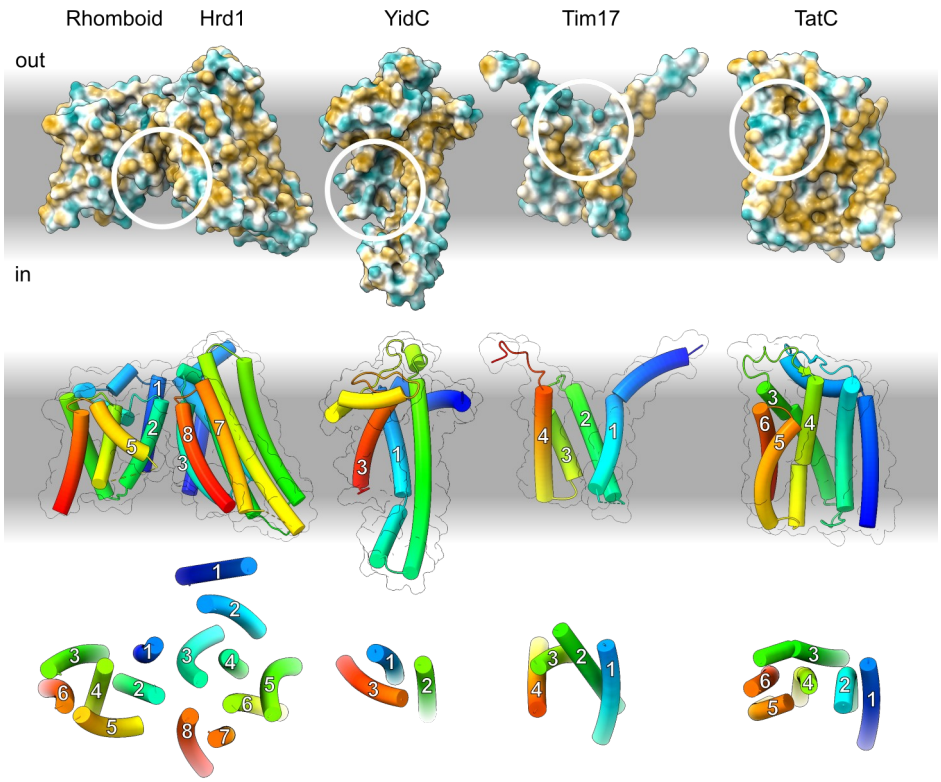

Figure 9. Structures of the known families of laterally open helical protein transporters. Top: Structural models are shown as solvent-excluded surfaces colour-coded by hydropathy (top), ranging from hydrophilic (dark cyan) to intermediate (white) to hydrophobic (dark goldenrod). The hydropathy of the lipidic and aqueous phases represented on a separate scale, ranging from hydrophilic (white) to hydrophobic (grey). White circles indicate intramembrane hydrophilic grooves. The hydropathy of the interfacial layers are approximated by linear gradients, each half the width of the hydrophobic layer. Middle: models are shown in tube representation, rainbow colour-coded by position. Transmembrane segments in the vicinity of the hydrophilic groove are numbered. Bottom: Axial views of each molecule, showing only transmembrane helices. From left to right, the models representing each family are as follows. Rhomboid: *S. cerevisiae* Der1 (PDB 6vjz, Wu et al., 2020), Hrd1: *S. cerevisiae* Hrd1 (6vjz), YidC: *M. jannaschii* Ylp1 (5c8j, Borowska et al., 2015, extended by structure prediction to include the originally unmodelled H2/3; see Figure 3-Figure supplement 1), Tim17: *S. cerevisiae* Tim22 (6lo8, Zhang et al., 2020), TatC: *Aquifex aeolicus* TatC (4b4a, Rollauer et al. 2012).

extending it to include the other laterally open channels, such as lipid scramblases (Brunner et al., 2014) and β-barrel assembly factors (Bakelar et al., 2016), would still yield no recurrences. Moreover, it is especially unlikely that the YidC fold evolved twice just in the limited time prior to the cenancestor, and then seemingly never again. Weighed against a null hypothesis that is supported by so little evidence and requires these several *ad hoc* assumptions, we conclude that the homology hypothesis is the most parsimonious.

Implications for the evolution of protein transport

Besides illuminating SecY's origins, identifying YidC as its progenitor also implies that YidC is the oldest known protein channel. This has implications for the evolution of IMPs generally, including other components of the general secretory pathway and YidC itself. The following is a stepwise model for the evolution of YidC and proto-SecY from a spontaneously inserting precursor (Figure 10a), which we propose parsimoniously explains the available data.

Step 1. The precursor to YidC was a membrane-anchored ribosome receptor. This simple function can be achieved with just two low-complexity domains: a hydrophobic anchor and a polybasic extension. Despite its simplicity, this receptor would function to reduce the deleterious aggregation of hydrophobic domains in the aqueous phase by creating a population of membrane-bound ribosomes, from which any nascent IMPs would be more likely to encounter the membrane. Similar polybasic C-terminal tails are known to occur in YidC and can compensate

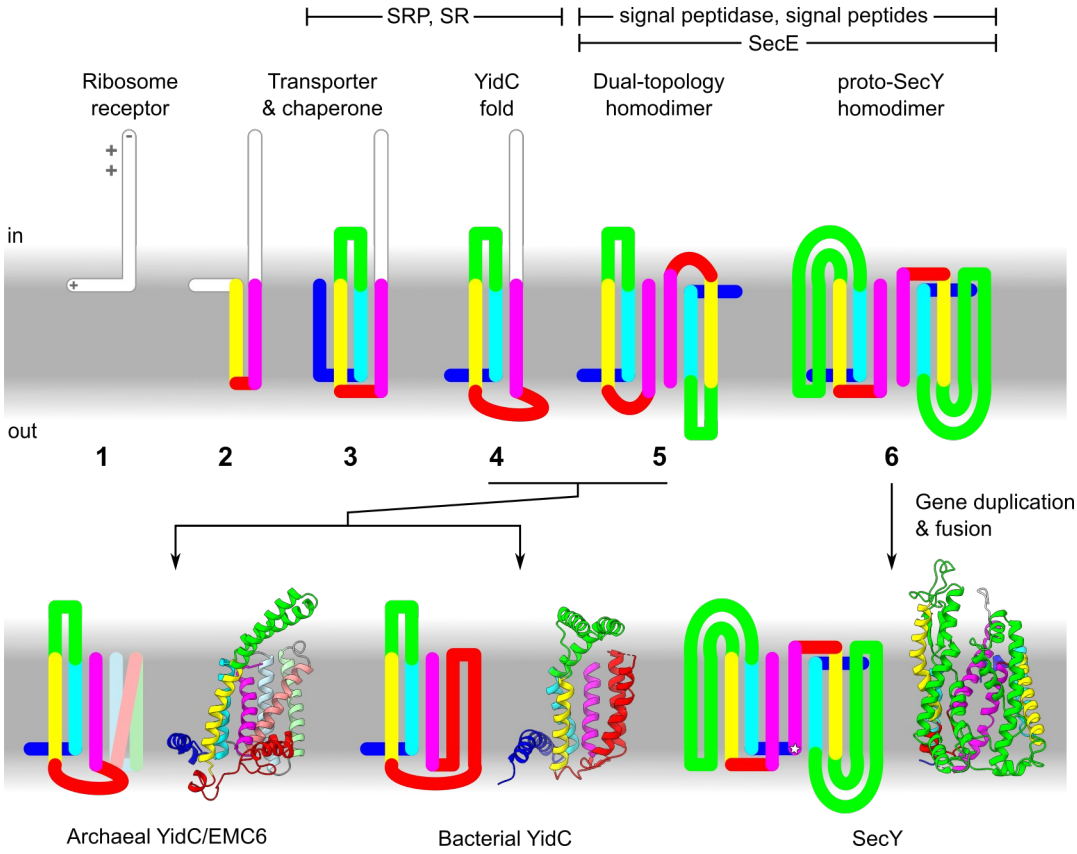

Figure 10. Model for the evolution of the SecY/YidC superfamily from a spontaneously-inserting precursor. **a** Intermediate stages. The spaces on either side of the membrane are marked 'in' and 'out,' but the membrane may not have been continuous enough to make this distinction meaningful at some of the early stages. Charged side chains and termini are indicated at stage **1** by grey symbols. At top, additional components of the secretory pathway label the stage at which they arise: the signal recognition particle (SRP), the SRP receptor (SR), signal peptidase, signal peptides, and SecE. **b** Archaeal YidC and EMC6, represented schematically at left and by structural models at right (*M. jannaschii* Ylp1 and Mj0606, as in Figure 7c). Archaeal YidC is connected via an arrow to either stage **4** or **5** of panel **a**. **c** Bacterial YidC (*B. halodurans* YidC2, PDB 3wo6). **d** SecY (*G. thermodenitrificans* SecY, 6itc).

for deletion of SRP or SRP receptor (SR) (Seitl et al., 2014; Szyrach et al., 2003). We assume that the initial anchor was a peripheral helix because a primitive membrane protein derived from a soluble protein or quasi-random sequence would initially lack the hydrophobicity to spontaneously insert (Mulkidjanian et al., 2009).

Step 2. The peripheral helix acquires a transmembrane hairpin, thereby integrating into the membrane. A hairpin is a likelier anchor than a single TMH, because in the absence of any protein transporters insertion would need to proceed spontaneously, and hairpins insert more efficiently than single TMHs (Engelman and Steitz, 1981). We assume that this insertion preceded SRP/SR-dependent targeting because it is simpler than SRP/SR, and at least as simple as any SRP/SR substrate. The proximity of this hairpin to nascent IMPs imposes a selective pressure on the hairpin to evolve features that facilitate IMP integration and folding, such as membrane-buried hydrophilic residues.

Step 3. Acquisition of a second transmembrane hairpin produces a 4-TMH protein containing the conserved three-helix bundle and hydrophilic groove. The loop between the first and second transmembrane hairpins becomes the cytosolic hairpin H2/3.

Step 4. The hydrophilic groove allows hydrophilic termini to efficiently translocate, including the N-terminus of YidC itself, which acquires a new position as the extracytoplasmic peripheral helix H0. Thus the full YidC fold is now acquired, as it is found in archaea. By this time, SRP/SR have evolved, and H2/3 evolves interactions with SR that will be retained in both SecY and YidC (Kuhn et al., 2011; Petriman et al., 2018).

Step 5. A subpopulation of YidC integrates with an inverted topology and forms antiparallel homodimers. We assume that antiparallel dimerisation precedes duplication because this is common in the evolution of antiparallel fusions (Lolkema et al., 2008; Rapp et al., 2006), and because both domains of SecY conserve the transmembrane insertion of H2/3, which appears to be an adaptation to antiparallel dimerisation. We assume that reorientation did not occur while the C-terminal tail was long and positively charged, because this tail would need to be translocated in the inverted orientation. In the presence of SRP/SR, a ribosome-binding tail is redundant and can be eliminated, facilitating reorientation.

Antiparallel homodimerisation positions hydrophilic grooves on both sides of the membrane, leaving at most a thin hydrophobic layer between them, as in the heterodimeric ERAD channel (Wu et al., 2020). This facilitates the translocation of IMPs with large soluble domains, including signal peptidase. In the presence of signal peptidase, signal-dependent secretion becomes possible, with the first cleavable signal peptides being the TMHs of IMPs which had previously anchored their now-secreted extracytoplasmic domains. Signal peptides' origin as TMHs explains why both engage SecY in a similar way.

At this stage or later, SecE is acquired and binds symmetrically to each half of the antiparallel homodimer. Its association would stabilise the dimer, particularly when they separate to accommodate substrates. In SecY, SecE remains the axis about which the N-domain pivots during engagement (Voorhees and Hegde, 2016). Evolution of SecE after YidC but before proto-SecY would explain why SecE integration is YidC-dependent (Yi et al., 2003).

Step 6. Transmembrane insertion of H2/3 creates a lateral gate. By inserting between the hydrophilic grooves and the membrane, H2/3 makes those grooves even deeper and more hydrophilic, tantamount to a pore when the lateral gate is closed. This more hydrophilic environment further facilitates translocation. As a secondary consequence, it also creates a more

hydrophilic environment for signal recognition. This allows cleavable signal peptides to become less hydrophobic than TMHs, and more easily distinguished by signal peptidase.

Duplication and fusion of the proto-SecY gene would allow each half of this initially symmetric protein to specialise for cytoplasmic and extracytoplasmic functions. For example, the C.h1h2 and C.h4h5 loops would continue to bind ribosomes, whereas these same loops in the N-domain atrophy. One such loop was repurposed as the plug domain.

At step 4 or 5, YidC diverges from SecY (Figure 10). It is not determinable *a priori* whether SecY and YidC diverged as orthologs or paralogs, since the cenancestor may have been capable of lateral gene transfer (Fournier, Andam, and Gogarten 2015). However, the essentiality of YidC suggests that the mutations required to generate proto-SecY would have been best tolerated if they occurred in a paralog, alongside YidC. Paralogous origin in a tandem duplication event is consistent with the commonly observed juxtaposition of YidC and SecY in prokaryotic genomes (Makarova, Galperin, and Koonin 2015). Once diverged from SecY, YidC in archaea evolves to heterodimerise via the distal face of H5 while this same surface in bacteria is covered by the h4h5 transmembrane hairpin.

Outlook

The novel EMC6 superfamily members identified here are intriguing subjects for further study. Details of their origin and distribution, and the context and function of their interactions with archaeal and eukaryotic YidC, remain to be determined. Bacterial YidC, despite being the most-studied YidC clade, also remains uncharacterised in certain important ways. Our analysis highlighted the fact that much of the prokaryotic membrane proteome has characteristics which suggest amenability to either SecY- or YidC-dependent integration, but it is not known how frequently IMPs use each pathway *in vivo,* and very few substrates have been characterised *in vitro*.

This analysis of distant evolutionary relationships was enabled by recent advances in structural techniques, both cryo-EM and structure prediction, which have generated a wealth of structural data about membrane proteins and complexes. The rapidly increasing availability of structures from diverse homologs will similarly facilitate the discernment of conserved structural features in other superfamilies. It may even yield new insight into other channels' origins, which have previously proved to be largely undetectable from sequence data (Hennerdal et al., 2010).

It is plausible, however, that the detectability of SecY's origins will prove to be unusual among fused channels, and is a result of the unusual properties of protein as a transport substrate. Unlike most substrates, protein can be sufficiently amphipathic to assist in its own translocation, making an incompletely penetrant channel such as YidC functionally sufficient. Moreover YidC is thought to serve a second function as a chaperone for IMP folding, which makes it non-redundant to SecY. The same hydrophilic groove used for transport is thought to mediate this chaperone function (Kumazaki et al., 2014b; Nagamori et al., 2004; Serdiuk et al., 2016). Other pre-fusion channel precursors may contain similar grooves for transport, but this secondary chaperone function is unique to protein substrates, because other substrates do not fold. Thus pre-fusion precursors to other channels may not be so well conserved.

Although theories about early evolutionary transitions are not experimentally testable, experimental reconstructions can at least demonstrate their plausibility. Toward that end, one could seek to engineer a pseudosymmetric channel from YidC. But it is uncontroversial that pseudosymmetric channels are formed from proteins like YidC, even if YidC duplication did not form SecY in particular. More intriguing is the implication that YidC supported the evolution of protocell membranes. Efforts to reconstruct protocells could capitalise on the synergy detailed above between

YidC and the putative primordial protein-secreting translocase (Mulkidjanian et al., 2007). This protein translocase is itself thought to have to descended from an RNA translocase, in part because its ATPase domain derives from an RNA helicase. By facilitating the integration of such an RNA translocase, YidC could indirectly facilitate the exchange of genetic information among a community of protocells, and thereby accelerate their evolution.

Acknowledgements

A.J.O.L. was supported by funding from the UK Medical Research Council and the Cambridge Commonwealth, European and International Trust. R.S.H is funded by the UK Medical Research Council (MC_ UP_A022_1007).

Methods

All-atom structural models for M. jannaschii Ylp1 (Figure 3-Figure supplement 1) were predicted using the trRosetta algorithm (Yang et al., 2020) as provided by the Yang lab's web server. This machine-learning algorithm was trained on a dataset of structures which included the incomplete crystallographic model of *M. jannaschii* Ylp1 (PDB 5c8j), and thus returns predictions for *M. jannaschii* Ylp1 which include the originally un-modelled atoms while minimally deviating from the modelled atoms, subject to certain constraints.

trRosetta was also used to generate structural models for the novel archaeal and eukaryotic EMC6 family proteins. NCBI RefSeq accession numbers for the protein sequences used are as follows: *P. horikoshii* WP_010885465.1, *Lokiarchaeum sp. GC14_75* KKK40543.1, *S. solfataricus* WP_009990433.1, *M. jannaschii* WP_010870110.1, *A. fulgidus* WP_010878056.1, *Methanosarcina* WP_011032380.1, *M. fervidus* WP_013413780.1, *T. acidophilum* WP_010900743.1, *H. jilantaiense* WP_089668789.1, *H. sapiens* NP_061328.1 (C20orf24 isoform a, a.k.a. UniParc isoform 2, Q9BUV8-2). Yeast and human EMC6 were used as homology templates (PDB 6wb9, 6z3w).

Heterodimeric contacts were predicted using the RaptorX ComplexContact algorithm (Zeng et al., 2018) as provided by the Xu group's web server.

All models were aligned and rendered in UCSF ChimeraX (Pettersen et al., 2020). Surface hydrophobicity was computed in ChimeraX by its default method: pyMLP (Broto et al., 1984; Laguerre et al., 1997) with Fauchere propagation and lipophilicity values from Ghose et al., 1998. Models depicted relative to a membrane are positioned and oriented according to the prediction algorithm provided by the Orientations of Proteins in Membranes server (Lomize et al. 2012).

Comparisons of profile hidden Markov models were performed using HHpred (Zimmermann et al., 2018) as provided by the Max Planck Institute for Developmental Biology's server, with the *H. sapiens* EMC6 (NP_001014764.1) or GET2 (NP_001736.1) or *Lokiarchaeum sp. GC14_75* Lokiarch_50810 (KKK40543.1) sequence as queries.

Hydropathy and charge were computed from protein sequences using EMBOSS Pepinfo (Madeira et al. 2019), topology predicted using TMHMM (Krogh et al. 2001), and plotted in Veusz. The 2-D histogram of IMP length vs TMH count was likewise computed and plotted in Veusz.

References

Andersson, H., and von Heijne, G. (1993). Sec dependent and sec independent assembly of E. coli inner membrane proteins: the topological rules depend on chain length. EMBO J. *12*, 683–691.

Anghel, S.A., McGilvray, P.T., Hegde, R.S., and Keenan, R.J. (2017). Identification of Oxa1 homologs operating in the eukaryotic endoplasmic reticulum. Cell Rep. *21*, 3708–3716.
Bai, L., You, Q., Feng, X., Kovach, A., and Li, H. (2020). Structure of the ER membrane complex, a transmembrane-domain insertase. Nature *584*, 475–478.
Bakelar, J., Buchanan, S.K., and Noinaj, N. (2016). The structure of the β-barrel assembly machinery complex. Science *351*, 180–186.
Bauer, M., Behrens, M., Esser, K., Michaelis, G., and Pratje, E. (1994). PET1402, a nuclear gene required for proteolytic processing of cytochrome oxidase subunit 2 in yeast. Mol. Gen. Genet. MGG *245*, 272–278.
Beltzer, J.P., Fiedler, K., Fuhrer, C., Geffen, I., Handschin, C., Wessels, H.P., and Spiess, M. (1991). Charged residues are major determinants of the transmembrane orientation of a signal-anchor sequence. J. Biol. Chem. *266*, 973–978.
Bonnefoy, N., Chalvet, F., Hamel, P., Slonimski, P.P., and Dujardin, G. (1994). OXA1, a Saccharomyces cerevisiae Nuclear Gene whose Sequence is Conserved form Prokaryotes to Eukaryotes Controls Cytochrome Oxidase Biogenesis. J. Mol. Biol. *239*, 201–212.
Borowska, M.T., Dominik, P.K., Anghel, S.A., Kossiakoff, A.A., and Keenan, R.J. (2015). A YidC-like Protein in the Archaeal Plasma Membrane. Structure *23*, 1715–1724.
Bowie, J.U. (1997). Helix packing in membrane proteins11Edited by G. von Heijne. J. Mol. Biol. *272*, 780–789.
Braunger, K., Pfeffer, S., Shrimal, S., Gilmore, R., Berninghausen, O., Mandon, E.C., Becker, T., Förster, F., and Beckmann, R. (2018). Structural basis for coupling protein transport and N-glycosylation at the mammalian endoplasmic reticulum. Science *360*, 215–219.
BROTO, P., MOREAU, G., and VANDYCKE, C. (1984). Molecular structures: perception, autocorrelation descriptor and sar studies: system of atomic contributions for the calculation of the n-octanol/water partition coefficients. Mol. Struct. Percept. Autocorrelation Descr. Sar Stud. Syst. At. Contrib. Calc. N-Octanolwater Partit. Coeff. *19*, 71–78.
Brown, J., Behnam, R., Coddington, L., Tervo, D.G.R., Martin, K., Proskurin, M., Kuleshova, E., Park, J., Phillips, J., Bergs, A.C.F., et al. (2018). Expanding the Optogenetics Toolkit by Topological Inversion of Rhodopsins. Cell *175*, 1131-1140.e11.
Brunner, J.D., Lim, N.K., Schenck, S., Duerst, A., and Dutzler, R. (2014). X-ray structure of a calcium-activated TMEM16 lipid scramblase. Nature *516*, 207–212.
Cao, T.B., and Saier, M.H. (2003). The general protein secretory pathway: phylogenetic analyses leading to evolutionary conclusions. Biochim. Biophys. Acta BBA - Biomembr. *1609*, 115–125.
Chen, Y., Capponi, S., Zhu, L., Gellenbeck, P., Freites, J.A., White, S.H., and Dalbey, R.E. (2017). YidC Insertase of Escherichia coli: Water Accessibility and Membrane Shaping. Structure *25*, 1403-1414.e3.
Engelman, D.M., and Steitz, T.A. (1981). The spontaneous insertion of proteins into and across membranes: the helical hairpin hypothesis. Cell *23*, 411–422.
Erlandson, K.J., Miller, S.B.M., Nam, Y., Osborne, A.R., Zimmer, J., and Rapoport, T.A. (2008). A role for the two-helix finger of the SecA ATPase in protein translocation. Nature *455*, 984–987.
Esser, L., Zhou, F., Yu, C.-A., and Xia, D. (2019). Crystal structure of bacterial cytochrome bc1 in complex with azoxystrobin reveals a conformational switch of the Rieske iron–sulfur protein subunit. J. Biol. Chem. *294*, 12007–12019.
Forrest, L.R. (2015). Structural Symmetry in Membrane Proteins. Annu. Rev. Biophys. *44*, 311–337.
Ghose, A.K., Viswanadhan, V.N., and Wendoloski, J.J. (1998). Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. J. Phys. Chem. A *102*, 3762–3772.
Gogala, M., Becker, T., Beatrix, B., Armache, J.-P., Barrio-Garcia, C., Berninghausen, O., and Beckmann, R. (2014). Structures of the Sec61 complex engaged in nascent peptide translocation or membrane insertion. Nature *506*, 107–110.
Guna, A., Volkmar, N., Christianson, J.C., and Hegde, R.S. (2018). The ER membrane protein complex is a transmembrane domain insertase. Science *359*, 470–473.
Guo, H., Suzuki, T., and Rubinstein, J.L. (2019). Structure of a bacterial ATP synthase. ELife *8*, e43128.
Gutiérrez-Fernández, J., Kaszuba, K., Minhas, G.S., Baradaran, R., Tambalo, M., Gallagher, D.T., and Sazanov, L.A. (2020). Key role of quinone in the mechanism of respiratory complex I. Nat. Commun. *11*, 4135.
He, H., Kuhn, A., and Dalbey, R.E. (2020). Tracking the Stepwise Movement of a Membrane-inserting Protein In Vivo. J. Mol. Biol. *432*, 484–496.
von Heijne, G. (1985). Signal sequences: The limits of variation. J. Mol. Biol. *184*, 99–105.
Hell, K., Herrmann, J.M., Pratje, E., Neupert, W., and Stuart, R.A. (1998). Oxa1p, an essential component of the N-tail protein export machinery in mitochondria. Proc. Natl. Acad. Sci. *95*, 2250–2255.
Hell, K., Neupert, W., and Stuart, R.A. (2001). Oxa1p acts as a general membrane insertion machinery for proteins encoded by mitochondrial DNA. EMBO J. *20*, 1281–1288.
Hennerdal, A., Falk, J., Lindahl, E., and Elofsson, A. (2010). Internal duplications in α-helical membrane protein topologies are common but the nonduplicated forms are rare. Protein Sci. *19*, 2305–2318.
Hennon, S.W., Soman, R., Zhu, L., and Dalbey, R.E. (2015). YidC/Alb3/Oxa1 Family of Insertases. J. Biol. Chem. *290*, 14866–14874.
Janouškovec, J., Tikhonenkov, D.V., Burki, F., Howe, A.T., Rohwer, F.L., Mylnikov, A.P., and Keeling, P.J. (2017). A New Lineage of Eukaryotes Illuminates Early Mitochondrial Genome Reduction. Curr. Biol. *27*, 3717-3724.e5.
Jonikas, M.C., Collins, S.R., Denic, V., Oh, E., Quan, E.M., Schmid, V., Weibezahn, J., Schwappach, B., Walter, P., Weissman, J.S., et al. (2009). Comprehensive Characterization of Genes Required for Protein Folding in the Endoplasmic Reticulum. Science *323*, 1693–1697.
Jungnickel, B., and Rapoport, T.A. (1995). A posttargeting signal sequence recognition event in the endoplasmic reticulum membrane. Cell *82*, 261–270.
Kater, L., Frieg, B., Berninghausen, O., Gohlke, H., Beckmann, R., and Kedrov, A. (2019). Partially inserted nascent chain unzips the lateral gate of the Sec translocon. EMBO Rep. *20*, e48191.
Koonin, E.V. (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat. Rev. Microbiol. *1*, 127–136.
Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L.L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. J. Mol. Biol. *305*, 567–580.
Kuhn, A. (1988). Alterations in the extracellular domain of M13 procoat protein make its membrane insertion dependent on secA and secY. Eur. J. Biochem. *177*, 267–271.
Kuhn, A., and Kiefer, D. (2017). Membrane protein insertase YidC in bacteria and archaea. Mol. Microbiol. *103*, 590–594.
Kuhn, P., Weiche, B., Sturm, L., Sommer, E., Drepper, F., Warscheid, B., Sourjik, V., and Koch, H.-G. (2011). The Bacterial SRP Receptor, SecA and the Ribosome Use Overlapping Binding Sites on the SecY Translocon. Traffic *12*, 563–578.
Kumazaki, K., Chiba, S., Takemoto, M., Furukawa, A., Nishiyama, K., Sugano, Y., Mori, T., Dohmae, N., Hirata, K., Nakada-Nakura, Y., et al. (2014a). Structural basis of Sec-independent membrane protein insertion by YidC. Nature *509*, 516–520.
Kumazaki, K., Kishimoto, T., Furukawa, A., Mori, H., Tanaka, Y., Dohmae, N., Ishitani, R., Tsukazaki, T., and Nureki, O. (2014b). Crystal structure of Escherichia coli YidC, a membrane protein chaperone and insertase. Sci. Rep. *4*, 1–6.
Laguerre, M., Saux, M., Dubost, J.P., and Carpy, A. (1997). MLPP: A Program for the Calculation of Molecular Lipophilicity Potential in Proteins. Pharm. Pharmacol. Commun. *3*, 217–222.
Lane, N., and Martin, W.F. (2012). The Origin of Membrane Bioenergetics. Cell *151*, 1406–1416.
Li, L., Park, E., Ling, J., Ingram, J., Ploegh, H., and Rapoport, T.A. (2016). Crystal structure of a substrate-engaged SecY protein-translocation channel. Nature *531*, 395.
Lolkema, J.S., Dobrowolski, A., and Slotboom, D.-J. (2008). Evolution of Antiparallel Two-domain Membrane Proteins: Tracing Multiple Gene Duplication Events in the DUF606 Family. J. Mol. Biol. *378*, 596–606.
Lombard, J., López-García, P., and Moreira, D. (2012). The early evolution of lipid membranes and the three domains of life. Nat. Rev. Microbiol. *10*, 507–515.
Lomize, M.A., Pogozheva, I.D., Joo, H., Mosberg, H.I., and Lomize, A.L. (2012). OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res. *40*, D370–D376.
Ma, C., Wu, X., Sun, D., Park, E., Catipovic, M.A., Rapoport, T.A., Gao, N., and Li, L. (2019). Structure of the substrate-engaged SecA-SecY protein translocation machine. Nat. Commun. *10*, 2872.
Madeira, F., Park, Y. mi, Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., Potter, S.C., Finn, R.D., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. *47*, W636–W641.
Makarova, K.S., Galperin, M.Y., and Koonin, E.V. (2015). Comparative genomic analysis of evolutionarily conserved but functionally uncharacterized membrane proteins in archaea: Prediction of novel components of secretion, membrane remodeling and glycosylation systems. Biochimie *118*, 302–312.
Mariappan, M., Mateja, A., Dobosz, M., Bove, E., Hegde, R.S., and Keenan, R.J. (2011). The mechanism of membrane-associated steps in tail-anchored protein insertion. Nature *477*, 61.

Matlack, K.E., Misselwitz, B., Plath, K., and Rapoport, T.A. (1999). BiP acts as a molecular ratchet during posttranslational transport of prepro-alpha factor across the ER membrane. Cell *97*, 553–564.
McDowell, M.A., Heimes, M., Fiorentino, F., Mehmood, S., Farkas, Á., Coy-Vergara, J., Wu, D., Bolla, J.R., Schmid, V., Heinze, R., et al. (2020). Structural Basis of Tail-Anchored Membrane Protein Biogenesis by the GET Insertase Complex. Mol. Cell.
McGilvray, P.T., Anghel, S.A., Sundaram, A., Zhong, F., Trnka, M.J., Fuller, J.R., Hu, H., Burlingame, A.L., and Keenan, R.J. (2020). An ER translocon for multi-pass membrane protein biogenesis. ELife *9*, e56889.
Miller-Vedam, L.E., Bräuning, B., Popova, K.D., Oakdale, N.T.S., Bonnar, J.L., Prabu, J.R., Boydston, E.A., Sevillano, N., Shurtleff, M.J., Stroud, R.M., et al. (2020). Structural and mechanistic basis of the EMC-dependent biogenesis of distinct transmembrane clients. BioRxiv 2020.09.02.280008.
Mothes, W., Prehn, S., and Rapoport, T.A. (1994). Systematic probing of the environment of a translocating secretory protein during translocation through the ER membrane. EMBO J. *13*, 3973–3982.
Mulkidjanian, A.Y., Makarova, K.S., Galperin, M.Y., and Koonin, E.V. (2007). Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. Nat. Rev. Microbiol. *5*, 892–899.
Mulkidjanian, A.Y., Galperin, M.Y., and Koonin, E.V. (2009). Co-evolution of primordial membranes and membrane proteins. Trends Biochem. Sci. *34*, 206–215.
Nagamori, S., Smirnova, I.N., and Kaback, H.R. (2004). Role of YidC in folding of polytopic membrane proteins. J. Cell Biol. *165*, 53–62.
O'Donnell, J.P., Phillips, B.P., Yagita, Y., Juszkiewicz, S., Wagner, A., Malinverni, D., Keenan, R.J., Miller, E.A., and Hegde, R.S. (2020). The architecture of EMC reveals a path for membrane protein insertion. ELife *9*, e57887.
Ooi, C.E., and Weiss, J. (1992). Bidirectional movement of a nascent polypeptide across microsomal membranes reveals requirements for vectorial translocation of proteins. Cell *71*, 87–96.
Paetzel, M., Karla, A., Strynadka, N.C.J., and Dalbey, R.E. (2002). Signal peptidases. Chem. Rev. *102*, 4549–4580.
Park, E., and Rapoport, T.A. (2012). Mechanisms of Sec61/SecY-mediated protein translocation across membranes. Annu. Rev. Biophys. *41*, 21–40.
Park, E., Ménétret, J.-F., Gumbart, J.C., Ludtke, S.J., Li, W., Whynot, A., Rapoport, T.A., and Akey, C.W. (2014). Structure of the SecY channel during initiation of protein translocation. Nature *506*, 102–106.
Peltier, J.-B., Emanuelsson, O., Kalume, D.E., Ytterberg, J., Friso, G., Rudella, A., Liberles, D.A., Söderberg, L., Roepstorff, P., Heijne, G. von, et al. (2002). Central Functions of the Lumenal and Peripheral Thylakoid Proteome of Arabidopsis Determined by Experimentation and Genome-Wide Prediction. Plant Cell *14*, 211–236.
Petersen, T.N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods *8*, 785–786.
Petriman, N.-A., Jauß, B., Hufnagel, A., Franz, L., Sachelaru, I., Drepper, F., Warscheid, B., and Koch, H.-G. (2018). The interaction network of the YidC insertase with the SecYEG translocon, SRP and the SRP receptor FtsY. Sci. Rep. *8*, 578.
Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., and Ferrin, T.E. (2020). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci. Publ. Protein Soc.
Plath, K., Mothes, W., Wilkinson, B.M., Stirling, C.J., and Rapoport, T.A. (1998). Signal Sequence Recognition in Posttranslational Protein Transport across the Yeast ER Membrane. Cell *94*, 795–807.
Pleiner, T., Tomaleri, G.P., Januszyk, K., Inglis, A.J., Hazu, M., and Voorhees, R.M. (2020). Structural basis for membrane insertion by the human ER membrane protein complex. Science *369*, 433–436.
Rapp, M., Granseth, E., Seppälä, S., and von Heijne, G. (2006). Identification and evolution of dual-topology membrane proteins. Nat. Struct. Mol. Biol. *13*, 112–116.
Rapp, M., Seppälä, S., Granseth, E., and Heijne, G. von (2007). Emulating Membrane Protein Evolution by Rational Design. Science *315*, 1282–1284.
Rollauer, S.E., Tarry, M.J., Graham, J.E., Jääskeläinen, M., Jäger, F., Johnson, S., Krehenbrink, M., Liu, S.-M., Lukey, M.J., Marcoux, J., et al. (2012). Structure of the TatC core of the twin-arginine protein transport system. Nature *492*, 210–214.
Sääf, A., Johansson, M., Wallin, E., and Heijne, G. von (1999). Divergent evolution of membrane protein topology: The Escherichia coli RnfA and RnfE homologues. Proc. Natl. Acad. Sci. *96*, 8540–8544.
Samuelson, J.C., Chen, M., Jiang, F., Möller, I., Wiedmann, M., Kuhn, A., Phillips, G.J., and Dalbey, R.E. (2000). YidC mediates membrane protein insertion in bacteria. Nature *406*, 637–641.
Saracco, S.A., and Fox, T.D. (2002). Cox18p Is Required for Export of the Mitochondrially Encoded Saccharomyces cerevisiae Cox2p C-Tail and Interacts with Pnt1p and Mss2p in the Inner Membrane. Mol. Biol. Cell *13*, 1122–1131.
Schmidt, A., Kochanowski, K., Vedelaar, S., Ahrné, E., Volkmer, B., Callipo, L., Knoops, K., Bauer, M., Aebersold, R., and Heinemann, M. (2016). The quantitative and condition-dependent Escherichia coli proteome. Nat. Biotechnol. *34*, 104–110.
Seitl, I., Wickles, S., Beckmann, R., Kuhn, A., and Kiefer, D. (2014). The C-terminal regions of YidC from Rhodopirellula baltica and Oceanicaulis alexandrii bind to ribosomes and partially substitute for SRP receptor function in Escherichia coli. Mol. Microbiol. *91*, 408–421.
Selosse, M.-A., Albert, B., and Godelle, B. (2001). Reducing the genome size of organelles favours gene transfer to the nucleus. Trends Ecol. Evol. *16*, 135–141.
Serdiuk, T., Balasubramaniam, D., Sugihara, J., Mari, S.A., Kaback, H.R., and Müller, D.J. (2016). YidC assists the stepwise and stochastic folding of membrane proteins. Nat. Chem. Biol. *12*, 911–917.
Serdiuk, T., Steudle, A., Mari, S.A., Manioglu, S., Kaback, H.R., Kuhn, A., and Müller, D.J. (2019). Insertion and folding pathways of single membrane proteins guided by translocases and insertases. Sci. Adv. *5*, eaau6824.
Shanmugam, S.K., Backes, N., Chen, Y., Belardo, A., Phillips, G.J., and Dalbey, R.E. (2019). New Insights into Amino-Terminal Translocation as Revealed by the Use of YidC and Sec Depletion Strains. J. Mol. Biol. *431*, 1025–1037.
Shaw, A.S., Rottier, P.J., and Rose, J.K. (1988). Evidence for the loop model of signal-sequence insertion into the endoplasmic reticulum. Proc. Natl. Acad. Sci. U. S. A. *85*, 7592–7596.
Stefer, S., Reitz, S., Wang, F., Wild, K., Pang, Y.-Y., Schwarz, D., Bomke, J., Hein, C., Löhr, F., Bernhard, F., et al. (2011). Structural Basis for Tail-Anchored Membrane Protein Biogenesis by the Get3-Receptor Complex. Science *333*, 758–762.
Sundberg, E., Slagter, J.G., Fridborg, I., Cleary, S.P., Robinson, C., and Coupland, G. (1997). ALBINO3, an Arabidopsis nuclear gene essential for chloroplast differentiation, encodes a chloroplast protein that shows homology to proteins present in bacterial membranes and yeast mitochondria. Plant Cell *9*, 717–730.
Svensson-Ek, M., Abramson, J., Larsson, G., Törnroth, S., Brzezinski, P., and Iwata, S. (2002). The X-ray Crystal Structures of Wild-type and EQ(I-286) Mutant Cytochrome c Oxidases from Rhodobacter sphaeroides. J. Mol. Biol. *321*, 329–339.
Szyrach, G., Ott, M., Bonnefoy, N., Neupert, W., and Herrmann, J.M. (2003). Ribosome binding to the Oxa1 complex facilitates co-translational protein insertion in mitochondria. EMBO J. *22*, 6448–6457.
The UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47*, D506–D515.
Van den Berg, B., Clemons Jr, W.M., Collinson, I., Modis, Y., Hartmann, E., Harrison, S.C., and Rapoport, T.A. (2004). X-ray structure of a protein-conducting channel. Nature *427*, 36.
Voorhees, R.M., and Hegde, R.S. (2016). Structure of the Sec61 channel opened by a signal sequence. Science *351*, 88–91.
Vothknecht, U.C., and Westhoff, P. (2001). Biogenesis and origin of thylakoid membranes. Biochim. Biophys. Acta BBA - Mol. Cell Res. *1541*, 91–101.
Wallin, E., and von Heijne, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci. *7*, 1029–1038.
Wang, F., Whynot, A., Tung, M., and Denic, V. (2011). The mechanism of tail-anchored protein insertion into the ER membrane. Mol. Cell *43*, 738–750.
Weiss, M.C., Sousa, F.L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., and Martin, W.F. (2016). The physiology and habitat of the last universal common ancestor. Nat. Microbiol. *1*, 1–8.
Welte, T., Kudva, R., Kuhn, P., Sturm, L., Braig, D., Müller, M., Warscheid, B., Drepper, F., and Koch, H.-G. (2011). Promiscuous targeting of polytopic membrane proteins to SecYEG or YidC by the Escherichia coli signal recognition particle. Mol. Biol. Cell *23*, 464–479.
Weng, T.-H., Steinchen, W., Beatrix, B., Berninghausen, O., Becker, T., Bange, G., Cheng, J., and Beckmann, R. (2020). Architecture of the active post-translational Sec translocon. EMBO J. e105643.
White, S.H., and von Heijne, G. (2005). Transmembrane helices before, during, and after insertion. Curr. Opin. Struct. Biol. *15*, 378–386.
Wiedemann, N., and Pfanner, N. (2017). Mitochondrial Machineries for Protein Import and Assembly. Annu. Rev. Biochem. *86*, 685–714.
Wu, X., Siggel, M., Ovchinnikov, S., Mi, W., Svetlov, V., Nudler, E., Liao, M., Hummer, G., and Rapoport, T.A. (2020). Structural basis of ER-associated protein degradation mediated by the Hrd1 ubiquitin ligase complex. Science *368*.
Xu, X., Ouyang, M., Lu, D., Zheng, C., and Zhang, L. (2020). Protein Sorting within Chloroplasts. Trends Cell Biol.

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. *117*, 1496–1503.
Yen, M.-R., Harley, K.T., Tseng, Y.-H., and Saier, M.H. (2001). Phylogenetic and structural analyses of the oxa1 family of protein translocases. FEMS Microbiol. Lett. *204*, 223–231.
Yi, L., Jiang, F., Chen, M., Cain, B., Bolhuis, A., and Dalbey, R.E. (2003). YidC Is Strictly Required for Membrane Insertion of Subunits a and c of the F1F0ATP Synthase and SecE of the SecYEG Translocase. Biochemistry *42*, 10537–10544.
Zeng, H., Wang, S., Zhou, T., Zhao, F., Li, X., Wu, Q., and Xu, J. (2018). ComplexContact: a web server for inter-protein contact prediction using deep learning. Nucleic Acids Res. *46*, W432–W437.
Zhang, Y., Ou, X., Wang, X., Sun, D., Zhou, X., Wu, X., Li, Q., and Li, L. (2020). Structure of the mitochondrial TIM22 complex from yeast. Cell Res. 1–3.
Zhang, Y.-J., Tian, H.-F., and Wen, J.-F. (2009). The evolution of YidC/Oxa/Alb3 family in the three domains of life: a phylogenomic analysis. BMC Evol. Biol. *9*, 137.
Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A.N., and Alva, V. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J. Mol. Biol. *430*, 2237–2243.

605

Supplementary Figures

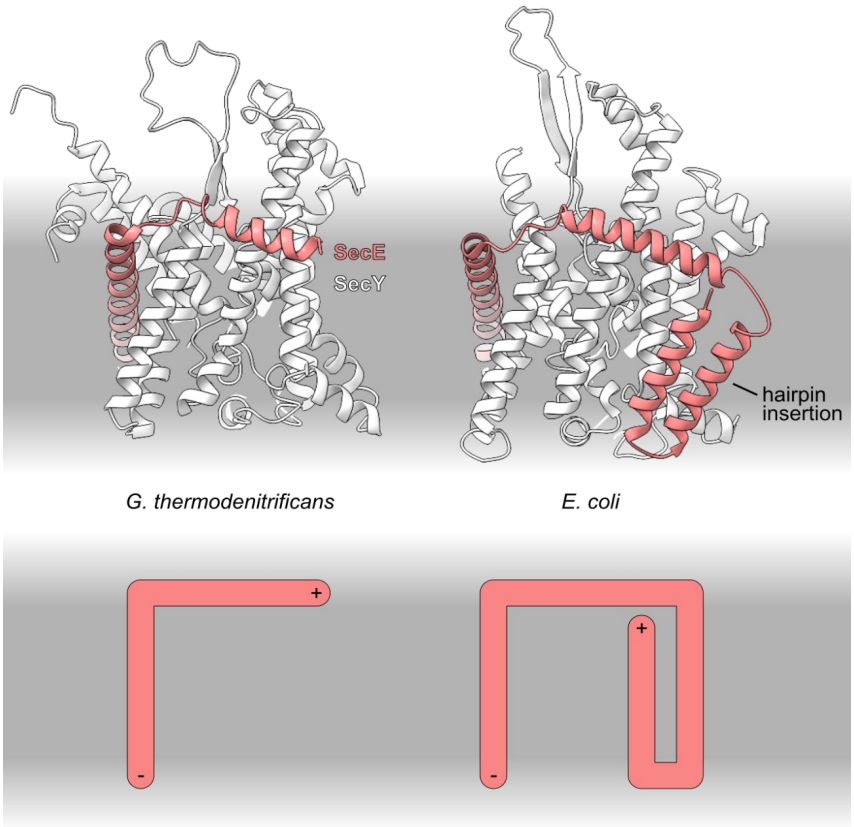

Figure 2-Figure supplement 1. Acquisition of a transmembrane hairpin in SecE. Left: *G. thermodenitrificans* SecYE (PDB 6itc; Ma et al. 2019). Right: *E. coli* SecYE (PDB 6r7L; Kater et al. 2019). Top: molecular models of SecY (white) and SecE (light coral). Bottom: diagram of SecE topology, in which plus and minus signs indicate the N- and C-termini.

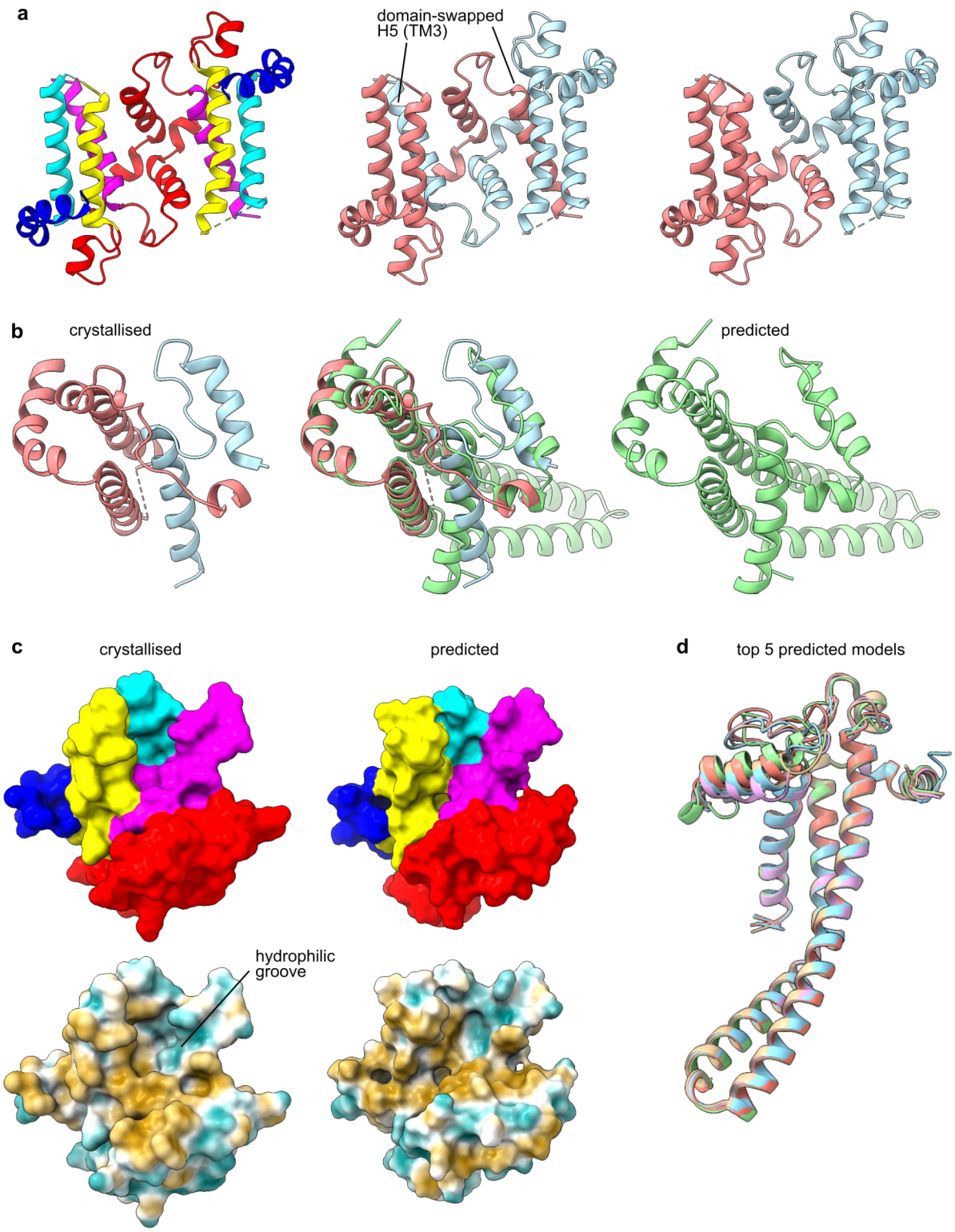

Figure 3-Figure supplement 1. Comparison of crystallised and predicted monomeric structures of Ylp1 from *M. jannaschii*. **a** Domain swapping in crystallised Ylp1 (PDB 5c8j). Left: colour-coded by consensus element, as in Figure 3: H0 (blue), H1 (cyan), H4 (yellow), H5 (magenta). H2/3 were relatively disordered and not modeled. Center: colour-coded by chain. Right: colour-coded by chain, except the domain-swapped parts of H4/5 have been recoloured to match the chain against which it packs. **b** Alignment of the crystal structure (left, red and blue chains) with the trRosetta-predicted monomeric structure (right, green). **c** Comparison of the solvent-excluded surfaces of the crystallised and predicted structures, colour-coded by consensus element (top) or by hydropathy (bottom). Hydropathy ranges from hydrophilic (dark cyan) to intermediate (white) to hydrophobic (dark goldenrod). **d** Alignment of the 5 highest-probability models built by trRosetta. The models were aligned by their first 40 amino acids.

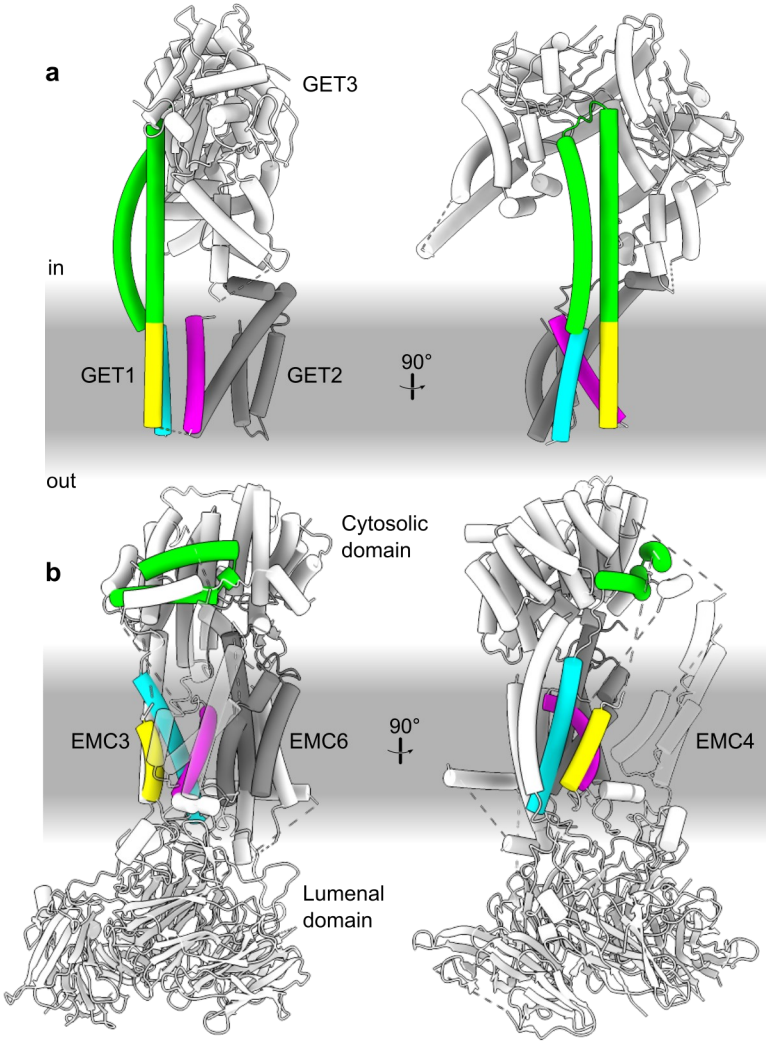

Figure 3-Figure supplement 2. Structures of the EMC and Get complexes. **a** Structure of the GET1/2/3 complex (*H. sapiens*, PDB 6so5). A second GET1/2 dimer present in the PDB model is not shown. **b** Structure of the EMC complex (*S. cerevisiae*, 6wb9) with EMC4 rendered partially transparent.

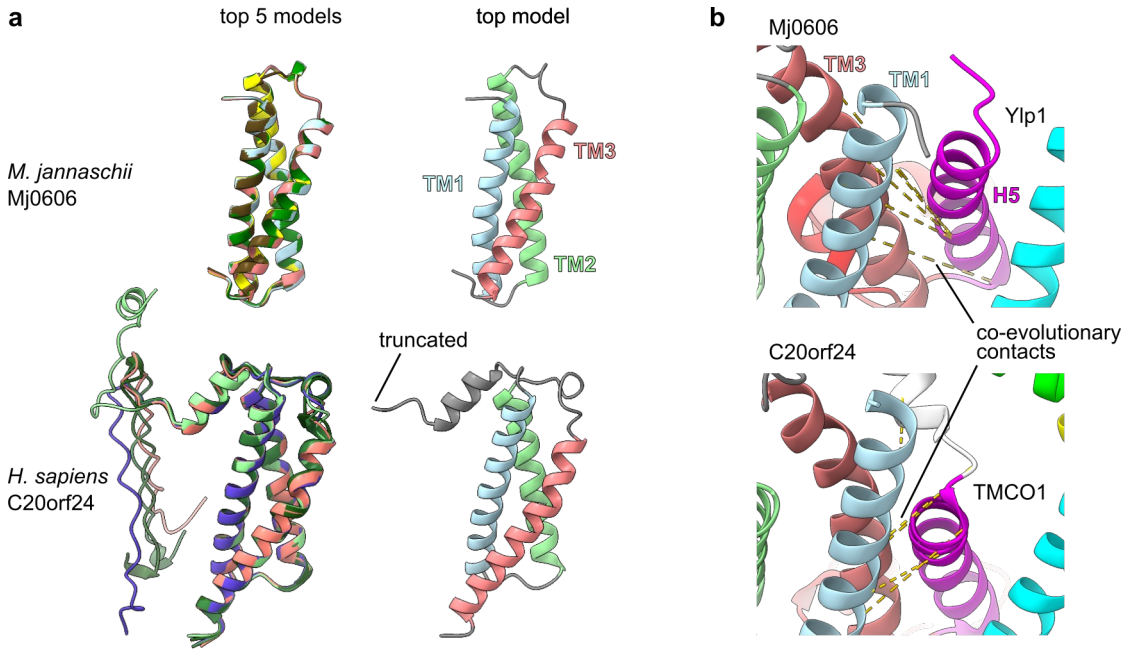

Figure 7-Figure supplement 1. Co-evolutionary structure and heterodimeric contact prediction for Mj0606 and C20orf24. **a** Alignment of the five highest-probability models determined by trRosetta (left) and an annotated view of the single best model (right), with disordered residues omitted. Models were predicted from the protein sequence of Mj0606 and C20orf24. **b** Close-up view of the five highest-probability contacts between Ylp1/Mj0606 and C20orf24/TMCO1, represented by dashed gold lines.

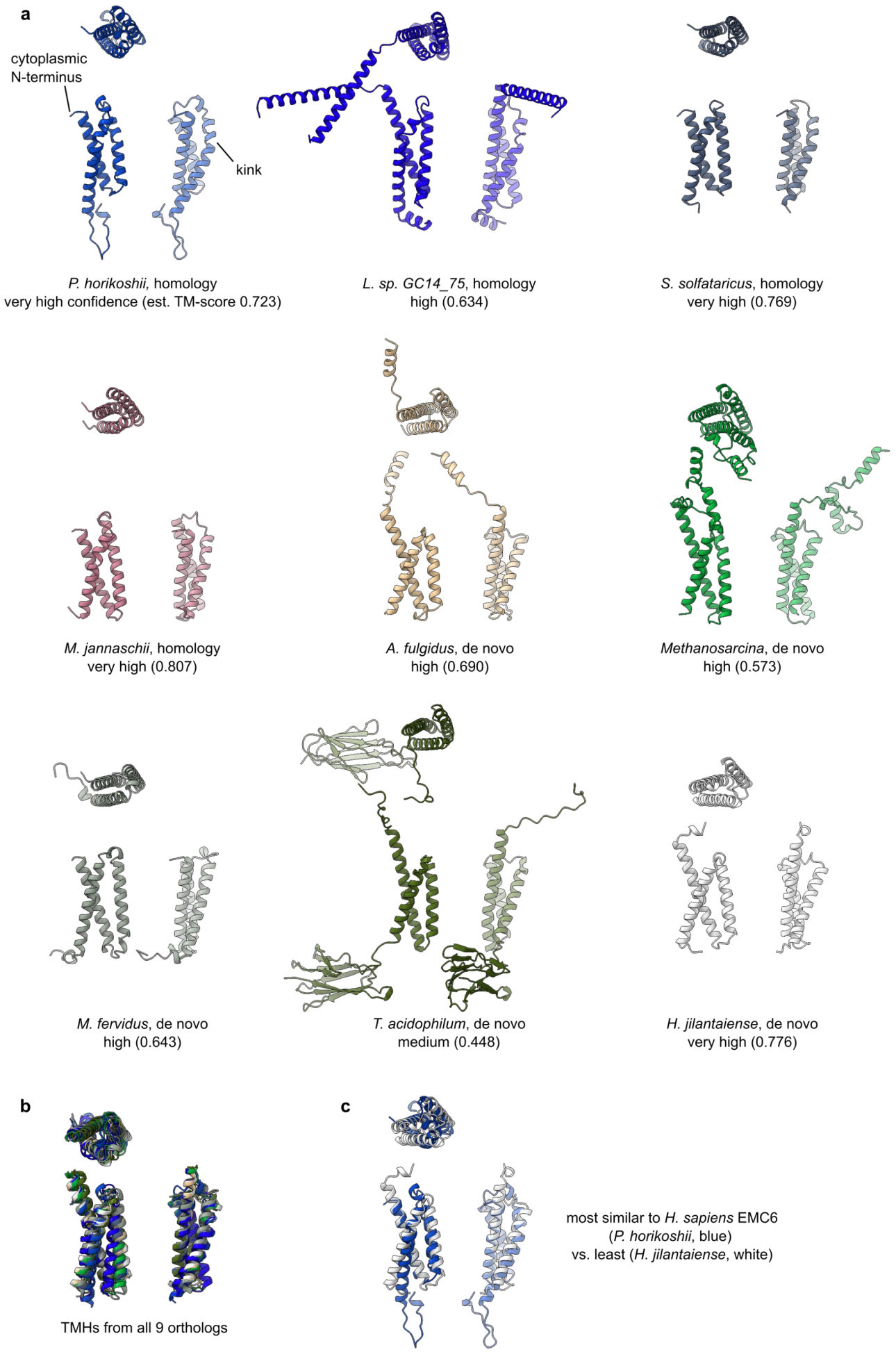

Figure 7-Figure supplement 2. Structural models for nine diverse archaeal EMC6 family proteins. **a** Front, top, and side views of each model, separated by 90° rotations. A kink in TM3 which is present in the homology models but absent from the *de novo* models is indicated. **b,c** Alignments.