Lewis and Hegde *BMC Biology* (2021) 19:266
https://doi.org/10.1186/s12915-021-01171-5

BMC Biology

RESEARCH ARTICLE Open Access

# A unified evolutionary origin for the ubiquitous protein transporters SecY and YidC

Aaron J. O. Lewis* and Ramanujan S. Hegde*

## Abstract

**Background:** Protein transporters translocate hydrophilic segments of polypeptide across hydrophobic cell membranes. Two protein transporters are ubiquitous and date back to the last universal common ancestor: SecY and YidC. SecY consists of two pseudosymmetric halves, which together form a membrane-spanning protein-conducting channel. YidC is an asymmetric molecule with a protein-conducting hydrophilic groove that partially spans the membrane. Although both transporters mediate insertion of membrane proteins with short translocated domains, only SecY transports secretory proteins and membrane proteins with long translocated domains. The evolutionary origins of these ancient and essential transporters are not known.

**Results:** The features conserved by the two halves of SecY indicate that their common ancestor was an antiparallel homodimeric channel. Structural searches with SecY's halves detect exceptional similarity with YidC homologs. The SecY halves and YidC share a fold comprising a three-helix bundle interrupted by a helical hairpin. In YidC, this hairpin is cytoplasmic and facilitates substrate delivery, whereas in SecY, it is transmembrane and forms the substrate-binding lateral gate helices. In both transporters, the three-helix bundle forms a protein-conducting hydrophilic groove delimited by a conserved hydrophobic residue. Based on these similarities, we propose that SecY originated as a YidC homolog which formed a channel by juxtaposing two hydrophilic grooves in an antiparallel homodimer. We find that archaeal YidC and its eukaryotic descendants use this same dimerisation interface to heterodimerise with a conserved partner. YidC's sufficiency for the function of simple cells is suggested by the results of reductive evolution in mitochondria and plastids, which tend to retain SecY only if they require translocation of large hydrophilic domains.

**Conclusions:** SecY and YidC share previously unrecognised similarities in sequence, structure, mechanism, and function. Our delineation of a detailed correspondence between these two essential and ancient transporters enables a deeper mechanistic understanding of how each functions. Furthermore, key differences between them help explain how SecY performs its distinctive function in the recognition and translocation of secretory proteins. The unified theory presented here explains the evolution of these features, and thus reconstructs a key step in the origin of cells.

**Keywords:** Oxa1 superfamily, Protein translocation, Membrane protein integration, Protocell evolution, SecY, YidC

* Correspondence: aaron.ohare.lewis@gmail.com; rhegde@mrc-lmb.cam.ac.uk
MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK

## Background

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 [1, 2]. Among those IMPs was SecY, a protein-conducting channel [3]. 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 typically requires that its hydrophilic translocation substrates be connected to a hydrophobic α-helix [4–6]. These hydrophobic helices serve as signals which open the SecY channel [7–9]. SecY is comprised of two separate halves [10] which open like a clamshell when a helix binds to the lipid interface between them (Fig. 1a). Spreading the halves apart destabilises a plug which sits between them, opening a hydrophilic pore that spans the width of the membrane. By binding at this site, the signal also threads one of its hydrophilic flanking regions through the hydrophilic pore, thereby initiating its translocation.

A set of conserved hydrophobic residues that line the narrowest part of the translocation pore form a gasket-like seal around the translocating chain [11]. These residues, which are contributed by both halves of SecY, are known collectively as the pore ring. They not only maintain the ion permeability barrier across the membrane [12, 13], but also bind the plug when the channel is closed [10].

The site between SecY's halves where signals bind is called the lateral gate. After binding and initiating translocation, sufficiently hydrophobic signals can diffuse away from the lateral gate into the surrounding hydrophobic membrane [14]. Many signals, particularly those at the N-terminus of secretory proteins, are ultimately cleaved off by signal peptidase, a membrane-anchored protease whose active site resides on the extracytoplasmic side of the membrane [15]. Longer and more hydrophobic signals that are not cleaved serve as the transmembrane helices (TMHs) of IMPs [16].

SecY is the only ubiquitous transporter for protein secretion. There is however a second ubiquitous superfamily of protein transporters which is specialised for IMP integration, Oxa1 [17]. The Oxa1 superfamily consists of four member families, each of which now has a known atomic structure. One, YidC, is found in the prokaryotic plasma membrane [18, 19], whereas the other three are paralogs located in the eukaryotic endoplasmic reticulum (ER): TMCO1 [20], EMC3 [21–24], and GET1 [25]. All share a conserved core of three TMHs and a cytoplasmic helical hairpin. With YidC also present in the plastid (Alb3 [26]) and mitochondrial inner membranes (Oxa1 [27, 28]), it appears that every membrane equivalent to the plasma membrane of the cenancestor contains Oxa1 superfamily proteins. As with SecY, archaeal and bacterial YidC are monophyletic and highly

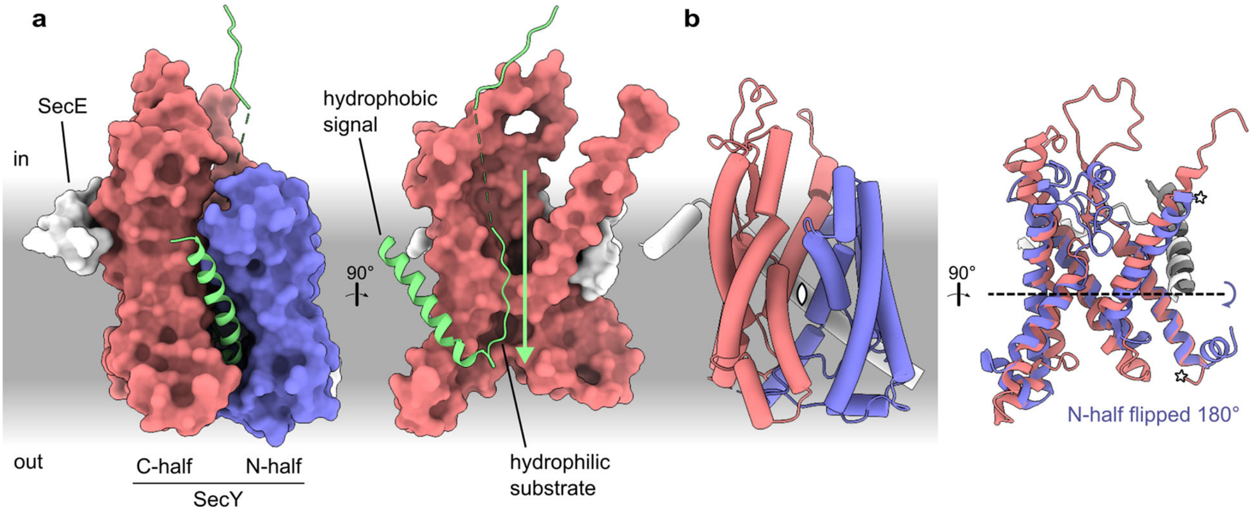


**Fig. 1** Structure and pseudosymmetry of the protein-conducting channel SecY. **a** Left: Structure of the channel engaged by a secretory substrate: *Geobacillus thermodenitrificans* SecYE engaged by proOmpA (Protein Data Bank ID [PDB] 6itc). The cytoplasmic ATPase SecA is present in the model but not shown. Right: Rotated view of only the SecY N-half and substrate. **b** Pseudosymmetry of the N- and C-halves. Left: SecYE shown as tubes with the pseudo-$C_2$ symmetry axis denoted by a pointed oval. Right: Rotated view in ribbon representation. The N-half has been rotated 180° around the pseudo-$C_2$ symmetry axis and aligned to the C-half. SecE is divided where it intersects the symmetry axis into N-terminal (white) and C-terminal (grey) segments. A dashed black line indicates the same pseudo-$C_2$ symmetry axis shown at left after a 90° rotation. Stars indicate where the halves were split

divergent [17, 29], suggesting that YidC was present alongside SecY in the cenancestor.

Like SecY, YidC facilitates IMP integration by translocating extracytoplasmic segments across the membrane [18, 30–32]. Unlike SecY substrates, however, YidC substrates are limited in the length of polypeptide translocated, typically to less than 30 amino acids [33]. This limitation may be due to YidC's lack of a membrane-spanning hydrophilic pore; instead, YidC structures show a membrane-exposed hydrophilic groove that only penetrates partway into the membrane [19]. YidC thus forms a partial channel and may also thin and distort the adjacent membrane [34].

The two halves of SecY are structurally similar and related by a two-fold rotational ($C_2$) pseudosymmetry axis parallel to the membrane plane (Fig. 1b) [10]. Such pseudosymmetry is common among membrane proteins and arises when the gene encoding an asymmetric progenitor undergoes duplication and fusion [35]. Channels are particularly likely to have a membrane-parallel $C_2$ axis of structural symmetry because they have the same axis of functional symmetry: they facilitate substrates' bidirectional diffusion across the membrane. Indeed, polypeptides can slide through SecY bidirectionally [36], with unidirectionality arising from other factors [37, 38]. Membrane-parallel $C_2$ pseudosymmetry requires that the two fused domains be antiparallel, and thus those domains typically derive from progenitors that existed as antiparallel homodimers [39, 40].

The ubiquity and essentiality of the SecY channel motivated us to investigate how it might have evolved. We identify several structural elements that are conserved between its two halves, which suggest that the SecY progenitor was an antiparallel homodimer featuring a symmetric pore ring at its dimerisation interface. Automated database searches for structures similar to the SecY halves show that they are uniquely similar to the Oxa1 superfamily, of which YidC is the prokaryotic member. Structural alignments indicate that key residues of YidC's hydrophilic groove and its capping hydrophobic residue are homologous to key residues in SecY's hydrophilic funnels and its pore ring, respectively.

In light of this new correspondence, we re-evaluate the extensive mechanistic literature on SecY and the Oxa1 superfamily, identifying surprising similarities and specific structural bases for their differences. Based on this analysis, we propose that SecY evolved from a dimeric Oxa1 superfamily member by gene duplication and fusion. We compare the range of substrates that can be translocated by YidC to the prokaryotic membrane proteome and find that a YidC-dependent, SecY-independent cell is plausible. We discuss the implications of this model for the evolution of YidC itself and other components of the general secretory pathway.

## Results

### Conserved pre-duplication features in SecY

Features shared by both of SecY's halves are likely to have been present in their last common ancestor, which we term proto-SecY. In an attempt to identify conserved sequence features of proto-SecY, we aligned the amino acid (a.a.) sequences of a set of N- and C-halves. However, their pairwise identities are just 12.5 ± 2.2% (s.d.), compared to 9.3 ± 4.3% between randomly shuffled sequences, an excess identity of only 6 a.a. per 200 a.a. half. By pairwise HHpred [41], the halves have similarity $p$ = 0.02, where $p$ estimates the likelihood of observing as much similarity between a random pair of unrelated proteins [42, 43]. For context, this means that in searching a typical whole-proteome database of ~ $10^4$ entries with one half of SecY, one would expect to find ~ 200 unrelated proteins just as similar as the other half of SecY. Reconstructing a cenancestral SecY sequence using methods previously successful for a different internally duplicated protein [44] yielded no increase in similarity between the SecY halves (see Additional file 1). Thus, the two halves of SecY have diverged too far from one another to reliably reconstruct proto-SecY's primary sequence.

Unlike primary sequence, a five-TMH tertiary structure is conserved by both halves of SecY (Fig. 1b [10]). To facilitate comparisons, we label these five consensus helices H1-H5 (Fig. 2a). A prefix of N or C is used when referring to a specific instance of a consensus element in the N- or C-half of SecY. For example, TM6 of SecY is labelled C.H1 because it is located in the C-half and corresponds to H1 of proto-SecY, as does TM1 (N.H1) in the N-half. 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 half, which we argue later was probably also present in proto-SecY, is named H0.

To identify more detailed conserved features, we pursued a precise structural alignment of the SecY halves. We collected a representative set of seven SecY structures: closed [45, 46], primed [47, 48], and open [9, 11] models for eukaryotic and bacterial SecY, and closed archaeal SecY [10]. Alignments between these halves generated by mTM-align [49] vary widely in accuracy and extent (Additional file 2: Figure S1a), but display one clear trend: the C-halves in the closed and primed structures are least like the N-halves of any structure. This is because closure induces symmetry-breaking tilts in C.H2 and C.H5 (Additional file 2: Figure S1b) whereas the N-halves remain relatively unchanged.

Intriguingly, the stability of this asymmetrical closed conformation depends on another asymmetrical feature, the plug (Fig. 2a [50]). This suggests that neither the

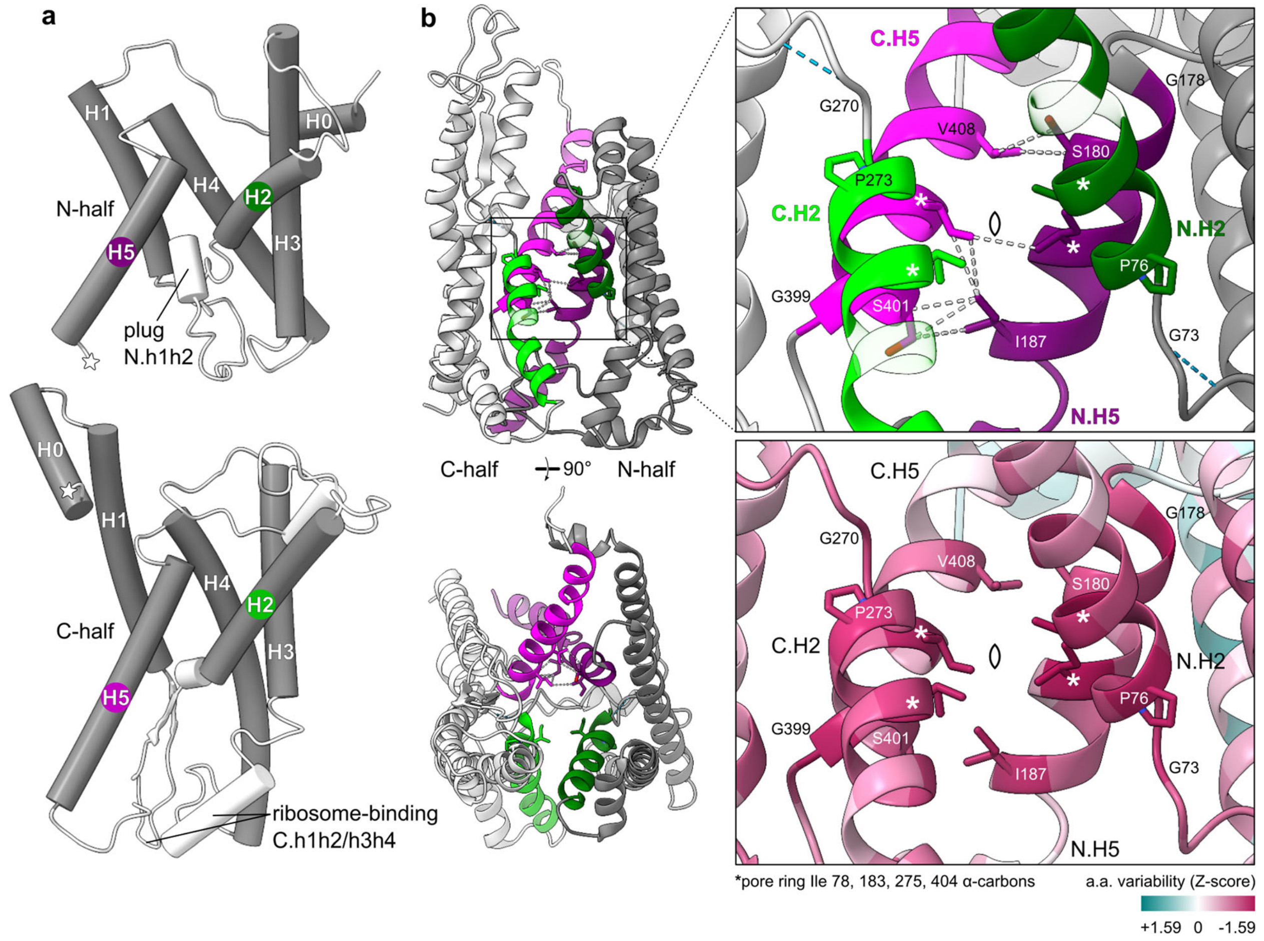


**Fig. 2** Features conserved between SecY's halves. **a** Consensus secondary structure elements (grey) in *Methanocaldococcus jannaschii* SecY (1rhz). Stars indicate where the halves were split. **b** Symmetric features in the most self-similar SecY structure (6itc). A pointed oval indicates the pseudo-$C_2$ symmetry axis. Dashed lines indicate hydrogen bonds (blue) and Van der Waals contacts (white). ConSurf variability scores for an alignment of the N- and C-half sequences are shown mapped onto each half's model. The colour scale encompasses the minimum but not maximum score. The most conserved residues are shown as sticks and labelled

plug nor the closed conformation may have been present in proto-SecY. A plugless proto-SecY is plausible, given that plug-deletion mutants of SecY are tolerated [13, 50, 51]. If proto-SecY did lack SecY-like gating or a plug, it would then more closely resemble other protein transporters like YidC or TatC, which are not gated [19, 52].

The most similar halves, 6fti N [47] and 6itc C [11], share a common core (< 4 Å deviation) of 121 a.a. across all 5 helices with 1.9 Å RMSD (Additional file 2: Figure S1c). This is precise enough that all 5 helices can be registered confidently. Their alignment shows that the four functionally important pore ring residues [12] are located at the same two homologous sites in each half. To identify other conserved sites, this structural alignment was then used to align a diverse set of N- and C-half sequences (Additional file 3). Their sequence conservation at each site was then scored and mapped to the most self-similar SecY structure (6itc).

This scored structure shows that the interface between halves is symmetrical and conserved (Fig. 2b). N.H5 and C.H5 contact each other via the H5 pore ring residue, which coincides with the symmetry axis, and also via residues −3 and +4 a.a. from the pore ring. Although SecY today is a pseudodimer, split mutants show that it remains able to form true dimers via this interface [53]. Less conserved than the pore ring but still notable are two helix-breaking residues which N-terminate H5 (glycine) and H2 (proline), and a glycine near H2 which bonds its α-hydrogen with the −3 backbone oxygen, thereby stabilising a small bulge. These conserved residues are all within 5 a.a. of the pore ring, underscoring the structural conservation of this central region. Altogether, these features suggest that while proto-SecY may not have had SecY-like gating or a plug, it did form antiparallel homodimers centred on a pore very similar to SecY's. Thus proto-SecY likely functioned as a protein-conducting channel.

### SecY is uniquely similar in structure to the Oxa1 superfamily

With this information about proto-SecY, we sought to identify distant homologs from before its duplication. For this, we used Dali, which measures structural similarity between protein backbones. Dali is competitive with other top methods for accurate homolog detection and outperforms them when the relationships in question are particularly distant [54]. Other methods construct 3-D superpositions with better geometric properties like RMSD, but Dali nonetheless outperforms them in detection accuracy [55]. Thus, we use Dali here, whereas a method optimised for 3-D superposition, mTM-align [49], was used above to align the SecY halves.

Queries of the PDB with the N- or C-half of SecY yielded a match correlation matrix [56] that indicates the possible presence of two separate subdomains (Fig. 3a). The three-helix bundle of H1/4/5 showed positive self-correlation, but anti-correlation with the H2/3 two-helix hairpin. Because Dali measures global similarity, including both subdomains in our searches would tend to obscure distant homologs which share only one subdomain [57]. We therefore performed searches with not only the whole N- and C-halves, but also the largest subdomain, H1/4/5 (Fig. 3b). We queried a non-redundant subset of the PDB filtered at 25% pairwise identity (PDB25).

After excluding SecY and soluble hits, the most consistently high-ranking hits were members of the Oxa1 superfamily (Fig. 3c; Additional file 4). Moreover, these hits link multiple Oxa1 families (GET1 and EMC3) to both SecY halves, suggesting that their similarity is due to conserved characteristics of the Oxa1 superfamily and proto-SecY rather than idiosyncrasies of any one structure. By contrast, almost all other hits were as highly ranked in only a single query.

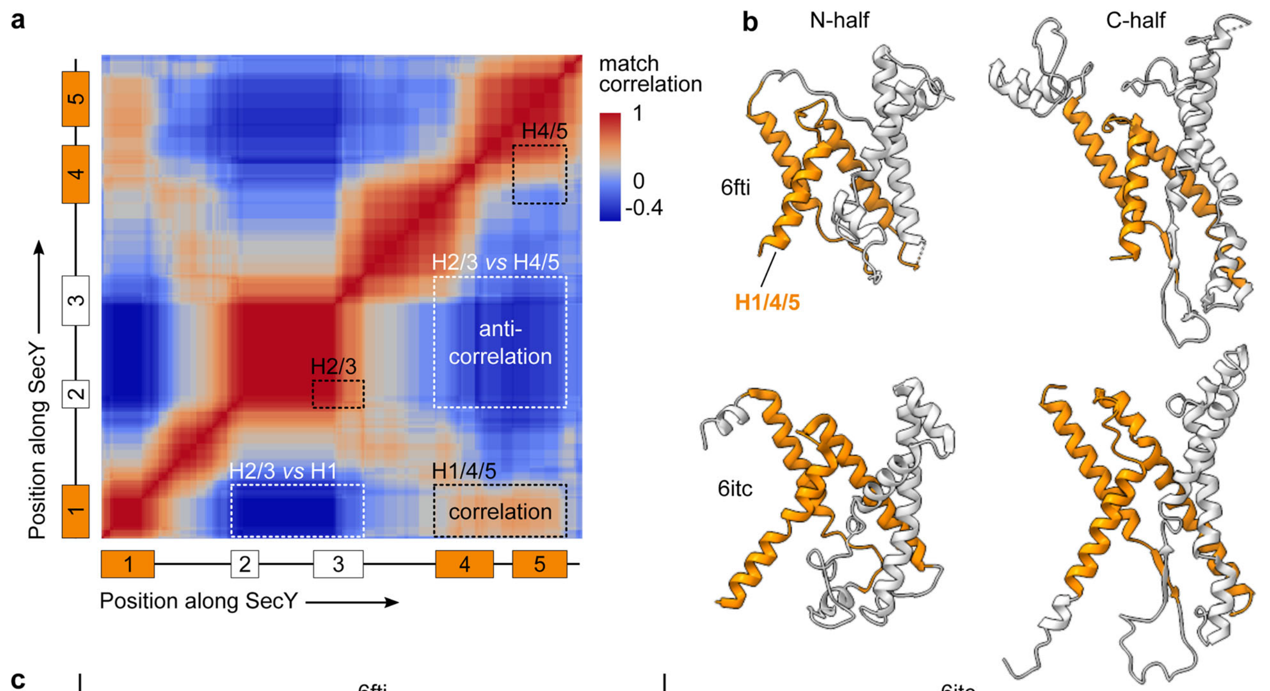
Manual review of these isolated hits shows them to be obviously dissimilar (Additional file 2: Figure S2a) due to features ignored by Dali's distance matrix metric, such as gaps, context, and handedness. There is one non-Oxa1 hit that tops multiple queries, APH-1 (5a63C; Additional file 2: Figure S2b). However, only two of the four aligned TMHs are conserved by the prokaryotic proteases from which APH-1 descends [58, 59], so the part of this alignment relevant to pre-cenancestral events is negligible. To test the sensitivity of these results to our choice of queries (6fti N and 6itc C), selected above for maximum symmetry, we repeated them with the opposite half of each structure (6fti C and 6itc N), with similar results. These results show that the SecY halves are more structurally similar to the Oxa1 superfamily than any other. This result is evident even if one considers only the queries with full N- and C-halves and thus does not depend on treating H1/4/5 as a subdomain.

In the H1/4/5 queries, the Oxa1 superfamily hits rank even higher than some SecY hits, and have Dali *Z*-scores 4.2 to 5.6 standard deviations above the mean, i.e. $p$ = 0.0081 to 0.0014. These $p$ values mean that Dali predicts one would find an unrelated cenancestral protein this similar if the cenancestor contained $0.0014^{-1} \approx 700$ or more homology candidates (multi-pass helical IMPs non-redundant at 25% identity). For scale, *E. coli* contains ~ 550 such proteins. Fewer such proteins can be confidently assigned to the cenancestor [1, 2], but the uncertainties involved are large. Thus, if one weighed no other comparisons between SecY and Oxa1 besides this Dali $p$ value, it alone may not provide strong evidence for homology. But as detailed in subsequent sections, SecY is uniquely similar to Oxa1 in several additional ways. In quantitative terms, this means that the Dali $p$ value can be combined with $p$ values expressing the rarity of these non-Dali similarities ($\sim10^{-4}$ in the PDB25). Thus, the totality of evidence provides stronger statistical support for homology than the Dali search results alone.

In addition to the structural similarities in H1/4/5, each consensus helix from proto-SecY can be matched to a consensus helix from the Oxa1 superfamily and linked with the same connectivity (Fig. 4, Table 1). The SecY/Oxa1 fold comprises a right-handed three-helix bundle (H1/4/5) interrupted after the first helix by a helical hairpin (H2/3) and prefixed by an N-terminal peripheral helix (H0) which abuts H4 (Fig. 4c). Thus, in addition to sharing a universally conserved core three-TMH bundle, proto-SecY and Oxa1 proteins share a similar composition and connectivity across their full ~200 a.a. lengths.

There is one conspicuous difference between these groups' structures: in SecY, the helical hairpin H2/3 is transmembrane, but in the Oxa1 superfamily, it is cytoplasmic (Fig. 4b). If SecY derived from an Oxa1 superfamily ancestor, this would suggest that an initially cytoplasmic H2/3 evolved to be transmembrane in the proto-SecY stem lineage. Transmembrane hairpins are indeed known to be acquired during membrane protein evolution; convenient examples are provided by the transmembrane hairpins in bacterial YidC h4h5 (Fig. 4) and in some SecE [60].

Starting from a YidC-like H2/3, more membrane-penetrating conformations could have been induced by hydrophobic substitution mutations around the hairpin tip, which lacks conserved hydrophilics (Additional file 2: Figure S3). SecY H2/3 could also derive to some degree from indel mutations, particularly since the segment between H1 and H4 is 10 a.a. longer in SecY N than in YidC, and 60 a.a. longer in



**c**

| | 6fti | | | | | | | | 6itc | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | | | | C | | | | N | | | | C | | | |
| Rank | Full | *Z* | H1/4/5 | *Z* | Full | *Z* | H1/4/5 | *Z* | Full | *Z* | H1/4/5 | *Z* | Full | *Z* | H1/4/5 | *Z* |
| 1 | 5a63C | 4.7 | **EMC3** | **4.2** | 6xjhA | 4.8 | **EMC3** | **5.6** | 5a63C | 4.4 | 6o84A | 4.8 | 5a63C | 5.8 | 6qp6A | 5.2 |
| 2 | 7d7nA* | 4.5 | | | 6u0oA | 4.8 | **GET1** | **5.1** | 6m49A | 4.0 | **GET1** | **4.6** | 4p6vD | 4.6 | **GET1** | **5.0** |
| 3 | 2witA | 4.4 | | | **EMC3** | **4.2** | | | 6z5sW | 3.7 | **EMC3** | **4.6** | 6l7oG | 4.4 | | |
| 4 | **EMC3** | **4.2** | | | | | | | **EMC3** | **3.7** | | | 4dveA | 4.2 | | |
| 5 | | | | | | | | | | | | | 7d7nA* | 4.1 | | |
| 6 | | | | | | | | | | | | | 5oqkA | 4.1 | | |
| 7 | | | | | | | | | | | | | **GET1** | **4.0** | | |
| SecY | 2ww9A | 15.0 | 2ww9A | 7.7 | 2ww9A | 10.8 | **3bo0A** | **4.0** | 5abbA | 14.3 | 2ww9A | 7.9 | 5abbA | 8.8 | 2ww9A | 5.6 |
| | 3bo0A | 13.8 | 3bo0A | 5.1 | 3bo0A | 10.4 | **2ww9A** | **3.1** | 2ww9A | 13.7 | 3bo0A | 6.1 | 2ww9A | 8.8 | **3bo0A** | **4.0** |
| | 5abbA | 11.8 | **5abbA** | **3.8** | 5abbA | 6.4 | **5abbA** | **2.2** | 3bo0A | 13.0 | **5abbA** | **4.6** | 3bo0A | 8.3 | **5abbA** | **3.1** |

**Fig. 3** SecY's halves are uniquely similar in structure to the Oxa1 superfamily. **a** Match correlation matrix returned by Dali for a half-SecY query (6itc C). The axes are labelled by a diagram of the SecY transmembrane helices. **b** The structural models used as Dali queries. The full models and the H1/4/5 subdomains (orange) were used. **c** Results from querying the PDB25. The top-ranking hits for each query are shown, and any lower-ranking hits that rank higher than the first Oxa1 superfamily hit. Asterisks mark 7d7nA because although it appears twice, those hits are with two non-overlapping parts of the model. Oxa1 superfamily hits are shown by name (EMC3, GET1) instead of PDB code (6ww7C, 6so5C). At bottom are the scores for the SecY hits, which were excluded from the ranking. SecY hits in boldface scored lower than an Oxa1 superfamily hit for that query

SecY C. Mutant H2/3 would readily sample membrane-penetrating conformations because it rests at the lipid-water interface [34] and is flexibly connected to H1/4/5, as evident in simulations [19, 34] and in the archaeal and bacterial crystal structures where H2/3 was too mobile to be modelled [18, 61]. Because SecY H2/3 is stabilised in the membrane by H1/4/5, it need not have become particularly hydrophobic; for example, most of the H2/3 helices in *G. thermodenitrificans* SecY are predicted to prefer the aqueous phase (N to C: $\Delta G_{app}$ = 1.7, 0.7, 1.0, −1.4 kcal/mol [62];).

Late acquisition of the transmembrane H2/3 would explain a curious feature of SecY's structure. H2/3 does not pack against H1 (Fig. 4c), despite the fact that during co-translational membrane insertion H1 would be exposed to H2/3 without competition. It is reasonable to expect that these elements would interact if their folding pathway had juxtaposed them throughout evolution. This is thought to be why most transmembrane helices

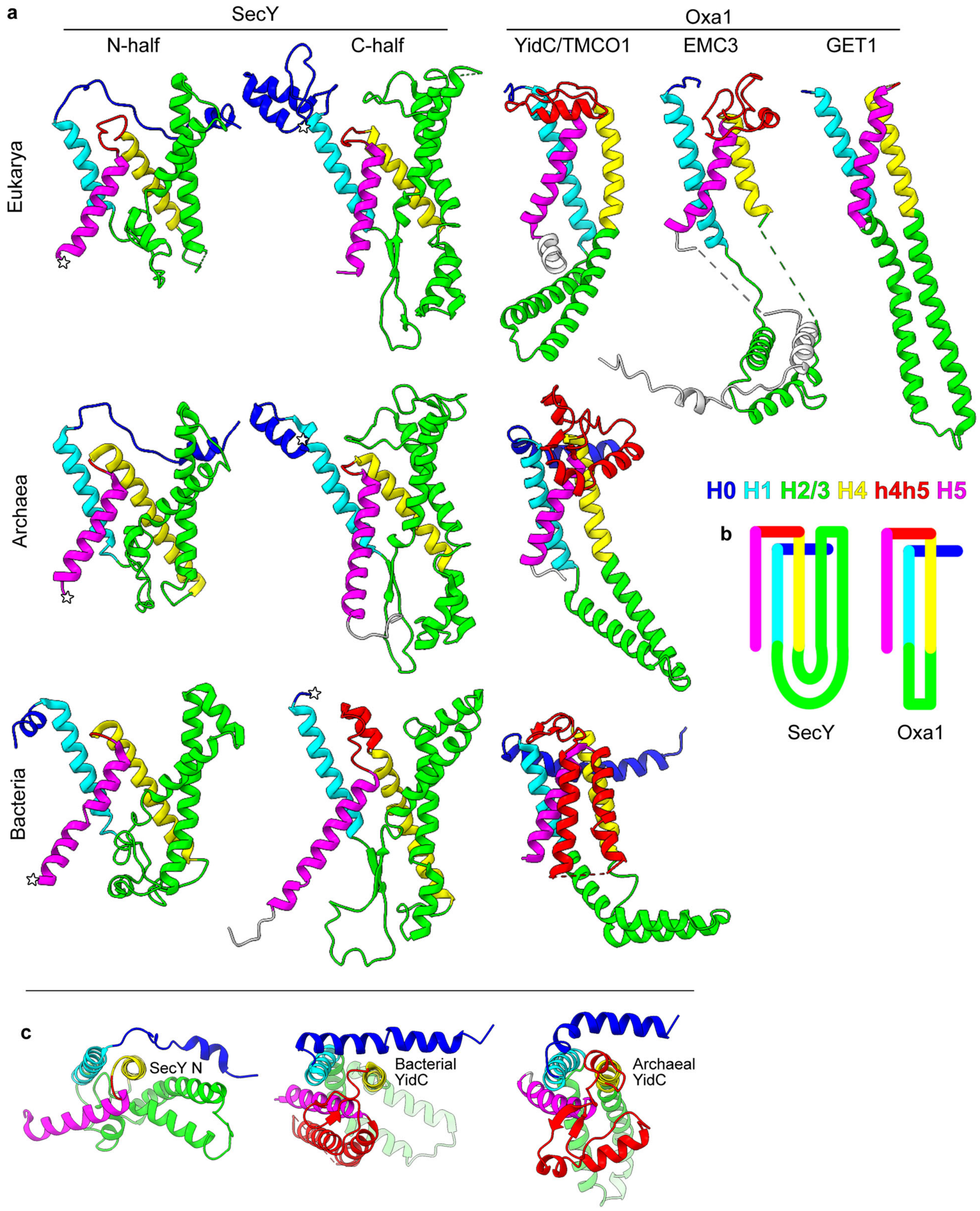


**Fig. 4** (See legend on next page.)

(See figure on previous page.)

**Fig. 4** Correspondence between structural elements of SecY and the Oxa1 superfamily. Consensus elements and the intervening element h4h5 are coloured according to the key shown. Other intervening elements are coloured to match a neighbouring consensus element, and flanking elements are coloured white. **a** The models are, from left to right and then top to bottom, *Canis lupus familiaris* SEC61A1 (6fti), *Homo sapiens* TMCO1 (6w6l), *H. sapiens* EMC3 (6ww7), *H. sapiens* GET1 (6so5), *M. jannaschii* SecY (1rhz), *M. jannaschii* MJ0480 (predicted, see Additional file 2: Figure S4), *G. thermodenitrificans* SecY (6itc), and *Bacillus halodurans* YidC2 (3wo6). **b** Topology diagrams. **c** Axial views of archaeal SecY N and prokaryotic YidC (models as in **a**)

pack sequentially against one another [63, 64]. In SecY, however, H1 and H2/3 are separated by H4/5. This suggests that H1/4/5 was the original, sequentially packed transmembrane bundle, and H2/3 only later became transmembrane and packed against its surface. The transmembrane hairpins in bacterial YidC h4h5 and SecE likewise break sequential packing, suggesting that a similar process of transmembrane hairpin acquisition may explain non-sequential TMH packing in other proteins. This is analogous to how RNA branch acquisition left structural fingerprints in the ribosome [65].

Thus, the SecY halves and the Oxa1 superfamily have backbone structures that not only are uniquely similar by standard measures, but also could plausibly descend from a common ancestor. This identifies the Oxa1 superfamily as the best candidate for the origin of SecY. The following sections analyse their similarities and differences in mechanistic and functional terms. We focus on archaeal and bacterial YidC and not their eukaryotic homologs, since eukaryotes derive from archaea [66].

## Like proto-SecY, YidC uses the distal face of H5 for dimerisation

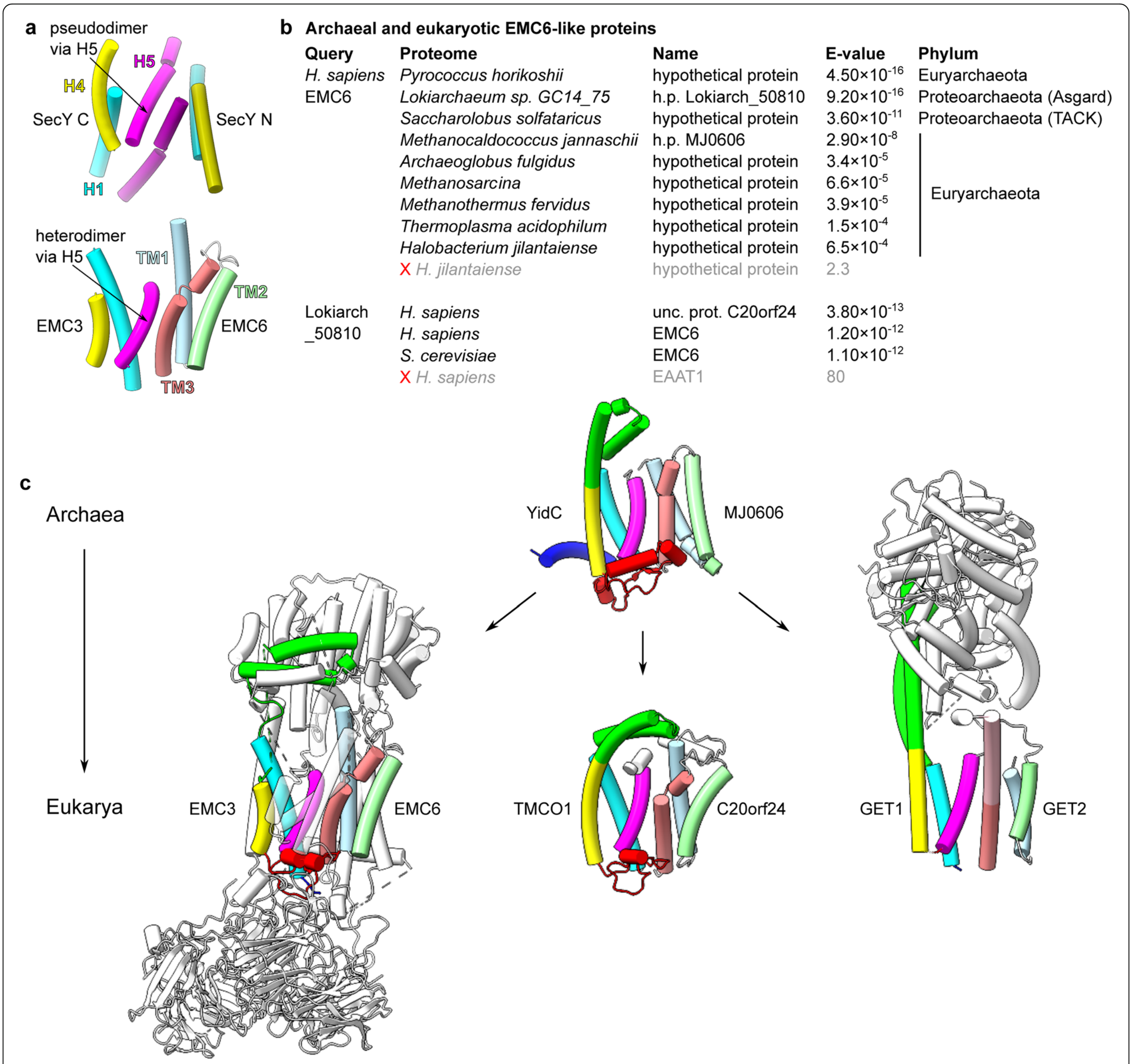
As shown above, proto-SecY formed antiparallel homodimers via the distal face of H5 (Fig. 5a); here, we consider whether this characteristic could have arisen in an ancient member of the Oxa1 superfamily. Antiparallel homodimerisation requires that the monomer possess two characteristics: a tendency to be produced in opposite topologies and an interface suitable for dimerisation. Although dual topology is not evident in the Oxa1 superfamily, distant ancestors could easily have had this property with relatively few changes. Making only a few changes to basic amino acids (especially lysine and arginine), flanking the first TMH of an IMP can influence its topology, and an inverted first TMH can invert an entire IMP containing several TMHs [40, 67–69]. Such changes in topology occur naturally in protein evolution [40, 70], and YidC does not contain any conserved basic residues in its soluble segments that would impede this evolutionary process (Additional file 2: Figure S3). Moreover, the lysine and arginine bias in extant YidC is no greater than that previously observed in proteins which acquired divergent orientations [70].

The other required characteristic, amenability to dimerisation via the distal face of H5, does indeed occur in some Oxa1 superfamily members. This interface is occupied by an intramolecular interaction with h4h5 in bacterial YidC, but it remains exposed in archaeal YidC and its eukaryotic descendants. There are no published data on YidC biochemistry in archaeal cells, but eukaryotic EMC3 and GET1 are known to form separate complexes, and structural models show that they use the distal face of H5 to do so (Fig. 5a [21–25]). These interactions via H5 are heterodimeric, rather than homodimeric, but nonetheless demonstrate that EMC3 and GET1 can dimerise (with EMC6 and GET2, respectively)

**Table 1** Consensus nomenclature for SecY and YidC

| SecY | | Consensus element | YidC | | |
|---|---|---|---|---|---|
| N-half | C-half | | Archaea-Eukarya | Bacteria (monoderms) | Bacteria (diderms) |
| | | h0 | | | TM1, P1 |
| | | H0 | EH1 | EH1 | EH1 |
| TM1 | TM6 | H1 | TM1 | TM1 | TM2 |
| TM2a (plug) | C4 (RBD) | h1h2 | | | |
| TM2b (LG) | TM7 (LG) | H2 | CH1 | CH1 | CH1 |
| | C5 (RBD) | h2h3 | | | |
| TM3 (LG) | TM8 (LG) | H3 | CH2 | CH2 | CH2 |
| TM4 | TM9 | H4 | TM2 | TM2 | TM3 |
| | | h4h5 | EH2 | TM3/4 | TM4/5 |
| TM5 | TM10 | H5 | TM3 | TM5 | TM6 |

*CH* cytoplasmic helix, *EH* extracytoplasmic helix, *P* periplasmic domain, *C* cytoplasmic domain, *LG* lateral gate, *RBD* ribosome-binding domain



**b Archaeal and eukaryotic EMC6-like proteins**

| Query | Proteome | Name | E-value | Phylum |
|---|---|---|---|---|
| *H. sapiens* EMC6 | *Pyrococcus horikoshii* | hypothetical protein | $4.50\times10^{-16}$ | Euryarchaeota |
| | *Lokiarchaeum sp. GC14_75* | h.p. Lokiarch_50810 | $9.20\times10^{-16}$ | Proteoarchaeota (Asgard) |
| | *Saccharolobus solfataricus* | hypothetical protein | $3.60\times10^{-11}$ | Proteoarchaeota (TACK) |
| | *Methanocaldococcus jannaschii* | h.p. MJ0606 | $2.90\times10^{-8}$ | Euryarchaeota |
| | *Archaeoglobus fulgidus* | hypothetical protein | $3.4\times10^{-5}$ | |
| | *Methanosarcina* | hypothetical protein | $6.6\times10^{-5}$ | |
| | *Methanothermus fervidus* | hypothetical protein | $3.9\times10^{-5}$ | |
| | *Thermoplasma acidophilum* | hypothetical protein | $1.5\times10^{-4}$ | |
| | *Halobacterium jilantaiense* | hypothetical protein | $6.5\times10^{-4}$ | |
| | X *H. jilantaiense* | hypothetical protein | 2.3 | |
| Lokiarch _50810 | *H. sapiens* | unc. prot. C20orf24 | $3.80\times10^{-13}$ | |
| | *H. sapiens* | EMC6 | $1.20\times10^{-12}$ | |
| | *S. cerevisiae* | EMC6 | $1.10\times10^{-12}$ | |
| | X *H. sapiens* | EAAT1 | 80 | |

**Fig. 5** Archaeal YidC and its descendants form dimers via the same interface as proto-SecY. **a** Comparison of the SecY (6itc) and EMC3/6 (6wb9) dimerisation interfaces. **b** Archaeal and eukaryotic HHpred hits for EMC6-like proteins. A red cross and grey text indicates the first rejected result. For sequence accession numbers, see Methods. **c** Structures of archaeal and eukaryotic complexes containing homologs of EMC3/6 (6wb9): *M. jannaschii* YidC a.k.a. MJ0480/MJ0606 (predicted, see Additional file 2: Figure S4), *H. sapiens* TMCO1/C20orf24 (predicted, see Additional file 2: Figure S4), and *H. sapiens* GET1/2/3 (6so5). A sequence insertion in the N-terminal half of GET2 TM3 is shown in pink

along the same interface as the proto-SecY homodimer without impeding their translocation activities.

To determine whether this propensity to dimerise via H5 is ancient or eukaryote-specific, we queried nine diverse archaeal proteomes for homologs of *H. sapiens* EMC6 or GET2 using HHpred [41]. Although none displayed significant similarity with GET2, every proteome queried contained exactly one protein similar to EMC6 (Fig. 5b). Among these archaeal proteins, those most similar to eukaryotic EMC6 tend to come from the species most closely related to eukaryotes: the Asgard archaean, then the TACK archaean, and then the euryarchaeans. This phylogenetic concordance indicates that the archaeal 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 Asgard EMC6-like protein (Lokiarch_50810) identified EMC6 in both cases as high-confidence hits. Unexpectedly, the *H. sapiens* search also identified an additional, even more similar hit, C20orf24 (Fig. 5b).

To determine if these EMC6 homologs bind to an Oxa1 superfamily member, we performed coevolutionary contact-restrained structure prediction using AlphaFold 2.0 [71, 72] for putative archaeal (*M. jannaschii* YidC/MJ0606) and eukaryotic (*H. sapiens* TMCO1/C20orf24) complexes. This yielded heterodimeric models with very high confidence scores for both the local structure of each protomer (Additional file 2: Figure S4a; Additional file 5) and those protomers' alignment in the heterodimer (Additional file 2: Figure S4b). This indicates that the distal face of H5 is used for heterodimerisation not only by eukaryotic EMC3 and GET1 but also by TMCO1 and archaeal YidC. TMCO1/C20orf24 interaction is consistent with the aforementioned absence of C20orf24 from *S. cerevisiae* (Fig. 5b) because *cerevisiae* also lacks TMCO1.

Although GET2 lacks strong sequence similarity with these EMC6 homologs, its structural similarity with EMC6 was immediately recognised [24, 25]. Our identification of archaeal EMC6 homologs reveals a plausible origin for GET2. Consistent with this, although our GET2 query of the lokiarchaean proteome did not identify any very high-similarity proteins, the most similar membrane protein was indeed an EMC6 homolog (Lokiarch_50810, HHpred $p$ = 0.0057). Moreover, the aligned columns between GET2 and Lokiarch_50810 correspond exactly to their structurally similar transmembrane domains. The single large gap in this alignment spans the cytoplasmic extension of GET2 TM3, which brings it into contact with GET3 (Fig. 5c). 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.

The absence of a similar heterodimer in bacteria suggests that it may have been acquired in archaea after divergence from bacteria, which instead acquired the H5-occluding transmembrane hairpin in h4h5 (Fig. 4). An archaeal origin for the EMC6-like proteins would be consistent with their genomic location, which is distant from the widely conserved cluster of cenancestral ribosomal genes, SecY and YidC [73]. In the period prior to heterodimerisation with EMC6-like proteins, a YidC homolog could have evolved to use H5 for homodimerisation, giving rise to proto-SecY. YidC's universal tendency to cover the distal face of H5 supports this possibility.

## Corresponding elements serve similar mechanistic roles using similar amino acids

In both YidC and SecY, the hydrophilic translocation interface is lined by H1/4/5 (Fig. 6a). Both three-TMH bundles have a right-handed twist, with H1 and H4 near parallel and H5 packing crossways against them. Of the three helices, it is this crossways H5 that makes the closest contacts with the translocating hydrophilic substrate in SecY (Fig. 6a) and in YidC [74]. Moreover, YidC's substrates initiate translocation as a hairpin with both termini in the cytoplasm [74], just as SecY's substrates do [75, 76]. From this intermediate state, some segments of the substrate can integrate into the membrane, and their propensity to do so is a similar function of the segment's sequence regardless of whether YidC or SecY is used [77].

The YidC and SecY H1/4/5 bundles are structurally similar enough that they can be aligned confidently (Fig. 6b). Across the 40 structurally aligned sites, YidC and SecY have 22.5% identical consensus sequences, compared to 30.0% between the SecY halves at these same sites. This alignment superimposes the pore ring residue in SecY H5 onto a conserved hydrophobic residue in YidC H5 that marks the end of the hydrophilic groove. In YidC and the SecY N-half this residue is positioned at a similar depth in the membrane (Fig. 6a), whereas the C-half is shifted. In bacterial YidC, the groove end residue is aromatic and intimately contacts the bacteria-specific h4h5 hairpin, but in archaeal YidC, this residue is aliphatic and most often an isoleucine, just as it is in SecY (Fig. 6c). Moreover, the same surrounding positions on H5 are polar (−3, +3, +7, +11) or polarisable aromatics (−1) in both YidC and SecY. Together with a conserved polar residue in H1, these comprise the entire hydrophilic groove of archaeal YidC, and thus that same groove is also hydrophilic in SecY. Finally, a conserved tryptophan is positioned at the lipid-water interface, tryptophan's preferred environment [78], where it is thought to stabilise YidC's particular transmembrane position [34].

This detailed similarity in both sequence and structure indicates that the residue at the end of YidC's hydrophilic groove is homologous to the pore ring residue at the end of SecY's hydrophilic funnel. Hydrophobic interactions between these residues in two antiparallel YidC-like monomers would have favoured dimers with a symmetry that juxtaposed them, allowing them to ultimately form the proto-SecY pore ring. Early dimers may have formed only transiently or been unable to open a full a membrane-spanning pore, but such channels can nonetheless be functional. For example, the channel for ER-associated degradation is a transient heterodimer of two protomers that contain hydrophilic grooves, one open to the cytosol and the other to the ER lumen [79]. Juxtaposing these two grooves thins the membrane enough that soluble proteins can be translocated across. Juxtaposing the grooves of two YidC homologs in an antiparallel homodimer could likewise have increased their

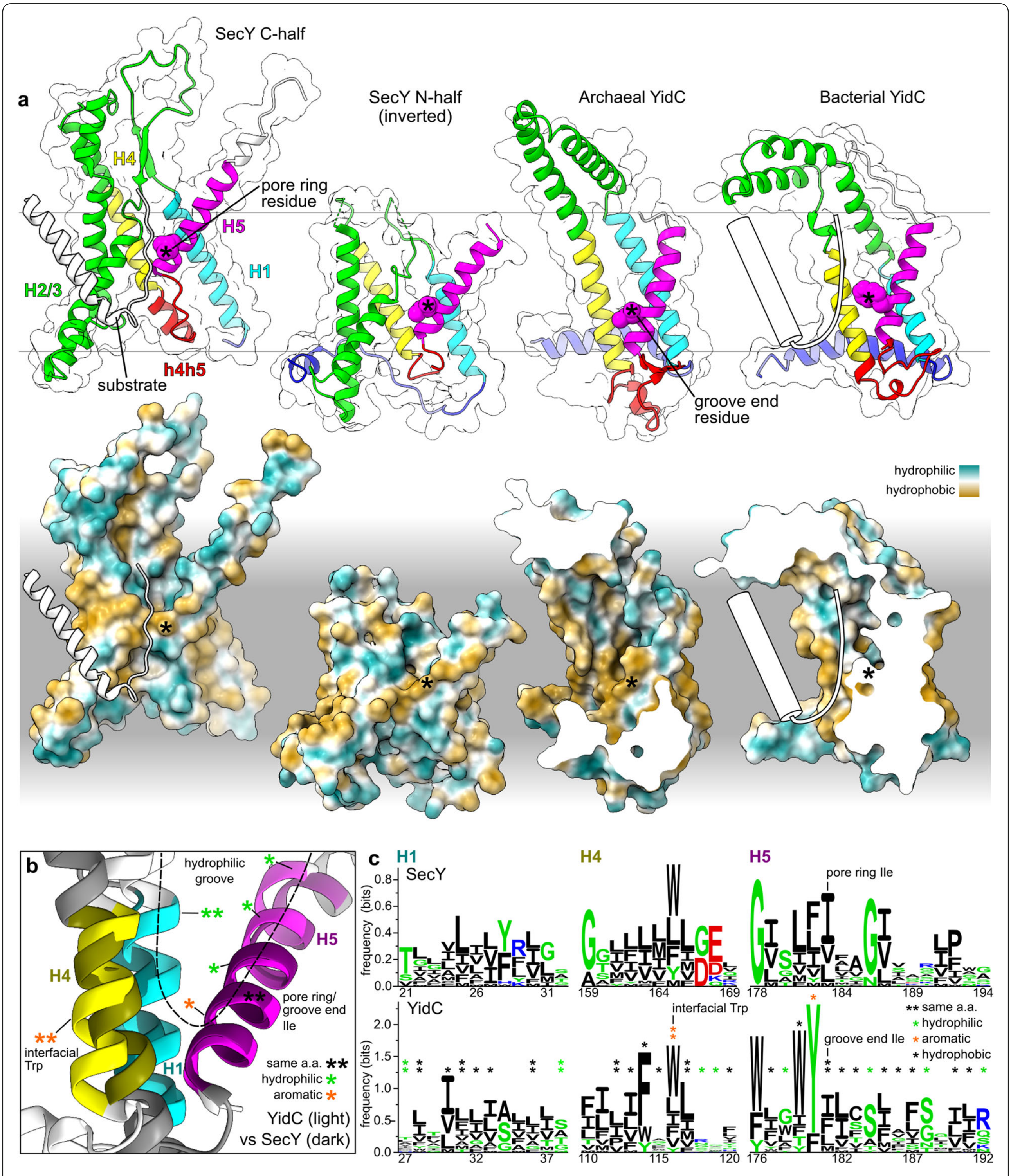


**Fig. 6** Corresponding elements of SecY and YidC serve similar mechanistic roles using similar amino acids. **a** SecY and YidC models aligned by fitting to a model membrane: *G. thermodenitrificans* SecY and substrate (6itc), *M. jannaschii* MJ0480 (predicted, see Additional file 2: Figure S4), and *B. halodurans* YidC2 (3wo7A). A cartoon substrate is superimposed on bacterial YidC to indicate the experimentally determined translocation interface and conformation [74]. YidC is shown clipped to allow a lateral view of the hydrophilic groove which would otherwise be occluded by h4h5, and likewise the plug loop in SecY N was removed. **b** Superposition of SecY and archaeal YidC (*G. thermodenitrificans* SecY N-half, 6itc, vs *M. jannaschii* MJ0480, 5c8j). Coloured segments correspond to the sequence logos in panel c. **c** Sequence logos for the structurally aligned regions of SecY and archaeal YidC. Column numbers correspond to the proteins in **a**

translocation activity and with subsequent adaptation yielded the membrane-spanning pore and pore ring of proto-SecY.

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

Whereas the conserved cores of SecY and YidC are similar, their differences are concentrated in regions which are hypervariable among the Oxa1 superfamily: h4h5 and H2/3 (Fig. 4). H2/3 forms a relatively compact cytoplasmic hairpin in YidC and TMCO1, is markedly elongated and rigid in GET1, and is tethered via long flexible loops in EMC3. By contrast, the H2/3 hairpin in SecY is folded back toward the H1/4/5 bundle and embedded in the membrane.

Despite their differences, H2/3 is a site for substrate signal recognition in both SecY (Fig. 7a) and the Oxa1 superfamily. In YidC, TMCO1, and EMC3, the membrane-facing side of H2/3 is thought to interact with substrate TMHs before they reach the hydrophilic groove [18–20, 24]. In contrast to direct TMH interaction, the rigid and elongated H2/3 coiled coil of GET1 [25] forms a binding site for the substrate targeting factor GET3 [80–82]. This adaptation may be due to the particularly hydrophobic TMHs inserted by this pathway [83], 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 (Fig. 7b). This allows SecY to create a 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 longer hydrophilic segments than those translocated by YidC.

As a secondary consequence, transmembrane insertion of H2/3 makes the site where signals initiate translocation more proteinaceous and hydrophilic (Fig. 6a [9, 84–87]). Because of this, translocation via SecY can be initiated via signals which are much less hydrophobic than the TMHs which initiate translocation via YidC [77]. This, too, is important for SecY's secretory function, because the signal peptides of secretory proteins are distinguished from TMHs by their relative hydrophilicity [4]. This biophysical difference allows signal peptidase to specifically recognise and cleave them [15]. 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 absent from SecY (Fig. 7b). Whereas the H2/3 transmembrane insertion differentiates how SecY and YidC receive and recognise hydrophobic domains, the absence of h4h5 clears the channel through which hydrophilic substrates translocate. As mentioned previously, h4h5 is, like H2/3, hypervariable in the Oxa1 superfamily, forming a peripheral helix in archaea and eukaryotes and a transmembrane hairpin in bacteria. If a more YidC-like h4h5 were present in proto-SecY, proto-SecY dimerisation would

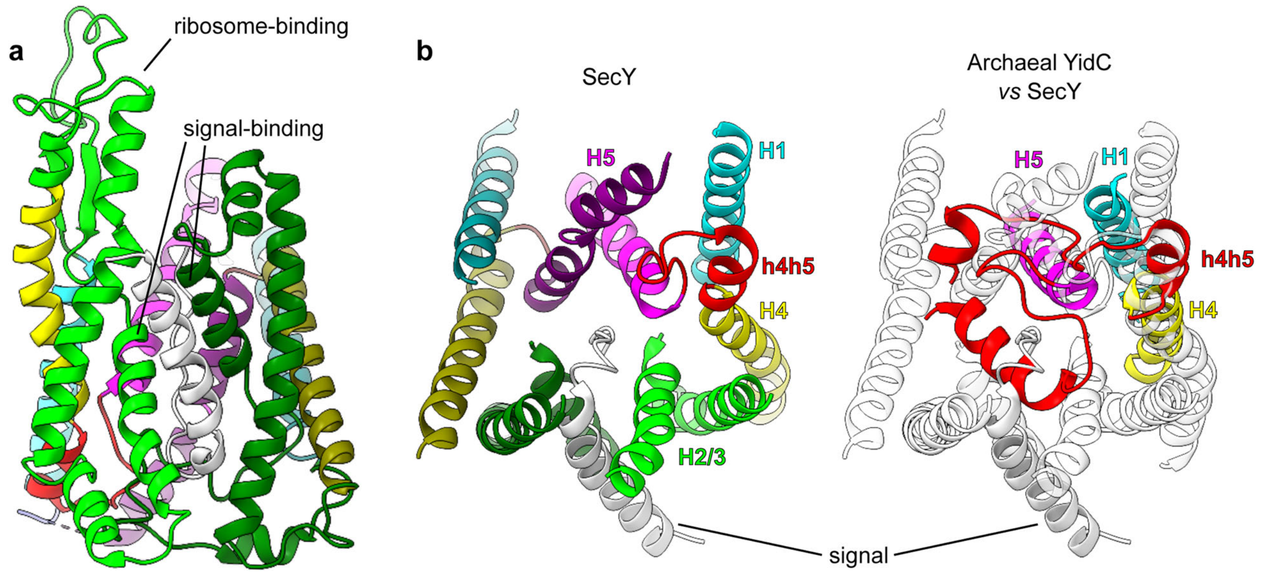


**Fig. 7** Structural features unique to SecY which enable signal binding and substrate translocation. SecY is *G. thermodenitrificans* SecY/proOmpA (6itc). **a** Signal-binding and ribosome-binding sites on SecY H2/3, viewed laterally. **b** The substrate translocation channel, viewed from its extracytoplasmic side. Only H1-5 and h4h5 of SecY are shown. SecY is colour-coded by consensus element as in Fig. 4 (left), or rendered transparent and superimposed by the corresponding elements of archaeal YidC *(M. jannaschii* MJ0480, 5c8j), aligned to the SecY C-half (right)

place h4h5 inside the hydrophilic groove of the opposite monomer, instead of in contact with the membrane. Thus, a YidC-like h4h5 would be selected against in SecY, to maintain a membrane-spanning hydrophilic pore and facilitate translocation.

### Reductive evolution in symbionts demonstrates the functional range of YidC

If proto-SecY originated in the YidC family, YidC might initially have been the cell's only transporter for the extracytoplasmic parts of IMPs. But some IMPs cannot be integrated by YidC and instead depend on SecY [88]. Thus, a cell with YidC and not SecY may have been constrained to express a more limited range of IMPs. The looser this constraint, the more plausible it is that such a cell would be viable, and that YidC could have preceded SecY.

Insight into this question of in vivo sufficiency can be obtained by inspection of the only cells known to have survived SecY deletion: the mitochondrial symbionts. SecY has been lost from all but one group of eukaryotes for which mitochondrial genome sequences are available, and it has not been observed to relocate to the nuclear genome [89]. The exceptional group is the jakobids, only a subset of which retain mitochondrial SecY. The incomplete presence of SecY in this group implies that SecY 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.

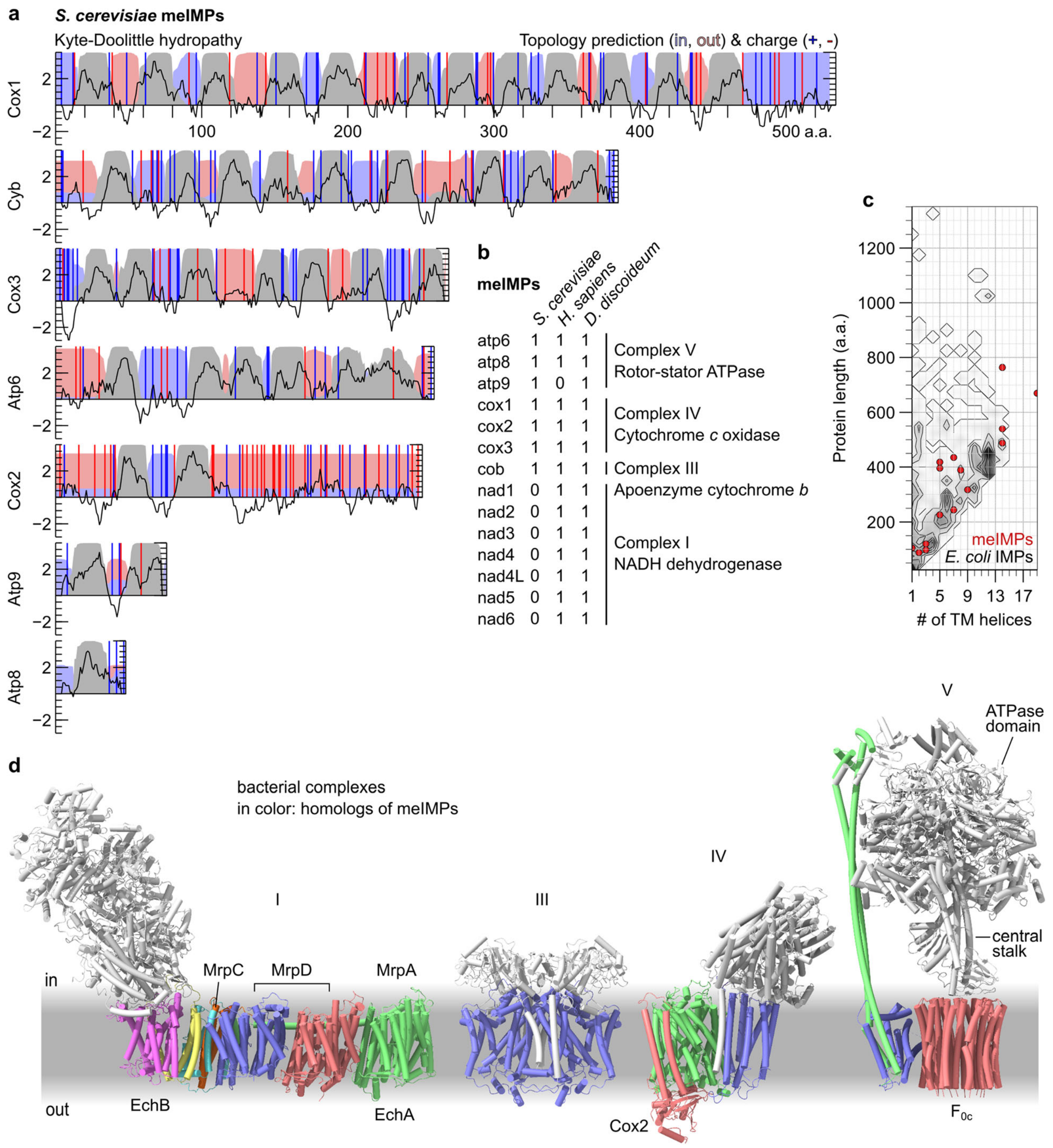
Mitochondria retain two YidC family proteins, Oxa1 and Oxa2 (Cox18), the genes for which relocated from the mitochondrial genome to the nuclear genome [27, 28]. As nuclear-encoded mitochondrial proteins, they are translated by cytoplasmic ribosomes and then imported into mitochondria via channels in the inner and outer mitochondrial membranes [90]. These channels are essential for the import of nuclear-encoded proteins, but are not known to function in the integration of mitochondrially encoded IMPs (meIMPs), which instead requires export from the matrix, where they are synthesised by mitochondrial ribosomes. This export is generally Oxa1-dependent [31].

The meIMPs have diverse properties, including 1 to 19 TMHs and exported parts of various sizes and charges (Fig. 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 [88, 91]. Ectopically expressed EMC3/6 can rescue meIMP integration in the absence of Oxa1, indicating that Oxa1's broad substrate spectrum is representative of the Oxa1 superfamily as a whole [92]. The only apparent constraint on the meIMPs is that they tend to have only 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 [33, 93, 94]. 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 [95].

This constraint on soluble segment length is less consequential than it may at first appear, because prokaryotic IMPs in general tend to have only short soluble segments (Fig. 8c [100]). 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 target nascent IMPs to either SecY or YidC [88], and YidC is present at a concentration 1–2× that of SecY [101]. By contrast, IMPs with large translocated domains became much more common in eukaryotes [100] concomitant with YidC's divergence into three niche paralogs, none of which are essential at the single-cell level [20, 102, 103].

Even without extrapolating from the meIMPs to other similar IMPs, it is clear that chemiosmotic complexes are amenable to YidC-dependent, SecY-independent biogenesis (Fig. 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 the oxygenation of Earth's atmosphere, they have homologs which enable chemiosmosis in anaerobes. In particular, chemiosmosis in methanogens and acetogens employs the rotor-stator ATPase, Mrp antiporters, and an energy-converting hydrogenase (Ech [104]), all of which have homologs of their IMP subunits among the meIMPs (Fig. 8d) and may have participated in primordial anaerobic metabolism [105].

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 evolution in chloroplasts, which retain both SecY (cpSecY) and YidC (Alb3) [106]. cpSecY imports soluble proteins across the chloroplast's third, innermost membrane, the thylakoid membrane [107]. 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 [108]. 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



| meIMPs | *S. cerevisiae* | *H. sapiens* | *D. discoideum* | |
|---|---|---|---|---|
| atp6 | 1 | 1 | 1 | Complex V<br>Rotor-stator ATPase |
| atp8 | 1 | 1 | 1 | |
| atp9 | 1 | 0 | 1 | |
| cox1 | 1 | 1 | 1 | Complex IV<br>Cytochrome *c* oxidase |
| cox2 | 1 | 1 | 1 | |
| cox3 | 1 | 1 | 1 | |
| cob | 1 | 1 | 1 | Complex III<br>Apoenzyme cytochrome *b* |
| nad1 | 0 | 1 | 1 | Complex I<br>NADH dehydrogenase |
| nad2 | 0 | 1 | 1 | |
| nad3 | 0 | 1 | 1 | |
| nad4 | 0 | 1 | 1 | |
| nad4L | 0 | 1 | 1 | |
| nad5 | 0 | 1 | 1 | |
| nad6 | 0 | 1 | 1 | |

**Fig. 8** (See legend on next page.)

(See figure on previous page.)
**Fig. 8** Substrates of the mitochondrial SecY-independent pathway for IMP integration. **a** Sequence characteristics of the mitochondrially-encoded IMPs (meIMPs) from *S. cerevisiae*. Kyte-Doolittle hydropathy (left axis) is averaged over a 9 a.a. moving window (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). Positive (blue) and negative (red) 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 TMHs in the meIMPs of a eukaryote (*D. discoideum*), superimposed on a contour plot and heat-map of all 910 IMPs from a proteobacterium (*E. coli*). Protein lengths were binned in 25 a.a. increments. Each contour represents an increase of 3 proteins per bin. **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; [96]), III, cytochrome *bc*1, (*Rhodobacter sphaeroides*, 6nhh; [97]), IV, cytochrome *c* oxidase (*R. sphaeroides,* 1m57; [98]), V, rotor-stator ATPase (*Bacillus* sp. PS3, 6n2y; [99]). 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

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 [109]. Translocases are transporters which use chemical reactions to drive translocation [110], such as the translocase formed when the SecA ATPase acts in tandem with the SecYEG channel [37]. The putative rotor-stator-like protein translocase used its ATPase subunit to unfold and feed substrates through the homo-oligomeric channel formed by $F_0$c, now occupied by the central stalk (Fig. 8d). The strict YidC-dependence of $F_0$c biogenesis in *E. coli* [111] hints that YidC and $F_0$c shared an early era of co-evolution, as a laterally closed channel for the secretion of soluble proteins ($F_0$c) and a laterally open channel for the integration of membrane proteins (YidC), including $F_0$c itself. The subsequent advent of a laterally gated channel, SecY, would have facilitated the biogenesis of a hybrid class of proteins: IMPs with large translocated domains.

## Discussion

By comparing structures of the SecY N- and C-halves, we identified a maximum-symmetry pair, and thus an estimate of the structure of their last common ancestor, proto-SecY. Their alignment identifies homologous sites in each half, revealing that both the hydrophobic pore ring and the interface between halves are symmetric. The conservation of these features indicates that they were also present in proto-SecY and thus that it formed antiparallel homodimers and functioned as a protein-conducting channel.

In automated database searches for structures similar to SecY's halves, the top hit is the Oxa1 superfamily, of which YidC is the prokaryotic member. The SecY/Oxa1 fold consists of a right-handed three-helix bundle (H1/4/5), interrupted after the first helix by a helical hairpin (H2/3), and prefixed by an N-terminal peripheral helix (H0) that abuts H4. The H2/3 hairpin is cytoplasmic in the Oxa1 superfamily but transmembrane in SecY, where it forms the lateral gate helices. This suggests that H2/3 was originally cytoplasmic, and then evolved to pack against the surface of H1/4/5 in the proto-SecY stem lineage. This sequence of events would explain the peculiar non-sequential packing arrangement of SecY's transmembrane helices.

This unexpected correspondence motivates a re-evaluation of the literature on SecY and YidC. In both, H1/4/5 buries a hydrophilic groove inside the membrane to facilitate the translocation of hydrophilic polypeptide. Juxtaposing two grooves, one on each side of the membrane, allows SecY to open a membrane-spanning pore, whereas YidC has only a cytoplasmic groove. Structural alignments superimpose the hydrophobic residue in H5 that rings the SecY pore onto the hydrophobic residue that ends the YidC groove, and likewise the surrounding polar and aromatic groove residues. Both SecY and YidC recognise hydrophobic helices in their substrates via binding at the protein-lipid interface, and in doing so induce a hairpin conformation in the substrate's hydrophilic flank which initiates its translocation. The SecY-specific lateral gate helices create a more hydrophilic environment for signal recognition and substrate translocation that is better suited to SecY's specific secretory function.

Whereas proto-SecY formed homodimers via the distal face of H5, two of the three eukaryotic Oxa1 member families are known to use this interface for heterodimerisation. Homology would predict that this is an ancient tendency. We indeed found indications that H5-mediated heterodimers are formed by the third eukaryotic Oxa1 superfamily member, TMCO1, and by archaeal YidC. In bacterial YidC, this interface instead makes intramolecular contacts with bacteria-specific TMHs. To gauge the plausibility of a YidC-dependent, SecY-independent primordial cell, we reviewed the range of substrates translocated by YidC in SecY-lacking mitochondria and found that it spans most of the diversity of the prokaryotic membrane proteome. The surprising conclusion of our study is that a YidC homolog could have both preceded and evolved into proto-SecY, whose gene duplication and fusion then originated the present-day SecY family.

### Evaluation of the homology hypothesis

It is important to consider whether the similarities between SecY and YidC could arise by convergent evolution under shared constraints (making them analogs), rather than divergent evolution from a common ancestor (making them homologs). Deciding between the analogy hypothesis and the homology hypothesis requires an assessment of whether any plausible constraints could explain their similarities [112]. We will weigh their functional, mechanistic, structural, and sequence similarities in turn.

Laterally open helical protein-conducting channels have arisen by functional convergence several times (Fig. 9). Thus, if the similarity between SecY and YidC were solely functional, the analogy hypothesis would be attractive. Analogy would also be plausible if the similarity between SecY and YidC were solely mechanistic, because their mechanism is common among amphiphile transporters. Many use a membrane-exposed hydrophilic groove to translocate the hydrophilic parts of an amphiphile while exposing its hydrophobic parts to the bilayer [113–115]. Moreover, the hairpin conformation which protein transporters induce in their substrates is a predictable result of physical constraints which disfavour head-first translocation [116].

Thus, there is precedent and a clear physical basis for SecY and YidC's functional and mechanistic similarities arising by convergence. But the same is not true of their structural similarities. First, it would be unprecedented for structural similarity to arise by convergence within this functional and mechanistic class, given that all other known amphiphile transporters are grossly dissimilar from one another, including all other laterally open helical protein-conducting channels (Fig. 9). This suggests that the space of mechanically sufficient folds is large, and thus the likelihood of convergence low.

Second, the extensive literature on SecY and YidC discussed throughout this paper suggests no physical reason why their mechanism would favour the SecY/Oxa1 fold. Thus, attributing their structural similarity to mechanistic constraints would require one to assume that such a constraint exists. On the contrary, structural convergence due to mechanistic constraints typically occurs in only those parts of a protein with clear mechanistic roles, such as the catalytic dyads and triads of enzymes. For example, a comprehensive survey of convergence in analogous enzymes identified 267 pairs with similar dyads or triads, but none with similar folds [118]. Fold space is evidently large enough that many folds are likely to be compatible with a given mechanism.

Perhaps the most extensive known case of structural convergence in functionally similar helical IMPs occurred among thiol oxidoreductases. Four analogous families all use four-helix bundles to bind their redox cofactors, despite two being IMPs and two being cytoplasmic [119]. But they are nonetheless easily distinguishable because they connect those four helices in different orders. This indicates that even an exceptionally tight constraint on the architecture of secondary structure elements does not comparably constrain the connectivity of those elements. Indeed, the seven TMHs of another IMP, rhodopsin, can be experimentally permuted while retaining activity [120]. Thirty-six such permutations are possible for proto-SecY H0-5. Although some permutations would be more likely to evolve than others, analogy would be as likely as homology only if all 35 other permutations were forbidden. Thus, even if the specific architectures of proto-SecY and YidC were favoured by some yet unknown mechanistic constraint, their identical connectivity would still weigh in favour of the homology hypothesis.

Without functional or mechanistic constraints, structural convergence can still occur in some cases due to folding constraints imposed by the intrinsic properties of polypeptide and solvent. One would expect such intrinsically preferred structures to occur frequently and in functionally unrelated contexts. For this reason, the phylogeny of ubiquitous and functionally diverse folds is challenging to discern [121]. But it is implausible that folding constraints strongly favour the SecY/Oxa1 fold because it is not found in other proteins, as our database queries show.

Finally, we consider the most detailed similarity between SecY and YidC, which is in their H1/4/5 sequence profiles. If this bundle was in the same transmembrane position and orientation in both proteins, one might imagine that their sequence similarity was a product of mechanistic constraints. However, this similarity occurs despite topological inversion (Fig. 6) and thus lends at least some weight to homology. Just how much weight is unclear. Ideally, one would compare SecY and YidC to analogous proteins with the same structure and function and see how exceptional their sequence similarity is among that set. Such a test is partly feasible for proteins with very common folds, like β-barrels [122], but impossible here, because our database queries find no other proteins with the SecY/Oxa1 fold.

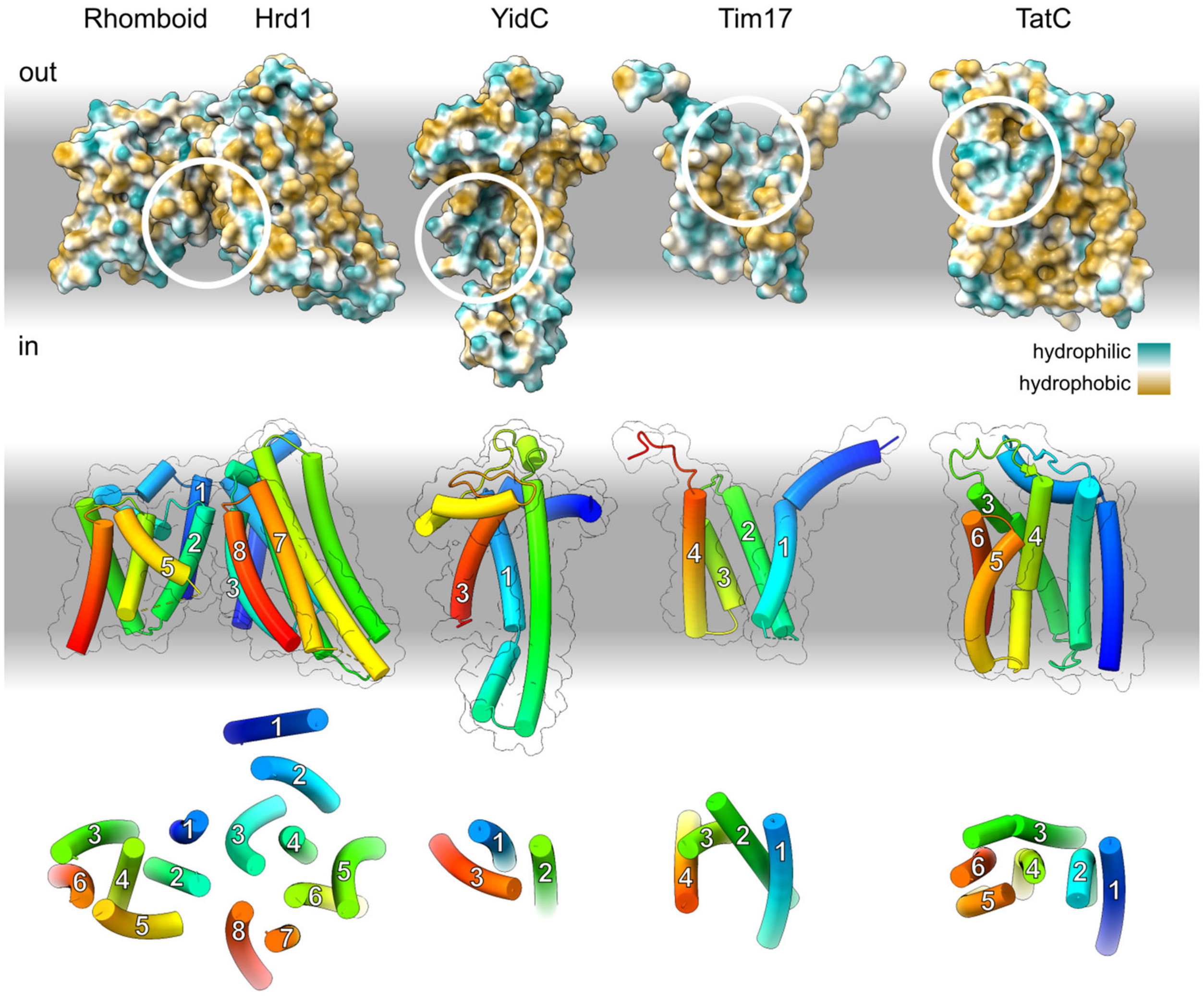


**Fig. 9** Structures of the known families of laterally open helical protein-conducting channels. Top: Structural models shown as solvent-excluded surfaces colour-coded by hydropathy. The hydropathy of the lipidic and aqueous phases represented on a separate scale, ranging from hydrophilic (white) to hydrophobic (grey). White circles indicate intramembrane hydrophilic grooves. Middle: models shown as tubes 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 (6vjz), Hrd1: *S. cerevisiae* Hrd1 (6vjz), YidC: *M. jannaschii* MJ0480 (predicted, see Additional file 2: Figure S4), Tim17: *S. cerevisiae* Tim22 (6lo8, [117]), TatC: *Aquifex aeolicus* TatC (4b4a, [52])

In sum, the dispositive evidence for homology between SecY and the Oxa1 superfamily is structural. It would be empirically unprecedented and theoretically improbable for their structural similarity to arise by convergence. We therefore conclude that they are more likely to be homologs than analogs, and describe them as homologs hereafter.

### Implications for the evolution of protein transport

Besides illuminating SecY's origins, identifying YidC as its progenitor implies that YidC is the oldest known channel. This has implications for the evolution of IMPs generally, including YidC itself, and other ancient components of the general secretory pathway [123–126]: SecEG (Additional file 2: Figure S5), signal peptidase, SRP and SRP receptor (SR). We propose that the following stepwise model (Fig. 10) is the simplest that is consistent with the available data.

**Step 1**. An ancestor of YidC was a membrane-peripheral ribosome receptor. This is parsimonious insofar as both YidC and SecY are ribosome receptors, and like all IMPs presumably descend from peripheral proteins [127]. Ribosome receptor function can be achieved with just two low-complexity domains: a weakly hydrophobic anchor and a polybasic extension. This receptor would reduce 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 for deletion of SRP or SR [128, 129].

**Step 2**. The peripheral helix acquires a transmembrane hairpin, thereby integrating into the membrane. Uncatalyzed insertion of a hairpin is more efficient than that of a single TMH [116], making a hairpin the more likely initial membrane anchor. We infer that SRP/SR-dependent targeting did not evolve until after this and other minimal IMPs existed for it to target. The proximity of this hairpin to nascent IMPs emerging from the bound ribosome imposes a selective pressure on the hairpin to evolve membrane-buried hydrophilic residues that can facilitate IMP integration. Substrates would engage this YidC ancestor in the same hairpin conformation that is favoured during uncatalysed translocation, and this conformation remains how substrates engage SecY and YidC today.

**Step 3**. Acquisition of a second transmembrane hairpin produces a four-TMH protein containing the conserved three-helix bundle and hydrophilic groove. The segment between the first and second transmembrane hairpins becomes the cytoplasmic hairpin H2/3. The additional TMHs allow YidC to form a hydrophilic groove in the membrane, thereby further facilitating substrate translocation.

**Step 4**. The hydrophilic groove allows hydrophilic termini to efficiently translocate, including the N-terminus of the YidC ancestor itself, which acquires a

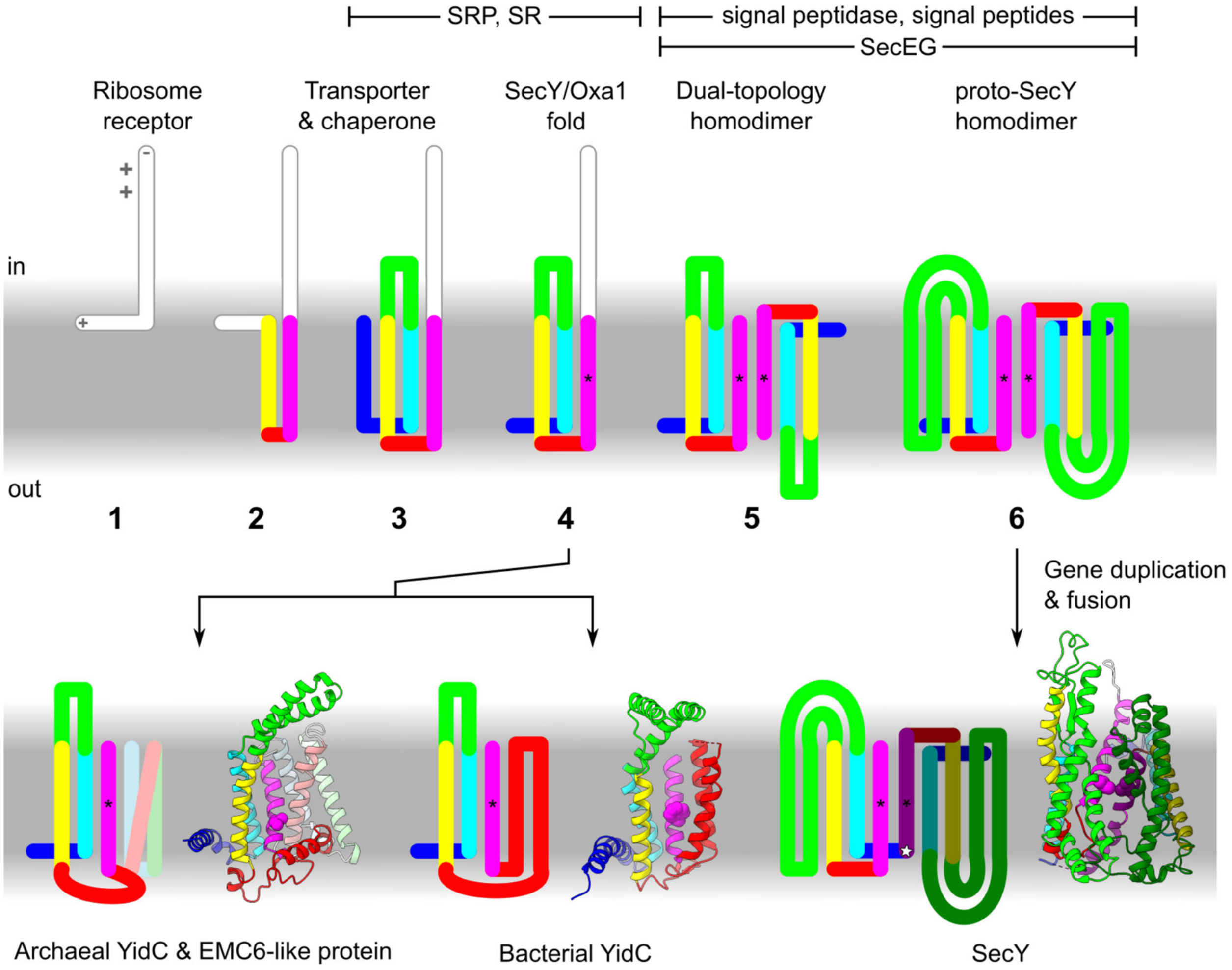


**Fig. 10** Model for the evolution of YidC and SecY. Charged side chains and termini are indicated only at stage 1, by grey symbols. Asterisks indicate the pore ring or groove end residue in H5. At top, additional components of the secretory pathway label a range of stages at which they may have arisen. Models show archaeal YidC and its partner EMC6-like protein (*M. jannaschii* MJ0480 and MJ0606), bacterial YidC (*B. halodurans* YidC2, 3wo6), and SecY (*G. thermodenitrificans* SecY, 6itc)

new position as the extracytoplasmic peripheral helix H0. Thus, the full SecY/Oxa1 fold is now attained. By this time, SRP/SR have evolved, and H2/3 evolves interactions with SR and the ribosome, features that are still evident in SecY, YidC, and TMCO1 [20, 130, 131]. At this stage, the YidC gene duplicated, allowing one paralog to seed the SecY lineage. Paralogous origin in a tandem duplication event would be consistent with the commonly observed juxtaposition of YidC and SecY in prokaryotic genomes [73].

**Step 5**. The original ribosome-binding tail is lost due to its redundancy with SRP/SR for targeting and H2/3 for docking. Loss of this element and genetic drift yields a subpopulation of inverted proteins. Antiparallel dimerisation of the two subpopulations would be favoured because the monomers prefer a similarly thinned membrane, especially near the distal face of H5. Hydrophobic interactions between the groove end residues would favour the particular dimer symmetry of SecY, which juxtaposes them. The non-SecY lineage of YidC (from step 4) evolves in archaea to heterodimerise with EMC6-like protein via the distal face of H5; in bacteria, this same surface becomes covered by the h4h5 transmembrane hairpin.

Antiparallel homodimerisation in the SecY lineage positions hydrophilic grooves on both sides of the membrane, leaving at most a thin hydrophobic layer between them, as in the heterodimeric channel used during ER-associated degradation [79]. 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 originating as TMHs would explain why both engage SecY in a similar way.

At this stage or later, SecEG are acquired. SecE's symmetrical binding to each half of the dimer would stabilise it, particularly when the monomers separate to accommodate substrates. Evolution of SecEG after YidC but before proto-SecY is consistent with evidence that their integration depends on other YidC homologs apart from SecY [102, 111].

**Step 6**. Transmembrane insertion of H2/3 creates a lateral gate, and thus the proto-SecY fold. By inserting between the hydrophilic grooves and the membrane, H2/3 makes those grooves deeper and more hydrophilic, further facilitating translocation. As a secondary consequence, it also creates a more hydrophilic site for signal recognition. This allows cleavable signal peptides to become less hydrophobic than TMHs and thus 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-half atrophy. One such loop was repurposed as the plug. We infer that gene duplication occurred after antiparallel dimerisation because this has precedent [39, 40] and because both halves of SecY conserve the transmembrane insertion of H2/3, which appears to be an adaptation to antiparallel dimerisation.

## Outlook

One might hope that the increasing diversity of known IMP structures will reveal the origins of other pseudosymmetric channels, which have been refractory to sequence searches [132]. But the detectability of SecY's origins may be due to the unusual properties of protein as a transport substrate. Unlike most substrates, protein can be sufficiently hydrophobic to assist in its own translocation, making a partial channel like 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 [19, 133, 134]. Other pre-fusion channel precursors may have exposed similar grooves for transport, but this non-redundant chaperone function is unique to protein substrates. Thus, pre-fusion homologs of other channels would have lacked this reason to be conserved.

Although theories about early evolutionary transitions are not experimentally testable, experimental reconstructions can at least demonstrate their plausibility. Efforts to reconstruct the earliest cells, called protocells, could capitalise on the synergy detailed above between YidC and the putative rotor-stator-like protein-secreting translocase [109]. This protein translocase is itself thought to descend from an RNA translocase, in part because its ATPase domain descends from an RNA helicase. By facilitating the integration of such an RNA translocase, YidC would have indirectly facilitated gene transfer among protocells, thereby allowing recombination to continue despite cellularisation and accelerating this stage of evolution.

Since early studies on protein transport, it has been theorised that protocells were preceded by inside-out precursors, called obcells, which arose when macromolecules colonised the surface of a vesicle [135]. The obcell's interior would then become the protocell's periplasm after an involution akin to gastrulation. This stage would be the earliest that could have hosted protein transporters, but may have featured only a rudimentary genetic code [136]. Consistent with such an early origin, the conserved pore and groove residues identified here (proline, glycine, serine, branched aliphatics) are all abiotically generated [137] and thought to be among the first encoded [138].

Moreover, even the simplest ancestors of YidC modelled here served functions that would be useful during the colonisation of a membrane. Thus, this stage is a reasonable early bound for the origin of YidC. More precise estimates may require more detailed contextual knowledge about protocells and their precursors.

## Conclusions

SecY and YidC have long been considered to be two qualitatively different types of protein transporter. SecY is a typical channel insofar as it transports its substrates through a membrane-spanning aqueous pore, whereas YidC is one of the few atypical channels thought to transport their substrates through a partially thinned membrane. But here, we showed that each half of SecY is in fact surprisingly similar to YidC. Among several previously unrecognised similarities, they share a unique fold whose universally conserved core is a hydrophilic groove through which protein can be transported. SecY is differentiated by its lateral gate helices and pseuodosymmetry, which serve to create a more enclosed, hydrophilic environment that is well suited for the translocation of long soluble domains. Conversely YidC's asymmetry leaves its hydrophilic groove exposed to the membrane, a distinctive feature that may explain its additional function as an intramembrane chaperone. Our analysis of SecY and YidC not only provides new insight into how they function in the present day, but also suggests that they descend from a common ancestor. We developed a unified theory to explain the evolution of their shared and differentiated features, thereby reconstructing a key step in the evolution of cells.

## Methods

### Sequence similarity measures, datasets, and queries

SecY sequence analyses used a recently published dataset of taxonomically diverse prokaryotic sequences [139]. To this dataset, we added the sequences for two structurally characterised SecY (*G. thermodenitrificans* and *M. jannaschii*) and removed 8 fully redundant sequences, 3 highly divergent Elusimicrobia sequences, and 4 N-terminally truncated sequences. The resulting alignment contains 342 sequences, 263 bacterial and 79 archaeal. Sequences were aligned with MAFFT L-INS-i. This and subsequent MAFFT alignments used default parameters, except for using the alternative gap extension penalty --ep 0.123 that is standard for sequences without domain-scale indels.

Pairwise sequence identities within groups of sequences were calculated by re-alignment with ClustalOmega [140] on the European Bioinformatics Institute server [141]. Clustal reports an all-against-all identity matrix and has previously been used to quantify long-term evolutionary trends in sequence identity [142]. Default parameters were used. The number of pairwise comparisons was $342^2$ for the SecY halves, 89 × 75 for the ComEA and UvrC $(HhH)_2$ families, and 79 × 263 for archaeal and bacterial SecY. Sequences were shuffled to estimate excess identity using the Sequence Manipulation Suite [143].

HHpred pairwise comparisons and database queries used the Max Planck Institute for Developmental Biology's server [41]. All used default parameters. The HHpred *p* between the full-length SecY halves was calculated using their subsequences from *M. jannaschii* SecY as input for automatic MSA generation. Database queries pertaining to EMC6/GET2 homologs used the *H. sapiens* EMC6 (NP_001014764.1), GET2 (NP_001736.1), or *Lokiarchaeum sp. GC14_75* Lokiarch_50810 (KKK40543.1) sequence.

The other identified EMC6-like proteins were as follows: *P. horikoshii* WP_010885465.1, *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, and *H. sapiens* NP_061328.1 (C20orf24 isoform a, a.k.a. UniParc isoform 2, Q9BUV8-2).

The N- and C-half sequences were aligned using the structurally similar regions of H1-5 as a seed alignment, to which the 684 N- and C-half sequences were added using MAFFT L-INS-i --seed. This alignment of halves was used as input to ConSurf [144] to score the conservation of each column across the two halves. Conservation scores for *B. halodurans* YidC2 were obtained from ConSurf-DB [145].

*E. coli* IMP annotations and sequences were fetched from UniProt [146]. The sequences for proteins annotated as multi-pass IMPs and not beta-stranded were filtered at 25% identity using MMseqs2 [147], yielding 554 sequences.

Archaeal YidC sequences were collected from Pfam family PF01956. All UniProt sequences assigned to PF01956 were retrieved, and non-archaeal sequences (EMC3 and TMCO1) were excluded. To speed subsequent alignment, the archaeal sequences were filtered at 80% sequence identity using MMseqs2, in target-coverage mode so as to preferentially eliminate fragments. The resulting 871 sequences were aligned by MAFFT L-INS-i. Sequence logos were computed for columns from this alignment and the SecY alignment using DTU Health Tech's Seq2Logo 2.0 server [148] with default parameters.

### Structural similarity measures, database queries, and predictions

Each SecY model was split into N- and C-halves at an arbitrary point in the poorly conserved loop between them close to the C-terminus of N.H5. The resulting half-SecY structures were multiply aligned and compared by TM-score using the Zhang group's mTM-align server [49, 149], which also reports their number of common core a.a. and RMSD.

Structural searches of the PDB25 were performed using the Holm group's Dali server [54]. The Dali PDB25 is a subset filtered at 25% maximum pairwise sequence identity and excludes some additional structures, including TMCO1 (6w6l), due to file format incompatibilities. It contained 21390 chains when queried. Results were manually reviewed to exclude hits with SecY proteins or regions that are not transmembrane. Dali *Z*-scores were equated to *p*-values by assuming an extreme value distribution of scores as in [150],

$$p = 1 - \exp\left(-\exp\left(\frac{\pi}{\sqrt{6}}Z - \gamma\right)\right)$$

where γ is Euler's constant.

The structures of MJ0480/MJ0606 and TMCO1/C20orf24 heterodimers were predicted using AlphaFold 2.0 [72] as implemented by ColabFold [151] and run in a Google Colab notebook. For each query, five models were generated and the top-scoring model by predicted TM-score was refined by AMBER relaxation. Full output and settings are included in Additional file 2.

The number of possible connectivities consistent with the architecture of proto-SecY was counted combinatorially, ($N_{\mathrm{in}}$ TMHs)! × ($N_{\mathrm{out}}$ TMHs)! × ($N_{\mathrm{in}}$ TMHs to which H0 could be prepended) = 3! × 2! × 3 = 36.

### Figure preparation

All models were aligned and rendered in UCSF ChimeraX [152]. Surface hydrophobicity was computed in ChimeraX by its default method: pyMLP [153, 154] with Fauchere propagation and lipophilicity values from [155]. Models depicted relative to a membrane are positioned and oriented according to the prediction algorithm provided by the Orientations of Proteins in Membranes server [156]. OPM does not account for any anisotropy which lipid-exposed hydrophilic residues may induce, and thus none is shown. The OPM-predicted midplane for YidC was adjusted 1.8 Å toward the cytoplasm to agree with molecular dynamics simulations in which a conserved arginine in H1 (homologous to *B. halodurans* YidC2 R72) sits at the bilayer midplane [34]. The membrane's interfacial layers are shown as linear gradients half the width the hydrophobic layer, to approximate experimentally determined polarity profiles [157].

Per-residue hydropathy and charge were computed from protein sequences using EMBOSS Pepinfo [141], topology predicted using TMHMM [5], and plotted in Veusz. The 2-D histogram of IMP length vs TMH count was likewise plotted in Veusz.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-021-01171-5.

**Additional file 1.** Ancestral SecY sequence reconstruction [158–168].

**Additional file 2. Figure S1.** Structural similarity and symmetry breaking between SecY halves. **Figure S2.** Structures of non-Oxa1 superfamily top Dali hits. **Figure S3.** Amino acid conservation in bacterial YidC. **Figure S4.** Structure and contact prediction for archaeal and human heterodimers homologous to EMC3/6. **Figure S5.** Similarity between archaeal and bacterial SecG.

**Additional file 3.** Structure-guided sequence alignment for the SecY halves.

**Additional file 4.** Dali search results.

**Additional file 5.** Predicted MJ0480/MJ0606 and TMCO1/C02orf24 structures.

### Acknowledgements

For critically reading the manuscript, we thank Robert J Keenan, John P O'Donnell, and Viswanathan Chandrasekaran. For insightful discussions, we thank Alexey Murzin.

### Authors' contributions

AJOL contributed to the conception and design of the study, data analysis and interpretation, and drafting and revising the paper. RSH contributed to the design of the study, data interpretation, and revising the paper. The authors read and approved the final manuscript.

### Funding

AJOL is funded by the UK Medical Research Council and the Cambridge Commonwealth, European and International Trust. RSH is funded by the UK Medical Research Council (MC_ UP_A022_1007).

### Availability of data and materials

All data generated or analysed during this study are included in this published article and its supplementary information files.

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

All authors read and approved the final manuscript.

### Competing interests

The authors declare that they have no competing interests.

Received: 17 September 2021 Accepted: 21 October 2021
Published online: 15 December 2021

### References

1. Coleman GA, Davín AA, Mahendrarajah TA, Szánthó LL, Spang A, Hugenholtz P, et al. A rooted phylogeny resolves early bacterial evolution. Science. 2021;372(6542). https://doi.org/10.1126/science.abe0511 Available from: https://science.sciencemag.org/content/372/6542/eabe0511.
2. Williams TA, Szöllősi GJ, Spang A, Foster PG, Heaps SE, Boussau B, et al. Integrative modeling of gene and genome evolution roots the archaeal tree of life. PNAS. 2017;114(23):E4602–11.
3. Park E, Rapoport TA. Mechanisms of Sec61/SecY-mediated protein translocation across membranes. Ann Rev Biophysics. 2012;41:21–40.
4. von Heijne G. Signal sequences: the limits of variation. J Mol Biol. 1985; 184(1):99–105.
5. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. J Mol Biol. 2001;305(3):567–80.
6. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8(10):785–6. https://doi.org/10.1038/nmeth.1701.

7. Jungnickel B, Rapoport TA. A posttargeting signal sequence recognition event in the endoplasmic reticulum membrane. Cell. 1995;82(2):261–70.
8. Li L, Park E, Ling J, Ingram J, Ploegh H, Rapoport TA. Crystal structure of a substrate-engaged SecY protein-translocation channel. Nature. 2016; 531(7594):395.
9. Voorhees RM, Hegde RS. Structure of the Sec61 channel opened by a signal sequence. Science. 2016;351(6268):88–91.
10. Van den Berg B, Clemons WM Jr, Collinson I, Modis Y, Hartmann E, Harrison SC, et al. X-ray structure of a protein-conducting channel. Nature. 2004; 427(6969):36.
11. Ma C, Wu X, Sun D, Park E, Catipovic MA, Rapoport TA, et al. Structure of the substrate-engaged SecA-SecY protein translocation machine. Nat Commun. 2019;10(1):2872. https://doi.org/10.1038/s41467-019-10918-2.
12. Dalal K, Duong F. The SecY complex forms a channel capable of ionic discrimination. EMBO Rep. 2009;10(7):762–8.
13. Park E, Rapoport TA. Preserving the membrane barrier for small molecules during bacterial protein translocation. Nature. 2011;473(7346):239–42.
14. Hessa T, Kim H, Bihlmaier K, Lundin C, Boekel J, Andersson H, et al. Recognition of transmembrane helices by the endoplasmic reticulum translocon. Nature. 2005;433(7024):377–81.
15. Paetzel M, Karla A, Strynadka NCJ, Dalbey RE. Signal peptidases. Chem Rev. 2002;102(12):4549–80.
16. White SH, von Heijne G. Transmembrane helices before, during, and after insertion. Curr Opin Struct Biol. 2005;15(4):378–86.
17. Anghel SA, McGilvray PT, Hegde RS, Keenan RJ. Identification of Oxa1 homologs operating in the eukaryotic endoplasmic reticulum. Cell Rep. 2017;21(13):3708–16.
18. Borowska MT, Dominik PK, Anghel SA, Kossiakoff AA, Keenan RJ. A YidC-like protein in the archaeal plasma membrane. Structure. 2015;23(9):1715–24.
19. Kumazaki K, Chiba S, Takemoto M, Furukawa A, Nishiyama K, Sugano Y, et al. Structural basis of Sec-independent membrane protein insertion by YidC. Nature. 2014;509(7501):516–20. https://doi.org/10.1038/nature13167.
20. McGilvray PT, Anghel SA, Sundaram A, Zhong F, Trnka MJ, Fuller JR, et al. An ER translocon for multi-pass membrane protein biogenesis. Frost A, Pfeffer SR, Frost A, Denic V, editors. eLife. 2020 21;9:e56889.
21. Bai L, You Q, Feng X, Kovach A, Li H. Structure of the ER membrane complex, a transmembrane-domain insertase. Nature. 2020;584(7821):475–8.
22. Miller-Vedam LE, Bräuning B, Popova KD, Schirle Oakdale NT, Bonnar JL, Prabu JR, et al. Structural and mechanistic basis of the EMC-dependent biogenesis of distinct transmembrane clients. Dötsch V, editor. eLife. 2020 Nov 25;9:e62611.
23. O'Donnell JP, Phillips BP, Yagita Y, Juszkiewicz S, Wagner A, Malinverni D, et al. The architecture of EMC reveals a path for membrane protein insertion. Dötsch V, Kuriyan J, Dötsch V, editors. eLife. 2020 May 27;9:e57887.
24. Pleiner T, Tomaleri GP, Januszyk K, Inglis AJ, Hazu M, Voorhees RM. Structural basis for membrane insertion by the human ER membrane protein complex. Science. 2020;369(6502):433–6.
25. McDowell MA, Heimes M, Fiorentino F, Mehmood S, Farkas Á, Coy-Vergara J, et al. Structural basis of tail-anchored membrane protein biogenesis by the GET insertase complex. Molecular Cell [Internet]. 2020 9 [cited 2020 Sep 15]; Available from: http://www.sciencedirect.com/science/article/pii/S109727652030575X
26. Sundberg E, Slagter JG, Fridborg I, Cleary SP, Robinson C, Coupland G. 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. 1997;9(5):717–30. https://doi.org/10.1105/tpc.9.5.717.
27. Bauer M, Behrens M, Esser K, Michaelis G, Pratje E. PET1402, a nuclear gene required for proteolytic processing of cytochrome oxidase subunit 2 in yeast. Molec Gen Genet. 1994;245(3):272–8.
28. Bonnefoy N, Chalvet F, Hamel P, Slonimski PP, Dujardin G. OXA1, a Saccharomyces cerevisiae nuclear gene whose sequence is conserved form prokaryotes to eukaryotes controls cytochrome oxidase biogenesis. J Mol Biol. 1994;239(2):201–12. https://doi.org/10.1006/jmbi.1994.1363.
29. Yen M-R, Harley KT, Tseng Y-H, Saier MH. Phylogenetic and structural analyses of the oxa1 family of protein translocases. FEMS Microbiol Lett. 2001;204(2):223–31.
30. Cymer F, von Heijne G, White SH. Mechanisms of integral membrane protein insertion and folding. J Mol Biol. 2015;427(5):999–1022.
31. Hell K, Neupert W, Stuart RA. Oxa1p acts as a general membrane insertion machinery for proteins encoded by mitochondrial DNA. EMBO J. 2001;20(6): 1281–8.
32. Samuelson JC, Chen M, Jiang F, Möller I, Wiedmann M, Kuhn A, et al. YidC mediates membrane protein insertion in bacteria. Nature. 2000;406(6796): 637–41. https://doi.org/10.1038/35020586.
33. Shanmugam SK, Backes N, Chen Y, Belardo A, Phillips GJ, Dalbey RE. New insights into amino-terminal translocation as revealed by the use of YidC and Sec depletion strains. J Mol Biol. 2019;431(5):1025–37. https://doi.org/10.1016/j.jmb.2019.01.006.
34. Chen Y, Capponi S, Zhu L, Gellenbeck P, Freites JA, White SH, et al. YidC insertase of Escherichia coli: water accessibility and membrane shaping. Structure. 2017;25, 1403(9):–1414.e3.
35. Forrest LR. Structural symmetry in membrane proteins. Annu Rev Biophys. 2015;44(1):311–37. https://doi.org/10.1146/annurev-biophys-051013-023008.
36. Ooi CE, Weiss J. Bidirectional movement of a nascent polypeptide across microsomal membranes reveals requirements for vectorial translocation of proteins. Cell. 1992;71(1):87–96. https://doi.org/10.1016/0092-8674(92)90268-H.
37. Erlandson KJ, Miller SBM, Nam Y, Osborne AR, Zimmer J, Rapoport TA. A role for the two-helix finger of the SecA ATPase in protein translocation. Nature. 2008;455(7215):984–7. https://doi.org/10.1038/nature07439.
38. Matlack KE, Mothes W, Rapoport TA. Protein translocation: tunnel vision. Cell. 1998;92(3):381–90. https://doi.org/10.1016/S0092-8674(00)80930-7.
39. Lolkema JS, Dobrowolski A, Slotboom D-J. Evolution of antiparallel two-domain membrane proteins: tracing multiple gene duplication events in the DUF606 family. J Mol Biol. 2008;378(3):596–606.
40. Rapp M, Granseth E, Seppälä S, von Heijne G. Identification and evolution of dual-topology membrane proteins. Nat Struct Mol Biol. 2006;13(2):112–6. https://doi.org/10.1038/nsmb1057.
41. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, et al. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. J Mol Biol. 2018;430(15):2237–43.
42. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2012; 9(2):173–5. https://doi.org/10.1038/nmeth.1818.
43. Söding J. Protein homology detection by HMM–HMM comparison. Bioinformatics. 2005;21(7):951–60.
44. Longo LM, Despotović D, Weil-Ktorza O, Walker MJ, Jabłońska J, Fridmann-Sirkis Y, et al. Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion. PNAS. 2020;117(27):15731–9. https://doi.org/10.1073/pnas.2001989117.
45. Itskanov S, Kuo KM, Gumbart JC, Park E. Stepwise gating of the Sec61 protein-conducting channel by Sec63 and Sec62. Nat Struct Mol Biol. 2021; 28(2):162–72.
46. Tanaka Y, Sugano Y, Takemoto M, Mori T, Furukawa A, Kusakizako T, et al. Crystal structures of SecYEG in lipidic cubic phase elucidate a precise resting and a peptide-bound state. Cell Rep. 2015;13(8):1561–8. https://doi.org/10.1016/j.celrep.2015.10.025.
47. Braunger K, Pfeffer S, Shrimal S, Gilmore R, Berninghausen O, Mandon EC, et al. Structural basis for coupling protein transport and N-glycosylation at the mammalian endoplasmic reticulum. Science. 2018;360(6385):215–9. https://doi.org/10.1126/science.aar7899.
48. Zimmer J, Nam Y, Rapoport TA. Structure of a complex of the ATPase SecA and the protein-translocation channel. Nature. 2008;455(7215):936–43.
49. Dong R, Peng Z, Zhang Y, Yang J. mTM-align: an algorithm for fast and accurate multiple protein structure alignment. Bioinformatics. 2018;34(10): 1719–25.
50. Li W, Schulman S, Boyd D, Erlandson K, Beckwith J, Rapoport TA. The plug domain of the SecY protein stabilizes the closed state of the translocation channel and maintains a membrane seal. Mol Cell. 2007;26(4):511–21. https://doi.org/10.1016/j.molcel.2007.05.002.
51. Junne T, Schwede T, Goder V, Spiess M. The plug domain of yeast Sec61p is important for efficient protein translocation, but is not essential for cell viability. Mol Biol Cell. 2006;17(9):4063–8. https://doi.org/10.1091/mbc.e06-03-0200.
52. Rollauer SE, Tarry MJ, Graham JE, Jääskeläinen M, Jäger F, Johnson S, et al. Structure of the TatC core of the twin-arginine protein transport system. Nature. 2012;492(7428):210–4. https://doi.org/10.1038/nature11683.
53. Wilkinson BM, Esnault Y, Craven RA, Skiba F, Fieschi J, Képès F, et al. Molecular architecture of the ER translocase probed by chemical crosslinking of Sss1p to complementary fragments of Sec61p. EMBO J. 1997; 16(15):4549–59. https://doi.org/10.1093/emboj/16.15.4549.
54. Holm L. DALI and the persistence of protein shape. Protein Sci. 2020;29(1): 128–40. https://doi.org/10.1002/pro.3749.

55. Kolodny R, Koehl P, Levitt M. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. J Mol Biol. 2005;346(4): 1173–88.
56. Tai C-H, Sam V, Gibrat J-F, Garnier J, Munson PJ, Lee B. Protein domain assignment from the recurrence of locally similar structures. Proteins. 2011; 79(3):853–66. https://doi.org/10.1002/prot.22923.
57. Holm L Using Dali for Protein Structure Comparison. In: Gáspári Z. (eds) Structural Bioinformatics. Methods in Molecular Biology, vol 2112. New York: Humana, NY. 2020. https://doi.org/10.1007/978-1-0716-0270-6_3.
58. Pei J, Mitchell DA, Dixon JE, Grishin NV. Expansion of type II CAAX proteases reveals evolutionary origin of γ-secretase subunit APH-1. J Mol Biol. 2011; 410(1):18–26. https://doi.org/10.1016/j.jmb.2011.04.066.
59. Schaeffer RD, Kinch L, Medvedev KE, Pei J, Cheng H, Grishin N. ECOD: identification of distant homology among multidomain and transmembrane domain proteins. BMC Mol Cell Biol. 2019;20(1):18.
60. van der Sluis EO, Driessen AJM. Stepwise evolution of the Sec machinery in Proteobacteria. Trends Microbiol. 2006;14(3):105–8. https://doi.org/10.1016/j.tim.2006.01.009.
61. Xin Y, Zhao Y, Zheng J, Zhou H, Zhang XC, Tian C, et al. Structure of YidC from Thermotoga maritima and its implications for YidC-mediated membrane protein insertion. FASEB J. 2018;32(5):2411–21. https://doi.org/10.1096/fj.201700893RR.
62. Hessa T, Meindl-Beinker NM, Bernsel A, Kim H, Sato Y, Lerch-Bader M, et al. Molecular code for transmembrane-helix recognition by the Sec61 translocon. Nature. 2007;450(7172):1026–30.
63. Bowie JU. Helix packing in membrane proteins11Edited by G. von Heijne. Journal of Molecular Biology. 1997;272(5):780–9.
64. Gimpelev M, Forrest LR, Murray D, Honig B. Helical packing patterns in membrane and soluble proteins. Biophys J. 2004;87(6):4075–86. https://doi.org/10.1529/biophysj.104.049288.
65. Petrov AS, Bernier CR, Hsiao C, Norris AM, Kovacs NA, Waterbury CC, et al. Evolution of the ribosome at atomic resolution. PNAS. 2014;111(28):10251–6.
66. Williams TA, Cox CJ, Foster PG, Szöllősi GJ, Embley TM. Phylogenomics provides robust support for a two-domains tree of life. Nat Ecol Evol. 2020; 4(1):138–47. https://doi.org/10.1038/s41559-019-1040-x.
67. Beltzer JP, Fiedler K, Fuhrer C, Geffen I, Handschin C, Wessels HP, et al. Charged residues are major determinants of the transmembrane orientation of a signal-anchor sequence. J Biol Chem. 1991;266(2):973–8.
68. Brown J, Behnam R, Coddington L, Tervo DGR, Martin K, Proskurin M, et al. Expanding the optogenetics toolkit by topological inversion of rhodopsins. Cell. 2018;175(4):1131–1140.e11.
69. Rapp M, Seppälä S, Granseth E, von Heijne G. Emulating membrane protein evolution by rational design. Science. 2007;315(5816):1282–4.
70. Sääf A, Johansson M, Wallin E, von Heijne G. Divergent evolution of membrane protein topology: the Escherichia coli RnfA and RnfE homologues. PNAS. 1999;96(15):8540–4.
71. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021;596(7873):590–6. https://doi.org/10.1038/s41586-021-03828-1.
72. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–589. https://doi.org/10.1038/s41586-021-03819-2.
73. Makarova KS, Galperin MY, Koonin EV. 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. 2015;118:302–12.
74. He H, Kuhn A, Dalbey RE. Tracking the stepwise movement of a membrane-inserting protein in vivo. J Mol Biol. 2020;432(2):484–96.
75. Mothes W, Prehn S, Rapoport TA. Systematic probing of the environment of a translocating secretory protein during translocation through the ER membrane. EMBO J. 1994;13(17):3973–82. https://doi.org/10.1002/j.1460-2075.1994.tb06713.x.
76. Shaw AS, Rottier PJ, Rose JK. Evidence for the loop model of signal-sequence insertion into the endoplasmic reticulum. Proc Natl Acad Sci U S A. 1988;85(20):7592–6.
77. Xie K, Hessa T, Seppälä S, Rapp M, von Heijne G, Dalbey RE. Features of transmembrane segments that promote the lateral release from the translocase into the lipid phase. Biochemistry. 2007;46(51):15153–61.
78. Yau W-M, Wimley WC, Gawrisch K, White SH. The preference of tryptophan for membrane interfaces. Biochemistry. 1998;37(42):14713–8.
79. Wu X, Siggel M, Ovchinnikov S, Mi W, Svetlov V, Nudler E, et al. Structural basis of ER-associated protein degradation mediated by the Hrd1 ubiquitin ligase complex. Science. 2020 24 368(6489):eaaz2449. Available from: https://science.sciencemag.org/content/368/6489/eaaz2449
80. Mariappan M, Mateja A, Dobosz M, Bove E, Hegde RS, Keenan RJ. The mechanism of membrane-associated steps in tail-anchored protein insertion. Nature. 2011;477(7362):61–6. https://doi.org/10.1038/nature10362.
81. Stefer S, Reitz S, Wang F, Wild K, Pang Y-Y, Schwarz D, et al. Structural basis for tail-anchored membrane protein biogenesis by the Get3-receptor complex. Science. 2011;333(6043):758–62. https://doi.org/10.1126/science.1207125.
82. Wang F, Whynot A, Tung M, Denic V. The mechanism of tail-anchored protein insertion into the ER membrane. Mol Cell. 2011;43(5):738–50.
83. Guna A, Hegde RS. Transmembrane domain recognition during membrane protein biogenesis and quality control. Curr Biol. 2018;28(8):R498–511.
84. Gogala M, Becker T, Beatrix B, Armache J-P, Barrio-Garcia C, Berninghausen O, et al. Structures of the Sec61 complex engaged in nascent peptide translocation or membrane insertion. Nature. 2014;506(7486):107–10. https://doi.org/10.1038/nature12950.
85. Park E, Ménétret J-F, Gumbart JC, Ludtke SJ, Li W, Whynot A, et al. Structure of the SecY channel during initiation of protein translocation. Nature. 2014; 506(7486):102–6.
86. Plath K, Mothes W, Wilkinson BM, Stirling CJ, Rapoport TA. Signal sequence recognition in posttranslational protein transport across the yeast ER membrane. Cell. 1998;94(6):795–807.
87. Weng T-H, Steinchen W, Beatrix B, Berninghausen O, Becker T, Bange G, et al. Architecture of the active post-translational Sec translocon. EMBO J. 2020;11:e105643.
88. Welte T, Kudva R, Kuhn P, Sturm L, Braig D, Müller M, et al. Promiscuous targeting of polytopic membrane proteins to SecYEG or YidC by the Escherichia coli signal recognition particle. MBoC. 2011;23(3):464–79.
89. Janouškovec J, Tikhonenkov DV, Burki F, Howe AT, Rohwer FL, Mylnikov AP, et al. A new lineage of eukaryotes illuminates early mitochondrial genome reduction. Curr Biol. 2017;27(23):3717–3724.e5.
90. Wiedemann N, Pfanner N. Mitochondrial machineries for protein import and assembly. Annu Rev Biochem. 2017;86(1):685–714.
91. Serdiuk T, Steudle A, Mari SA, Manioglu S, Kaback HR, Kuhn A, et al. Insertion and folding pathways of single membrane proteins guided by translocases and insertases. Sci Adv. 2019;5(1):eaau6824.
92. Güngör B, Flohr T, Garg SG, Herrmann JM. The ER transmembrane complex (EMC) can functionally replace the Oxa1 insertase in mitochondria. bioRxiv. 2021 2021.08.02.454725.
93. Andersson H, von Heijne G. Sec dependent and sec independent assembly of E. coli inner membrane proteins: the topological rules depend on chain length. EMBO J. 1993;12(2):683–91.
94. Kuhn A. Alterations in the extracellular domain of M13 procoat protein make its membrane insertion dependent on secA and secY. Eur J Biochem. 1988;177(2):267–71. https://doi.org/10.1111/j.1432-1033.1988.tb14371.x.
95. Saracco SA, Fox TD. Cox18p is required for export of the mitochondrially encoded Saccharomyces cerevisiae Cox2p C-Tail and interacts with Pnt1p and Mss2p in the inner membrane. MBoC. 2002;13(4):1122–31.
96. Gutiérrez-Fernández J, Kaszuba K, Minhas GS, et al. Key role of quinone in the mechanism of respiratory complex I. Nat Commun. 2020;11:4135. https://doi.org/10.1038/s41467-020-17957-0.
97. Esser L, Zhou F, Yu C-A, Xia D. Crystal structure of bacterial cytochrome bc1 in complex with azoxystrobin reveals a conformational switch of the Rieske iron–sulfur protein subunit. 2019;294(32):12007–19. ISSN: 0021-9258.
98. Svensson-Ek M, Abramson J, Larsson G, Törnroth S, Brzezinski P, Iwata S. The X-ray Crystal Structures of Wild-type and EQ(I-286) Mutant Cytochrome c Oxidases from Rhodobacter sphaeroides. J Mol Biol. 2002;321(2):Pages 329–39. ISSN 0022-2836. https://doi.org/10.1016/S0022-2836(02)00619-8.
99. Guo H, Suzuki T, Rubinstein JL. Structure of a bacterial ATP synthase. 2019. https://doi.org/10.7554/eLife.43128.001.
100. Wallin E, von Heijne G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci. 1998;7(4):1029–38. https://doi.org/10.1002/pro.5560070420.
101. Schmidt A, Kochanowski K, Vedelaar S, Ahrné E, Volkmer B, Callipo L, et al. The quantitative and condition-dependent Escherichia coli proteome. Nat Biotechnol. 2016;34(1):104–10.
102. Guna A, Volkmar N, Christianson JC, Hegde RS. The ER membrane protein complex is a transmembrane domain insertase. Science. 2018;359(6374):470–3.
103. Jonikas MC, Collins SR, Denic V, Oh E, Quan EM, Schmid V, et al. Comprehensive characterization of genes required for protein folding in the

endoplasmic reticulum. Science. 2009;323(5922):1693–7. https://doi.org/10.1126/science.1167983.

104. Lane N, Martin WF. The origin of membrane bioenergetics. Cell. 2012;151(7):1406–16.

105. Weiss MC, Sousa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S, et al. The physiology and habitat of the last universal common ancestor. Nat Microbiol. 2016;1(9):1–8.

106. Xu X, Ouyang M, Lu D, Zheng C, Zhang L. Protein sorting within chloroplasts. Trends in Cell Biology [Internet]. 2020 26 [cited 2020 Dec 10]; Available from: http://www.sciencedirect.com/science/article/pii/S0962892420301914

107. Peltier J-B, Emanuelsson O, Kalume DE, Ytterberg J, Friso G, Rudella A, et al. Central functions of the lumenal and peripheral thylakoid proteome of arabidopsis determined by experimentation and genome-wide prediction. Plant Cell. 2002;14(1):211–36.

108. Vothknecht UC, Westhoff P. Biogenesis and origin of thylakoid membranes. Biochim Biophys Acta. 2001;1541(1):91–101.

109. Mulkidjanian AY, Makarova KS, Galperin MY, Koonin EV. Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. Nat Rev Microbiol. 2007;5(11):892–9. https://doi.org/10.1038/nrmicro1767.

110. Tipton K. Translocases (EC 7): A new EC Class [Internet]. International Union for Biochemistry and Molecular Biology; 2018 Aug [cited 2021 Jul 14]. (Enzyme Nomenclature News). Available from: https://iubmb.org/wp-content/uploads/sites/10116/2018/10/Translocases-EC-7.pdf

111. Yi L, Jiang F, Chen M, Cain B, Bolhuis A, Dalbey RE. YidC is strictly required for membrane insertion of subunits a and c of the F1F0ATP synthase and SecE of the SecYEG translocase. Biochemistry. 2003;42(35):10537–44.

112. Doolittle RF. Convergent evolution: the need to be explicit. Trends Biochem Sci. 1994;19(1):15–8.

113. Bakelar J, Buchanan SK, Noinaj N. The structure of the β-barrel assembly machinery complex. Science. 2016;351(6269):180–6. https://doi.org/10.1126/science.aad3460.

114. Brunner JD, Lim NK, Schenck S, Duerst A, Dutzler R. X-ray structure of a calcium-activated TMEM16 lipid scramblase. Nature. 2014;516(7530):207–12. https://doi.org/10.1038/nature13984.

115. McKenna MJ, Sim SI, Ordureau A, Wei L, Harper JW, Shao S, et al. The endoplasmic reticulum P5A-ATPase is a transmembrane helix dislocase. Science. 2020;369(6511):eabc5809 Available from: https://science.sciencemag.org/content/369/6511/eabc5809.

116. Engelman DM, Steitz TA. The spontaneous insertion of proteins into and across membranes: the helical hairpin hypothesis. Cell. 1981;23(2):411–22.

117. Zhang Y, Ou X, Wang X, et al. Structure of the mitochondrial TIM22 complex from yeast. Cell Res. 2021;31:366–8. https://doi.org/10.1038/s41422-020-00399-0.

118. Gherardini PF, Wass MN, Helmer-Citterich M, Sternberg MJE. Convergent evolution of enzyme active sites is not a rare phenomenon. J Mol Biol. 2007;372(3):817–45.

119. Li S, Shen G, Li W. Intramembrane thiol oxidoreductases: evolutionary convergence and structural controversy. Biochemistry. 2018;57(3):258–66.

120. Mackin KA, Roy RA, Theobald DL. An empirical test of convergent evolution in rhodopsins. Mol Biol Evol. 2014;31(1):85–95.

121. Lupas AN, Ponting CP, Russell RB. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol. 2001;134(2):191–203.

122. Remmert M, Biegert A, Linke D, Lupas AN, Söding J. Evolution of outer membrane β-barrels from an ancestral ββ hairpin. Mol Biol Evol. 2010;27(6):1348–58. https://doi.org/10.1093/molbev/msq017.

123. Cao TB, Saier MH. The general protein secretory pathway: phylogenetic analyses leading to evolutionary conclusions. Biochim Biophys Acta. 2003;1609(1):115–25. https://doi.org/10.1016/S0005-2736(02)00662-4.

124. Kinch LN, Saier J, Milton H, Grishin NV. Sec61β – a component of the archaeal protein secretory system. Trends Biochem Sci. 2002;27(4):170–1. https://doi.org/10.1016/S0968-0004(01)02055-2.

125. Rawlings ND, Bateman A. Origins of peptidases. Biochimie. 166:4–18. https://doi.org/10.1016/j.biochi.2019.07.026.

126. Gribaldo S, Cammarano P. The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery. J Mol Evol. 1998;47(5):508–16. https://doi.org/10.1007/PL00006407.

127. Mulkidjanian AY, Galperin MY, Koonin EV. Co-evolution of primordial membranes and membrane proteins. Trends Biochem Sci. 2009;34(4):206–15.

128. Seitl I, Wickles S, Beckmann R, Kuhn A, Kiefer D. 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. 2014;91(2):408–21. https://doi.org/10.1111/mmi.12465.

129. Szyrach G, Ott M, Bonnefoy N, Neupert W, Herrmann JM. Ribosome binding to the Oxa1 complex facilitates co-translational protein insertion in mitochondria. EMBO J. 2003;22(24):6448–57.

130. Kuhn P, Weiche B, Sturm L, Sommer E, Drepper F, Warscheid B, et al. The bacterial SRP receptor, SecA and the ribosome use overlapping binding sites on the SecY translocon. Traffic. 2011;12(5):563–78.

131. Petriman N-A, Jauß B, Hufnagel A, Franz L, Sachelaru I, Drepper F, et al. The interaction network of the YidC insertase with the SecYEG translocon, SRP and the SRP receptor FtsY. Sci Rep. 2018;8(1):578.

132. Hennerdal A, Falk J, Lindahl E, Elofsson A. Internal duplications in α-helical membrane protein topologies are common but the nonduplicated forms are rare. Protein Sci. 2010;19(12):2305–18.

133. Nagamori S, Smirnova IN, Kaback HR. Role of YidC in folding of polytopic membrane proteins. J Cell Biol. 2004;165(1):53–62. https://doi.org/10.1083/jcb.200402067.

134. Serdiuk T, Balasubramaniam D, Sugihara J, Mari SA, Kaback HR, Müller DJ. YidC assists the stepwise and stochastic folding of membrane proteins. Nat Chem Biol. 2016;12(11):911–7.

135. Blobel G. Intracellular protein topogenesis. PNAS. 1980;77(3):1496–500.

136. Cavalier-Smith T. Obcells as proto-organisms: membrane heredity, lithophosphorylation, and the origins of the genetic code, the first cells, and photosynthesis. J Mol Evol. 2001;53(4–5):555–95.

137. Weber AL, Miller SL. Reasons for the occurrence of the twenty coded protein amino acids. J Mol Evol. 1981;17(5):273–84.

138. Trifonov EN. Consensus temporal order of amino acids and evolution of the triplet code. Gene. 2000;261(1):139–51. https://doi.org/10.1016/S0378-1119(00)00476-5.

139. Harris AJ, Goldman AD. The very early evolution of protein translocation across membranes. PLoS Comput Biol. 2021;17(3):e1008623.

140. Sievers F, Higgins DG. The Clustal Omega Multiple Alignment Package. In: The clustal omega multiple alignment package. In: Multiple Sequence Alignment. Springer; 2021. p. 3–16.

141. Madeira F, Park Y. mi, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019;47(W1):W636–41.

142. Konaté MM, Plata G, Park J, Usmanova DR, Wang H, Vitkup D. Molecular function limits divergent protein evolution on planetary timescales. Elife. 2019;8:e39705. https://doi.org/10.7554/eLife.39705.

143. Stothard P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. BioTechniques. 2000;28(6):1102–4. https://doi.org/10.2144/00286ir01.

144. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res. 2016;44(W1):W344–50.

145. Chorin AB, Masrati G, Kessel A, Narunsky A, Sprinzak J, Lahav S, et al. ConSurf-DB: an accessible repository for the evolutionary conservation patterns of the majority of PDB proteins. Protein Sci. 2020;29(1):258–67.

146. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47(D1):D506–15. https://doi.org/10.1093/nar/gky1049.

147. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026–8.

148. Thomsen MCF, Nielsen M. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. Nucleic Acids Res. 2012;40(W1):W281–7. https://doi.org/10.1093/nar/gks469.

149. Dong R, Pan S, Peng Z, Zhang Y, Yang J. mTM-align: a server for fast protein structure database search and multiple protein structure alignment. Nucleic Acids Res. 2018;46(W1):W380–6.

150. Sierk ML, Pearson WR. Sensitivity and selectivity in protein structure comparison. Protein Sci. 2004;13(3):773–85. https://doi.org/10.1110/ps.03328504.

151. Mirdita M, Ovchinnikov S, Steinegger M. ColabFold - Making protein folding accessible to all [Internet]. 2021 Aug [cited 2021 Oct 6] p. 2021.08.15.456425. Available from: https://www.biorxiv.org/content/10.1101/2021.08.15.456425v1

152. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci. 2020

153. BROTO P, MOREAU G, VANDYCKE C. Molecular structures: perception, autocorrelation descriptor and sar studies: system of atomic contributions

for the calculation of the n-octanol/water partition coefficients. Eur J Med Chem. 1984;19(1):71–8.
154. Laguerre M, Saux M, Dubost JP, Carpy A. MLPP: a program for the calculation of molecular lipophilicity potential in proteins. Pharm Pharmacol Commun. 1997;3(5–6):217–22.
155. Ghose AK, Viswanadhan VN, Wendoloski JJ. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. J Phys Chem A. 1998;102(21): 3762–72. https://doi.org/10.1021/jp980230o.
156. Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res. 2012;40(D1):D370–6.
157. White SH, Wimley WC. Membrane protein folding and stability: physical principles. Annu Rev Biophys Biomol Struct. 1999;28(1):319–65.
158. Trifinopoulos J, Nguyen L-T, von Haeseler A, Minh BQ. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res. 2016;44(W1):W232–5.
159. Naser-Khdour S, Minh BQ, Zhang W, Stone EA, Lanfear R. The prevalence and impact of model violations in phylogenetic analysis. Genome Biol Evol. 2019 Dec 1;11(12):3341–52.
160. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14(6):587–9.
161. Strimmer K, von Haeseler A. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. PNAS. 1997;94(13): 6815–9. https://doi.org/10.1073/pnas.94.13.6815.
162. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37(5):1530–4.
163. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol. 2018;35(2):518–22.
164. Moody ERR, Mahendrarajah TA, Dombrowski N, Clark JW, Petitjean C, Offre P, et al. Universal markers support a long inter-domain branch between Archaea and Bacteria. bioRxiv. 2021 20;2021.01.19.427276.
165. Gouy R, Baurain D, Philippe H. Rooting the tree of life: the phylogenetic jury is still out. Philos Trans R Soc Lond B Biol Sci. 2015;370(1678):20140329. https://doi.org/10.1098/rstb.2014.0329.
166. Jones DT, Taylor WR, Thornton JM. A mutation data matrix for transmembrane proteins. FEBS Lett. 1994;339(3):269–75. https://doi.org/10.1016/0014-5793(94)80429-X.
167. Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol Biol Evol. 2008;25(7):1307–20.
168. Minh BQ, Dang CC, Vinh LS, Lanfear R. QMaker: Fast and accurate method to estimate empirical models of protein evolution. Systematic Biology. 2021.

## Publisher's Note

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