ANCESTRAL MITOCHONDRIAL PROTEIN SECRETION MACHINERY

Lenka Horváthová[1*], Vojtěch Žárský[1*], Tomáš Pánek[2*], Romain Derelle[3], Jan Pyrih[4], Alžběta Motyčková[1], Veronika Klápšťová[1], Vladimír Klimeš[2], Markéta Petrů[1], Zuzana Vaitová[1], Ivan Čepička[5], Karel Harant[6], Michael W. Gray[7], Ingrid Guilvout[8], Olivera Francetic[8], B. Franz Lang[9], Čestmír Vlček[10], Anastasios D. Tsaousis[4], Marek Eliáš[2#], Pavel Doležal[1#]

[1]Department of Parasitology, Faculty of Science, Charles University, BIOCEV, Průmyslová 595, Vestec, 252 42, Czech Republic
[2]Department of Biology and Ecology, Faculty of Science, University of Ostrava, Chittussiho 10, 710 00 Ostrava, Czech Republic
[3]School of Biosciences, University of Birmingham, Edgbaston, B15 2TT, UK
[4]Laboratory of Molecular & Evolutionary Parasitology, RAPID group, School of Biosciences, University of Kent, Canterbury, CT2 7NZ, UK
[5]Department of Zoology, Faculty of Science, Charles University, Viničná 7, Prague 2, 128 44, Czech Republic
[6]Proteomic core facility, Faculty of Science, Charles University , BIOCEV, Průmyslová 595, Vestec, 252 42, Czech Republic
[7]Department of Biochemistry and Molecular Biology and Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, NS B3H 4R2, Canada
[8]Institut Pasteur, Biochemistry of Macromolecular Interactions Unit, Department of Structural Biology and Chemistry, CNRS UMR3528, 75015 Paris, France
[9]Robert Cedergren Centre for Bioinformatics and Genomics, Département de Biochimie, Université de Montréal, Montreal, QC, Canada H3T 1J4
[10]Institute of Molecular Genetics, Czech Academy of Sciences, 142 20 Prague 4, Czech Republic

*these authors equally contributed to the study
#corresponding authors: Marek Eliáš (marek.elias@osu.cz), Pavel Doležal (pavel.dolezal@natur.cuni.cz)

Abstract

Modern mitochondria have preserved few traits of the original bacterial endosymbiont. Unexpectedly, we find that certain representatives of heteroloboseans, jakobids and malawimonads possess homologues of four core components of the type 2 secretion system (T2SS) so far restricted to eubacteria. We show that these components are localized to the mitochondrion, and their behaviour in functional assays is consistent with the formation of a mitochondrial T2SS-derived protein secretion system. We additionally identified 23 protein families exactly co-occurring in eukaryotes with the T2SS. Seven of these proteins could be directly linked to the core T2SS by functional data and/or sequence features, whereas others may represent different parts of a broader functional pathway, possibly linking the mitochondrion with the peroxisome. Its distribution in eukaryotes and phylogenetic evidence indicate that the whole mitochondrial T2SS-centred pathway is an ancestral eukaryotic trait. Our findings thus have direct implications for the functional properties of the early mitochondrion.

Introduction

Mitochondria of all eukaryotes arose from the same Alphaproteobacteria-related endosymbiotic bacterium[1,2]. New functions have been incorporated into the bacterial blueprint during mitochondrial evolution, while many ancestral traits have been lost. Importantly, in some cases, these losses occurred independently in different lineages of eukaryotes, resulting in a patchy distribution of the respective ancestral mitochondrial traits in extant eukaryotes. A good example of this is the ancestral mitochondrial division apparatus (including homologues of bacterial Min proteins) retained in several distantly related protist lineages[3,4]. It is likely that additional pieces of the ancestral bacterial cell physiology will be discovered in mitochondria of poorly studied eukaryotes.

An apparent significant difference between the mitochondrion and bacteria (including those living as endosymbionts of eukaryotes) lies in the directionality of protein transport across their envelope. All bacteria export specific proteins from the cell via the plasma membrane using the Sec or Tat machineries[5], and many diderm (Gram-negative) bacteria exhibit specialized systems mediating further protein translocation across the outer membrane (OM)[6]. In contrast, the mitochondrion depends on a newly evolved protein import system spanning both envelope membranes and enabling import of proteins encoded by the nuclear genome[7]. The capacity of mitochondria to secrete proteins seems to be limited. Mitochondrial homologues of Tat translocase subunits occur in some eukaryotic taxa, but their role in protein secretion has not been established[8]. A mitochondrial homologue of the SecY protein (a Sec translocase subunit) has been described only in jakobids[9,10], but its function remains elusive[11]. No dedicated machinery for protein export from the mitochondrion across the outer mitochondrial membrane has been described.

One of the best characterized bacterial protein translocation machineries is the so-called type 2 secretion system (T2SS)[12,13]. The T2SS belongs to a large bacterial superfamily of type 4 pili (T4P)-related molecular machines, most of which secrete long extracellular filaments (pili) for motility, adhesion, or DNA uptake[14–16]. Using building blocks homologous to components of the other members of the T4P superfamily, the T2SS constitutes a specialized secretion apparatus, whose filament (pseudopilus) remains in the periplasm[12,13]. It is composed of 12-15 conserved components, commonly referred to as general secretion pathway (Gsp) proteins, which assemble into four main subcomplexes (Fig. 1A). The OM pore is formed by an oligomer of 15-16 molecules of the GspD protein[17]. The subcomplex in the inner membrane (IM) is called the assembly platform and consists of the central multispanning membrane protein GspF surrounded by single-pass membrane proteins GspC, GspL, and GspM. GspC links the assembly platform to the OM pore by interacting with the periplasmic N-terminal domain of GspD[18]. The third subcomplex, called the pseudopilus, is a helical filament formed mainly of GspG subunits, with minor pseudopilins (GspH, GspI, GspJ and GspK) assembled at its tip[15]. The pseudopilus is assembled at the assembly platform and its growth is believed to push the periplasmic T2SS substrate through the OM pore. The energy for pseudopilus assembly is provided by the fourth subcomplex, the hexameric ATPase GspE, interacting with the assembly platform from the cytoplasmic side[16].

Substrates for T2SS-mediated secretion are first transported by the Tat (as folded proteins) or the Sec (in an unfolded form) system across the IM into the periplasm, where they undergo maturation and/or folding. The folded substrates are finally loaded onto the pseudopilus for the release outside the cell via the OM pore. The known T2SS substrates differ between taxa and no common sequence features have been identified for them. Proteins transported by the T2SS in different species include catabolic enzymes (such as lipases, proteases or phosphatases) and, in the case of bacterial pathogens, toxins[12]. A recent survey of bacterial genomes showed that the T2SS is mainly present in Proteobacteria[19]. Crucially, neither the T2SS nor other systems of the T4P superfamily have been reported from eukaryotes[6,12,20].

Here we show that certain distantly related eukaryotes unexpectedly contain homologues of key T2SS subunits representing all four functional T2SS subcomplexes. We provide evidence for mitochondrial localization of these eukaryotic Gsp homologues and describe experimental results supporting the idea that they constitute a system similar to the bacterial T2SS. Furthermore, we point to the existence of 23 proteins with a perfect taxonomic co-occurrence with the eukaryotic Gsp homologues. Some of these co-occurring proteins seem to be additional components of the mitochondrial T2SS-related machinery, whereas others are candidates for components of a broader functional pathway linking the mitochondrion with other parts of the cell. Given its phylogenetic distribution we propose that the newly discovered pathway was ancestrally present in eukaryotes. Its further characterization may provide fundamental new insights into the evolutionary conversion of the protomitochondrion into the mitochondrial organelle.

Results

Certain protist lineages code for a conserved set of homologues of T2SS core components

While searching the genome of the heterolobosean *Naegleria gruberi* for proteins of bacterial origin with a possible mitochondrial role, we surprisingly discovered homologues of four core subunits of the bacterial T2SS, specifically GspD, GspE, GspF, and GspG (Fig. 1A; Supplementary Table 1). Using genomic and transcriptomic data from public repositories and our on-going sequencing projects for several protist species of key evolutionary interest, we mapped the distribution of these four components in eukaryotes. All four genes were found in the following characteristic set of taxa (Fig. 1B, Supplementary Table 1): three additional heteroloboseans (*Naegleria fowleri*, *Neovahlkampfia damariscottae*, *Pharyngomonas kirbyi*), two jakobids (*R. americana* and *Andalucia godoyi*), and two malawimonads (*Malawimonas jakobiformis* and *Gefionella okellyi*). In addition, three separate representatives of the heterolobosean genus *Percolomonas* (Supplementary Fig. 1) each exhibited a homologue of GspD, but not of the remaining Gsp proteins, in the available transcriptomic data. In contrast, all four genes were missing in sequence data from all other eukaryotes investigated, including the genome and transcriptome of another malawimonad ("Malawimonas californiana") and deeply-sequenced transcriptomes of a third jakobid (*Stygiella incarcerata*) and four additional heteroloboseans (*Creneis carolina*, "Dactylomonas venusta", *Harpagon schusteri*, and the undescribed strain Heterolobosea sp. BB2).

Probing *N. gruberi* nuclei with fluorescent *in situ* hybridization ruled out an unidentified bacterial endosymbiont as the source of the Gsp genes (Supplementary Fig. 2). Moreover, the eukaryotic Gsp genes usually have introns and constitute robustly supported monophyletic groups well separated from bacterial homologues (Fig. 1C; Supplementary Fig. 3), ruling out bacterial contamination in all cases. In an attempt to illuminate the origin of the eukaryotic Gsp proteins we carried out systematic phylogenetic analyses based on progressively expanded datasets of prokaryotic homologues and for each tree inferred the taxonomic identity of the bacterial ancestor of the eukaryotic branch (see Methods for details on the procedure). The results, summarized in Supplementary Fig. 3, showed that the inference is highly unstable depending on the dataset analysed, and no specific bacterial group can be identified as an obvious donor of the eukaryotic Gsp genes. This result probably stems from a combination of factors, including the long branches separating the eukaryotic and bacterial Gsp sequences, the length of Gsp proteins restricting the amount of the phylogenetic signal retained, and perhaps also rampant horizontal gene transfer of the T2SS system genes between bacterial taxa. The eukaryotic Gsp genes are in fact so divergent that some of them could not be unambiguously classified as specific homologs of T2SS components (as opposed to the related machineries of the T4P superfamily) when analysed using models developed for the bacterial genomes[19] (Supplementary Fig. 3).

Heteroloboseans, jakobids and malawimonads have been classified in the hypothetical supergroup Excavata[21]. However, recent phylogenomic analyses indicate that excavates are non-monophyletic and even suggest that malawimonads are separated from heteroloboseans and jakobids by the root of the eukaryote phylogeny[22–25]. Hence, the current phylogenetic distribution of the Gsp homologues in eukaryotes may reflect their presence in the last eukaryotic common ancestor (LECA) followed by multiple independent losses (Fig. 1C). Heteroloboseans and malawimonads have two GspG paralogues, but the phylogenetic analyses did not resolve whether this is due to multiple independent GspG gene duplications or one ancestral eukaryotic duplication followed by loss of one of the paralogues in jakobids (Supplementary Fig. 3D; Supplementary Table 1).

The eukaryotic Gsp proteins localize to the mitochondrion

We hypothesized that the eukaryotic homologues of the four Gsp proteins are parts of a functional T2SS-related system localized to the mitochondrion. This notion was supported by the presence of predicted N-terminal mitochondrial targeting sequences (MTSs) in some of the eukaryotic Gsp proteins (Supplementary Table 1). The prediction algorithms identified putative N-terminal MTSs for proteins from jakobids and malawimonads but failed to recognize them in the orthologues from heteroloboseans, which, however, carry the longest N-terminal extensions (Supplementary Fig. 4). We assumed that these extensions might still function as MTSs in heteroloboseans. Indeed, labelling of *N. gruberi* cells using specific polyclonal antibodies showed that GspD, GspF and GspG1 are present in mitochondria (Fig. 2A). Moreover, the atypical MTSs of *N. gruberi* Gsp proteins were efficiently recognized by the yeast mitochondrial import machinery (Supplementary Fig. 5). Analogously, three Gsp proteins from *G. okellyi* were all localized to mitochondria when expressed in yeast (Fig. 2B).

In order to further confirm the mitochondrial localization of the Gsp proteins in *N. gruberi*, we analysed the mitochondrial proteome of this species by partial purification of the organelle and identification of resident proteins by mass spectrometry. A mitochondria-enriched fraction was obtained from a cellular lysate by several steps of differential centrifugation and further separated by OptiPrep gradient centrifugation. Three sub-fractions of different densities were collected (Supplementary Fig. 6A) and subjected to proteomic analysis. The relative amount of each protein in the gradient was determined by label-free quantification and the proteins were grouped by a multicomponent analysis (for details see Methods) according to their distributions across the gradient (Fig. 3). A set of marker proteins (homologs of well characterized typical mitochondrial proteins from other species) was used to identify a cluster of mitochondrial proteins. Due to the partial co-purification of peroxisomes with mitochondria, a peroxisome-specific cluster was defined analogously. As a result, 946 putative mitochondrial and 78 putative peroxisomal proteins were identified among the total of 4,198 proteins detected. Encouragingly, the putative mitochondrial proteome of *N. gruberi* is dominated by proteins expected to be mitochondrial or whose mitochondrial localization is not unlikely (Supplementary Fig. 6B, Supplementary Table 2A). On the other hand, the putative peroxisomal proteome seems to be contaminated by mitochondrial proteins (owing to the presence of several mitochondrial ribosomal proteins; Supplementary Table 2B). Importantly, all five Gsp proteins (including both GspG paralogs) were identified in the putative mitochondrial but not peroxisomal proteome of *N. gruberi*.

The properties of the eukaryotic Gsp proteins support the existence of a mitochondrial T2SS-related machinery

The foregoing experiments support the idea that all four eukaryotic Gsp homologues localize to and function in the mitochondrion. However, direct *in vivo* demonstration of the existence of a functional mitochondrial T2SS-related machinery is currently not feasible, because none of the Gsp homologue-carrying eukaryotes represents a tractable genetic system. We thus used *in vitro* approaches and heterologous expression systems to test the key properties of the eukaryotic Gsp proteins.

Crucial for the T2SS function is the formation of the OM pore, which is a β-barrel formed by the oligomerization of the C-domain of the GspD protein[26]. The actual assembly of the bacterial pore requires the interaction of the very C-terminal domain of GspD (S-domain) with the outer membrane lipoprotein GspS[27]. In addition, the bacterial GspD carries four short N-terminal domains exposed to the periplasm, called N0 to N3, of which N1 to N3 share a similar fold[28] (Fig. 4A). While the N3 domain has been shown to participate in the pore assembly, N0 interacts with GspC of the assembly platform[18]. However, sequence analysis of the mitochondrial GspD homologue revealed that it, in fact, corresponds to only a C-terminal part of the bacterial GspD β -barrel C-domain, whereas the N-terminal domains N0 to N3, the N-terminal part of the C-domain, and the S-domain are missing (Fig. 4A). This finding raised a question whether the mitochondrial GspD homologue has retained the ability to form a membrane pore. Nevertheless, homology modelling of GspD from *G. okellyi* (*Go*GspD) using *Vibrio cholerae* GspD[29] as a template indicated that the protein could be fitted into solved structure of the pentadecameric pore complex with the conserved amphipathic helical loop (AHL)(Fig. 4B).

Testing the function of *Go*GspD in bacteria was impossible due to its high toxicity leading to rapid cell death upon induction of protein expression (Fig. 4C), which is a typical behaviour of pore-forming proteins. The protein toxicity was less pronounced in the yeast two-hybrid (Y2H) system, which indicated strong self-interaction of *Go*GspD (Fig. 4D), and hence its ability to oligomerize. Indeed, radioactively labelled *Go*GspD assembled into a high-molecular-weight complex in an experimental membrane in an *in vitro* translation assay (Fig. 4E). The formation of the complex was dependent on the presence of the membrane and the complex was resistant to 2M urea treatment, which would remove nonspecific protein aggregates. These results showed that the mitochondrial GspD, despite being significantly truncated when compared to its bacterial homologues, has retained the capability to form membrane pores, characteristic for the secretins of the T2SS [30]. Compared to the bacterial GspD, the predicted *Go*GspD structure suggests a unique biogenesis pathway, where the secretin pore-forming domain may be directly inserted in the mitochondrial outer membrane, bypassing the membrane transport essential for its bacterial counterparts.

The secretion mechanism of the T2SS relies on assembly of pseudopilus made up of GspG subunits[15]. A possible assembly of mitochondrial GspG from *G. okellyi* (*Go*GspG1) into the pseudopilus structure was indicated by modelling the protein sequence into the recently obtained cryoEM reconstruction of the PulG complex from *Klebsiella oxytoca*[20] (Supplementary Fig. 7). The actual interaction properties of *Go*GspG1 were followed by the bacterial two-hybrid assay (B2H). When expressed in bacteria (in a truncated form with the MTS region removed, see

Fig. 5A), the mitochondrial *Go*GspG1 interacted with itself (Fig. 5B), which is a prerequisite for pseudopilus formation. An analogous B2H assays of *N. gruberi* Gsp proteins also showed GspG1 self-interaction (data not shown). In addition, *Go*GspG1 showed positive interaction with *Go*GspF, the IM component believed to participate in transfer of energy for the pseudopilus assembly from GspE (Fig. 1A). Moreover, the mitochondrial *Go*GspF and *Go*GspE each self-interacted in the B2H assay (Fig. 5B). These interactions are in agreement with the role of both proteins as T2SS components, as GspF forms dimers within the IM complex and GspE assembles into an active hexameric ATPase. Furthermore, B2H assay has identified the same interactions between the GspG and GspF homologues in the bacterial T2SS [31]. Tests of all other possible interactions of *G. okellyi* Gsp proteins were negative.

The *in silico* analyses and experiments described above are consistent with the hypothesized existence of a functional mitochondrial secretion machinery derived from the bacterial T2SS. However, the mitochondrial subunits identified would assemble only a minimalist version of the secretion system, reduced to the functional core of the four subcomplexes of the bacterial T2SS, i.e. the luminal ATPase (GspE), the IM pseudopilus assembly platform (GspF), the intermembrane space pseudopilus (GspG), and the OM pore (truncated GspD). Despite using sensitive HMM-based searches, we did not detect homologues of other conserved T2SS subunits in any of the eukaryotes possessing GspD to GspG proteins. One of the missing subunits is GspC, which connects the assembly platform with the N0 domain of GspD pore[18,32]. Thus, the absence of GspC in eukaryotes correlates with the lack of the N0 domain in the eukaryotic GspD. Analogously, the absence of the C-terminal S-domain in the mitochondrial GspD (Fig. 4A), known to be missing also from some bacterial GspD proteins, rationalizes the lack of a eukaryotic homologue of the bacterial OM component GspS that binds to GspD via the S-domain during the pore assembly[27].

The mitochondrial system also apparently lacks a homologue of GspO, a bifunctional enzyme that is essential for GspG maturation. Despite this absence, eukaryotic GspG homologues have conserved all the characteristic sequence features required for GspG maturation (the polar anchor and the trans-membrane domain with a conserved glutamate residue at the +5 position relative to the processing site) (Fig. 5A, Supplementary Fig. 4D). Notably, all the *Ng*GspG1 and *Ng*GspG2-derived peptides detected in our proteomic analysis come from the region of the protein downstream of the conserved processing site (Fig.5C), and an anti-*Ng*GspG1 antibody detected a specific band of a much smaller size than expected for the full-length protein (around 44 kDa) on a western blot of electrophoretically separated *N. gruberi* proteins (Fig.5D). However, the theoretical Mw of the *Ng*GspG1 processed at the conserved site is 25.5 kDa, whereas the protein detected by the immunoblot is even smaller, with a size similar to that of bacterial pseudopilins. Hence, the actual nature of the mitochondrial GspG maturation needs to be studied further.

New putative components of the mitochondrial T2SS-based functional pathway identified by phylogenetic profiling

Since none of the eukaryotes with the Gsp homologues is currently amenable to functional studies, we tried to further illuminate the role of the mitochondrial T2SS system using a comparative genomic approach. Specifically, we reasoned that possible additional components of the machinery, as well as its actual substrate(s), might show the same phylogenetic distribution as the originally identified four subunits. Using a combination of an automated identification of candidate protein families and subsequent manual scrutiny by exhaustive searches of available eukaryote sequence data (for details of the procedure see Methods), we identified 23 proteins (more precisely, groups of orthologues) that proved to exhibit precisely the same phylogenetic distribution in eukaryotes as the four core T2SS components. Specifically, all 23 proteins were represented in each of the heterolobosean, jakobid, and malawimonad species possessing all four core Gsp proteins, whereas only seven of them were found in the transcriptomic data from the *Percolomonas* lineage that possesses only GspD (Fig. 1B; Supplementary Table 3). Except for two presumably Gsp-positive jakobids represented by incomplete EST surveys and a case of a likely contamination (Supplementary Table 4), no orthologues of any of these proteins were found in any other eukaryote (including the Gsp-lacking members of heteroloboseans, jakobids and malawimonads). The sequences of these 23 proteins were analysed by various *in silico* approaches, including sensitive homology-detection methods (HMM-HMM comparisons with HHpred[33] and protein modelling using the Phyre2 server[34]) to assess their possible function (Fig. 6A; Supplementary Table 3).

These analyses revealed that seven of the families have a direct link to the T2SS suggested by discerned homology to known T2SS components. One of them represents an additional, more divergent homologue of the C-terminal part of the bacterial GspD. Hence, the protein has been marked as GspDL (GspD-like). Three other families, referred to as GspDN1 to GspDN3, proved to be homologous to the Secretin_N domain (Pfam family PF03958), present in the bacterial GspD protein in three copies as the domains N1, N2, and N3 (Fig. 4A). The N1-N3 array protrudes into the periplasmic space, where it oligomerizes to form three stacked rings[35]. As mentioned above, the initially identified eukaryotic GspD homologues lack the N-terminal region, suggesting that the gene was split into multiple parts in eukaryotes. Unfortunately, high sequence divergence makes it impossible to identify potential specific correspondence between the N1 to N3 domains of the bacterial GspD and the eukaryotic GspDN1 to GspDN3 proteins. Importantly, an initial Y2H assay indicated that the two separate polypeptides GspD and GspDN1 of *N. gruberi* may interact *in vivo* (Fig. 4F), perhaps forming a larger mitochondrial complex. In addition, we identified most of the newly discovered GspD-related proteins (GspDL and GspDN) in the *N. gruberi* mitochondrial proteome (the exception being GspDN1, which was not detected in a sufficient number of replicates to be included in the downstream analysis; Supplementary Table 2A).

The final three proteins linked to the T2SS based on their sequence features represent three divergent paralogues of the GspE subunit (GspE-like) here denoted GspEL1 to GspEL3. However, abrogation of ATPase-specific motifs in these paralogues suggests the loss of the ATPase activity (Supplementary Fig. 4B). GspEL2 and GspEL3 were identified among *N. gruberi* mitochondrial proteins in the proteomic analysis, whereas GspEL1 was found in the cluster of putative peroxisomal proteins.

The remaining sixteen proteins co-occurring with the core eukaryotic T2SS subunits, hereafter referred to as Gcp (Gsp-co-occurring proteins), were divided into three categories. The first comprises four proteins that constitute novel paralogues within broader common eukaryotic (super)families (Fig. 6B). Three of them (Gcp1 to Gcp3) belong to the WD40 superfamily, in which they form a single clade together with the peroxisomal protein import co-receptor Pex7 (Fig. 6B; Supplementary Fig. 8). None of these proteins has any putative N-terminal targeting sequence, but interestingly, the peroxisomal targeting signal 1 (PTS1) could be predicted on most Gcp1 and some Gcp2 proteins (Supplementary Table 3). However, these predictions are not fully consistent with the results of our proteomic analysis: *Ng*Gcp1 was found among the mitochondrial proteins and *Ng*Gcp2 in the cluster of putative peroxisomal proteins (Supplementary Table 2), but PTS1 is predicted to be present in the *Ng*Gcp1 protein (Supplementary Table 3). The fourth Gcp protein (Gcp4) is a novel paralogue of the ubiquitin-like superfamily, distinctly different from the previously characterized members including ubiquitin, SUMO, NEDD8 and others (Supplementary Fig. 9).

The second Gcp category comprises eleven proteins (Gcp5 to Gcp15) well conserved at the sequence level among the Gsp-containing eukaryotes, yet lacking any discernible homologues in other eukaryotes or in prokaryotes. Two of these proteins (Gcp8, Gcp15) were not identified in the proteomic analysis of *N. gruberi* (Supplementary Table 3). Of those identified, several (Gcp5, Gcp6, Gcp13) were found among the mitochondrial proteins, whereas some others (Gcp9, Gcp10, Gcp11) clustered with peroxisomal markers. Specific localization of the three remaining proteins (Gcp7, Gcp12, and Gcp14) could not be determined due to their presence at the boundaries of the mitochondrial or peroxisomal clusters. No homology to other proteins or domains could be discerned for the Gsp5 to Gsp15 proteins even when sensitive homology-detection algorithms were employed. However, four of them are predicted as single-pass membrane proteins, with the transmembrane segment in the N- (Gcp7, Gcp11, Gcp15) or C-terminus (Gcp5) (Fig. 6A; Supplementary Fig. 10). Interestingly, Gcp6 and Gcp12 proteins contain multiple absolutely conserved cysteine or histidine residues (Fig. 6A; Supplementary Fig. 11).

Finally, Gcp16 constitutes a category of its own. It typifies a family of predicted membrane proteins with non-eukaryotic representatives restricted to bacteria of the PVC superphylum (Supplementary Fig. 12), some of which are known to have the T2SS[36]. Interestingly, Gcp16 proteins from *Neochlamydia* spp. are fused to the N-terminus of a protein from the Lactamase_B_2 (PF12706) family that generally occurs as an independent protein widely conserved in various bacteria. Phylogenetic analyses confirmed that the eukaryotic members of the family are of the same origin rather than acquisitions by independent HGT events into different lineages of eukaryotes (Supplementary Fig. 13). Most eukaryotic Gcp16 proteins exhibit an N-terminal extension compared to the bacterial homologues (Supplementary Fig. 12), but only some of these extensions are recognized as putative MTSs and the *N. gruberi* Gcp16 was not identified either in putative mitochondrial or peroxisomal proteome.

Discussion

Our analyses revealed that a subset of species belonging to three eukaryotic lineages share a set of at least 27 proteins (or families of orthologues) absent from other eukaryotes for which genomic or transcriptomic data are currently available (Fig. 1C). At least eleven of these proteins (the Gsp proteins) are evolutionarily related to components of the bacterial T2SS, although seven of them are so divergent that their evolutionary connection to the T2SS could be recognized only retrospectively after their identification based on their characteristic phylogenetic profile. For the sixteen remaining proteins (Gcp1 to Gcp16) no other evolutionary or functional link to the T2SS is evident apart from the same phyletic pattern as exhibited by the T2SS subunit homologues. Nevertheless, similar phylogenetic profiles are generally a strong indication for proteins being parts of the same functional system or pathway, and have enabled identification of new components of different cellular structures or pathways (e.g. refs[37,38]). Is it, therefore, possible that the 27 Gsp/Gcp proteins similarly belong to a single functional pathway?

The phylogenetic profile shared by the eukaryotic Gsp and Gcp proteins is not trivial, as it implies independent gene losses in a specific set of multiple eukaryotic branches (Fig. 1B). The likelihood of a chance emergence of the same taxonomic distribution of these proteins is thus low. Nevertheless, false positives cannot be completely excluded among the Gcp proteins and their list may be revised when a more comprehensive sampling of eukaryote genomes or transcriptomes becomes available. It is also possible that the currently inferred phylogenetic profile of some of the Gsp/Gcp proteins is inaccurate due to incomplete sampling of the actual gene repertoire of species represented by transcriptome assemblies only. An interesting case in point is the heterolobosean *Percolomonas* lineage. Transcriptomic data from three different members revealed only the presence of GspD, GspDL, the three GspDN variants, and four Gcp proteins (Fig. 1B, Supplementary Tables 1 and 3), which may reflect incomplete data. However, the relatively coherent pattern of Gsp/Gcp protein occurrence in the three independently sequenced transcriptomes and the fact that in other Gsp/Gcp - containing eukaryotes all 27 families are always represented in the respective transcriptome assembly (Supplementary Tables 1 and 3) suggest that the *Percolomonas* lineage has preserved only a subset of Gsp/Gcp families. Genome sequencing is required to test this possibility.

All uncertainties notwithstanding, our data favour the idea that a hitherto unknown complex functional pathway exists in some eukaryotic cells, underpinned by most, if not all, of the 27 Gsp/Gcp proteins and possibly others yet to be discovered. Direct biochemical and cell biological investigations are required for testing its very existence and the actual cellular role. Nevertheless, we integrated the experimental data gathered so far with the insights from bioinformatic analyses to propose a hypothetical working model (Fig. 7).

Our main proposition is that the eukaryotic homologues of the bacterial Gsp proteins assemble a functional transport system, here denoted miT2SS, that spans the mitochondrial OM and mediates the export of specific substrate proteins from the mitochondrion. Although the actual architecture of the miT2SS needs to be determined, the available data suggest that it departs in detail from the canonical bacterial T2SS organization, as homologues of some of the important bacterial T2SS components are apparently missing. Most notable is the absence of GspC, presumably related to the modified structure of its interacting partner GspD, which in eukaryotes is split into multiple polypeptides and seems to completely lack the N0 domain involved in GspC binding. It thus remains unclear whether and how the IM assembly platform and the OM pore interact in mitochondria. One possible explanation is that GspC has been replaced by an unrelated protein. It is notable that three Gcp proteins (Gcp7, Gcp11, and Gcp15) have the same general architecture as GspC: they possess a transmembrane segment at the N-terminus and a (predicted) globular domain at the C-terminus (Fig. 6A). Testing possible interactions between these proteins and T2SS core subunits (particularly GspF and GspDN) using B2H or Y2H assays will be of future interest.

Future investigations also must address the question of whether the mitochondrial GspG is processed analogously to the bacterial homologues and how such processing occurs in the absence of discernible homologues of GspO (see above). The mitochondrial GspG is presumably inserted into the IM by the Tim22 or Tim23 complex, resulting in a GspG precursor with the N-terminus, including the MTS, protruding into the matrix. It is possible that N-terminal cleavage by matrix processing peptidase serves not only to remove the transit peptide, but at the same time to generate the mature N-terminus of the processed GspG form, ready for recruitment into the pseudopilus.

In parallel with its apparent simplification, the miT2SS may have been specifically elaborated compared to the ancestral bacterial machinery. This possibility is suggested by the existence of the three divergent, possibly ATPase activity-deficient GspE paralogues (GspEL1 to GspEL3) that we discovered in all miT2SS-containing eukaryotes but not elsewhere. We can only speculate as to the function of these proteins, but they may interact with and regulate the catalytically active GspE protein. The fact that the bacterial GspE assembles into a homohexamer raises the possibility that in eukaryotes GspEL proteins are included in a heterooligomer with GspE, a situation analogous to the presence of catalytically active and inactive paralogous subunits in some well known protein complexes (e.g. refs[39,40]). The co-occurrence of two different paralogues of the GspD C-domain, one (GspDL) being particularly divergent, suggests a eukaryote-specific elaboration of the putative pore in the mitochondrial OM.

An unanswered key question is what is the actual substrate (or substrates) exported from the mitochondrion by the miT2SS. No bioinformatic tool for T2SS substrate prediction is available due to the enigmatic nature of the mechanism of substrate recognition by the pathway[12], so at the moment we can only speculate. It is notable that no protein encoded by the mitochondrial genomes of jakobids, heteroloboseans and malawimonads stands out as an obvious candidate for the miT2SS substrate, since they either have well-established roles in the mitochondrion or are hypothetical proteins with a restricted (genus-specific) distribution. Therefore, we hypothesize that the substrate is encoded by the nuclear genome and imported into the mitochondrion to undergo a specific processing step. This may include addition of a prosthetic group – a scenario modelled on the process of cytochrome *c* or Rieske protein maturation[41,42]. Interestingly, the proteins Gcp6 and Gcp12, each exhibiting an array of absolutely conserved cysteine and histidine residues (Supplementary Fig. 11), are good candidates for proteins that are loaded with a specific prosthetic group, so any of them may well be the sought-after miT2SS substrate. Some of the other Gcp proteins may then represent components of the hypothetical machinery responsible for the substrate modification. The putative functionalization step may occur either in the mitochondrial matrix or in the intermembrane space (IMS), but we note that the former localization would necessitate a mechanism of protein translocation across the mitochondrial IM in the direction from the matrix to the IMS, which has not been demonstrated yet. Regardless, the modified protein would eventually be translocated across the mitochondrial OM by the T2SS system to the cytoplasm.

However, this may not be the end of the journey, since there are hints of a link between the miT2SS-associated pathway and peroxisomes. First, three Gcp proteins, namely Gcp1 to Gcp3, are specifically related to Pex7, a protein mediating import of peroxisomal proteins characterized by the peroxisomal targeting signal 2 (PTS2)[43]. Second, some of the Gcp proteins (Gcp1, Gcp2, Gcp13) have at the C-terminus a predicted PTS1 signal (at least in some species; Supplementary Table 3). Third, several Gcp proteins (Gcp2, Gcp9, Gcp10, and Gcp11) and GspEL1 were assigned to the putative peroxisomal proteome in our proteomic analysis (Supplementary Table 2B). We note the discrepancy between the PTS1 signal predictions and the actual set of experimentally defined peroxisomal proteins, which might be due to an incomplete separation of peroxisome and mitochondria by our purification procedure, but may also reflect protein shuttling between the two organelles.

We thus hypothesize that upon its export from the mitochondrion, the miT2SS substrate is eventually delivered to the peroxisome. This is possibly mediated by the Gcp1/2/3 trio, but other Gcp proteins might participate as well. One such protein might be the ubiquitin-related protein Gcp4. Ubiquitination and deubiquitination of several components of the peroxisome protein import machinery is a critical part of the import mechanism[43] and Gcp4 could serve as an analogous peptide modifier in the hypothetical novel peroxisome import pathway functionally linked to the miT2SS.

Altogether, our data suggest the existence of a novel elaborate functional pathway combining components of bacterial origin with newly evolved eukaryote-specific proteins. The modern phylogenetic distribution of the pathway is sparse, but our current understanding of eukaryote phylogeny suggests that it was ancestrally present in eukaryotes and for some reason dispensed with, multiple times during evolution. Although we could not define a specific bacterial group as the actual source of the eukaryotic Gsp genes, it is tempting to speculate that the T2SS was introduced into eukaryotes by the bacterial progenitor of mitochondria and that it was involved in delivering specific proteins from the endosymbiont into the host cell, as is known in the case of current intracellular bacteria[36]. Elucidating the actual role of this communication route in establishing the endosymbiont as a fully integrated organelle requires understanding the cellular function of the modern miT2SS-associated pathways, which is a challenge for future research.

Methods

Sequence data and homology searches

Homologues of relevant genes/proteins were searched in sequence databases accessible via the National Center for Biotechnology Information BLAST server (https://blast.ncbi.nlm.nih.gov/Blast.cgi), including the nucleotide and protein non-redundant (nr) databases, whole-genome shotgun assemblies (WGAs), expressed sequence tags (ESTs), and transcriptome shotgun assemblies (TSAs). Additional public databases searched included the data provided by the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP[44]) comprising TSAs from hundreds of diverse protists (https://www.imicrobe.us/#/projects/104), the OneKP project[45] (https://sites.google.com/a/ualberta.ca/onekp/) comprising TSAs from hundreds of plants and algae, and individual WGAs and TSAs deposited at various on-line repositories (Supplementary Table 5). Non-public sequence data analysed included genome and/or transcriptome assemblies from several heteroloboseans, jakobids and malawimonads generated in our laboratories using standard sequencing technologies (454 and or Illumina) and sequence assembly programs (Supplementary Table 5). Details on the sequencing and assembly and full analyses of these genomes and transcriptomes will be published elsewhere.

Homology searches were done using BLAST[46] (blastp or tblastn, depending on the database queried) and HMMER[47] using profile HMMs built from sequence alignments of proteins of interest. Hits were evaluated by BLAST (blastp or blastx) searches against the nr protein dataset at NCBI to distinguish orthologues of Gsp and Gcp proteins from paralogous proteins or non-specific matches. This was facilitated by a high degree of conservation of individual eukaryotic Gsp/Gcp proteins among different species (see also Supplementary Figs 4 and 10-12) and in most cases by the lack of other close homologues in eukaryotic genomes (the exceptions being members of broader protein families, including the ATPase GspE, the WD40 superfamily proteins Gcp1 to Gcp3, and the ubiquitin related protein Gcp4). All identified eukaryotic Gsp and Gcp sequences were carefully manually curated to ensure maximal accuracy and completeness of the data, which included correction of existing gene models, extension of truncated sequences by manual analysis of raw sequencing reads, and correction of assembly errors (for details see Supplementary Methods). All newly predicted or curated Gsp and Gcp sequences are provided in Supplementary Tables 1 and 3, respectively; additional Gsp and Gcp sequences from non-target species are listed in Supplementary Table 4.

Phylogenetic profiling

In order to identify genes with the same phylogenetic distribution as the eukaryotic homologues of the four core T2SS components, we carried out two partially overlapping analyses based on defining groups of putative orthologous genes in select Gsp-positive species and phylogenetically diverse Gsp-negative eukaryotic species. The list of taxa included is provided in Supplementary Table 6. The first analysis was based on 18 species, including three Gsp-positive ones (*N. gruberi*, *A. godoyi* and *M. jakobiformis*), for the second analysis the set was expanded by adding one additional Gsp-positive species (*G. okellyi*) and one Gsp-negative species (*Monocercomonoides* sp. PA203). Briefly, the protein sequences of a given species were compared to those of all other species using blastp followed by fast phylogenetic analyses and orthologous relationships between proteins were then inferred from this set of phylogenetic trees using a reference-species-tree-independent approach. This procedure was repeated for each species and all resulting sets of orthologous relationships, also known as phylomes[48], were combined in a dense network of orthologous relationships. This network was finally trimmed in several successive steps to remove week or spurious connections and to account for (genuine or artificial) gene fusions, with the first analysis being less restrictive than the second. Details of this pipeline are provided in Supplementary Methods.

For each of the two analyses, the final set of defined groups of orthologs (orthogroups) was parsed to identify those comprising genes from at least two Gsp-positive species yet lacking genes from any Gsp-negative species. The orthogroups passing this criterion were further analysed manually by blastp and tblastn searches against various public and private sequence repositories (see the section "Sequence data and homology searches") to exclude those orthogroups with obvious orthologs in Gsp-negative species. *Percolomonas* spp. exhibiting only GspD and jakobids represented by incomplete EST surveys (these species are likely to possess the miT2SS system) were not considered as Gsp-negative. The orthogroups that remained were then evaluated for their conservation in Gsp-positive species and those that proved to have a representative in all these species (*N. gruberi*, *N. fowleri*, *N. damariscottae*, *P. kirbyi*, *A. godoyi*, *R. americana*, *M. jakobiformis*, *G. okellyi*) were considered as bona fide Gcp (Gsp-co-occurring protein) candidates. It is of note that some of these proteins are short and were missed by the automated annotation of some of the genomes, so using relaxed criteria for the initial consideration of candidate orthogroups (i.e. allowing for their absence from some of the Gsp-positive species) proved critical for decreasing the number of false-negative identifications.

Sequence analyses and phylogenetic inference

The presence of N-terminal mitochondrial transit peptides and peroxisomal targeting signal 1 (PTS1) in the Gsp and Gcp proteins was evaluated using MitoFates[49] (http://mitf.cbrc.jp/MitoFates/cgi-bin/top.cgi) and PTS1 predictor[50] (http://mendel.imp.ac.at/pts1/), respectively. Transmembrane domains were predicted using TMHMM[51] (http://www.cbs.dtu.dk/services/TMHMM/). Homology of Gsp and Gcp protein families to other proteins was evaluated by searches against Pfam v. 31 (ref.[52]; http://pfam.xfam.org/) and Superfamily 1.75 database[53] (http://supfam.org/SUPERFAMILY/index.html) and by using HHpred[33] (https://toolkit.tuebingen.mpg.de/#/tools/hhpred) and the Phyre2 server[34] (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index). The relative position of the Gcp4 family among Ubiquitin-like proteins was analysed by a cluster analysis using CLANS[54] (https://www.eb.tuebingen.mpg.de/protein-evolution/software/clans/); for the analysis the Gcp4 family was combined with all 59 defined families included in the clan Ubiquitin (CL0072) as defined in the Pfam database (each family was represented by sequences from the respective seed alignments stored in the Pfam database). For further details on the procedure see the legend of Supplementary Fig. 9A. Multiple sequence alignments used for presentation of the conservation and specific sequence features of Gsp and Gcp families were built using MUSCLE[55] and shaded using BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html)

In order to obtain datasets for the phylogenetic analyses of eukaryotic GspD to GspG proteins, the protein sequences were aligned using MAFFT[56] and trimmed manually. Profile hidden Markov models (HMMs) built on the basis of the respective alignments were used as queries to search the UniProt database using HMMER. All recovered sequences were assigned to components of the T4P superfamily machineries using HMMER searches against a collection of profile HMMs reported by Abby et al. (ref.[19]). For each GspD to GspG proteins, a series of alignments was built by progressively expanding the sequence set by including more distant homologues (as retrieved by the HMMER searches). Specifically, the different sets of sequences were defined by the HMMER score based on the formula $\text{score}_{\text{cutoff}} = c{*}\text{score}_{\text{best prokaryotic hit}}$, with the coefficient c decreasing from 0.99 to 0.70 incrementally by 0.01. The sequences were then aligned using MAFFT, trimmed with BMGE[57] and the phylogenies were computed with IQ-TREE[58] using the best-fit model (selected by the program from standard protein evolution models and the mixture models[59] offered). The topologies were tested using 10,000 ultra-fast bootstraps. The resulting trees were systematically analyzed for support of the monophyly of eukaryotic sequences and for the taxonomic assignment of the parental prokaryotic node of the eukaryotic subtree. The assignment was done using the following procedure. The tree was artificially rooted between the eukaryotic and prokaryotic sequences. From sub-leaf nodes to the deepest node of the prokaryotic subtree, the taxonomic affiliation of each node was assigned by proportionally considering the known or inferred taxonomic affiliations (at the phylum or class level) of the descending nodes. See the legend to Supplementary Fig. 3 for further details.

The phylogenetic analysis of the WD40 superfamily including Gcp1 to Gcp3 proteins was performed as follows. The starting dataset was prepared by a combination of two different approaches: 1) each identified sequence of Gcp1 to Gcp3 proteins was used as a query in a blastp search against the non-redundant (nr) NCBI protein database and the 500 best hits for each sequence were kept; 2) protein sequences of each the Gcp1 to Gcp3 family were aligned using MAFFT and the multiple alignment was used as a query in a HMMER3 search (https://toolkit.tuebingen.mpg.de/#/tools/hmmer) against the UniProt database. Best hits (E-value cutoff 1e-50) from all three searches were pooled and de-duplicated, and the resulting sequence set (including Gcp1 to Gcp3 sequences) was aligned using MAFFT and trimmed manually to remove poorly conserved regions. Because WD40 proteins are extremely diversified, sequences that were too divergent were removed from the starting dataset during three subsequent rounds of sequence removal, based on a manual inspection of the alignment and phylogenetic trees computed by IQ-TREE (using the best-fit model as described above). The final dataset was enriched by adding PEX7 and WDR24 orthologues from eukaryotes known to possess miT2SS components. The final phylogenetic tree was computed using IQ-TEE as described in the legend to Supplementary Fig. 8. IQ-TREE was used also for inferring trees of the heterolobosean 18S rRNA gene sequences (Supplementary Fig. 1), ubiquitin-related proteins (Supplementary Fig. 9B) and the Gcp16 family (Supplementary Fig. 13); details on the analyses are provided in legends to the respective figures.

Homology modelling

The PDB database was searched by the SWISS-MODEL server[60] for structural homologues of *Go*GspD and *Go*GspG1. *V. cholerae* GspD[35] (PDB entry 5Wq9) and *K. oxytoca* PulG[20] pseudopilus (PDB entry 5wda) were selected as the top matches, respectively. Models were built based on the target-template alignment using ProMod3[60]. Coordinates that were conserved between the target and the template were copied from the template to the model. Insertions and deletions were remodelled using a fragment library, followed by rebuilding side chains. Finally, the geometry of the resulting model was regularized by using a force field. In the case of loop modelling with ProMod3 fails, an alternative model was built with PROMOD-II[61]. The quaternary structure annotation of the template was used to model the target sequence in its oligomeric form[62].

Cultivation and fractionation of *N. gruberi* and proteomic analysis

Naegleria gruberi str. NEG-M was axenically cultured in M7 medium with PenStrep (100 U/mL of penicillin and 100 μg/mL of streptomycin) at 27°C in vented tissue culture flasks. Mitochondria of *N. gruberi* were isolated in seven independent experiments, which were analyzed individually (see below). Each time ~1x10^9 *N. gruberi* cells were resuspended in 2 mL of SM buffer (250 mM sucrose, 20 mM MOPS, pH 7.4) supplemented with DNase I (40 ug/mL) and Roche cOmplete™ EDTA-free Protease Inhibitor Cocktail and homogenized by eight passages through a 33-gauge hypodermic needle (Sigma Aldrich). The resulting cell homogenate was then cleaned of cellular debris using differential centrifugation and separated by a 2-hr centrifugation in a discontinuous density OptiPrep gradient (10%, 15%, 20%, 30% and 50%) as described previously[63]. Three visually identifiable fractions corresponding to 10-15% (OPT-1015), 15-20% (OPT-1520) and 20-30% (OPT-2023) OptiPrep densities were collected (each in five biological replicates) and washed with SM buffer.

Proteins extracted from these samples were then digested with trypsin and peptides were separated by nanoflow liquid chromatography and analyzed by tandem mass spectrometry (nLC-MS2) on a Thermo Orbitrap Fusion (q-OT-IT) instrument as described elsewhere[64]. The quantification of mass spectrometry data in the MaxQuant software[65] provided normalized intensity values for 4,198 proteins in all samples and all three fractions. These values were further processed using the Perseus software[66]. Data were filtered and only proteins with at least two valid values in one fraction were kept. Imputation of missing values, which represent low-abundance measurements, was performed with random distribution around the value of instrument sensitivity using default settings of Perseus software[66].

The data were analyzed by principle component analysis (PCA). The first two loadings of the PCA were used to plot a two-dimensional graph. Based on a set of marker proteins (376 mitochondrial and 26 peroxisomal, Supplementary Table 2), clusters of proteins co-fractionating with mitochondria and peroxisomes were defined and the proteins within the clusters were further analyzed. This workflow was set up on the basis of the LOPIT protocol[67]. As a result, out of the 4,198 proteins detected, 946 putative mitochondrial and 78 putative peroxisomal proteins were defined. All proteins were subjected to *in silico* predictions concerning their function (BLAST, HHpred[33]) and subcellular localization (Psort II, https://psort.hgc.jp/form2.html; TargetP, http://www.cbs.dtu.dk/services/TargetP/; MultiLoc2, https://abi.inf.uni-tuebingen.de/Services/MultiLoc2). The mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium via the PRIDE[68] partner repository with the dataset identifier PXD007764.

Fluorescence *in situ* hybridization (FISH)

The PCR products of the *Ng*GspE and *Ng*GspF genes were labelled by alkali-stable digoxigenin-11-dUTP (Roche) using DecaLabel DNA Labeling Kit (Thermo Scientific). Labelled probes were purified on columns of QIAquick Gel Extraction Kit (Qiagen, 28704) in a final volume of 50 µL. Labelling efficiencies were tested by dot blotting with anti-digoxigenin alkaline phosphatase conjugate and CSPD chemiluminescence substrate for alkaline phosphatase from DIG High Prime DNA Labeling and Detection Starter Kit II (Roche) according to the manufacturer`s protocol. FISH with digoxigenin-labelled probes was performed essentially according to the procedure described in Zubacova at al. (ref.[69]) with some modifications. *N. gruberi* cells were pelleted by centrifugation for 10 min at 2,000 x *g* at 4°C. Cells were placed in hypotonic solution, fixed twice with a freshly prepared mixture of methanol and acetic acid (3:1) and dropped on superfrost microscope slides (ThermoScientific). Preparations for hybridizations were treated with RNase A, 20 µg in 100 µL 2 x SSC, for 1 hr at 37°C, washed twice in 2 x SSC for 5 min, dehydrated in a methanol series and air-dried. Slides were treated with 50% acetic acid followed by pepsin treatment and postfixation with 2% paraformaldehyde. Endogenous peroxidase activity of the cell remnants (undesirable for tyramide signal amplification) was inactivated by incubation in 1% hydrogen peroxide, followed by dehydration in a graded methanol series. All slides were denatured together with 2 µL (25 ng) of the probe in 50 µL of hybridization mixture containing 50% deionised formamide (Sigma) in 2 x SSC for 5 min at 82°C. Hybridizations were carried out overnight. Slides were incubated with tyramide reagent for 7 min. Preparations were counterstained with DAPI in VectaShield and observed under an Olympus IX81 microscope equipped with a Hamamatsu Orca-AG digital camera using the Cell^R imaging software.

Heterologous gene expression, preparation of antibodies, and immunodetection of Gsp proteins

The selected Gsp genes from *G. okellyi* and *N. gruberi* were amplified from commercially synthesized templates (Genscript) (for primers used for PCR amplification of the coding sequences see Supplementary Table 7) and cloned into the pUG35 vector. The constructs were introduced into *S. cerevisiae* strain YPH499 by lithium acetate/PEG method. The positive colonies grown on SD-URA plates were incubated with MitoTracker Red CMXRos (Thermo Fisher Scientific) and observed for GFP and MitoTracker fluorescence (using the same equipment as used for FISH, see above). For bacterial protein expression, *N. gruberi* GspD, GspE, GspF and GspG genes were amplified from commercially synthesized templates and cloned into pET42b vector (for primers used for PCR amplification of the coding sequences, see Supplementary Table 6). The constructs were introduced into chemically-competent *E. coli* strain BL21(DE3) and their expression induced by 1 mM IPTG. The recombinant proteins were purified under denaturing conditions on Ni-NTA agarose (Qiagen). The purified proteins were used for rat immunization in an in-house animal facility at Charles University.

The sera obtained were used for immunodetection of Gsp proteins in *N. gruberi* cells. Briefly, cells were fixed for 5 min in methanol (-20°C) and permeabilized for 5 min by acetone (-20°C). The slides were incubated in blocking buffer (BB) (PBS supplemented by 0.25% BSA, 0.05% TWEEN® 20 and 0.25% gelatin) for 1 hr at room temperature. The slides were incubated overnight at 4°C with primary antibodies diluted in BB and washed three times in PBS for 10 min. Slides were then incubated for 1 hr with an anti-rat antibody conjugated with Alexa488 (Thermo Fisher Scientific) diluted in. After washing three times for 10 min in PBS, the slides were mounted in VectaShield DAPI solution and observed as above. For mitochondrial labelling, the cells were incubated with MitoTracker Red CMXRos for 30 min before fixation.

In vitro protein translation

The *Go*GspD gene was amplified from the commercially synthesized template (for primers used for PCR amplification of the coding sequences, see Supplementary Table 6) and cloned into pDHFR vector provided in the PURExpress *In Vitro* Protein Synthesis Kit (NEB). The 25 µl translation reaction contained 10 µL of solution A, 7.5 µL of solution B, 250 ng of pDHFR plasmid carrying *Go*GspD gene, 1 µL of an RNase inhibitor (RNAsin, Promega), radioactively labelled ^{35}S-methionine, and 50 µg of lecithin liposomes. The liposomes were prepared from a stock solution of soybean L-α-lecithin in chloroform by evaporating the chloroform under a nitrogen flow, resuspending the lipid film in dH_20, and subsequent sonication in a waterbath sonicator. The translation reaction was incubated for 2 hr at 37°C and then centrifuged for 45 min at 13,000 x *g*. The pellet was resuspended in 50 mM sodium phosphate buffer (pH = 8) with 2 M urea, centrifuged, and then washed in clear 50 mM sodium phosphate buffer. The output was analyzed by Blue Native PAGE using 2% digitonin and NativePAGE Novex 4-16% Bis-Tris Protein Gel (Thermo Fisher Scientific).

Testing protein interactions using two-hybrid systems

Bacterial two-hybrid system (B2H) analysis was performed as described in ref.[70]. Gsp genes were amplified for commercially synthesized DNA and cloned into pKT25 and pUT18c plasmids. *E. coli* strain DHT1 competent cells were co-transformed with two plasmids with different combinations of Gsp genes. Co-transformants were selected on LB plates with ampicillin (100 μg/mL) and kanamycin (25 μg/mL). Colonies were grown at 30°C for 48 to 96 hr. From each plate three colonies were picked, transferred to 1 mL of LB medium with ampicillin and kanamycin, and grown overnight at 30°C with shaking. Next day precultures (0.25 mL) were inoculated to 5 mL of LB medium with ampicillin, kanamycin and 1 mM IPTG. Cultures were grown with shaking at 30°C to OD_{600} of about 1-1.5. Bacteria (0.5 mL) were mixed with 0.5 mL of Z buffer and subjected to the β-galactosidase assay[71].

The yeast two-hybrid system (Y2H) was employed as described in ref.[72]. Cells of *S. cerevisiae* strain AH109 were co-transformed with two plasmids (pGADT7, pGBKT7) with different combinations of Gsp genes. Co-transformants were selected on double-dropout SD-Leu/-Trp and triple-dropout SD-Leu/-Trp/-His plates. The colonies were grown for a few days. Positive colonies from the triple dropout were grown overnight at 30°C with shaking and then the serial dilution test was performed on double- and triple-dropout plates.

Data availability

All newly reported sequences of Gsp and Gcp proteins are provided in Supplementary Table 1 and were deposited at GenBank with accession numbers ######. Other relevant data (e.g. multiple sequence alignments used for phylogenetic analyses) are available from the authors upon request.

References

1. Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. The origin and diversification of mitochondria. *Curr. Biol.* **27,** R1177–R1192 (2017).
2. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557,** 101–105 (2018).
3. Leger, M. M. *et al.* An ancestral bacterial division system is widespread in eukaryotic mitochondria. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 10239–46 (2015).
4. Beech, P. L. Mitochondrial FtsZ in a chromophyte alga. *Science (80-.).* **287,** 1276–1279 (2000).
5. Natale, P., Brüser, T. & Driessen, A. J. M. Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membrane—Distinct translocases and mechanisms. *Biochim. Biophys. Acta - Biomembr.* **1778,** 1735–1756 (2008).
6. Costa, T. R. D. *et al.* Secretion systems in Gram-negative bacteria: structural and mechanistic insights. *Nat. Rev. Microbiol.* **13,** 343–359 (2015).
7. Dolezal, P., Likic, V., Tachezy, J. & Lithgow, T. Evolution of the molecular machines for protein import into mitochondria. *Science* **313,** 314–8 (2006).
8. Palmer, T. & Berks, B. C. The twin-arginine translocation (Tat) protein export pathway. *Nat. Rev. Microbiol.* **10,** 483–96 (2012).
9. Lang, B. F. *et al.* An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* **387,** 493–7 (1997).
10. Burger, G., Gray, M. W., Forget, L. & Lang, B. F. Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. *Genome Biol.*

Evol. **5,** 418–38 (2013).

11. Tong, J. *et al.* Ancestral and derived protein import pathways in the mitochondrion of *Reclinomonas america*. *Mol. Biol. Evol.* **28,** 1581–91 (2011).
12. Korotkov, K. V., Sandkvist, M. & Hol, W. G. J. The type II secretion system: biogenesis, molecular architecture and mechanism. *Nat. Rev. Microbiol.* **10,** 336–51 (2012).
13. Thomassin, J.-L., Santos Moreno, J., Guilvout, I., Tran Van Nhieu, G. & Francetic, O. The trans-envelope architecture and function of the type 2 secretion system: new insights raising new questions. *Mol. Microbiol.* **105,** 211–226 (2017).
14. Berry, J.-L. & Pelicic, V. Exceptionally widespread nanomachines composed of type IV pilins: the prokaryotic Swiss Army knives. *FEMS Microbiol. Rev.* **39,** 134–54 (2015).
15. Nivaskumar, M. & Francetic, O. Type II secretion system: a magic beanstalk or a protein escalator. *Biochim. Biophys. Acta* **1843,** 1568–77 (2014).
16. Peabody, C. R. *et al.* Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella. *Microbiology* **149,** 3051–72 (2003).
17. d'Enfert, C., Reyss, I., Wandersman, C. & Pugsley, A. P. Protein secretion by gram-negative bacteria. Characterization of two membrane proteins required for pullulanase secretion by *Escherichia coli* K-12. *J. Biol. Chem.* **264,** 17462–8 (1989).
18. Wang, X. *et al.* Cysteine scanning mutagenesis and disulfide mapping analysis of arrangement of GspC and GspD protomers within the type 2 secretion system. *J. Biol. Chem.* **287,** 19082–93 (2012).
19. Abby, S. S. *et al.* Identification of protein secretion systems in bacterial genomes. *Sci. Rep.* **6,** 23080 (2016).
20. López-Castilla, A. *et al.* Structure of the calcium-dependent type 2 secretion pseudopilus. *Nat. Microbiol.* **2,** 1686–1695 (2017).
21. Adl, S. M. *et al.* The revised classification of eukaryotes. *J. Eukaryot. Microbiol.* **59,** 429–514 (2012).
22. Derelle, R. *et al.* Bacterial proteins pinpoint a single eukaryotic root. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E693-9 (2015).
23. Karnkowska, A. *et al.* A eukaryote without a mitochondrial organelle. *Curr. Biol.* **26,** 1274–84 (2016).
24. Heiss, A. A. *et al.* Combined morphological and phylogenomic re-examination of malawimonads, a critical taxon for inferring the evolutionary history of eukaryotes. *R. Soc. Open Sci.* **5,** 171707 (2018).
25. Brown, M. W. *et al.* Phylogenomics places orphan protistan lineages in a novel eukaryotic super-group. *Genome Biol. Evol.* **10,** 427–433 (2018).
26. Nouwen, N. *et al.* Secretin PulD: Association with pilot PulS, structure, and ion-conducting channel formation. *Proc. Natl. Acad. Sci.* **96,** 8173–8177 (1999).
27. Hardie, K. R., Lory, S. & Pugsley, A. P. Insertion of an outer membrane protein in *Escherichia coli* requires a chaperone-like protein. *EMBO J.* **15,** 978–88 (1996).
28. Korotkov, K. V, Pardon, E., Steyaert, J. & Hol, W. G. J. Crystal structure of the N-

terminal domain of the secretin GspD from ETEC determined with the assistance of a nanobody. *Structure* **17,** 255–265 (2009).

29. Yin, M., Yan, Z. & Li, X. Structural insight into the assembly of the Type II secretion system pilotin-Secretin complex from enterotoxigenic *Escherichia coli*. *Nat Microbiol* (2018). doi:10.2210/PDB5ZDH/PDB

30. Guilvout, I. *et al.* In vitro multimerization and membrane insertion of bacterial outer membrane secretin PulD. *J. Mol. Biol.* **382,** 13–23 (2008).

31. Nivaskumar, M. *et al.* Pseudopilin residue E5 is essential for recruitment by the type 2 secretion system assembly platform. *Mol. Microbiol.* **101,** 924–41 (2016).

32. Korotkov, K. V. *et al.* Structural and functional studies on the interaction of GspC and GspD in the type II secretion system. *PLoS Pathog.* **7,** e1002228 (2011).

33. Alva, V., Nam, S.-Z., Söding, J. & Lupas, A. N. The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucleic Acids Res.* **44,** W410–W415 (2016).

34. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10,** 845–858 (2015).

35. Yan, Z., Yin, M., Xu, D., Zhu, Y. & Li, X. Structural insights into the secretin translocation channel in the type II secretion system. *Nat. Struct. Mol. Biol.* **24,** 177–183 (2017).

36. Nguyen, B. D. & Valdivia, R. H. Virulence determinants in the obligate intracellular pathogen *Chlamydia trachomatis* revealed by forward genetic approaches. *Proc. Natl. Acad. Sci.* **109,** 1263–1268 (2012).

37. Tabach, Y. *et al.* Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. *Nature* **493,** 694–698 (2012).

38. Nevers, Y. *et al.* Insights into ciliary genes and evolution from multi-level phylogenetic profiling. *Mol. Biol. Evol.* **34,** 2016–2034 (2017).

39. Okuno, D., Iino, R. & Noji, H. Rotation and structure of FoF1-ATP synthase. *J. Biochem.* **149,** 655–664 (2011).

40. Tomko, R. J. & Hochstrasser, M. Molecular architecture and assembly of the eukaryotic proteasome. *Annu. Rev. Biochem.* **82,** 415–445 (2013).

41. Babbitt, S. E., Sutherland, M. C., San Francisco, B., Mendez, D. L. & Kranz, R. G. Mitochondrial cytochrome c biogenesis: no longer an enigma. *Trends Biochem. Sci.* **40,** 446–55 (2015).

42. Hartl, F. U., Schmidt, B., Wachter, E., Weiss, H. & Neupert, W. Transport into mitochondria and intramitochondrial sorting of the Fe/S protein of ubiquinol-cytochrome c reductase. *Cell* **47,** 939–51 (1986).

43. Francisco, T. *et al.* Protein transport into peroxisomes: Knowns and unknowns. *BioEssays* **39,** 1700047 (2017).

44. Keeling, P. J. *et al.* The marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.* **12,** e1001889 (2014).

45. Matasci, N. *et al.* Data access for the 1,000 Plants (1KP) project. *Gigascience* **3,**

17 (2014).

46. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–402 (1997).

47. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39,** W29-37 (2011).

48. Huerta-Cepas, J., Dopazo, H., Dopazo, J. & Gabaldón, T. The human phylome. *Genome Biol.* **8,** R109 (2007).

49. Fukasawa, Y. *et al.* MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites. *Mol. Cell. Proteomics* **14,** 1113–26 (2015).

50. Neuberger, G., Maurer-Stroh, S., Eisenhaber, B., Hartig, A. & Eisenhaber, F. Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *J. Mol. Biol.* **328,** 581–92 (2003).

51. Käll, L., Krogh, A. & Sonnhammer, E. L. L. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res.* **35,** W429-32 (2007).

52. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44,** D279-85 (2016).

53. de Lima Morais, D. A. *et al.* SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic Acids Res.* **39,** D427-34 (2011).

54. Frickey, T. & Lupas, A. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20,** 3702–3704 (2004).

55. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–7 (2004).

56. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).

57. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10,** 210 (2010).

58. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32,** 268–74 (2015).

59. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21,** 1095–109 (2004).

60. Bienert, S. *et al.* The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res.* **45,** D313–D319 (2017).

61. Guex, N., Peitsch, M. C. & Schwede, T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis* **30,** S162–S173 (2009).

62. Biasini, M. *et al.* SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **42,** W252-8 (2014).

63. Jedelský, P. L. *et al.* The minimal proteome in the reduced mitochondrion of the parasitic protist *Giardia intestinalis*. *PLoS One* **6,** e17285 (2011).

64. Černá, M., Kuntová, B., Talacko, P., Stopková, R. & Stopka, P. Differential

regulation of vaginal lipocalins (OBP, MUP) during the estrous cycle of the house mouse. *Sci. Rep.* **7,** 11674 (2017).

65. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13,** 2513–2526 (2014).
66. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13,** 731–740 (2016).
67. Dunkley, T. P. J., Watson, R., Griffin, J. L., Dupree, P. & Lilley, K. S. Localization of organelle proteins by isotope tagging (LOPIT). *Mol. Cell. Proteomics* **3,** 1128–1134 (2004).
68. Vizcaíno, J. A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44,** 11033–11033 (2016).
69. Zubáčová, Z., Krylov, V. & Tachezy, J. Fluorescence in situ hybridization (FISH) mapping of single copy genes on *Trichomonas vaginalis* chromosomes. *Mol. Biochem. Parasitol.* **176,** 135–137 (2011).
70. Battesti, A. & Bouveret, E. The bacterial two-hybrid system based on adenylate cyclase reconstitution in *Escherichia coli*. *Methods* **58,** 325–334 (2012).
71. Miller, J. H. *Experiments in molecular genetics*. (Cold Spring Harbor Laboratory, 1972).
72. Fields, S. & Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340,** 245–246 (1989).

Acknowledgements

We would like to thank Michelle Leger and Alastair Simpson for granting us an access, prior to publication, to their transcriptomic data from *A. godoyi* and *G. okellyi*, respectively, which were instrumental in annotating the Gsp and Gcp genes in our genome assemblies. This work was supported by Czech Science Foundation grants 13-29423S to PD and 18-18699S to ME and the KONTAKT II grant LH15253 provided by Ministry of Education, Youth and Sports of CR (MEYS) to PD. This work was also supported by MEYS within the National Sustainability Program II (Project BIOCEV-FAR, LQ1604) the project BIOCEV (CZ.1.05/1.1.00/02.0109), and the project "Centre for research of pathogenicity and virulence of parasites" (No. CZ.02.1.01/0.0/0.0/16_019/0000759) funded by European Regional Development Fund (ERDF) and MEYS. The work in OF laboratory was funded by the ANR-14-CE09-0004 grant. JP was supported by a grant from the Gordon and Betty Moore Foundation to ADT. This work was supported by The Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project „IT4Innovations National Supercomputing Center – LM2015070".

Author information

Affiliations

Department of Parasitology, Faculty of Science, Charles University, BIOCEV, Průmyslová 595, Vestec, 252 42, Czech Republic
L. Horváthová, V. Žárský, A. Krupičková, V. Klápšťová, M. Petrů, Z. Vaitová & P. Doležal

Department of Biology and Ecology, Faculty of Science, University of Ostrava, Chittussiho 10, 710 00 Ostrava, Czech Republic
T. Pánek, V. Klimeš & M. Eliáš

School of Biosciences, University of Birmingham, Edgbaston, B15 2TT, UK
R. Derelle

Laboratory of Molecular & Evolutionary Parasitology, RAPID group, School of Biosciences, University of Kent, Canterbury, CT2 7NZ, UK
J. Pyrih & A. D. Tsaousis

Department of Zoology, Faculty of Science, Charles University, Viničná 7, Prague 2, 128 44, Czech Republic
I. Čepička

Proteomic core facility, Faculty of Science, Charles University , BIOCEV, Průmyslová 595, Vestec, 252 42, Czech Republic
K. Harant

Department of Biochemistry and Molecular Biology and Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, NS B3H 4R2, Canada
M. W. Gray

Institut Pasteur, Biochemistry of Macromolecular Interactions Unit, Department of Structural Biology and Chemistry, CNRS UMR3528, 75015 Paris, France
I. Guilvout & O. Francetic

Robert Cedergren Centre for Bioinformatics and Genomics, Département de Biochimie, Université de Montréal, Montreal, QC, Canada H3T 1J4
B. F. Lang

Institute of Molecular Genetics, Czech Academy of Sciences, 142 20 Prague 4, Czech Republic
Č. Vlček

Author contributions

L.H. planned and carried out the experiments, V.Ž. conceived the original idea and carried out the bioinformatics analyses, T.P. carried out the genome and bioinformatic analyses, R.D. designed and carried out comparative genomic analyses, J.P. planned and carried out the experiments on *N. gruberi* mitochondrial proteome and analysed the data, A.M. carried out the experiments, Ve.K. carried out the experiments, Vl.K. participated in genome sequencing and analysis, M.P. planned carried out the experiments, I.Č. participated in genome data acquisition, Z.V. carried out the experiments, K.H. analysed the proteome of *N. gruberi* mitochondria, M.W.G. contributed to the interpretation of the results and manuscript preparation, I.G. designed and planned the experiments, O.F. designed, planned and carried out the experiments and analyzed the data, B.F.L. provided the genome data and analyses and contributed to manuscript preparation, Č.V. participated in genome data acquisition, A.D.T. designed and planned the experiments, M.E. conceived the idea, performed genomic analyses and wrote the manuscript, P.D. conceived the idea, designed and performed experiments and wrote the manuscript.

Competing interests

None declared.

Corresponding author

Correspondence to M. Eliáš or P. Doležal.

Figure Legends

Fig. 1 Some eukaryotes harbour homologues of core components of the bacterial T2SS machinery. (A) Schematic representation of the complete bacterial T2SS; subunits having identified eukaryotic homologues are highlighted in colour. (B) Phylogenetic distribution of eukaryotic homologues of bacterial T2SS subunits (Gsp proteins) and co-occurring proteins (Gcp). Core T2SS components (cyan), eukaryote-specific T2SS components (dark blue), Gcp proteins carrying protein domains found in eukaryotes (magenta), and Gcp proteins without discernible homologues or with homologues only in prokaryotes (orange). Coloured sections indicate proteins found to be present in genome or transcriptome data; white sections, proteins absent from complete genome data; grey sections, proteins absent from transcriptome data. The asterisk indicates the presence of the particular protein in at least two of three species of *Percolomonas* analyzed. The two species names in parentheses have not been yet been formally published. Sequence IDs and additional details on the eukaryotic Gsp and Gcp proteins are provided in Supplementary Table 1. (C) Maximum likelihood (ML) phylogenetic tree of eukaryotic and selected bacterial GspF proteins demonstrating the monophyletic origin of the eukaryotic GspF proteins and their separation from bacterial homologues by a long branch (the tree inferred using IQ-TREE). Branch support (bootstrap / posterior probability values) was assessed by ML ultrafast bootstrapping and is shown only for branches where > 50.

Fig. 2 Eukaryotic T2SS components are localized in mitochondria. (A) *N. gruberi* cells labelled with specific polyclonal antibodies raised against GspD, GspF and

GspG1, and co-stained with MitoTracker red CMX ROS show mitochondrial localization of the proteins; scale bar, 10 μm. (B) *S. cerevisiae* expressing *G. okellyi* T2SS components as C-terminal GFP fusions co-stained with MitoTracker red CMX ROS; scale bar, 10 μm.

Fig. 3 Analysis of the *N. gruberi* mitochondrial proteome. PCA analysis of 4198 proteins identified in the proteomic analysis of *N. gruberi* mitochondria. The cluster of mitochondrial proteins was defined on the basis of 376 mitochondrial markers. The boundaries of the cluster of co-purified peroxisomal proteins were defined by 26 peroxisomal markers.

Fig. 4. Mitochondrial GspD oligomerizes towards the formation of membrane pores. (A) Domain architecture of the canonical bacterial GspD protein and eukaryotic proteins homologous to its different parts. (B) Structural model of *Go*GspD built by ProMod3 on the *Vibrio cholerae* GspD template. Top and side view of a cartoon and a transparent surface representation of the *Go*GspD pentadecamer model is shown in blue. The amphipathic helical loop (AHL), the signature of the secretin family, is highlighted and coloured according to the secondary structure with strands in magenta, helices in cyan and loops in light brown. The C-terminal GpsD residues are highlighted as spheres. The detailed view of the AHL region shows the essential residues V162 and F166 pointing towards the membrane surface. (C) Expression of the mitochondrial *Go*GspD quickly induces cell death in bacteria. (D) Y2H assay shows the self-interaction of the mitochondrial *Go*GspD. (E) *In vitro* translation and assembly of mitochondrial *Go*GspD into a high-molecular-weight complex; lipo – liposomes added, urea – extraction by 2M urea. (F) Y2H assay suggests the interaction of *Ng*GspDN1 with itself and with *Ng*GspD.

Fig. 5 Structure, maturation, and interactions of the mitochondrial GspG. (A) Domain architecture of the bacterial and the mitochondrial pseudopilin GspG. The arrow indicates the processing site of the bacterial GspG during protein maturation. MTS – mitochondria targeting sequence, + – polar anchor, TMD – transmembrane domain. (B) Positive interactions between the mitochondrial GspG protein and other T2SS subunits were determined by the B2H assays. (C) Peptides specific to *Ng*GspG1 retrieved from the proteomic analysis of *N. gruberi* mitochondria. The arrow indicates the position of the processing site of bacterial GspG proteins. (D) Immunodetection of *Ng*GspG1 in *N. gruberi* cellular fractions. The arrow marks the *Ng*GspG1-specific band.

Fig. 6 Proteins with the same phylogenetic profile as the originally identified mitochondrial Gsp homologues. (A) Schematic domain representation of 23 proteins occurring in heteroloboseans, jakobids and malawimonads with the core T2SS subunits but not in other eukaryotes analyzed. Proteins with a functional link to the T2SS suggested by sequence homology are shown in royal blue, proteins representing novel paralogues within broader (super)families are shown in red, and proteins without discernible homologues or with homologues only in prokaryotes are shown in yellow. The presence of conserved protein domains or characteristic

structural motifs is shown if detected in the given protein. Grey block – predicted transmembrane domain (see also Supplementary Fig. 10); “C H C H” – the presence of absolutely conserved cysteine and histidine residues (see also Supplementary Fig. 11) that may mediate binding of a prosthetic group. The length of the rectangles corresponds to the relative size of the proteins. (B) Evolutionary relationships among Gcp1 to Gcp3 proteins and other members of the WD40 superfamily. The schematic phylogenetic tree was drawn on the basis of a ML phylogenetic tree available as Supplementary Fig. 8.

Fig. 7 A hypothetical novel eukaryotic functional pathway including a mitochondrial version of the T2SS (miT2SS) and connecting the mitochondrion with the peroxisome. A nucleus-encoded protein (magenta) is imported via the TOM complex into the mitochondrial inner membrane space, where it is modified by addition of a specific prosthetic group catalysed by certain Gcp proteins. After folding it becomes a substrate of the miT2SS machinery and is exported from the mitochondrion. Finally it is imported into the peroxisome by the action of a dedicated import system including other Gcp proteins. OMM – outer mitochondrial membrane, IMS – intermembrane space, IMM – inner mitochondrial membrane, MM – mitochondrial matrix.

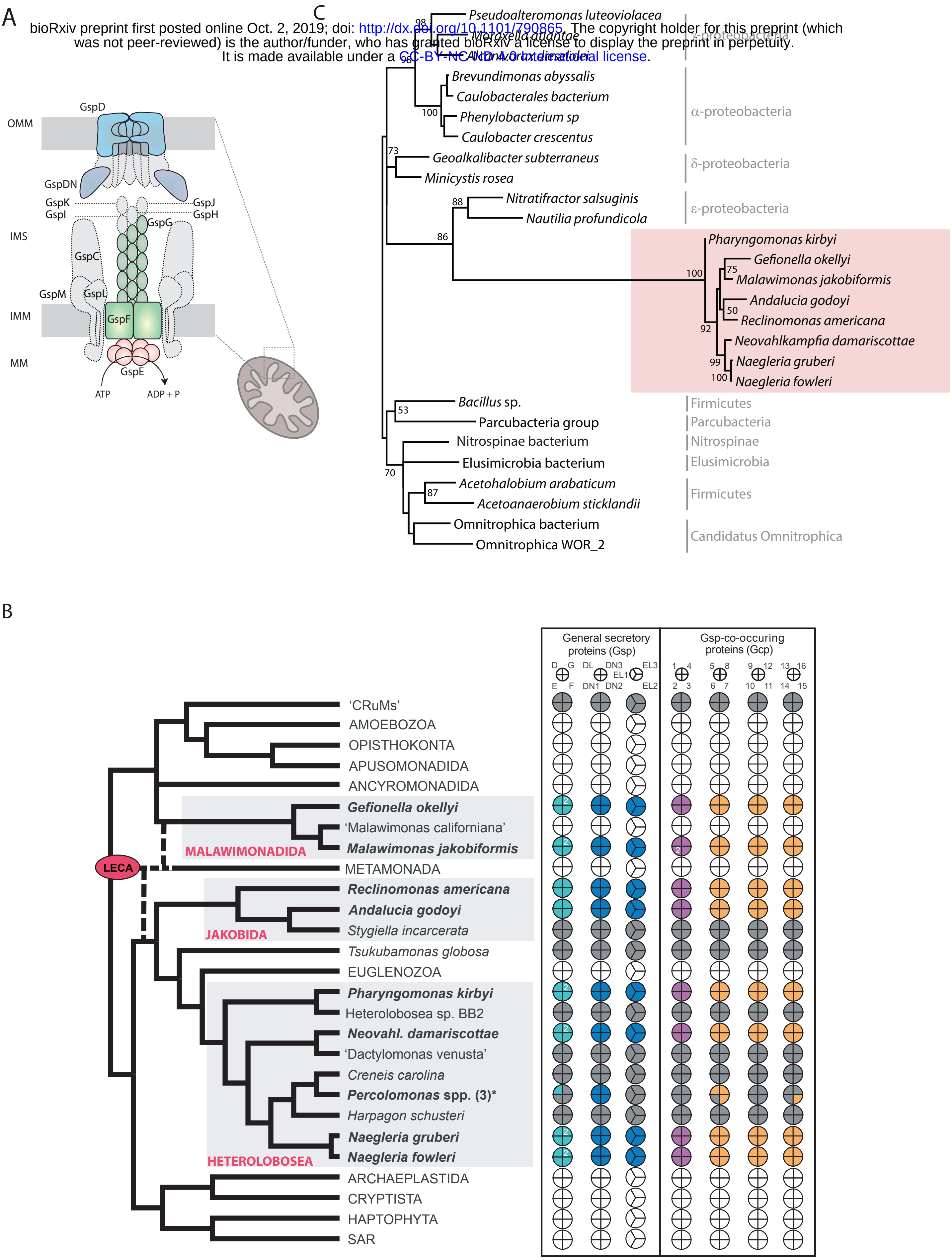
A
GspD
OMM
GspDN
GspK
GspI
GspJ
GspH
GspG
IMS
GspC
GspM
GspL
GspF
IMM
MM
GspE
ATP
ADP + P
C
Pseudoalteromonas luteoviolacea
98
Moraxella atlantae
98
Acinetobacter calcoaceticus
γ-proteobacteria
Brevundimonas abyssalis
Caulobacterales bacterium
100
Phenylobacterium sp
Caulobacter crescentus
α-proteobacteria
73
Geoalkalibacter subterraneus
Minicystis rosea
δ-proteobacteria
88
Nitratifractor salsuginis
Nautilia profundicola
ε-proteobacteria
86
Pharyngomonas kirbyi
Gefionella okellyi
75
100
Malawimonas jakobiformis
Andalucia godoyi
50
Reclinomonas americana
92
Neovahlkampfia damariscottae
99
Naegleria gruberi
100
Naegleria fowleri
Bacillus sp.
53
Parcubacteria group
Firmicutes
Parcubacteria
Nitrospinae bacterium
Nitrospinae
Elusimicrobia bacterium
Elusimicrobia
70
Acetohalobium arabaticum
87
Acetoanaerobium sticklandii
Firmicutes
Omnitrophica bacterium
Omnitrophica WOR_2
Candidatus Omnitrophica
B
General secretory proteins (Gsp)
Gsp-co-occuring proteins (Gcp)
D G E F
DL DN3 DN1 DN2
EL3 EL1 EL2
1 4 2 3
5 8 6 7
9 12 10 11
13 16 14 15
'CRuMs'
AMOEBOZOA
OPISTHOKONTA
APUSOMONADIDA
ANCYROMONADIDA
Gefionella okellyi
'Malawimonas californiana'
MALAWIMONADIDA
Malawimonas jakobiformis
LECA
METAMONADA
Reclinomonas americana
Andalucia godoyi
JAKOBIDA
Stygiella incarcerata
Tsukubamonas globosa
EUGLENOZOA
Pharyngomonas kirbyi
Heterolobosea sp. BB2
Neovahl. damariscottae
'Dactylomonas venusta'
Creneis carolina
Percolomonas spp. (3)*
Harpagon schusteri
Naegleria gruberi
HETEROLOBOSEA
Naegleria fowleri
ARCHAEPLASTIDA
CRYPTISTA
HAPTOPHYTA
SAR
2

Figure 1

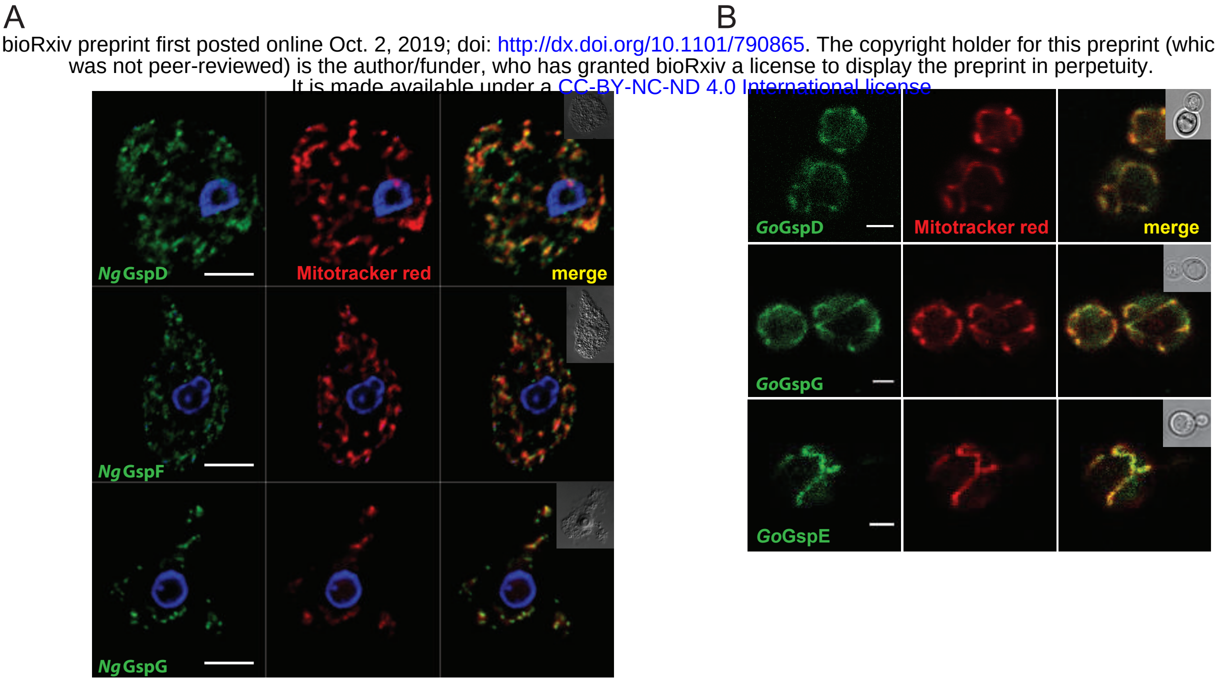
A
B
Ng GspD
Mitotracker red
merge
Ng GspF
Ng GspG
*Go*GspD
Mitotracker red
merge
*Go*GspG
*Go*GspE

Figure 2

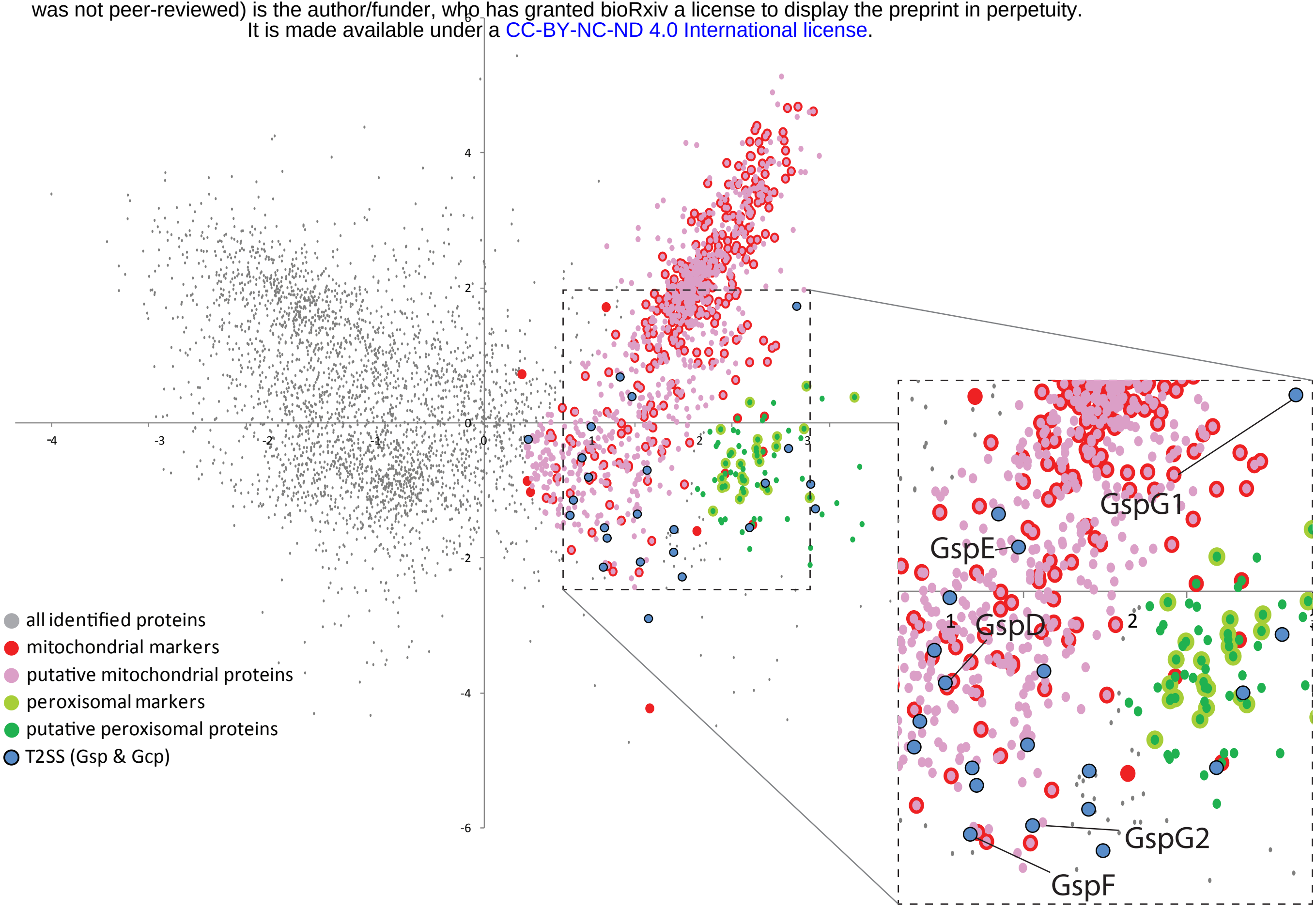
6
4
2
0
-2
-4
-6
-4
-3
-2
-1
0
1
2
3
● all identified proteins
● mitochondrial markers
● putative mitochondrial proteins
● peroxisomal markers
● putative peroxisomal proteins
● T2SS (Gsp & Gcp)
GspG1
GspE
1
GspD
2
GspG2
GspF

Figure 3

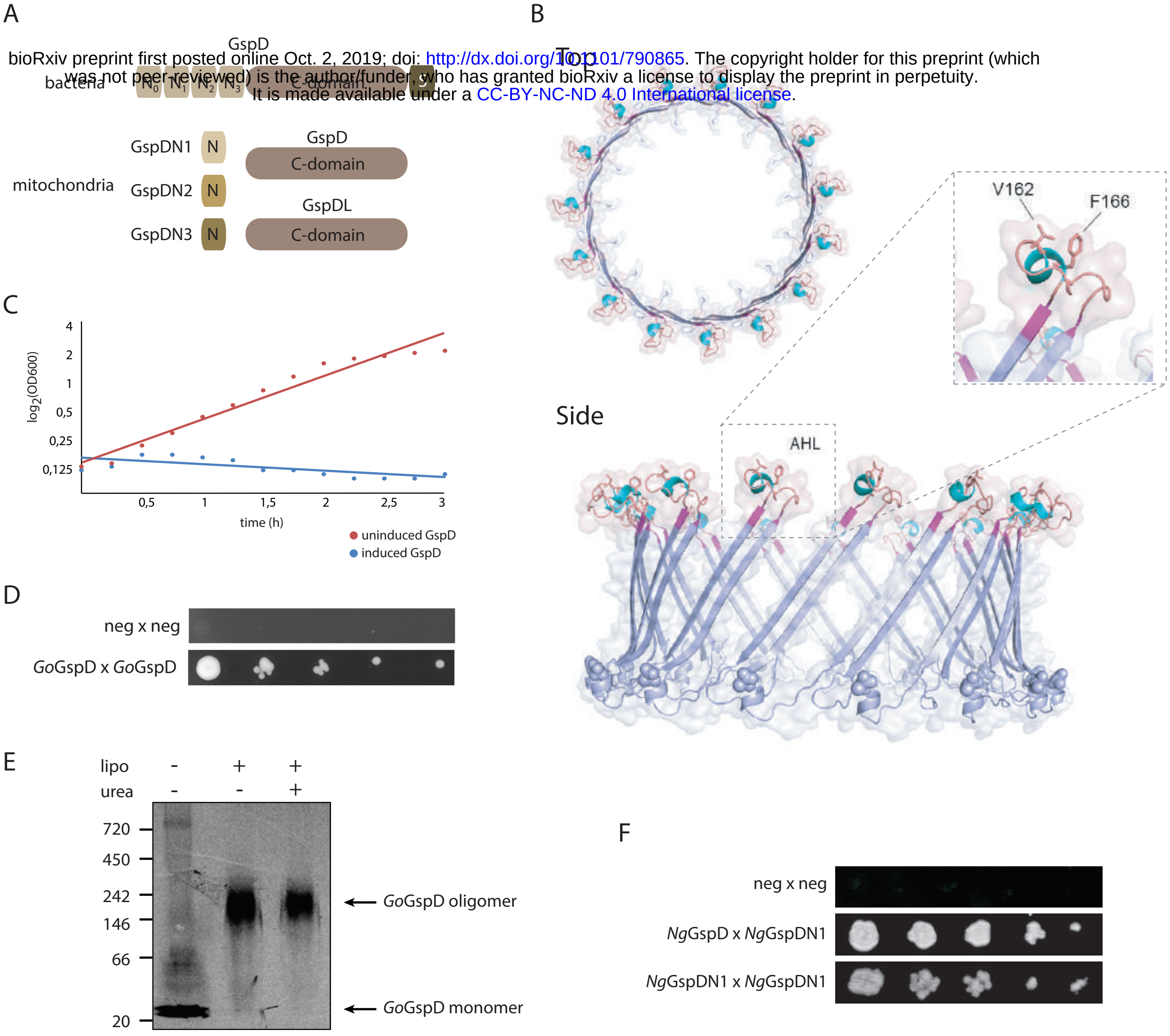
A
GspD
bacteria
N_0
N_1
N_2
N_3
C-domain
S
GspDN1
N
GspD
C-domain
mitochondria
GspDN2
N
GspDL
GspDN3
N
C-domain
B
Top
V162
F166
Side
AHL
C
$\log_2$(OD600)
4
2
1
0,5
0,25
0,125
0,5
1
1,5
2
2,5
3
time (h)
● uninduced GspD
● induced GspD
D
neg x neg
*Go*GspD x *Go*GspD
E
lipo - + +
urea - - +
720
450
242
146
66
20
*Go*GspD oligomer
*Go*GspD monomer
F
neg x neg
*Ng*GspD x *Ng*GspDN1
*Ng*GspDN1 x *Ng*GspDN1

Figure 4

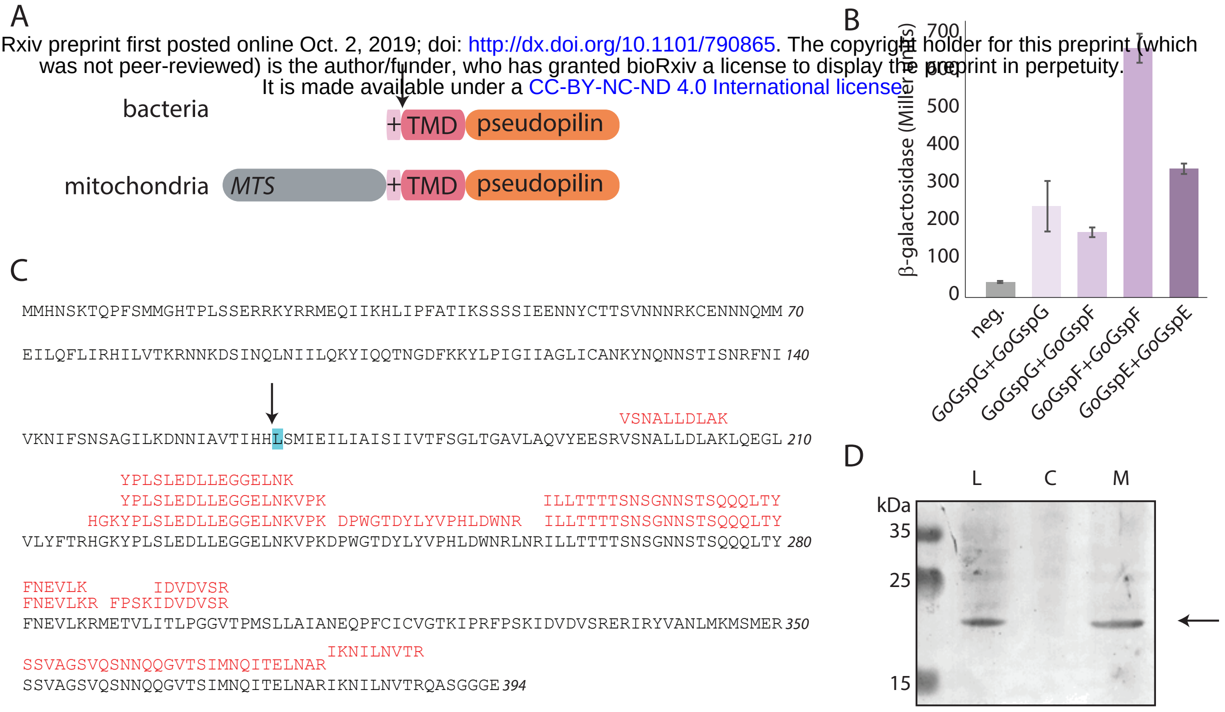
A
bacteria
+
TMD
pseudopilin
mitochondria
MTS
+
TMD
pseudopilin
B
β-galactosidase (Miller units)
0
100
200
300
400
500
600
700
neg.
GoGspG+GoGspG
GoGspG+GoGspF
GoGspF+GoGspF
GoGspE+GoGspE
C
MMHNSKTQPFSMMGHTPLSSERRKYRRMEQIIKHLIPFATIKSSSSIEENNYCTTSVNNNRKCENNNQMM 70
EILQFLIRHILVTKRNNKDSINQLNIILQKYIQQTNGDFKKYLPIGIIAGLICANKYNQNNSTISNRFNI 140
VSNALLDLAK
VKNIFSNSAGILKDNNIAVTIHHLSMIEILIAISIIVTFSGLTGAVLAQVYEESRVSNALLDLAKLQEGL 210
YPLSLEDLLEGGELNK
YPLSLEDLLEGGELNKVPK
ILLTTTTSNSGNNSTSQQQLTY
HGKYPLSLEDLLEGGELNKVPK
DPWGTDYLYVPHLDWNR
ILLTTTTSNSGNNSTSQQQLTY
VLYFTRHGKYPLSLEDLLEGGELNKVPKDPWGTDYLYVPHLDWNRLNRILLTTTTSNSGNNSTSQQQLTY 280
FNEVLK
IDVDVSR
FNEVLKR
FPSKIDVDVSR
FNEVLKRMETVLITLPGGVTPMSLLAIANEQPFCICVGTKIPRFPSKIDVDVSRERIRYVANLMKMSMER 350
IKNILNVTR
SSVAGSVQSNNQQGVTSIMNQITELNAR
SSVAGSVQSNNQQGVTSIMNQITELNARIKNILNVTRQASGGGE 394
D
L
C
M
kDa
35
25
15

Figure 5

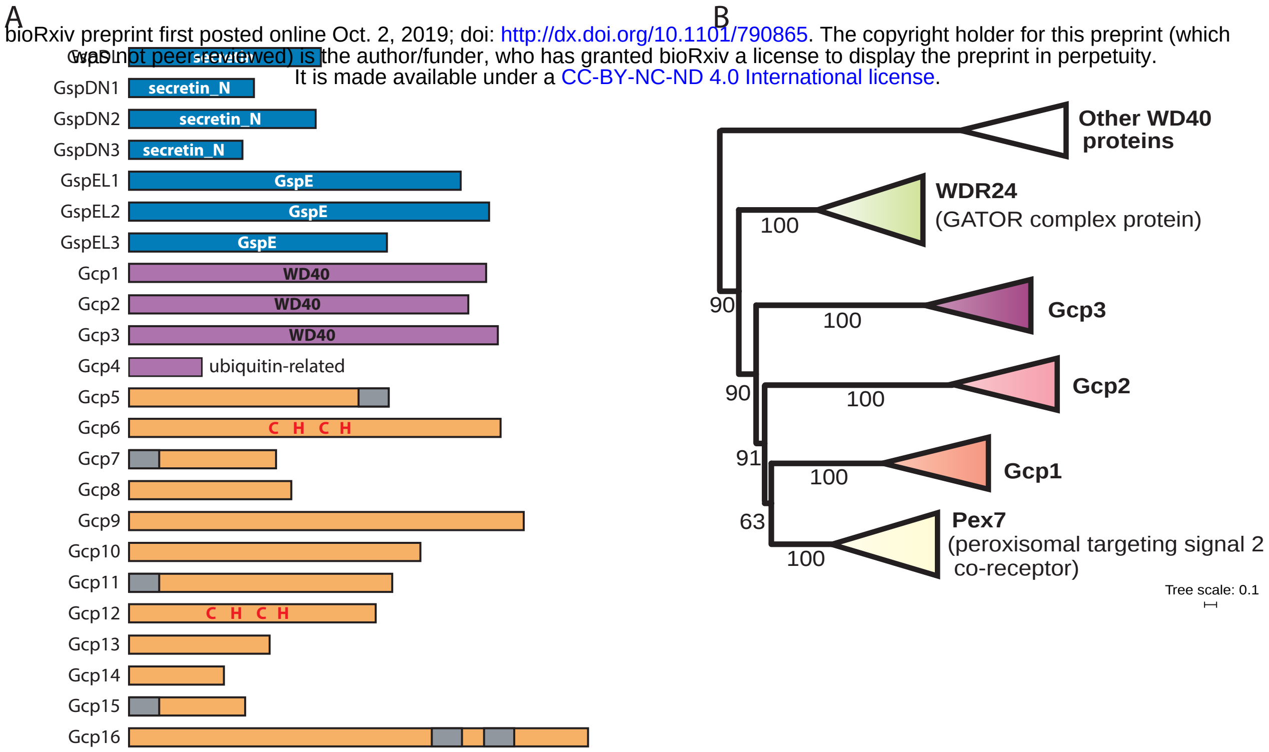
A
GspDL
secretin
GspDN1
secretin_N
GspDN2
secretin_N
GspDN3
secretin_N
GspEL1
GspE
GspEL2
GspE
GspEL3
GspE
Gcp1
WD40
Gcp2
WD40
Gcp3
WD40
Gcp4
ubiquitin-related
Gcp5
Gcp6
C H C H
Gcp7
Gcp8
Gcp9
Gcp10
Gcp11
Gcp12
C H C H
Gcp13
Gcp14
Gcp15
Gcp16
B
Other WD40 proteins
WDR24
(GATOR complex protein)
100
90
Gcp3
100
90
Gcp2
100
91
Gcp1
100
63
Pex7
(peroxisomal targeting signal 2 co-receptor)
100
Tree scale: 0.1

Figure 6

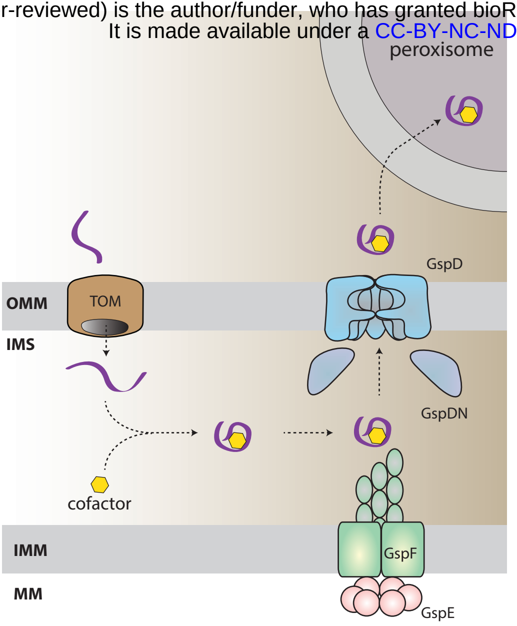

Figure 7