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

185 malawimonads have two GspG paralogues, but the phylogenetic analyses did not
186 resolve whether this is due to multiple independent GspG gene duplications or one
187 ancestral eukaryotic duplication followed by loss of one of the paralogues in
188 jakobids (Supplementary Fig. 3D; Supplementary Table 1).
189
190 ***The eukaryotic Gsp proteins localize to the mitochondrion***
191 We hypothesized that the eukaryotic homologues of the four Gsp proteins are parts
192 of a functional T2SS-related system localized to the mitochondrion. This notion was
193 supported by the presence of predicted N-terminal mitochondrial targeting
194 sequences (MTSs) in some of the eukaryotic Gsp proteins (Supplementary Table 1).
195 The prediction algorithms identified putative N-terminal MTSs for proteins from
196 jakobids and malawimonads but failed to recognize them in the orthologues from
197 heteroloboseans, which, however, carry the longest N-terminal extensions
198 (Supplementary Fig. 4). We assumed that these extensions might still function as
199 MTSs in heteroloboseans. Indeed, labelling of *N. gruberi* cells using specific
200 polyclonal antibodies showed that GspD, GspF and GspG1 are present in
201 mitochondria (Fig. 2A). Moreover, the atypical MTSs of *N. gruberi* Gsp proteins were
202 efficiently recognized by the yeast mitochondrial import machinery (Supplementary
203 Fig. 5). Analogously, three Gsp proteins from *G. okellyi* were all localized to
204 mitochondria when expressed in yeast (Fig. 2B).
205 In order to further confirm the mitochondrial localization of the Gsp proteins
206 in *N. gruberi*, we analysed the mitochondrial proteome of this species by partial
207 purification of the organelle and identification of resident proteins by mass
208 spectrometry. A mitochondria-enriched fraction was obtained from a cellular lysate
209 by several steps of differential centrifugation and further separated by OptiPrep
210 gradient centrifugation. Three sub-fractions of different densities were collected
211 (Supplementary Fig. 6A) and subjected to proteomic analysis. The relative amount
212 of each protein in the gradient was determined by label-free quantification and the
213 proteins were grouped by a multicomponent analysis (for details see Methods)
214 according to their distributions across the gradient (Fig. 3). A set of marker proteins
215 (homologs of well characterized typical mitochondrial proteins from other species)
216 was used to identify a cluster of mitochondrial proteins. Due to the partial co-
217 purification of peroxisomes with mitochondria, a peroxisome-specific cluster was
218 defined analogously. As a result, 946 putative mitochondrial and 78 putative
219 peroxisomal proteins were identified among the total of 4,198 proteins detected.
220 Encouragingly, the putative mitochondrial proteome of *N. gruberi* is dominated by
221 proteins expected to be mitochondrial or whose mitochondrial localization is not
222 unlikely (Supplementary Fig. 6B, Supplementary Table 2A). On the other hand, the
223 putative peroxisomal proteome seems to be contaminated by mitochondrial
224 proteins (owing to the presence of several mitochondrial ribosomal proteins;
225 Supplementary Table 2B). Importantly, all five Gsp proteins (including both GspG
226 paralogs) were identified in the putative mitochondrial but not peroxisomal
227 proteome of *N. gruberi*.
228
229 ***The properties of the eukaryotic Gsp proteins support the existence of a***
230 ***mitochondrial T2SS-related machinery***

231 The foregoing experiments support the idea that all four eukaryotic Gsp homologues
232 localize to and function in the mitochondrion. However, direct *in vivo* demonstration
233 of the existence of a functional mitochondrial T2SS-related machinery is currently
234 not feasible, because none of the Gsp homologue-carrying eukaryotes represents a
235 tractable genetic system. We thus used *in vitro* approaches and heterologous
236 expression systems to test the key properties of the eukaryotic Gsp proteins.
237 Crucial for the T2SS function is the formation of the OM pore, which is a β-
238 barrel formed by the oligomerization of the C-domain of the GspD protein[26].
239 The actual assembly of the bacterial pore requires the interaction of the very C-
240 terminal domain of GspD (S-domain) with the outer membrane lipoprotein GspS[27].
241 In addition, the bacterial GspD carries four short N-terminal domains exposed to the
242 periplasm, called N0 to N3, of which N1 to N3 share a similar fold[28] (Fig. 4A). While
243 the N3 domain has been shown to participate in the pore assembly, N0 interacts
244 with GspC of the assembly platform[18]. However, sequence analysis of the
245 mitochondrial GspD homologue revealed that it, in fact, corresponds to only a C-
246 terminal part of the bacterial GspD β -barrel C-domain, whereas the N-terminal
247 domains N0 to N3, the N-terminal part of the C-domain, and the S-domain are
248 missing (Fig. 4A). This finding raised a question whether the mitochondrial GspD
249 homologue has retained the ability to form a membrane pore. Nevertheless,
250 homology modelling of GspD from *G. okellyi* (*Go*GspD) using *Vibrio cholerae* GspD[29]
251 as a template indicated that the protein could be fitted into solved structure of the
252 pentadecameric pore complex with the conserved amphipathic helical loop
253 (AHL)(Fig. 4B).
254 Testing the function of *Go*GspD in bacteria was impossible due to its high
255 toxicity leading to rapid cell death upon induction of protein expression (Fig. 4C),
256 which is a typical behaviour of pore-forming proteins. The protein toxicity was less
257 pronounced in the yeast two-hybrid (Y2H) system, which indicated strong self-
258 interaction of *Go*GspD (Fig. 4D), and hence its ability to oligomerize. Indeed,
259 radioactively labelled *Go*GspD assembled into a high-molecular-weight complex in
260 an experimental membrane in an *in vitro* translation assay (Fig. 4E). The formation
261 of the complex was dependent on the presence of the membrane and the complex
262 was resistant to 2M urea treatment, which would remove nonspecific protein
263 aggregates. These results showed that the mitochondrial GspD, despite being
264 significantly truncated when compared to its bacterial homologues, has retained the
265 capability to form membrane pores, characteristic for the secretins of the T2SS [30].
266 Compared to the bacterial GspD, the predicted *Go*GspD structure suggests a unique
267 biogenesis pathway, where the secretin pore-forming domain may be directly
268 inserted in the mitochondrial outer membrane, bypassing the membrane transport
269 essential for its bacterial counterparts.
270 The secretion mechanism of the T2SS relies on assembly of pseudopilus
271 made up of GspG subunits[15]. A possible assembly of mitochondrial GspG from *G.*
272 *okellyi* (*Go*GspG1) into the pseudopilus structure was indicated by modelling the
273 protein sequence into the recently obtained cryoEM reconstruction of the PulG
274 complex from *Klebsiella oxytoca*[20] (Supplementary Fig. 7). The actual interaction
275 properties of *Go*GspG1 were followed by the bacterial two-hybrid assay (B2H).
276 When expressed in bacteria (in a truncated form with the MTS region removed, see

277 Fig. 5A), the mitochondrial *Go*GspG1 interacted with itself (Fig. 5B), which is a
278 prerequisite for pseudopilus formation. An analogous B2H assays of *N. gruberi* Gsp
279 proteins also showed GspG1 self-interaction (data not shown). In addition, *Go*GspG1
280 showed positive interaction with *Go*GspF, the IM component believed to participate
281 in transfer of energy for the pseudopilus assembly from GspE (Fig. 1A). Moreover,
282 the mitochondrial *Go*GspF and *Go*GspE each self-interacted in the B2H assay (Fig.
283 5B). These interactions are in agreement with the role of both proteins as T2SS
284 components, as GspF forms dimers within the IM complex and GspE assembles into
285 an active hexameric ATPase. Furthermore, B2H assay has identified the same
286 interactions between the GspG and GspF homologues in the bacterial T2SS [31]. Tests
287 of all other possible interactions of *G. okellyi* Gsp proteins were negative.
288 The *in silico* analyses and experiments described above are consistent with
289 the hypothesized existence of a functional mitochondrial secretion machinery
290 derived from the bacterial T2SS. However, the mitochondrial subunits identified
291 would assemble only a minimalist version of the secretion system, reduced to the
292 functional core of the four subcomplexes of the bacterial T2SS, i.e. the luminal
293 ATPase (GspE), the IM pseudopilus assembly platform (GspF), the intermembrane
294 space pseudopilus (GspG), and the OM pore (truncated GspD). Despite using
295 sensitive HMM-based searches, we did not detect homologues of other conserved
296 T2SS subunits in any of the eukaryotes possessing GspD to GspG proteins. One of the
297 missing subunits is GspC, which connects the assembly platform with the N0 domain
298 of GspD pore[18,32]. Thus, the absence of GspC in eukaryotes correlates with the lack of
299 the N0 domain in the eukaryotic GspD. Analogously, the absence of the C-terminal S-
300 domain in the mitochondrial GspD (Fig. 4A), known to be missing also from some
301 bacterial GspD proteins, rationalizes the lack of a eukaryotic homologue of the
302 bacterial OM component GspS that binds to GspD via the S-domain during the pore
303 assembly[27].
304 The mitochondrial system also apparently lacks a homologue of GspO, a
305 bifunctional enzyme that is essential for GspG maturation. Despite this absence,
306 eukaryotic GspG homologues have conserved all the characteristic sequence
307 features required for GspG maturation (the polar anchor and the trans-membrane
308 domain with a conserved glutamate residue at the +5 position relative to the
309 processing site) (Fig. 5A, Supplementary Fig. 4D). Notably, all the *Ng*GspG1 and
310 *Ng*GspG2-derived peptides detected in our proteomic analysis come from the region
311 of the protein downstream of the conserved processing site (Fig.5C), and an anti-
312 *Ng*GspG1 antibody detected a specific band of a much smaller size than expected for
313 the full-length protein (around 44 kDa) on a western blot of electrophoretically
314 separated *N. gruberi* proteins (Fig.5D). However, the theoretical Mw of the *Ng*GspG1
315 processed at the conserved site is 25.5 kDa, whereas the protein detected by the
316 immunoblot is even smaller, with a size similar to that of bacterial pseudopilins.
317 Hence, the actual nature of the mitochondrial GspG maturation needs to be studied
318 further.
319
320 ***New putative components of the mitochondrial T2SS-based functional pathway***
321 ***identified by phylogenetic profiling***

322 Since none of the eukaryotes with the Gsp homologues is currently amenable to
323 functional studies, we tried to further illuminate the role of the mitochondrial T2SS
324 system using a comparative genomic approach. Specifically, we reasoned that
325 possible additional components of the machinery, as well as its actual substrate(s),
326 might show the same phylogenetic distribution as the originally identified four
327 subunits. Using a combination of an automated identification of candidate protein
328 families and subsequent manual scrutiny by exhaustive searches of available
329 eukaryote sequence data (for details of the procedure see Methods), we identified
330 23 proteins (more precisely, groups of orthologues) that proved to exhibit precisely
331 the same phylogenetic distribution in eukaryotes as the four core T2SS components.
332 Specifically, all 23 proteins were represented in each of the heterolobosean, jakobid,
333 and malawimonad species possessing all four core Gsp proteins, whereas only seven
334 of them were found in the transcriptomic data from the *Percolomonas* lineage that
335 possesses only GspD (Fig. 1B; Supplementary Table 3). Except for two presumably
336 Gsp-positive jakobids represented by incomplete EST surveys and a case of a likely
337 contamination (Supplementary Table 4), no orthologues of any of these proteins
338 were found in any other eukaryote (including the Gsp-lacking members of
339 heteroloboseans, jakobids and malawimonads). The sequences of these 23 proteins
340 were analysed by various *in silico* approaches, including sensitive homology-
341 detection methods (HMM-HMM comparisons with HHpred[33] and protein modelling
342 using the Phyre2 server[34]) to assess their possible function (Fig. 6A; Supplementary
343 Table 3).
344 These analyses revealed that seven of the families have a direct link to the
345 T2SS suggested by discerned homology to known T2SS components. One of them
346 represents an additional, more divergent homologue of the C-terminal part of the
347 bacterial GspD. Hence, the protein has been marked as GspDL (GspD-like). Three
348 other families, referred to as GspDN1 to GspDN3, proved to be homologous to the
349 Secretin_N domain (Pfam family PF03958), present in the bacterial GspD protein in
350 three copies as the domains N1, N2, and N3 (Fig. 4A). The N1-N3 array protrudes
351 into the periplasmic space, where it oligomerizes to form three stacked rings[35]. As
352 mentioned above, the initially identified eukaryotic GspD homologues lack the N-
353 terminal region, suggesting that the gene was split into multiple parts in eukaryotes.
354 Unfortunately, high sequence divergence makes it impossible to identify potential
355 specific correspondence between the N1 to N3 domains of the bacterial GspD and
356 the eukaryotic GspDN1 to GspDN3 proteins. Importantly, an initial Y2H assay
357 indicated that the two separate polypeptides GspD and GspDN1 of *N. gruberi* may
358 interact *in vivo* (Fig. 4F), perhaps forming a larger mitochondrial complex. In
359 addition, we identified most of the newly discovered GspD-related proteins (GspDL
360 and GspDN) in the *N. gruberi* mitochondrial proteome (the exception being GspDN1,
361 which was not detected in a sufficient number of replicates to be included in the
362 downstream analysis; Supplementary Table 2A).
363 The final three proteins linked to the T2SS based on their sequence features
364 represent three divergent paralogues of the GspE subunit (GspE-like) here denoted
365 GspEL1 to GspEL3. However, abrogation of ATPase-specific motifs in these
366 paralogues suggests the loss of the ATPase activity (Supplementary Fig. 4B). GspEL2
367 and GspEL3 were identified among *N. gruberi* mitochondrial proteins in the

368 proteomic analysis, whereas GspEL1 was found in the cluster of putative
369 peroxisomal proteins.
370 The remaining sixteen proteins co-occurring with the core eukaryotic T2SS
371 subunits, hereafter referred to as Gcp (Gsp-co-occurring proteins), were divided
372 into three categories. The first comprises four proteins that constitute novel
373 paralogues within broader common eukaryotic (super)families (Fig. 6B). Three of
374 them (Gcp1 to Gcp3) belong to the WD40 superfamily, in which they form a single
375 clade together with the peroxisomal protein import co-receptor Pex7 (Fig. 6B;
376 Supplementary Fig. 8). None of these proteins has any putative N-terminal targeting
377 sequence, but interestingly, the peroxisomal targeting signal 1 (PTS1) could be
378 predicted on most Gcp1 and some Gcp2 proteins (Supplementary Table 3). However,
379 these predictions are not fully consistent with the results of our proteomic analysis:
380 *Ng*Gcp1 was found among the mitochondrial proteins and *Ng*Gcp2 in the cluster of
381 putative peroxisomal proteins (Supplementary Table 2), but PTS1 is predicted to be
382 present in the *Ng*Gcp1 protein (Supplementary Table 3). The fourth Gcp protein
383 (Gcp4) is a novel paralogue of the ubiquitin-like superfamily, distinctly different
384 from the previously characterized members including ubiquitin, SUMO, NEDD8 and
385 others (Supplementary Fig. 9).
386 The second Gcp category comprises eleven proteins (Gcp5 to Gcp15) well
387 conserved at the sequence level among the Gsp-containing eukaryotes, yet lacking
388 any discernible homologues in other eukaryotes or in prokaryotes. Two of these
389 proteins (Gcp8, Gcp15) were not identified in the proteomic analysis of *N. gruberi*
390 (Supplementary Table 3). Of those identified, several (Gcp5, Gcp6, Gcp13) were
391 found among the mitochondrial proteins, whereas some others (Gcp9, Gcp10,
392 Gcp11) clustered with peroxisomal markers. Specific localization of the three
393 remaining proteins (Gcp7, Gcp12, and Gcp14) could not be determined due to their
394 presence at the boundaries of the mitochondrial or peroxisomal clusters.
395 No homology to other proteins or domains could be discerned for the Gsp5 to Gsp15
396 proteins even when sensitive homology-detection algorithms were employed.
397 However, four of them are predicted as single-pass membrane proteins, with the
398 transmembrane segment in the N- (Gcp7, Gcp11, Gcp15) or C-terminus (Gcp5) (Fig.
399 6A; Supplementary Fig. 10). Interestingly, Gcp6 and Gcp12 proteins contain multiple
400 absolutely conserved cysteine or histidine residues (Fig. 6A; Supplementary Fig. 11).
401 Finally, Gcp16 constitutes a category of its own. It typifies a family of
402 predicted membrane proteins with non-eukaryotic representatives restricted to
403 bacteria of the PVC superphylum (Supplementary Fig. 12), some of which are known
404 to have the T2SS[36]. Interestingly, Gcp16 proteins from *Neochlamydia* spp. are fused
405 to the N-terminus of a protein from the Lactamase_B_2 (PF12706) family that
406 generally occurs as an independent protein widely conserved in various bacteria.
407 Phylogenetic analyses confirmed that the eukaryotic members of the family are of
408 the same origin rather than acquisitions by independent HGT events into different
409 lineages of eukaryotes (Supplementary Fig. 13). Most eukaryotic Gcp16 proteins
410 exhibit an N-terminal extension compared to the bacterial homologues
411 (Supplementary Fig. 12), but only some of these extensions are recognized as
412 putative MTSs and the *N. gruberi* Gcp16 was not identified either in putative
413 mitochondrial or peroxisomal proteome.

414
415 **Discussion**
416 Our analyses revealed that a subset of species belonging to three eukaryotic lineages
417 share a set of at least 27 proteins (or families of orthologues) absent from other
418 eukaryotes for which genomic or transcriptomic data are currently available (Fig.
419 1C). At least eleven of these proteins (the Gsp proteins) are evolutionarily related to
420 components of the bacterial T2SS, although seven of them are so divergent that their
421 evolutionary connection to the T2SS could be recognized only retrospectively after
422 their identification based on their characteristic phylogenetic profile. For the sixteen
423 remaining proteins (Gcp1 to Gcp16) no other evolutionary or functional link to the
424 T2SS is evident apart from the same phyletic pattern as exhibited by the T2SS
425 subunit homologues. Nevertheless, similar phylogenetic profiles are generally a
426 strong indication for proteins being parts of the same functional system or pathway,
427 and have enabled identification of new components of different cellular structures
428 or pathways (e.g. refs[37,38]). Is it, therefore, possible that the 27 Gsp/Gcp proteins
429 similarly belong to a single functional pathway?
430 The phylogenetic profile shared by the eukaryotic Gsp and Gcp proteins is
431 not trivial, as it implies independent gene losses in a specific set of multiple
432 eukaryotic branches (Fig. 1B). The likelihood of a chance emergence of the same
433 taxonomic distribution of these proteins is thus low. Nevertheless, false positives
434 cannot be completely excluded among the Gcp proteins and their list may be revised
435 when a more comprehensive sampling of eukaryote genomes or transcriptomes
436 becomes available. It is also possible that the currently inferred phylogenetic profile
437 of some of the Gsp/Gcp proteins is inaccurate due to incomplete sampling of the
438 actual gene repertoire of species represented by transcriptome assemblies only. An
439 interesting case in point is the heterolobosean *Percolomonas* lineage.
440 Transcriptomic data from three different members revealed only the presence of
441 GspD, GspDL, the three GspDN variants, and four Gcp proteins (Fig. 1B,
442 Supplementary Tables 1 and 3), which may reflect incomplete data. However, the
443 relatively coherent pattern of Gsp/Gcp protein occurrence in the three
444 independently sequenced transcriptomes and the fact that in other Gsp/Gcp -
445 containing eukaryotes all 27 families are always represented in the respective
446 transcriptome assembly (Supplementary Tables 1 and 3) suggest that the
447 *Percolomonas* lineage has preserved only a subset of Gsp/Gcp families. Genome
448 sequencing is required to test this possibility.
449 All uncertainties notwithstanding, our data favour the idea that a hitherto
450 unknown complex functional pathway exists in some eukaryotic cells, underpinned
451 by most, if not all, of the 27 Gsp/Gcp proteins and possibly others yet to be
452 discovered. Direct biochemical and cell biological investigations are required for
453 testing its very existence and the actual cellular role. Nevertheless, we integrated
454 the experimental data gathered so far with the insights from bioinformatic analyses
455 to propose a hypothetical working model (Fig. 7).
456 Our main proposition is that the eukaryotic homologues of the bacterial Gsp
457 proteins assemble a functional transport system, here denoted miT2SS, that spans
458 the mitochondrial OM and mediates the export of specific substrate proteins from
459 the mitochondrion. Although the actual architecture of the miT2SS needs to be

460 determined, the available data suggest that it departs in detail from the canonical
461 bacterial T2SS organization, as homologues of some of the important bacterial T2SS
462 components are apparently missing. Most notable is the absence of GspC,
463 presumably related to the modified structure of its interacting partner GspD, which
464 in eukaryotes is split into multiple polypeptides and seems to completely lack the
465 N0 domain involved in GspC binding. It thus remains unclear whether and how the
466 IM assembly platform and the OM pore interact in mitochondria. One possible
467 explanation is that GspC has been replaced by an unrelated protein. It is notable that
468 three Gcp proteins (Gcp7, Gcp11, and Gcp15) have the same general architecture as
469 GspC: they possess a transmembrane segment at the N-terminus and a (predicted)
470 globular domain at the C-terminus (Fig. 6A). Testing possible interactions between
471 these proteins and T2SS core subunits (particularly GspF and GspDN) using B2H or
472 Y2H assays will be of future interest.
473 Future investigations also must address the question of whether the
474 mitochondrial GspG is processed analogously to the bacterial homologues and how
475 such processing occurs in the absence of discernible homologues of GspO (see
476 above). The mitochondrial GspG is presumably inserted into the IM by the Tim22 or
477 Tim23 complex, resulting in a GspG precursor with the N-terminus, including the
478 MTS, protruding into the matrix. It is possible that N-terminal cleavage by matrix
479 processing peptidase serves not only to remove the transit peptide, but at the same
480 time to generate the mature N-terminus of the processed GspG form, ready for
481 recruitment into the pseudopilus.
482 In parallel with its apparent simplification, the miT2SS may have been
483 specifically elaborated compared to the ancestral bacterial machinery. This
484 possibility is suggested by the existence of the three divergent, possibly ATPase
485 activity-deficient GspE paralogues (GspEL1 to GspEL3) that we discovered in all
486 miT2SS-containing eukaryotes but not elsewhere. We can only speculate as to the
487 function of these proteins, but they may interact with and regulate the catalytically
488 active GspE protein. The fact that the bacterial GspE assembles into a homohexamer
489 raises the possibility that in eukaryotes GspEL proteins are included in a
490 heterooligomer with GspE, a situation analogous to the presence of catalytically
491 active and inactive paralogous subunits in some well known protein complexes (e.g.
492 refs[39,40]). The co-occurrence of two different paralogues of the GspD C-domain, one
493 (GspDL) being particularly divergent, suggests a eukaryote-specific elaboration of
494 the putative pore in the mitochondrial OM.
495 An unanswered key question is what is the actual substrate (or substrates)
496 exported from the mitochondrion by the miT2SS. No bioinformatic tool for T2SS
497 substrate prediction is available due to the enigmatic nature of the mechanism of
498 substrate recognition by the pathway[12], so at the moment we can only speculate. It
499 is notable that no protein encoded by the mitochondrial genomes of jakobids,
500 heteroloboseans and malawimonads stands out as an obvious candidate for the
501 miT2SS substrate, since they either have well-established roles in the
502 mitochondrion or are hypothetical proteins with a restricted (genus-specific)
503 distribution. Therefore, we hypothesize that the substrate is encoded by the nuclear
504 genome and imported into the mitochondrion to undergo a specific processing step.
505 This may include addition of a prosthetic group – a scenario modelled on the

506 process of cytochrome *c* or Rieske protein maturation[41,42]. Interestingly, the
507 proteins Gcp6 and Gcp12, each exhibiting an array of absolutely conserved cysteine
508 and histidine residues (Supplementary Fig. 11), are good candidates for proteins
509 that are loaded with a specific prosthetic group, so any of them may well be the
510 sought-after miT2SS substrate. Some of the other Gcp proteins may then represent
511 components of the hypothetical machinery responsible for the substrate
512 modification. The putative functionalization step may occur either in the
513 mitochondrial matrix or in the intermembrane space (IMS), but we note that the
514 former localization would necessitate a mechanism of protein translocation across
515 the mitochondrial IM in the direction from the matrix to the IMS, which has not been
516 demonstrated yet. Regardless, the modified protein would eventually be
517 translocated across the mitochondrial OM by the T2SS system to the cytoplasm.
518 However, this may not be the end of the journey, since there are hints of a
519 link between the miT2SS-associated pathway and peroxisomes. First, three Gcp
520 proteins, namely Gcp1 to Gcp3, are specifically related to Pex7, a protein mediating
521 import of peroxisomal proteins characterized by the peroxisomal targeting signal 2
522 (PTS2)[43]. Second, some of the Gcp proteins (Gcp1, Gcp2, Gcp13) have at the C-
523 terminus a predicted PTS1 signal (at least in some species; Supplementary Table 3).
524 Third, several Gcp proteins (Gcp2, Gcp9, Gcp10, and Gcp11) and GspEL1 were
525 assigned to the putative peroxisomal proteome in our proteomic analysis
526 (Supplementary Table 2B). We note the discrepancy between the PTS1 signal
527 predictions and the actual set of experimentally defined peroxisomal proteins,
528 which might be due to an incomplete separation of peroxisome and mitochondria by
529 our purification procedure, but may also reflect protein shuttling between the two
530 organelles.
531 We thus hypothesize that upon its export from the mitochondrion, the
532 miT2SS substrate is eventually delivered to the peroxisome. This is possibly
533 mediated by the Gcp1/2/3 trio, but other Gcp proteins might participate as well.
534 One such protein might be the ubiquitin-related protein Gcp4. Ubiquitination and
535 deubiquitination of several components of the peroxisome protein import
536 machinery is a critical part of the import mechanism[43] and Gcp4 could serve as an
537 analogous peptide modifier in the hypothetical novel peroxisome import pathway
538 functionally linked to the miT2SS.
539 Altogether, our data suggest the existence of a novel elaborate functional
540 pathway combining components of bacterial origin with newly evolved eukaryote-
541 specific proteins. The modern phylogenetic distribution of the pathway is sparse,
542 but our current understanding of eukaryote phylogeny suggests that it was
543 ancestrally present in eukaryotes and for some reason dispensed with, multiple
544 times during evolution. Although we could not define a specific bacterial group as
545 the actual source of the eukaryotic Gsp genes, it is tempting to speculate that the
546 T2SS was introduced into eukaryotes by the bacterial progenitor of mitochondria
547 and that it was involved in delivering specific proteins from the endosymbiont into
548 the host cell, as is known in the case of current intracellular bacteria[36]. Elucidating
549 the actual role of this communication route in establishing the endosymbiont as a
550 fully integrated organelle requires understanding the cellular function of the
551 modern miT2SS-associated pathways, which is a challenge for future research.

552
553 **Methods**
554
555 **Sequence data and homology searches**
556 Homologues of relevant genes/proteins were searched in sequence databases
557 accessible via the National Center for Biotechnology Information BLAST server
558 (https://blast.ncbi.nlm.nih.gov/Blast.cgi), including the nucleotide and protein non-
559 redundant (nr) databases, whole-genome shotgun assemblies (WGAs), expressed
560 sequence tags (ESTs), and transcriptome shotgun assemblies (TSAs). Additional
561 public databases searched included the data provided by the Marine Microbial
562 Eukaryote Transcriptome Sequencing Project (MMETSP[44]) comprising TSAs from
563 hundreds of diverse protists (https://www.imicrobe.us/#/projects/104), the
564 OneKP project[45] (https://sites.google.com/a/ualberta.ca/onekp/) comprising TSAs
565 from hundreds of plants and algae, and individual WGAs and TSAs deposited at
566 various on-line repositories (Supplementary Table 5). Non-public sequence data
567 analysed included genome and/or transcriptome assemblies from several
568 heteroloboseans, jakobids and malawimonads generated in our laboratories using
569 standard sequencing technologies (454 and or Illumina) and sequence assembly
570 programs (Supplementary Table 5). Details on the sequencing and assembly and full
571 analyses of these genomes and transcriptomes will be published elsewhere.
572 Homology searches were done using BLAST[46] (blastp or tblastn, depending
573 on the database queried) and HMMER[47] using profile HMMs built from sequence
574 alignments of proteins of interest. Hits were evaluated by BLAST (blastp or blastx)
575 searches against the nr protein dataset at NCBI to distinguish orthologues of Gsp
576 and Gcp proteins from paralogous proteins or non-specific matches. This was
577 facilitated by a high degree of conservation of individual eukaryotic Gsp/Gcp
578 proteins among different species (see also Supplementary Figs 4 and 10-12) and in
579 most cases by the lack of other close homologues in eukaryotic genomes (the
580 exceptions being members of broader protein families, including the ATPase GspE,
581 the WD40 superfamily proteins Gcp1 to Gcp3, and the ubiquitin related protein
582 Gcp4). All identified eukaryotic Gsp and Gcp sequences were carefully manually
583 curated to ensure maximal accuracy and completeness of the data, which included
584 correction of existing gene models, extension of truncated sequences by manual
585 analysis of raw sequencing reads, and correction of assembly errors (for details see
586 Supplementary Methods). All newly predicted or curated Gsp and Gcp sequences are
587 provided in Supplementary Tables 1 and 3, respectively; additional Gsp and Gcp
588 sequences from non-target species are listed in Supplementary Table 4.
589
590 **Phylogenetic profiling**
591 In order to identify genes with the same phylogenetic distribution as the eukaryotic
592 homologues of the four core T2SS components, we carried out two partially
593 overlapping analyses based on defining groups of putative orthologous genes in
594 select Gsp-positive species and phylogenetically diverse Gsp-negative eukaryotic
595 species. The list of taxa included is provided in Supplementary Table 6. The first
596 analysis was based on 18 species, including three Gsp-positive ones (*N. gruberi*, *A.*
597 *godoyi* and *M. jakobiformis*), for the second analysis the set was expanded by adding

598 one additional Gsp-positive species (*G. okellyi*) and one Gsp-negative species
599 (*Monocercomonoides* sp. PA203). Briefly, the protein sequences of a given species
600 were compared to those of all other species using blastp followed by fast
601 phylogenetic analyses and orthologous relationships between proteins were then
602 inferred from this set of phylogenetic trees using a reference-species-tree-
603 independent approach. This procedure was repeated for each species and all
604 resulting sets of orthologous relationships, also known as phylomes[48], were
605 combined in a dense network of orthologous relationships. This network was finally
606 trimmed in several successive steps to remove week or spurious connections and to
607 account for (genuine or artificial) gene fusions, with the first analysis being less
608 restrictive than the second. Details of this pipeline are provided in Supplementary
609 Methods.
610 For each of the two analyses, the final set of defined groups of orthologs
611 (orthogroups) was parsed to identify those comprising genes from at least two Gsp-
612 positive species yet lacking genes from any Gsp-negative species. The orthogroups
613 passing this criterion were further analysed manually by blastp and tblastn searches
614 against various public and private sequence repositories (see the section "Sequence
615 data and homology searches") to exclude those orthogroups with obvious orthologs
616 in Gsp-negative species. *Percolomonas* spp. exhibiting only GspD and jakobids
617 represented by incomplete EST surveys (these species are likely to possess the
618 miT2SS system) were not considered as Gsp-negative. The orthogroups that
619 remained were then evaluated for their conservation in Gsp-positive species and
620 those that proved to have a representative in all these species (*N. gruberi*, *N. fowleri*,
621 *N. damariscottae*, *P. kirbyi*, *A. godoyi*, *R. americana*, *M. jakobiformis*, *G. okellyi*) were
622 considered as bona fide Gcp (Gsp-co-occurring protein) candidates. It is of note that
623 some of these proteins are short and were missed by the automated annotation of
624 some of the genomes, so using relaxed criteria for the initial consideration of
625 candidate orthogroups (i.e. allowing for their absence from some of the Gsp-positive
626 species) proved critical for decreasing the number of false-negative identifications.
627
628 **Sequence analyses and phylogenetic inference**
629 The presence of N-terminal mitochondrial transit peptides and peroxisomal
630 targeting signal 1 (PTS1) in the Gsp and Gcp proteins was evaluated using
631 MitoFates[49] (http://mitf.cbrc.jp/MitoFates/cgi-bin/top.cgi) and PTS1 predictor[50]
632 (http://mendel.imp.ac.at/pts1/), respectively. Transmembrane domains were
633 predicted using TMHMM[51] (http://www.cbs.dtu.dk/services/TMHMM/). Homology
634 of Gsp and Gcp protein families to other proteins was evaluated by searches against
635 Pfam v. 31 (ref.[52]; http://pfam.xfam.org/) and Superfamily 1.75 database[53]
636 (http://supfam.org/SUPERFAMILY/index.html) and by using HHpred[33]
637 (https://toolkit.tuebingen.mpg.de/#/tools/hhpred) and the Phyre2 server[34]
638 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index). The relative
639 position of the Gcp4 family among Ubiquitin-like proteins was analysed by a cluster
640 analysis using CLANS[54] (https://www.eb.tuebingen.mpg.de/protein-
641 evolution/software/clans/); for the analysis the Gcp4 family was combined with all
642 59 defined families included in the clan Ubiquitin (CL0072) as defined in the Pfam
643 database (each family was represented by sequences from the respective seed

644 alignments stored in the Pfam database). For further details on the procedure see
645 the legend of Supplementary Fig. 9A. Multiple sequence alignments used for
646 presentation of the conservation and specific sequence features of Gsp and Gcp
647 families were built using MUSCLE[55] and shaded using BioEdit
648 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html)
649 In order to obtain datasets for the phylogenetic analyses of eukaryotic GspD
650 to GspG proteins, the protein sequences were aligned using MAFFT[56] and trimmed
651 manually. Profile hidden Markov models (HMMs) built on the basis of the respective
652 alignments were used as queries to search the UniProt database using HMMER. All
653 recovered sequences were assigned to components of the T4P superfamily
654 machineries using HMMER searches against a collection of profile HMMs reported
655 by Abby et al. (ref.[19]). For each GspD to GspG proteins, a series of alignments was
656 built by progressively expanding the sequence set by including more distant
657 homologues (as retrieved by the HMMER searches). Specifically, the different sets of
658 sequences were defined by the HMMER score based on the formula $score_{cutoff}$ =
659 $c*score_{best\ prokaryotic\ hit}$, with the coefficient c decreasing from 0.99 to 0.70
660 incrementally by 0.01. The sequences were then aligned using MAFFT, trimmed
661 with BMGE[57] and the phylogenies were computed with IQ-TREE[58] using the best-fit
662 model (selected by the program from standard protein evolution models and the
663 mixture models[59] offered). The topologies were tested using 10,000 ultra-fast
664 bootstraps. The resulting trees were systematically analyzed for support of the
665 monophyly of eukaryotic sequences and for the taxonomic assignment of the
666 parental prokaryotic node of the eukaryotic subtree. The assignment was done
667 using the following procedure. The tree was artificially rooted between the
668 eukaryotic and prokaryotic sequences. From sub-leaf nodes to the deepest node of
669 the prokaryotic subtree, the taxonomic affiliation of each node was assigned by
670 proportionally considering the known or inferred taxonomic affiliations (at the
671 phylum or class level) of the descending nodes. See the legend to Supplementary Fig.
672 3 for further details.
673 The phylogenetic analysis of the WD40 superfamily including Gcp1 to Gcp3
674 proteins was performed as follows. The starting dataset was prepared by a
675 combination of two different approaches: 1) each identified sequence of Gcp1 to
676 Gcp3 proteins was used as a query in a blastp search against the non-redundant (nr)
677 NCBI protein database and the 500 best hits for each sequence were kept; 2) protein
678 sequences of each the Gcp1 to Gcp3 family were aligned using MAFFT and the
679 multiple alignment was used as a query in a HMMER3 search
680 (https://toolkit.tuebingen.mpg.de/#/tools/hmmer) against the UniProt database.
681 Best hits (E-value cutoff 1e-50) from all three searches were pooled and de-
682 duplicated, and the resulting sequence set (including Gcp1 to Gcp3 sequences) was
683 aligned using MAFFT and trimmed manually to remove poorly conserved regions.
684 Because WD40 proteins are extremely diversified, sequences that were too
685 divergent were removed from the starting dataset during three subsequent rounds
686 of sequence removal, based on a manual inspection of the alignment and
687 phylogenetic trees computed by IQ-TREE (using the best-fit model as described
688 above). The final dataset was enriched by adding PEX7 and WDR24 orthologues
689 from eukaryotes known to possess miT2SS components. The final phylogenetic tree

690 was computed using IQ-TEE as described in the legend to Supplementary Fig. 8. IQ-
691 TREE was used also for inferring trees of the heterolobosean 18S rRNA gene
692 sequences (Supplementary Fig. 1), ubiquitin-related proteins (Supplementary Fig.
693 9B) and the Gcp16 family (Supplementary Fig. 13); details on the analyses are
694 provided in legends to the respective figures.
695
696 **Homology modelling**
697 The PDB database was searched by the SWISS-MODEL server[60] for structural
698 homologues of *Go*GspD and *Go*GspG1. *V. cholerae* GspD[35] (PDB entry 5Wq9) and
699 *K. oxytoca* PulG[20] pseudopilus (PDB entry 5wda) were selected as the top matches,
700 respectively. Models were built based on the target-template alignment using
701 ProMod3[60]. Coordinates that were conserved between the target and the template
702 were copied from the template to the model. Insertions and deletions were
703 remodelled using a fragment library, followed by rebuilding side chains. Finally, the
704 geometry of the resulting model was regularized by using a force field. In the case of
705 loop modelling with ProMod3 fails, an alternative model was built with PROMOD-
706 II[61]. The quaternary structure annotation of the template was used to model the
707 target sequence in its oligomeric form[62].
708
709 **Cultivation and fractionation of *N. gruberi* and proteomic analysis**
710 *Naegleria gruberi* str. NEG-M was axenically cultured in M7 medium with PenStrep
711 (100 U/mL of penicillin and 100 µg/mL of streptomycin) at 27°C in vented tissue
712 culture flasks. Mitochondria of *N. gruberi* were isolated in seven independent
713 experiments, which were analyzed individually (see below). Each time ~$1x10^9$ *N.*
714 *gruberi* cells were resuspended in 2 mL of SM buffer (250 mM sucrose, 20 mM MOPS,
715 pH 7.4) supplemented with DNase I (40 ug/mL) and Roche cOmplete™ EDTA-free
716 Protease Inhibitor Cocktail and homogenized by eight passages through a 33-gauge
717 hypodermic needle (Sigma Aldrich). The resulting cell homogenate was then
718 cleaned of cellular debris using differential centrifugation and separated by a 2-hr
719 centrifugation in a discontinuous density OptiPrep gradient (10%, 15%, 20%, 30%
720 and 50%) as described previously[63]. Three visually identifiable fractions
721 corresponding to 10-15% (OPT-1015), 15-20% (OPT-1520) and 20-30% (OPT-
722 2023) OptiPrep densities were collected (each in five biological replicates) and
723 washed with SM buffer.
724 Proteins extracted from these samples were then digested with trypsin and
725 peptides were separated by nanoflow liquid chromatography and analyzed by
726 tandem mass spectrometry (nLC-MS2) on a Thermo Orbitrap Fusion (q-OT-IT)
727 instrument as described elsewhere[64]. The quantification of mass spectrometry data
728 in the MaxQuant software[65] provided normalized intensity values for 4,198
729 proteins in all samples and all three fractions. These values were further processed
730 using the Perseus software[66]. Data were filtered and only proteins with at least two
731 valid values in one fraction were kept. Imputation of missing values, which
732 represent low-abundance measurements, was performed with random distribution
733 around the value of instrument sensitivity using default settings of Perseus
734 software[66].

735 The data were analyzed by principle component analysis (PCA). The first two
736 loadings of the PCA were used to plot a two-dimensional graph. Based on a set of
737 marker proteins (376 mitochondrial and 26 peroxisomal, Supplementary Table 2),
738 clusters of proteins co-fractionating with mitochondria and peroxisomes were
739 defined and the proteins within the clusters were further analyzed. This workflow
740 was set up on the basis of the LOPIT protocol[67]. As a result, out of the 4,198 proteins
741 detected, 946 putative mitochondrial and 78 putative peroxisomal proteins were
742 defined. All proteins were subjected to *in silico* predictions concerning their function
743 (BLAST, HHpred[33]) and subcellular localization (Psort II,
744 https://psort.hgc.jp/form2.html; TargetP,
745 http://www.cbs.dtu.dk/services/TargetP/; MultiLoc2, https://abi.inf.uni-
746 tuebingen.de/Services/MultiLoc2). The mass spectrometry proteomics data have
747 been deposited in the ProteomeXchange Consortium via the PRIDE[68] partner
748 repository with the dataset identifier PXD007764.
749
750 **Fluorescence *in situ* hybridization (FISH)**
751 The PCR products of the *Ng*GspE and *Ng*GspF genes were labelled by alkali-stable
752 digoxigenin-11-dUTP (Roche) using DecaLabel DNA Labeling Kit (Thermo
753 Scientific). Labelled probes were purified on columns of QIAquick Gel Extraction Kit
754 (Qiagen, 28704) in a final volume of 50 µL. Labelling efficiencies were tested by dot
755 blotting with anti-digoxigenin alkaline phosphatase conjugate and CSPD
756 chemiluminescence substrate for alkaline phosphatase from DIG High Prime DNA
757 Labeling and Detection Starter Kit II (Roche) according to the manufacturer`s
758 protocol. FISH with digoxigenin-labelled probes was performed essentially
759 according to the procedure described in Zubacova at al. (ref.[69]) with some
760 modifications. *N. gruberi* cells were pelleted by centrifugation for 10 min at 2,000 x
761 *g* at 4°C. Cells were placed in hypotonic solution, fixed twice with a freshly prepared
762 mixture of methanol and acetic acid (3:1) and dropped on superfrost microscope
763 slides (ThermoScientific). Preparations for hybridizations were treated with RNase
764 A, 20 µg in 100 µL 2 x SSC, for 1 hr at 37°C, washed twice in 2 x SSC for 5 min,
765 dehydrated in a methanol series and air-dried. Slides were treated with 50% acetic
766 acid followed by pepsin treatment and postfixation with 2% paraformaldehyde.
767 Endogenous peroxidase activity of the cell remnants (undesirable for tyramide
768 signal amplification) was inactivated by incubation in 1% hydrogen peroxide,
769 followed by dehydration in a graded methanol series. All slides were denatured
770 together with 2 µL (25 ng) of the probe in 50 µL of hybridization mixture containing
771 50% deionised formamide (Sigma) in 2 x SSC for 5 min at 82°C. Hybridizations
772 were carried out overnight. Slides were incubated with tyramide reagent for 7 min.
773 Preparations were counterstained with DAPI in VectaShield and observed under an
774 Olympus IX81 microscope equipped with a Hamamatsu Orca-AG digital camera
775 using the Cell^R imaging software.
776
777 **Heterologous gene expression, preparation of antibodies, and**
778 **immunodetection of Gsp proteins**
779 The selected Gsp genes from *G. okellyi* and *N. gruberi* were amplified from
780 commercially synthesized templates (Genscript) (for primers used for PCR

781 amplification of the coding sequences see Supplementary Table 7) and cloned into
782 the pUG35 vector. The constructs were introduced into *S. cerevisiae* strain YPH499
783 by lithium acetate/PEG method. The positive colonies grown on SD-URA plates were
784 incubated with MitoTracker Red CMXRos (Thermo Fisher Scientific) and observed
785 for GFP and MitoTracker fluorescence (using the same equipment as used for FISH,
786 see above). For bacterial protein expression, *N. gruberi* GspD, GspE, GspF and GspG
787 genes were amplified from commercially synthesized templates and cloned into
788 pET42b vector (for primers used for PCR amplification of the coding sequences, see
789 Supplementary Table 6). The constructs were introduced into chemically-competent
790 *E. coli* strain BL21(DE3) and their expression induced by 1 mM IPTG. The
791 recombinant proteins were purified under denaturing conditions on Ni-NTA
792 agarose (Qiagen). The purified proteins were used for rat immunization in an in-
793 house animal facility at Charles University.
794 The sera obtained were used for immunodetection of Gsp proteins in *N.*
795 *gruberi* cells. Briefly, cells were fixed for 5 min in methanol (-20°C) and
796 permeabilized for 5 min by acetone (-20°C). The slides were incubated in blocking
797 buffer (BB) (PBS supplemented by 0.25% BSA, 0.05% TWEEN® 20 and 0.25%
798 gelatin) for 1 hr at room temperature. The slides were incubated overnight at 4°C
799 with primary antibodies diluted in BB and washed three times in PBS for 10 min.
800 Slides were then incubated for 1 hr with an anti-rat antibody conjugated with
801 Alexa488 (Thermo Fisher Scientific) diluted in. After washing three times for 10 min
802 in PBS, the slides were mounted in VectaShield DAPI solution and observed as above.
803 For mitochondrial labelling, the cells were incubated with MitoTracker Red CMXRos
804 for 30 min before fixation.
805
806 ***In vitro* protein translation**
807 The *Go*GspD gene was amplified from the commercially synthesized template (for
808 primers used for PCR amplification of the coding sequences, see Supplementary
809 Table 6) and cloned into pDHFR vector provided in the PURExpress *In Vitro* Protein
810 Synthesis Kit (NEB). The 25 µl translation reaction contained 10 µL of solution A, 7.5
811 µL of solution B, 250 ng of pDHFR plasmid carrying *Go*GspD gene, 1 µL of an RNase
812 inhibitor (RNAsin, Promega), radioactively labelled ^{35}S-methionine, and 50 µg of
813 lecithin liposomes. The liposomes were prepared from a stock solution of soybean
814 L-α-lecithin in chloroform by evaporating the chloroform under a nitrogen flow,
815 resuspending the lipid film in dH_20, and subsequent sonication in a waterbath
816 sonicator. The translation reaction was incubated for 2 hr at 37°C and then
817 centrifuged for 45 min at 13,000 x *g*. The pellet was resuspended in 50 mM sodium
818 phosphate buffer (pH = 8) with 2 M urea, centrifuged, and then washed in clear 50
819 mM sodium phosphate buffer. The output was analyzed by Blue Native PAGE using
820 2% digitonin and NativePAGE Novex 4-16% Bis-Tris Protein Gel (Thermo Fisher
821 Scientific).
822
823 **Testing protein interactions using two-hybrid systems**
824 Bacterial two-hybrid system (B2H) analysis was performed as described in ref.[70].
825 Gsp genes were amplified for commercially synthesized DNA and cloned into pKT25
826 and pUT18c plasmids. *E. coli* strain DHT1 competent cells were co-transformed with

827 two plasmids with different combinations of Gsp genes. Co-transformants were
828 selected on LB plates with ampicillin (100 μg/mL) and kanamycin (25 μg/mL).
829 Colonies were grown at 30°C for 48 to 96 hr. From each plate three colonies were
830 picked, transferred to 1 mL of LB medium with ampicillin and kanamycin, and
831 grown overnight at 30°C with shaking. Next day precultures (0.25 mL) were
832 inoculated to 5 mL of LB medium with ampicillin, kanamycin and 1 mM IPTG.
833 Cultures were grown with shaking at 30°C to OD_{600} of about 1-1.5. Bacteria (0.5 mL)
834 were mixed with 0.5 mL of Z buffer and subjected to the β-galactosidase assay[71].
835 The yeast two-hybrid system (Y2H) was employed as described in ref.[72]. Cells
836 of *S. cerevisiae* strain AH109 were co-transformed with two plasmids (pGADT7,
837 pGBKT7) with different combinations of Gsp genes. Co-transformants were selected
838 on double-dropout SD-Leu/-Trp and triple-dropout SD-Leu/-Trp/-His plates. The
839 colonies were grown for a few days. Positive colonies from the triple dropout were
840 grown overnight at 30°C with shaking and then the serial dilution test was
841 performed on double- and triple-dropout plates.
842
843 **Data availability**
844 All newly reported sequences of Gsp and Gcp proteins are provided in
845 Supplementary Table 1 and were deposited at GenBank with accession numbers
846 ######. Other relevant data (e.g. multiple sequence alignments used for
847 phylogenetic analyses) are available from the authors upon request.
848
849 **References**
850
851 1. Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. The origin and diversification
852 of mitochondria. *Curr. Biol.* **27,** R1177–R1192 (2017).
853 2. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial
854 origin outside the sampled alphaproteobacteria. *Nature* **557,** 101–105 (2018).
855 3. Leger, M. M. *et al.* An ancestral bacterial division system is widespread in
856 eukaryotic mitochondria. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 10239–46 (2015).
857 4. Beech, P. L. Mitochondrial FtsZ in a chromophyte alga. *Science (80-.).* **287,**
858 1276–1279 (2000).
859 5. Natale, P., Brüser, T. & Driessen, A. J. M. Sec- and Tat-mediated protein
860 secretion across the bacterial cytoplasmic membrane—Distinct translocases
861 and mechanisms. *Biochim. Biophys. Acta - Biomembr.* **1778,** 1735–1756
862 (2008).
863 6. Costa, T. R. D. *et al.* Secretion systems in Gram-negative bacteria: structural
864 and mechanistic insights. *Nat. Rev. Microbiol.* **13,** 343–359 (2015).
865 7. Dolezal, P., Likic, V., Tachezy, J. & Lithgow, T. Evolution of the molecular
866 machines for protein import into mitochondria. *Science* **313,** 314–8 (2006).
867 8. Palmer, T. & Berks, B. C. The twin-arginine translocation (Tat) protein export
868 pathway. *Nat. Rev. Microbiol.* **10,** 483–96 (2012).
869 9. Lang, B. F. *et al.* An ancestral mitochondrial DNA resembling a eubacterial
870 genome in miniature. *Nature* **387,** 493–7 (1997).
871 10. Burger, G., Gray, M. W., Forget, L. & Lang, B. F. Strikingly bacteria-like and
872 gene-rich mitochondrial genomes throughout jakobid protists. *Genome Biol.*

873 *Evol.* **5,** 418–38 (2013).
874 11. Tong, J. *et al.* Ancestral and derived protein import pathways in the
875 mitochondrion of *Reclinomonas america*. *Mol. Biol. Evol.* **28,** 1581–91 (2011).
876 12. Korotkov, K. V., Sandkvist, M. & Hol, W. G. J. The type II secretion system:
877 biogenesis, molecular architecture and mechanism. *Nat. Rev. Microbiol.* **10,**
878 336–51 (2012).
879 13. Thomassin, J.-L., Santos Moreno, J., Guilvout, I., Tran Van Nhieu, G. & Francetic,
880 O. The trans-envelope architecture and function of the type 2 secretion
881 system: new insights raising new questions. *Mol. Microbiol.* **105,** 211–226
882 (2017).
883 14. Berry, J.-L. & Pelicic, V. Exceptionally widespread nanomachines composed of
884 type IV pilins: the prokaryotic Swiss Army knives. *FEMS Microbiol. Rev.* **39,**
885 134–54 (2015).
886 15. Nivaskumar, M. & Francetic, O. Type II secretion system: a magic beanstalk or
887 a protein escalator. *Biochim. Biophys. Acta* **1843,** 1568–77 (2014).
888 16. Peabody, C. R. *et al.* Type II protein secretion and its relationship to bacterial
889 type IV pili and archaeal flagella. *Microbiology* **149,** 3051–72 (2003).
890 17. d'Enfert, C., Reyss, I., Wandersman, C. & Pugsley, A. P. Protein secretion by
891 gram-negative bacteria. Characterization of two membrane proteins required
892 for pullulanase secretion by *Escherichia coli* K-12. *J. Biol. Chem.* **264,** 17462–8
893 (1989).
894 18. Wang, X. *et al.* Cysteine scanning mutagenesis and disulfide mapping analysis
895 of arrangement of GspC and GspD protomers within the type 2 secretion
896 system. *J. Biol. Chem.* **287,** 19082–93 (2012).
897 19. Abby, S. S. *et al.* Identification of protein secretion systems in bacterial
898 genomes. *Sci. Rep.* **6,** 23080 (2016).
899 20. López-Castilla, A. *et al.* Structure of the calcium-dependent type 2 secretion
900 pseudopilus. *Nat. Microbiol.* **2,** 1686–1695 (2017).
901 21. Adl, S. M. *et al.* The revised classification of eukaryotes. *J. Eukaryot. Microbiol.*
902 **59,** 429–514 (2012).
903 22. Derelle, R. *et al.* Bacterial proteins pinpoint a single eukaryotic root. *Proc. Natl.*
904 *Acad. Sci. U. S. A.* **112,** E693-9 (2015).
905 23. Karnkowska, A. *et al.* A eukaryote without a mitochondrial organelle. *Curr.*
906 *Biol.* **26,** 1274–84 (2016).
907 24. Heiss, A. A. *et al.* Combined morphological and phylogenomic re-examination
908 of malawimonads, a critical taxon for inferring the evolutionary history of
909 eukaryotes. *R. Soc. Open Sci.* **5,** 171707 (2018).
910 25. Brown, M. W. *et al.* Phylogenomics places orphan protistan lineages in a novel
911 eukaryotic super-group. *Genome Biol. Evol.* **10,** 427–433 (2018).
912 26. Nouwen, N. *et al.* Secretin PulD: Association with pilot PulS, structure, and
913 ion-conducting channel formation. *Proc. Natl. Acad. Sci.* **96,** 8173–8177
914 (1999).
915 27. Hardie, K. R., Lory, S. & Pugsley, A. P. Insertion of an outer membrane protein
916 in *Escherichia coli* requires a chaperone-like protein. *EMBO J.* **15,** 978–88
917 (1996).
918 28. Korotkov, K. V, Pardon, E., Steyaert, J. & Hol, W. G. J. Crystal structure of the N-

919 terminal domain of the secretin GspD from ETEC determined with the
920 assistance of a nanobody. *Structure* **17,** 255–265 (2009).
921 29. Yin, M., Yan, Z. & Li, X. Structural insight into the assembly of the Type II
922 secretion system pilotin-Secretin complex from enterotoxigenic *Escherichia*
923 *coli*. *Nat Microbiol* (2018). doi:10.2210/PDB5ZDH/PDB
924 30. Guilvout, I. *et al.* In vitro multimerization and membrane insertion of bacterial
925 outer membrane secretin PulD. *J. Mol. Biol.* **382,** 13–23 (2008).
926 31. Nivaskumar, M. *et al.* Pseudopilin residue E5 is essential for recruitment by
927 the type 2 secretion system assembly platform. *Mol. Microbiol.* **101,** 924–41
928 (2016).
929 32. Korotkov, K. V. *et al.* Structural and functional studies on the interaction of
930 GspC and GspD in the type II secretion system. *PLoS Pathog.* **7,** e1002228
931 (2011).
932 33. Alva, V., Nam, S.-Z., Söding, J. & Lupas, A. N. The MPI bioinformatics Toolkit as
933 an integrative platform for advanced protein sequence and structure analysis.
934 *Nucleic Acids Res.* **44,** W410–W415 (2016).
935 34. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The
936 Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.*
937 **10,** 845–858 (2015).
938 35. Yan, Z., Yin, M., Xu, D., Zhu, Y. & Li, X. Structural insights into the secretin
939 translocation channel in the type II secretion system. *Nat. Struct. Mol. Biol.* **24,**
940 177–183 (2017).
941 36. Nguyen, B. D. & Valdivia, R. H. Virulence determinants in the obligate
942 intracellular pathogen *Chlamydia trachomatis* revealed by forward genetic
943 approaches. *Proc. Natl. Acad. Sci.* **109,** 1263–1268 (2012).
944 37. Tabach, Y. *et al.* Identification of small RNA pathway genes using patterns of
945 phylogenetic conservation and divergence. *Nature* **493,** 694–698 (2012).
946 38. Nevers, Y. *et al.* Insights into ciliary genes and evolution from multi-level
947 phylogenetic profiling. *Mol. Biol. Evol.* **34,** 2016–2034 (2017).
948 39. Okuno, D., Iino, R. & Noji, H. Rotation and structure of FoF1-ATP synthase. *J.*
949 *Biochem.* **149,** 655–664 (2011).
950 40. Tomko, R. J. & Hochstrasser, M. Molecular architecture and assembly of the
951 eukaryotic proteasome. *Annu. Rev. Biochem.* **82,** 415–445 (2013).
952 41. Babbitt, S. E., Sutherland, M. C., San Francisco, B., Mendez, D. L. & Kranz, R. G.
953 Mitochondrial cytochrome c biogenesis: no longer an enigma. *Trends Biochem.*
954 *Sci.* **40,** 446–55 (2015).
955 42. Hartl, F. U., Schmidt, B., Wachter, E., Weiss, H. & Neupert, W. Transport into
956 mitochondria and intramitochondrial sorting of the Fe/S protein of ubiquinol-
957 cytochrome c reductase. *Cell* **47,** 939–51 (1986).
958 43. Francisco, T. *et al.* Protein transport into peroxisomes: Knowns and
959 unknowns. *BioEssays* **39,** 1700047 (2017).
960 44. Keeling, P. J. *et al.* The marine microbial eukaryote transcriptome sequencing
961 project (MMETSP): illuminating the functional diversity of eukaryotic life in
962 the oceans through transcriptome sequencing. *PLoS Biol.* **12,** e1001889
963 (2014).
964 45. Matasci, N. *et al.* Data access for the 1,000 Plants (1KP) project. *Gigascience* **3,**

965 17 (2014).
966 46. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein
967 database search programs. *Nucleic Acids Res.* **25,** 3389–402 (1997).
968 47. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence
969 similarity searching. *Nucleic Acids Res.* **39,** W29-37 (2011).
970 48. Huerta-Cepas, J., Dopazo, H., Dopazo, J. & Gabaldón, T. The human phylome.
971 *Genome Biol.* **8,** R109 (2007).
972 49. Fukasawa, Y. *et al.* MitoFates: improved prediction of mitochondrial targeting
973 sequences and their cleavage sites. *Mol. Cell. Proteomics* **14,** 1113–26 (2015).
974 50. Neuberger, G., Maurer-Stroh, S., Eisenhaber, B., Hartig, A. & Eisenhaber, F.
975 Prediction of peroxisomal targeting signal 1 containing proteins from amino
976 acid sequence. *J. Mol. Biol.* **328,** 581–92 (2003).
977 51. Käll, L., Krogh, A. & Sonnhammer, E. L. L. Advantages of combined
978 transmembrane topology and signal peptide prediction--the Phobius web
979 server. *Nucleic Acids Res.* **35,** W429-32 (2007).
980 52. Finn, R. D. *et al.* The Pfam protein families database: towards a more
981 sustainable future. *Nucleic Acids Res.* **44,** D279-85 (2016).
982 53. de Lima Morais, D. A. *et al.* SUPERFAMILY 1.75 including a domain-centric
983 gene ontology method. *Nucleic Acids Res.* **39,** D427-34 (2011).
984 54. Frickey, T. & Lupas, A. CLANS: a Java application for visualizing protein
985 families based on pairwise similarity. *Bioinformatics* **20,** 3702–3704 (2004).
986 55. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and
987 high throughput. *Nucleic Acids Res.* **32,** 1792–7 (2004).
988 56. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software
989 version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,**
990 772–780 (2013).
991 57. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with
992 Entropy): a new software for selection of phylogenetic informative regions
993 from multiple sequence alignments. *BMC Evol. Biol.* **10,** 210 (2010).
994 58. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and
995 effective stochastic algorithm for estimating maximum-likelihood phylogenies.
996 *Mol. Biol. Evol.* **32,** 268–74 (2015).
997 59. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site
998 heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21,**
999 1095–109 (2004).
1000 60. Bienert, S. *et al.* The SWISS-MODEL Repository-new features and functionality.
1001 *Nucleic Acids Res.* **45,** D313–D319 (2017).
1002 61. Guex, N., Peitsch, M. C. & Schwede, T. Automated comparative protein
1003 structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical
1004 perspective. *Electrophoresis* **30,** S162–S173 (2009).
1005 62. Biasini, M. *et al.* SWISS-MODEL: modelling protein tertiary and quaternary
1006 structure using evolutionary information. *Nucleic Acids Res.* **42,** W252-8
1007 (2014).
1008 63. Jedelský, P. L. *et al.* The minimal proteome in the reduced mitochondrion of
1009 the parasitic protist *Giardia intestinalis*. *PLoS One* **6,** e17285 (2011).
1010 64. Černá, M., Kuntová, B., Talacko, P., Stopková, R. & Stopka, P. Differential

1011 regulation of vaginal lipocalins (OBP, MUP) during the estrous cycle of the
1012 house mouse. *Sci. Rep.* **7,** 11674 (2017).
1013 65. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed
1014 normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell.*
1015 *Proteomics* **13,** 2513–2526 (2014).
1016 66. Tyanova, S. *et al.* The Perseus computational platform for comprehensive
1017 analysis of (prote)omics data. *Nat. Methods* **13,** 731–740 (2016).
1018 67. Dunkley, T. P. J., Watson, R., Griffin, J. L., Dupree, P. & Lilley, K. S. Localization
1019 of organelle proteins by isotope tagging (LOPIT). *Mol. Cell. Proteomics* **3,**
1020 1128–1134 (2004).
1021 68. Vizcaíno, J. A. *et al.* 2016 update of the PRIDE database and its related tools.
1022 *Nucleic Acids Res.* **44,** 11033–11033 (2016).
1023 69. Zubáčová, Z., Krylov, V. & Tachezy, J. Fluorescence in situ hybridization (FISH)
1024 mapping of single copy genes on *Trichomonas vaginalis* chromosomes. *Mol.*
1025 *Biochem. Parasitol.* **176,** 135–137 (2011).
1026 70. Battesti, A. & Bouveret, E. The bacterial two-hybrid system based on
1027 adenylate cyclase reconstitution in *Escherichia coli*. *Methods* **58,** 325–334
1028 (2012).
1029 71. Miller, J. H. *Experiments in molecular genetics*. (Cold Spring Harbor Laboratory,
1030 1972).
1031 72. Fields, S. & Song, O. A novel genetic system to detect protein-protein
1032 interactions. *Nature* **340,** 245–246 (1989).
1033
1034
1035 **Acknowledgements**
1036 We would like to thank Michelle Leger and Alastair Simpson for granting us an
1037 access, prior to publication, to their transcriptomic data from *A. godoyi* and *G. okellyi*,
1038 respectively, which were instrumental in annotating the Gsp and Gcp genes in our
1039 genome assemblies. This work was supported by Czech Science Foundation grants
1040 13-29423S to PD and 18-18699S to ME and the KONTAKT II grant LH15253
1041 provided by Ministry of Education, Youth and Sports of CR (MEYS) to PD. This work
1042 was also supported by MEYS within the National Sustainability Program II (Project
1043 BIOCEV-FAR, LQ1604) the project BIOCEV (CZ.1.05/1.1.00/02.0109), and the
1044 project "Centre for research of pathogenicity and virulence of parasites" (No.
1045 CZ.02.1.01/0.0/0.0/16_019/0000759) funded by European Regional Development
1046 Fund (ERDF) and MEYS. The work in OF laboratory was funded by the ANR-14-
1047 CE09-0004 grant. JP was supported by a grant from the Gordon and Betty Moore
1048 Foundation to ADT. This work was supported by The Ministry of Education, Youth
1049 and Sports from the Large Infrastructures for Research, Experimental Development
1050 and Innovations project „IT4Innovations National Supercomputing Center –
1051 LM2015070".
1052
1053
1054 **Author information**
1055
1056 **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

1103 and analysed the data, A.M. carried out the experiments, Ve.K. carried out the
1104 experiments, Vl.K. participated in genome sequencing and analysis, M.P. planned
1105 carried out the experiments, I.Č. participated in genome data acquisition, Z.V.
1106 carried out the experiments, K.H. analysed the proteome of *N. gruberi* mitochondria,
1107 M.W.G. contributed to the interpretation of the results and manuscript preparation,
1108 I.G. designed and planned the experiments, O.F. designed, planned and carried out
1109 the experiments and analyzed the data, B.F.L. provided the genome data and
1110 analyses and contributed to manuscript preparation, Č.V. participated in genome
1111 data acquisition, A.D.T. designed and planned the experiments, M.E. conceived the
1112 idea, performed genomic analyses and wrote the manuscript, P.D. conceived the
1113 idea, designed and performed experiments and wrote the manuscript.
1114
1115
1116 **Competing interests**
1117 None declared.
1118
1119 **Corresponding author**
1120 Correspondence to M. Eliáš or P. Doležal.
1121
1122
1123
1124
1125 **Figure Legends**
1126 **Fig. 1** Some eukaryotes harbour homologues of core components of the bacterial
1127 T2SS machinery. (A) Schematic representation of the complete bacterial T2SS;
1128 subunits having identified eukaryotic homologues are highlighted in colour. (B)
1129 Phylogenetic distribution of eukaryotic homologues of bacterial T2SS subunits (Gsp
1130 proteins) and co-occurring proteins (Gcp). Core T2SS components (cyan),
1131 eukaryote-specific T2SS components (dark blue), Gcp proteins carrying protein
1132 domains found in eukaryotes (magenta), and Gcp proteins without discernible
1133 homologues or with homologues only in prokaryotes (orange). Coloured sections
1134 indicate proteins found to be present in genome or transcriptome data; white
1135 sections, proteins absent from complete genome data; grey sections, proteins absent
1136 from transcriptome data. The asterisk indicates the presence of the particular
1137 protein in at least two of three species of *Percolomonas* analyzed. The two species
1138 names in parentheses have not been yet been formally published. Sequence IDs and
1139 additional details on the eukaryotic Gsp and Gcp proteins are provided in
1140 Supplementary Table 1. (C) Maximum likelihood (ML) phylogenetic tree of
1141 eukaryotic and selected bacterial GspF proteins demonstrating the monophyletic
1142 origin of the eukaryotic GspF proteins and their separation from bacterial
1143 homologues by a long branch (the tree inferred using IQ-TREE). Branch support
1144 (bootstrap / posterior probability values) was assessed by ML ultrafast
1145 bootstrapping and is shown only for branches where > 50.
1146
1147 **Fig. 2** Eukaryotic T2SS components are localized in mitochondria. (A) *N. gruberi*
1148 cells labelled with specific polyclonal antibodies raised against GspD, GspF and

1149 GspG1, and co-stained with MitoTracker red CMX ROS show mitochondrial
1150 localization of the proteins; scale bar, 10 μm. (B) *S. cerevisiae* expressing *G. okellyi*
1151 T2SS components as C-terminal GFP fusions co-stained with MitoTracker red CMX
1152 ROS; scale bar, 10 μm.
1153
1154 **Fig. 3** Analysis of the *N. gruberi* mitochondrial proteome. PCA analysis of 4198
1155 proteins identified in the proteomic analysis of *N. gruberi* mitochondria. The cluster
1156 of mitochondrial proteins was defined on the basis of 376 mitochondrial markers.
1157 The boundaries of the cluster of co-purified peroxisomal proteins were defined by
1158 26 peroxisomal markers.
1159
1160 **Fig. 4.** Mitochondrial GspD oligomerizes towards the formation of membrane pores.
1161 (A) Domain architecture of the canonical bacterial GspD protein and eukaryotic
1162 proteins homologous to its different parts. (B) Structural model of *Go*GspD built by
1163 ProMod3 on the *Vibrio cholerae* GspD template. Top and side view of a cartoon and a
1164 transparent surface representation of the *Go*GspD pentadecamer model is shown in
1165 blue. The amphipathic helical loop (AHL), the signature of the secretin family, is
1166 highlighted and coloured according to the secondary structure with strands in
1167 magenta, helices in cyan and loops in light brown. The C-terminal GpsD residues are
1168 highlighted as spheres. The detailed view of the AHL region shows the essential
1169 residues V162 and F166 pointing towards the membrane surface. (C) Expression of
1170 the mitochondrial *Go*GspD quickly induces cell death in bacteria. (D) Y2H assay
1171 shows the self-interaction of the mitochondrial *Go*GspD. (E) *In vitro* translation and
1172 assembly of mitochondrial *Go*GspD into a high-molecular-weight complex; lipo –
1173 liposomes added, urea – extraction by 2M urea. (F) Y2H assay suggests the
1174 interaction of *Ng*GspDN1 with itself and with *Ng*GspD.
1175
1176 **Fig. 5** Structure, maturation, and interactions of the mitochondrial GspG. (A)
1177 Domain architecture of the bacterial and the mitochondrial pseudopilin GspG. The
1178 arrow indicates the processing site of the bacterial GspG during protein maturation.
1179 MTS – mitochondria targeting sequence, + – polar anchor, TMD – transmembrane
1180 domain. (B) Positive interactions between the mitochondrial GspG protein and other
1181 T2SS subunits were determined by the B2H assays. (C) Peptides specific to *Ng*GspG1
1182 retrieved from the proteomic analysis of *N. gruberi* mitochondria. The arrow
1183 indicates the position of the processing site of bacterial GspG proteins. (D)
1184 Immunodetection of *Ng*GspG1 in *N. gruberi* cellular fractions. The arrow marks the
1185 *Ng*GspG1-specific band.
1186
1187 **Fig. 6** Proteins with the same phylogenetic profile as the originally identified
1188 mitochondrial Gsp homologues. (A) Schematic domain representation of 23 proteins
1189 occurring in heteroloboseans, jakobids and malawimonads with the core T2SS
1190 subunits but not in other eukaryotes analyzed. Proteins with a functional link to the
1191 T2SS suggested by sequence homology are shown in royal blue, proteins
1192 representing novel paralogues within broader (super)families are shown in red, and
1193 proteins without discernible homologues or with homologues only in prokaryotes
1194 are shown in yellow. The presence of conserved protein domains or characteristic

1195 structural motifs is shown if detected in the given protein. Grey block – predicted
1196 transmembrane domain (see also Supplementary Fig. 10); "C H C H" – the presence
1197 of absolutely conserved cysteine and histidine residues (see also Supplementary Fig.
1198 11) that may mediate binding of a prosthetic group. The length of the rectangles
1199 corresponds to the relative size of the proteins. (B) Evolutionary relationships
1200 among Gcp1 to Gcp3 proteins and other members of the WD40 superfamily. The
1201 schematic phylogenetic tree was drawn on the basis of a ML phylogenetic tree
1202 available as Supplementary Fig. 8.
1203
1204 **Fig. 7** A hypothetical novel eukaryotic functional pathway including a mitochondrial
1205 version of the T2SS (miT2SS) and connecting the mitochondrion with the
1206 peroxisome. A nucleus-encoded protein (magenta) is imported via the TOM complex
1207 into the mitochondrial inner membrane space, where it is modified by addition of a
1208 specific prosthetic group catalysed by certain Gcp proteins. After folding it becomes
1209 a substrate of the miT2SS machinery and is exported from the mitochondrion.
1210 Finally it is imported into the peroxisome by the action of a dedicated import system
1211 including other Gcp proteins. OMM – outer mitochondrial membrane, IMS –
1212 intermembrane space, IMM – inner mitochondrial membrane, MM – mitochondrial
1213 matrix.
1214

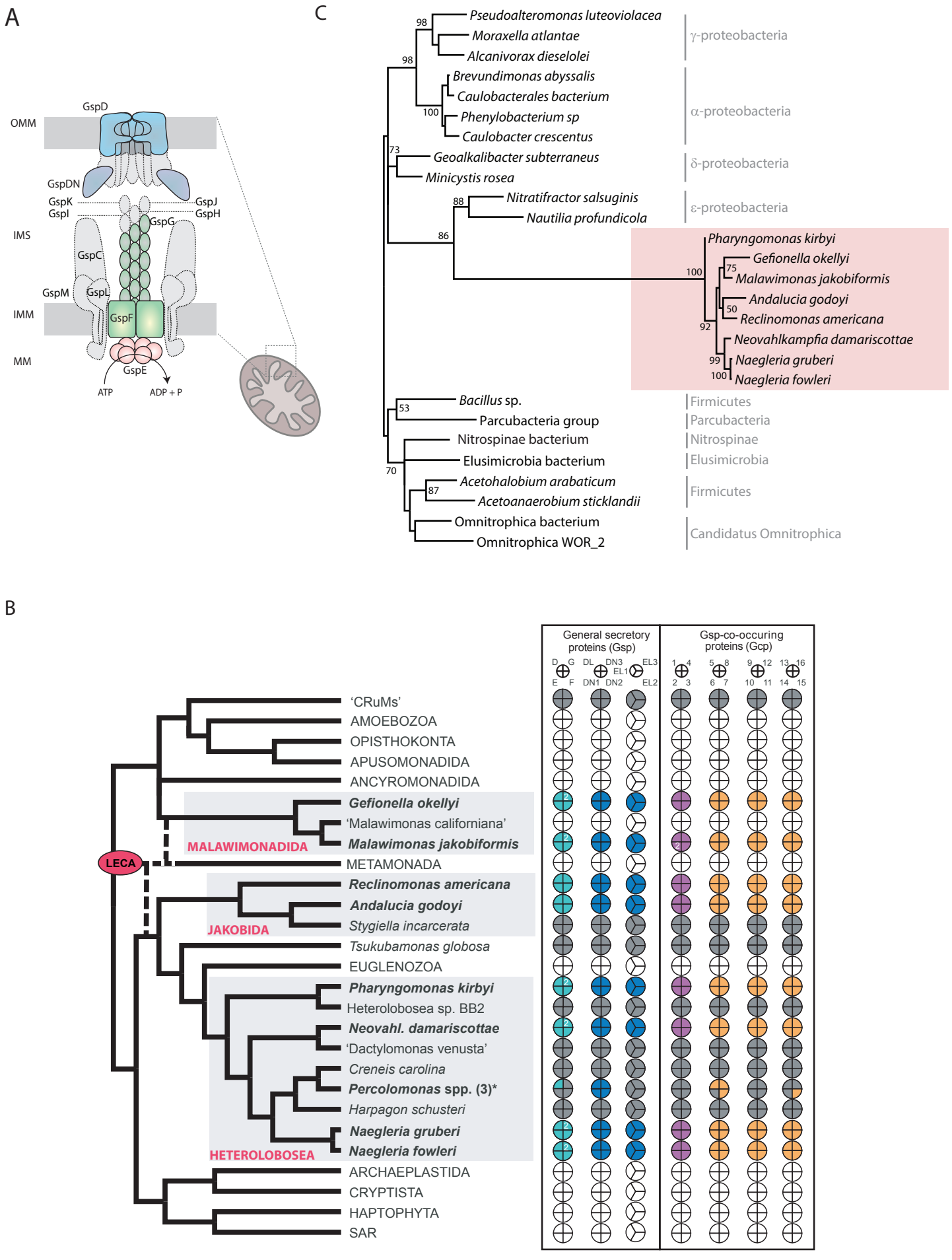
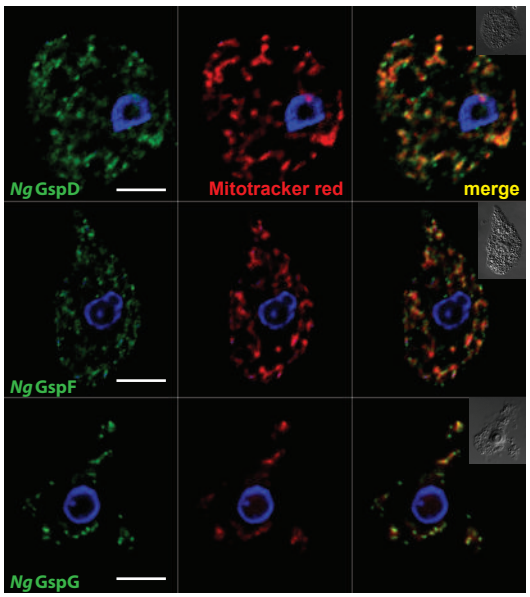

A
OMM
GspD
GspDN
GspK
GspI
GspJ
GspH
GspG
IMS
GspC
GspM
GspL
IMM
GspF
MM
GspE
ATP
ADP + P
C
Pseudoalteromonas luteoviolacea
Moraxella atlantae
Alcanivorax dieselolei
Brevundimonas abyssalis
Caulobacterales bacterium
Phenylobacterium sp
Caulobacter crescentus
Geoalkalibacter subterraneus
Minicystis rosea
Nitratifractor salsuginis
Nautilia profundicola
Pharyngomonas kirbyi
Gefionella okellyi
Malawimonas jakobiformis
Andalucia godoyi
Reclinomonas americana
Neovahlkampfia damariscottae
Naegleria gruberi
Naegleria fowleri
Bacillus sp.
Parcubacteria group
Nitrospinae bacterium
Elusimicrobia bacterium
Acetohalobium arabaticum
Acetoanaerobium sticklandii
Omnitrophica bacterium
Omnitrophica WOR_2
γ-proteobacteria
α-proteobacteria
δ-proteobacteria
ε-proteobacteria
Firmicutes
Parcubacteria
Nitrospinae
Elusimicrobia
Firmicutes
Candidatus Omnitrophica
B
LECA
'CRuMs'
AMOEBOZOA
OPISTHOKONTA
APUSOMONADIDA
ANCYROMONADIDA
Gefionella okellyi
'Malawimonas californiana'
Malawimonas jakobiformis
MALAWIMONADIDA
METAMONADA
Reclinomonas americana
Andalucia godoyi
Stygiella incarcerata
JAKOBIDA
Tsukubamonas globosa
EUGLENOZOA
Pharyngomonas kirbyi
Heterolobosea sp. BB2
Neovahl. damariscottae
'Dactylomonas venusta'
Creneis carolina
Percolomonas spp. (3)*
Harpagon schusteri
Naegleria gruberi
Naegleria fowleri
HETEROLOBOSEA
ARCHAEPLASTIDA
CRYPTISTA
HAPTOPHYTA
SAR
General secretory proteins (Gsp)
Gsp-co-occuring proteins (Gcp)

Figure 1

A


Ng GspD
Mitotracker red
merge
Ng GspF
Ng GspG

B

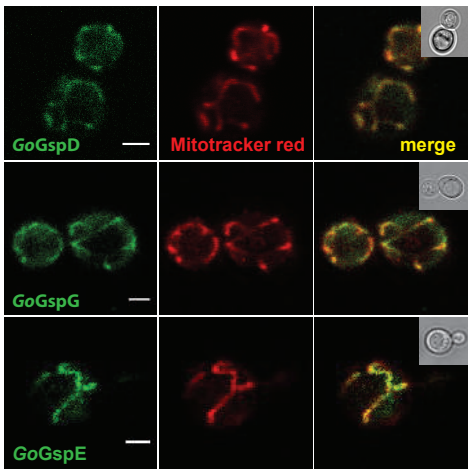

GoGspD
Mitotracker red
merge
GoGspG
GoGspE

Figure 2

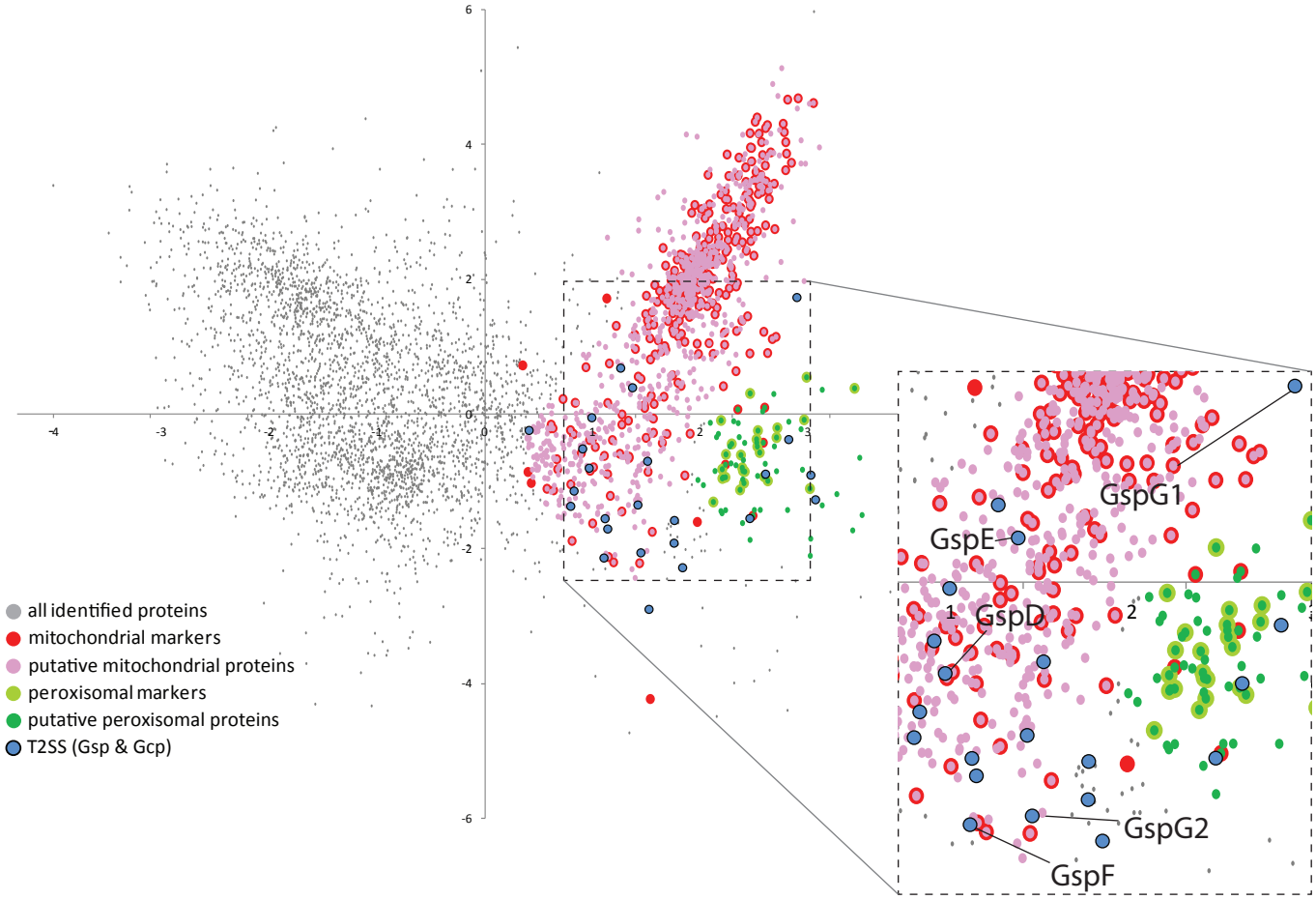

all identified proteins
mitochondrial markers
putative mitochondrial proteins
peroxisomal markers
putative peroxisomal proteins
T2SS (Gsp & Gcp)
GspG1
GspE
GspD
GspG2
GspF

Figure 3

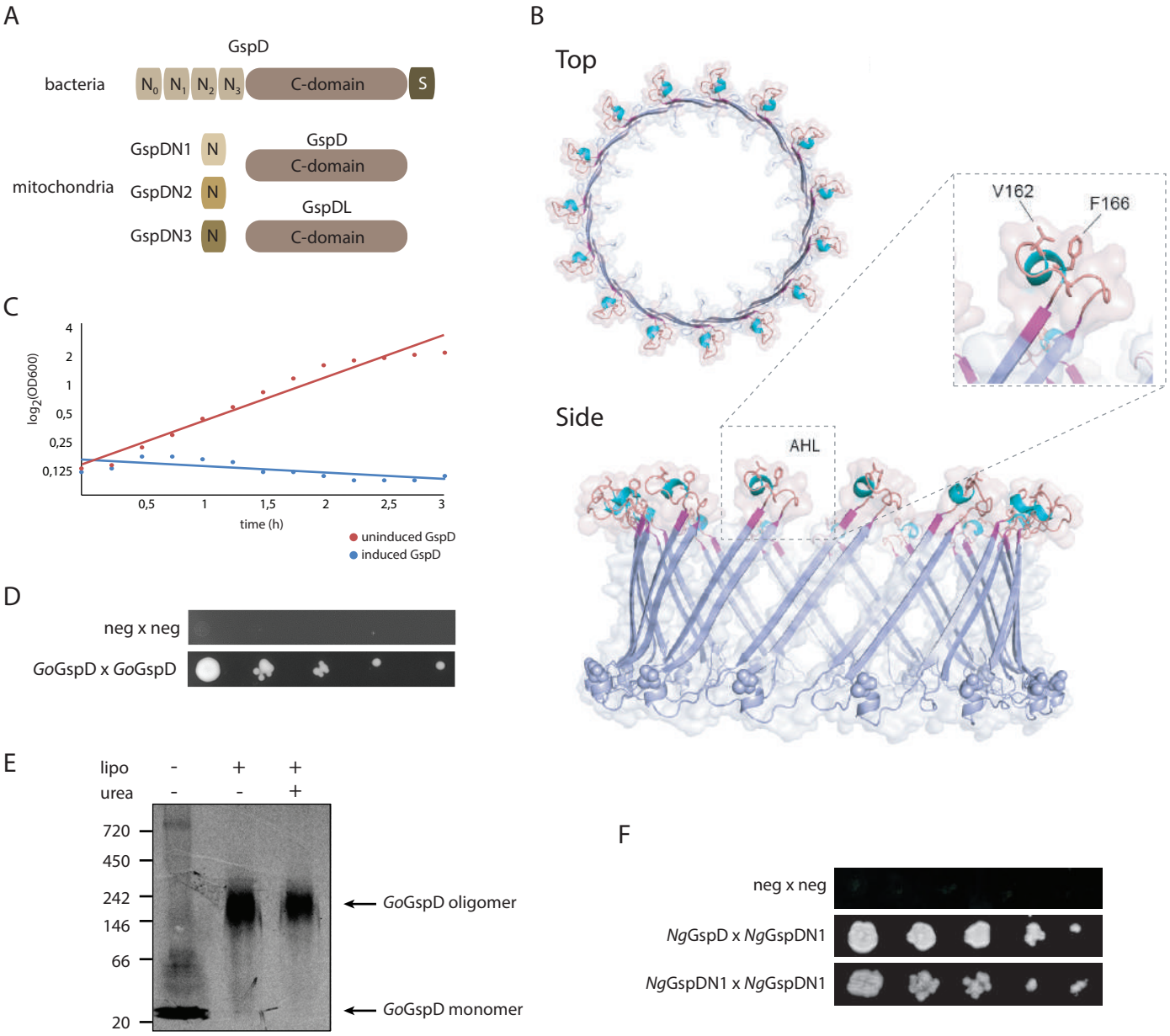

A
GspD
bacteria
N0 N1 N2 N3
C-domain
S
mitochondria
GspDN1
N
GspD
C-domain
GspDN2
N
GspDL
GspDN3
N
C-domain
B
Top
V162
F166
Side
AHL
C
log2(OD600)
4
2
1
0,5
0,25
0,125
0,5
1
1,5
2
2,5
3
time (h)
uninduced GspD
induced GspD
D
neg x neg
GoGspD x GoGspD
E
lipo - + +
urea - - +
720
450
242
146
66
20
GoGspD oligomer
GoGspD monomer
F
neg x neg
NgGspD x NgGspDN1
NgGspDN1 x NgGspDN1

Figure 4

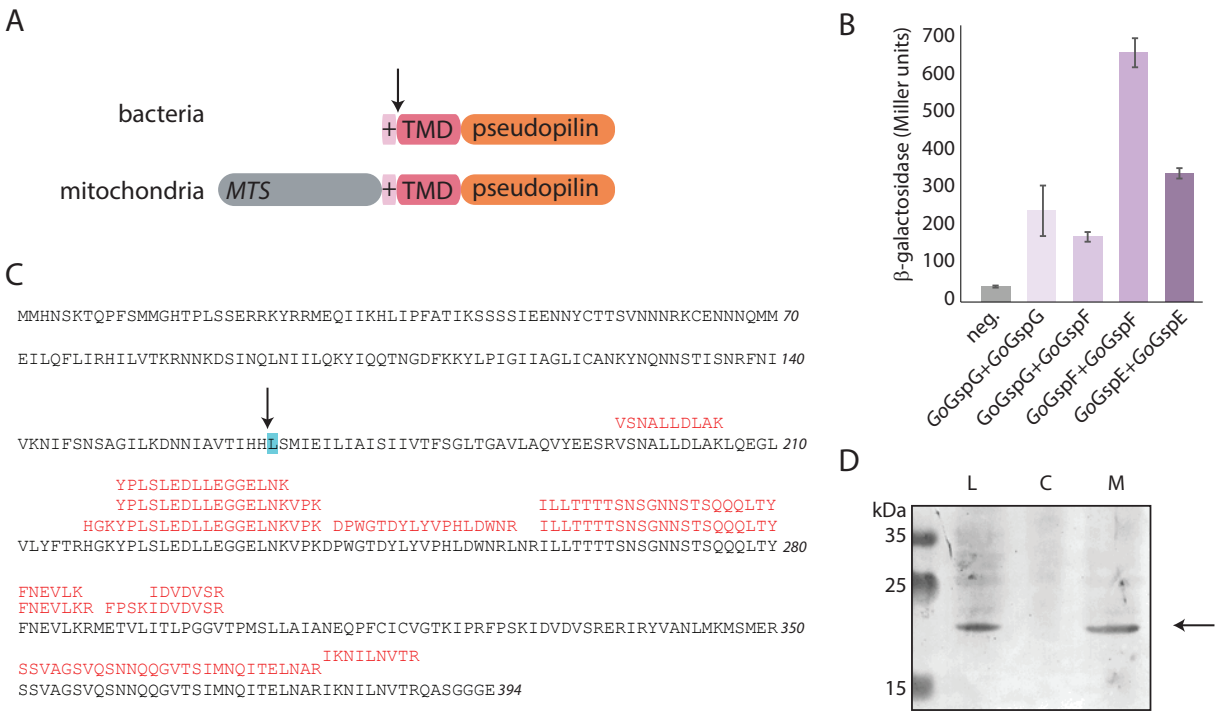

A
bacteria
+ TMD pseudopilin
mitochondria MTS + TMD pseudopilin
B
β-galactosidase (Miller units)
700
600
500
400
300
200
100
0
neg.
GoGspG+GoGspG
GoGspG+GoGspF
GoGspF+GoGspF
GoGspE+GoGspE
C
MMHNSKTQPFSMMGHTPLSSERRKYRRMEQIIKHLIPFATIKSSSSIEENNYCTTSVNNNRKCENNNQMM 70
EILQFLIRHILVTKRNNKDSINQLNIILQKYIQQTNGDFKKYLPIGIIAGLICANKYNQNNSTISNRFNI 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

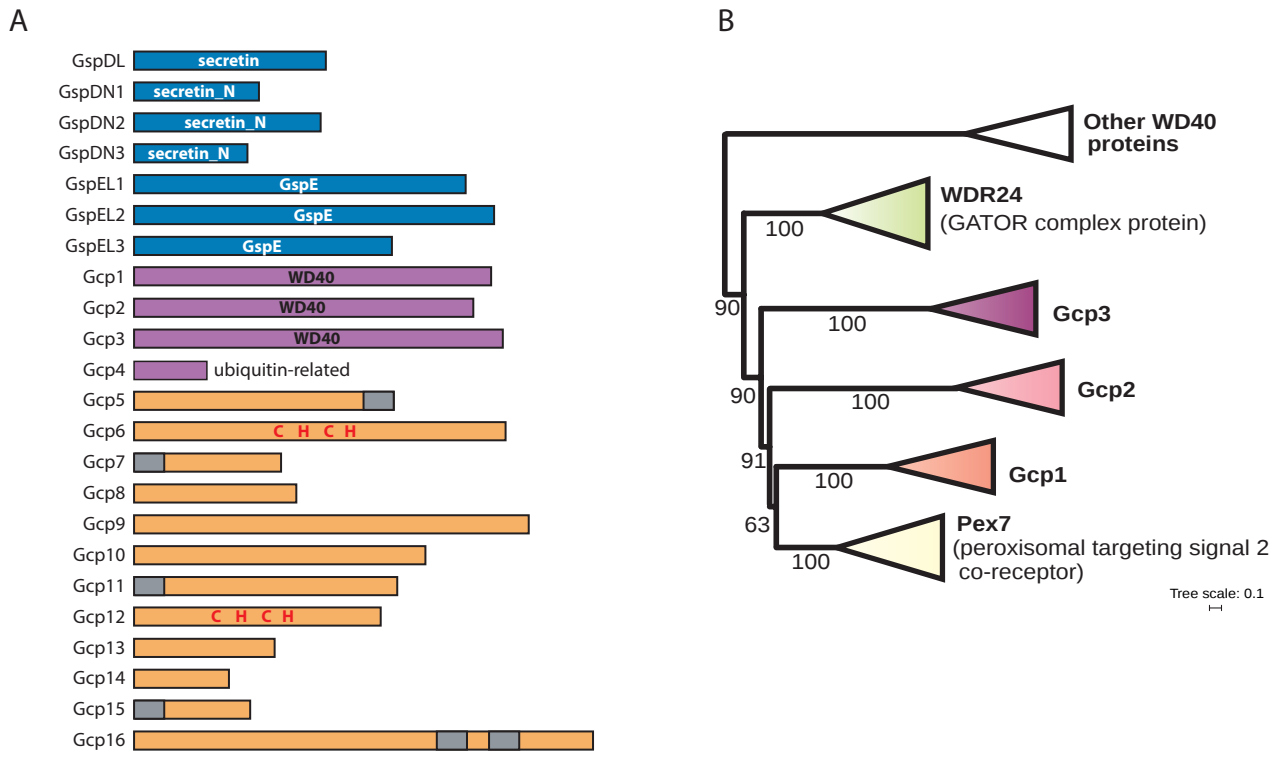

A
GspDL
secretin
GspDN1
secretin_N
GspDN2
secretin_N
GspDN3
secretin_N
GspEL1
GspE
GspEL2
GspE
GspEL3
GspE
Gcp1
WD40
Gcp2
WD40
Gcp3
WD40
Gcp4
ubiquitin-related
Gcp5
Gcp6
C H C H
Gcp7
Gcp8
Gcp9
Gcp10
Gcp11
Gcp12
C H C H
Gcp13
Gcp14
Gcp15
Gcp16
B
Other WD40 proteins
WDR24
(GATOR complex protein)
100
90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

Figure 6

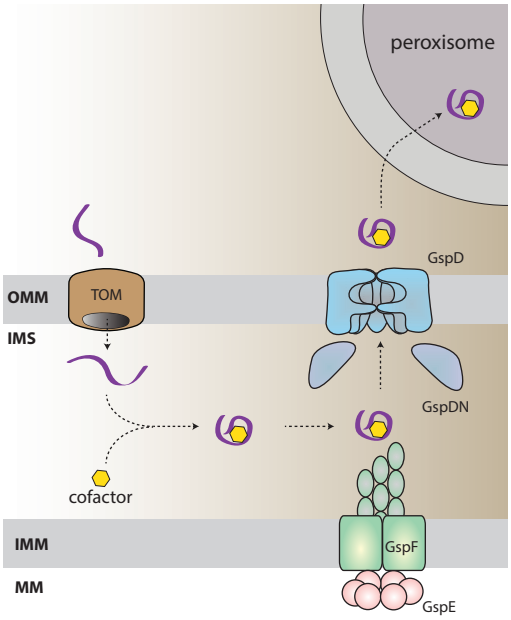


Figure 7