Analysis of diverse eukaryotes suggests the existence of an ancestral mitochondrial apparatus derived from the bacterial type II secretion system.

Lenka Horváthová[1*], Vojtěch Žárský[1*], Tomáš Pánek[2*$], Romain Derelle[3], Jan Pyrih[4], Alžběta Motyčková[1], Veronika Klápšťová[1], Martina Vinopalová[1], Lenka Marková[1], Luboš Voleman[1], Vladimír Klimeš[2], Markéta Petrů[1], Zuzana Vaitová[1], Ivan Čepička[5], Klára Hryzáková[6], Karel Harant[7], Michael W. Gray[8], Mohamed Chami[9], Ingrid Guilvout[10], Olivera Francetic[10], B. Franz Lang[11], Čestmír Vlček[12], 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]Department of Genetics and Microbiology, Faculty of Science, Charles University, Viničná 7, Prague 2, 128 44, Czech Republic[7]Proteomic core facility, Faculty of Science, Charles University, BIOCEV, Průmyslová 595, Vestec, 252 42, Czech Republic
[8]Department of Biochemistry and Molecular Biology and Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, NS B3H 4R2, Canada
[9]Center for Cellular Imaging and NanoAnalytics, University of Basel, Mattenstrasse 26, CH-4058 Basel, Switzerland
[10]Institut Pasteur, Biochemistry of Macromolecular Interactions Unit, Department of Structural Biology and Chemistry, CNRS UMR3528, 75015 Paris, France
[11]Robert Cedergren Centre for Bioinformatics and Genomics, Département de Biochimie, Université de Montréal, Montreal, QC, Canada H3T 1J4
[12]Institute of Molecular Genetics, Czech Academy of Sciences, 142 20 Prague 4, Czech Republic
[$]Current address: Department of Zoology, Faculty of Science, Charles University, Viničná 7, Prague 2, 128 44, 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

The type 2 secretion system (T2SS) is present in some Gram-negative eubacteria and used to secrete proteins across the outer membrane. Here we report that certain representative heteroloboseans, jakobids, malawimonads and hemimastigotes unexpectedly possess homologues of core T2SS components. We show that at least some of them are present in mitochondria, and their behaviour in biochemical assays is consistent with the presence of a mitochondrial T2SS-derived system (miT2SS). We additionally identified 23 protein families co-occurring with miT2SS in eukaryotes. Seven of these proteins could be directly linked to the core miT2SS by functional data and/or sequence features, whereas others may represent different parts of a broader functional pathway, possibly also involving the peroxisome. Its distribution in eukaryotes and phylogenetic evidence together indicate that the miT2SS-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. Examples are the ancestral mitochondrial division apparatus (including homologues of bacterial Min proteins), the aerobic-type rubrerythrin system, or the tmRNA-SmpB complex, each retained in different subsets of distantly related protist lineages[3–5]. 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[6], and many diderm (Gram-negative) bacteria exhibit specialized systems mediating further protein translocation across the outer membrane (OM)[7]. 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[8]. 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[9,10]. A mitochondrial homologue of the SecY protein (a Sec translocase subunit) has been described only in jakobids[11,12] and its function remains elusive[13]. 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)[14,15]. 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[16–18]. The T2SS constitutes a specialized secretion apparatus, whose filament (pseudopilus) remains in the

93 periplasm[14,15]. It is composed of 12-15 conserved components, commonly referred to as
94 general secretion pathway (Gsp) proteins, which assemble into four main subcomplexes
95 (Fig. 1A). The OM pore is formed by an oligomer of 15-16 molecules of the GspD
96 protein[19,20]. The subcomplex in the inner membrane (IM) is called the assembly platform
97 and consists of the central polytopic membrane protein GspF surrounded by single-pass
98 membrane proteins GspC, GspL, and GspM[21]. GspC links the assembly platform to the
99 OM pore by interacting with the periplasmic N-terminal domain of GspD[22,23]. The third
100 subcomplex, called the pseudopilus, is a helical filament formed mainly of GspG
101 subunits, with minor pseudopilins (GspH, GspI, GspJ and GspK) assembled at its tip[24].
102 Pseudopilus assembly from its inner membrane base is believed to push the periplasmic
103 T2SS substrate through the OM pore. The energy for pseudopilus assembly is provided
104 by the fourth subcomplex, the hexameric ATPase GspE, interacting with the assembly
105 platform from the cytoplasmic side[25,26]. Substrates for T2SS-mediated secretion are first
106 transported by the Tat (as folded proteins) or the Sec (in an unfolded form) system across
107 the IM into the periplasm, where they undergo maturation and/or folding. The folded
108 substrates are finally loaded onto the pseudopilus for the release outside the cell via the
109 OM pore. The known T2SS substrates differ among taxa and share no common sequence
110 or structural features. Proteins transported by the T2SS in different species include
111 catabolic enzymes (such as lipases, proteases or phosphatases) and, in the case of
112 bacterial pathogens, toxins[14]. A recent survey of bacterial genomes showed that the T2SS
113 is mainly present in Proteobacteria[18]. Crucially, neither the T2SS nor other systems of the
114 T4P superfamily have been reported from eukaryotes[7,14,18].
115 Here we show that certain distantly related eukaryotes unexpectedly contain
116 homologues of key T2SS subunits representing all four functional T2SS subcomplexes.
117 We provide evidence for mitochondrial localization of some of these eukaryotic Gsp
118 protein homologues and describe experimental results supporting the idea that they
119 constitute a system similar to the bacterial T2SS. Furthermore, we point to the existence
120 of 23 proteins with a perfect taxonomic co-occurrence with the eukaryotic Gsp
121 homologues. Some of these co-occurring proteins seem to be additional components of
122 the mitochondrial T2SS-related machinery, whereas others are candidates for components
123 of a broader functional pathway linking the mitochondrion with other parts of the cell.
124 Given its phylogenetic distribution, we propose that the discovered pathway was
125 ancestrally present in eukaryotes. Its further characterization may provide fundamental
126 insights into the evolutionary conversion of the protomitochondrion into the
127 mitochondrial organelle.
128
129 **Results**
130 ***Certain protist lineages code for a conserved set of homologues of T2SS core***
131 ***components***
132 While searching the genome of the heterolobosean *Naegleria gruberi* for proteins of
133 bacterial origin with a possible mitochondrial role, we surprisingly discovered
134 homologues of four core subunits of the bacterial T2SS, specifically GspD, GspE, GspF,
135 and GspG (Fig. 1A; Supplementary Data 1). Using genomic and transcriptomic data from
136 public repositories and our on-going sequencing projects for several protist species of key
137 evolutionary interest (see Methods), we mapped the distribution of these four components
138 in eukaryotes. All four genes were found in the following characteristic set of taxa (Fig.

139 1B, Supplementary Data 1): three additional heteroloboseans (*Naegleria fowleri*,
140 *Neovahlkampfia damariscottae*, *Pharyngomonas kirbyi*), two jakobids (*R. americana* and
141 *Andalucia godoyi*), and two malawimonads (*Malawimonas jakobiformis* and *Gefionella*
142 *okellyi*). In addition, single-cell transcriptomes from two species of Hemimastigophora
143 (hemimastigotes[27]) revealed the presence of homologues of GspD and GspG (*Hemimastix*
144 *kukwesjijk*) and GspG only (*Spironema* cf. *mulitciliatum*), possibly reflecting
145 incompleteness of the data. Finally, three separate representatives of the heterolobosean
146 genus *Percolomonas* (Supplementary Fig. 1) each exhibited a homologue of GspD, but
147 not of the remaining Gsp proteins, in the available transcriptomic data. In contrast, all
148 four genes were missing in sequence data from all other eukaryotes investigated,
149 including the genome and transcriptome of another malawimonad ("Malawimonas
150 californiana") and deeply-sequenced transcriptomes of a third jakobid (*Stygiella*
151 *incarcerata*) and four additional heteroloboseans (*Creneis carolina*, *Dactylomonas*
152 *venusta*, *Harpagon schusteri*, and the undescribed strain Heterolobosea sp. BB2).
153 Probing *N. gruberi* nuclei with fluorescence *in situ* hybridization ruled out an
154 unidentified bacterial endosymbiont as the source of the *Gsp* genes (Supplementary Fig.
155 2). Moreover, the eukaryotic *Gsp* genes usually have introns and constitute robustly
156 supported monophyletic groups well separated from bacterial homologues (Fig. 2;
157 Supplementary Fig. 3), ruling out bacterial contamination in all cases. In an attempt to
158 illuminate the origin of the eukaryotic Gsp proteins we carried out systematic
159 phylogenetic analyses based on progressively expanded datasets of prokaryotic
160 homologues and for each tree inferred the taxonomic identity of the bacterial ancestor of
161 the eukaryotic branch (see Methods for details on the procedure). The results,
162 summarized in Supplementary Fig. 3, showed that the inference is highly unstable
163 depending on the dataset analysed, and no specific bacterial group can be identified as an
164 obvious donor of the eukaryotic Gsp genes. This result probably stems from a
165 combination of factors, including the long branches separating the eukaryotic and
166 bacterial Gsp sequences, the length of Gsp proteins restricting the amount of the
167 phylogenetic signal retained, and perhaps also rampant horizontal gene transfer (HGT) of
168 the T2SS system genes between bacterial taxa. The eukaryotic Gsp genes are in fact so
169 divergent that some of them could not be unambiguously classified as specific
170 homologues of T2SS components (rather than the related machineries of the T4P
171 superfamily) when analysed using models developed for the bacterial genomes[18]
172 (Supplementary Fig. 3).
173 Heteroloboseans, jakobids and malawimonads have been classified in the
174 supergroup Excavata[28]. However, recent phylogenomic analyses indicate that excavates
175 are non-monophyletic and even suggest that malawimonads are separated from
176 heteroloboseans and jakobids by the root of the eukaryote phylogeny[29–32]. Together with
177 the presence of at least some Gsp proteins in hemimastigotes, which constitute an
178 independent eukaryotic supergroup[27], it is likely that the T2SS-related proteins were
179 present in the last eukaryotic common ancestor (LECA) but lost in most eukaryote
180 lineages (Fig. 1B). Heteroloboseans and malawimonads have two GspG paralogues, but
181 the phylogenetic analyses did not resolve whether this is due to multiple independent
182 GspG gene duplications or one ancestral eukaryotic duplication followed by loss of one
183 of the paralogues in jakobids (Fig. 2; Supplementary Fig. 3D; Supplementary Data 1).
184

The eukaryotic Gsp proteins localize to the mitochondrion

We hypothesized that the eukaryotic homologues of the four Gsp proteins are parts of a functional T2SS-related system localized to the mitochondrion. This notion was supported by the presence of predicted N-terminal mitochondrial targeting sequences (MTSs) in some of the eukaryotic Gsp proteins (Supplementary Data 1). The prediction algorithms identified putative N-terminal MTSs for proteins from jakobids and malawimonads but failed to recognize them in the orthologues from heteroloboseans, which, however, carry the longest N-terminal extensions (Supplementary Fig. 4).

In order to test if Gsp proteins carry functional mitochondrial targeting information, we expressed *GspD*, *GspE* and *GspG1* genes from *N. gruberi* (*Ng*) and *G. okellyi* (*Go*) in *Saccharomyces cerevisiae*; no expression of *GspF* was achieved. All proteins were specifically localised in mitochondria, as confirmed by their co-localisation with Mitotracker red CMX Ros (Fig. 3A). Additionally, we attempted to express these genes in *Trypanosoma brucei*, which represents the evolutionarily closest experimental model to the eukaryotes carrying the *Gsp* genes. Of all the proteins tested, only *Go*GspD and *Go*GspG2 were detected (Fig. 3B), in addition to a weak signal for *Ng*GspG1 (Supplementary Fig. 5A). While both GspG proteins could be found specifically in the mitochondrion of *T. brucei*, *Go*GspD was found in a different membrane compartment, perhaps due to mistargeting. Finally, we tested if the atypically long N-terminal extension of *Ng*GspG1 targets the protein to the *T. brucei* mitochondrion. Indeed, the 160 N-terminal amino acid residues of *Ng*GspG1 were able to deliver the marker (mNeonGreen) into the organelle (Supplementary Fig. 5B). As a complementary approach, we raised specific polyclonal antibodies against *Ng*GspG1 and probed *N. gruberi* cellular fractions (Fig. 3C; Supplementary Fig. 6). *Ng*GspG1 and *Ng*GspEN2A co-fractionated with the mitochondrial markers including alternative oxidase (AOX) and TatC[9], but not with the cytosolic protein hemerythrin[33]. Immunofluorescence microscopy of *N. gruberi* with the anti-*Ng*GspG1 antibody provided further evidence that *Ng*GspG1 is targeted to mitochondria (Fig. 3D, Supplementary Fig. 7).

In order to further confirm the mitochondrial localization of the Gsp proteins in *N. gruberi*, we analysed the mitochondrial proteome of this organism by partial purification of the organelle and identification of resident proteins by mass spectrometry. A mitochondria-enriched fraction was obtained from a cellular lysate by several steps of differential centrifugation and OptiPrep gradient centrifugation. Three sub-fractions of different densities, named accordingly as Opt1015, Opt1520, Opt2030, were collected (Supplementary Fig. 8A, see Methods for more details), with the densest one being most enriched in mitochondria. The sub-fractions were subjected to proteomic analysis. The relative amount of each protein in each sub-fraction was determined by label-free quantification and the proteins were grouped by a multicomponent analysis (for details see Methods) according to their distributions across the gradient (Fig. 3E). A set of marker proteins (homologues of well-characterized typical mitochondrial proteins from other species) was used to identify a cluster of mitochondrial proteins. Due to the partial co-purification of peroxisomes with mitochondria, a peroxisome-specific cluster was defined analogously. As a result, 946 putative mitochondrial and 78 putative peroxisomal proteins were identified among the total of 4,198 proteins detected in all three sub-fractions combined. Encouragingly, the putative mitochondrial proteome of *N. gruberi* is dominated by proteins expected to be mitochondrial or whose mitochondrial localization

231 is not unlikely (Supplementary Fig. 8B, Supplementary Data 2). On the other hand, the
232 putative peroxisomal proteome seems to be contaminated by mitochondrial proteins
233 (owing to the presence of several mitochondrial ribosomal proteins; Supplementary
234 Data). Importantly, all five Gsp proteins (including both GspG paralogues) were
235 identified in the putative mitochondrial but not peroxisomal proteome of *N. gruberi*.
236
237 ***The properties of the eukaryotic Gsp proteins support the existence of a mitochondrial***
238 ***T2SS-related machinery***
239 The foregoing experiments support the idea that all four eukaryotic Gsp homologues
240 localize to and function in the mitochondrion. However, direct *in vivo* demonstration of
241 the existence of a functional mitochondrial T2SS-related machinery is currently not
242 feasible, because none of the Gsp homologue-carrying eukaryotes represents a tractable
243 genetic system. We thus used *in vitro* approaches and heterologous expression systems to
244 test the key properties of the eukaryotic Gsp proteins.
245 Crucial for the T2SS function is the formation of the OM pore, which is a β-barrel
246 formed by the oligomerization of the C-domain of the GspD protein[34]. The actual
247 assembly of the bacterial pore requires GspD targeting to the outer membrane through
248 interaction of its very C-terminal domain (S-domain) with the outer membrane
249 lipoprotein GspS[35]. GspD forms a pre-pore multimer, whose OM membrane insertion is
250 independent on the β□barrel assembly machinery (BAM) complex[36,37]. In addition, the
251 bacterial GspD carries four short N-terminal domains exposed to the periplasm, called N0
252 to N3, of which N1 to N3 share a similar fold[38] (Fig. 4A). While the N3 domain is
253 required for the pore assembly[39], N0 interacts with GspC of the assembly platform[22,40].
254 Sequence analysis of the mitochondrial GspD homologue revealed that it corresponds to
255 a C-terminal part of the bacterial GspD C-domain, whereas the N-terminal domains N0 to
256 N3, the N-terminal part of the C-domain, and the S-domain are missing (Fig. 4A;
257 Supplementary Fig. 4A).
258 According to our hypothesis, the mitochondrial GspD should be present in the
259 outer mitochondrial membrane. In order to test this localization, *N. gruberi* GspD
260 (*Ng*GspD) was *in vitro* imported into yeast mitochondria. The import reaction was also
261 performed with the widely used synthetic substrate Su9-DHFR destined to mitochondrial
262 matrix, which is composed of 69 amino acid residues of F_o-ATPase subunit 9 from
263 *Neurospora crassa* fused to the mouse DHFR[41]. Both proteins accumulated in the
264 mitochondria in a time-dependent manner, but only the import of Su9-DHFR could be
265 inhibited by the addition of ionophores, which dissipate membrane potential ($\Delta\Psi$) (Fig.
266 4B). This result showed that the import of GspD is independent of $\Delta\Psi$, which is a typical
267 feature of mitochondrial outer membrane proteins. The key question was if the
268 mitochondrial GspD homologue has retained the ability to assemble into an oligomeric
269 pore-forming complex. This possibility was supported by homology modelling of GspD
270 from *G. okellyi* (*Go*GspD) using *Vibrio cholerae* GspD[42] as a template, which indicated
271 that the protein has the same predicted fold as the typical secretin C-domain. Remarkably,
272 the two transmembrane β-strands of *Go*GspD are separated by the highly conserved
273 amphipathic helical loop (AHL) (Fig. 4C) essential for secretin membrane insertion,
274 suggesting its capability to form a pentadecameric pore complex. Indeed, radioactively
275 labelled *Go*GspD assembled into a high-molecular-weight complex of around 200 kDa in
276 an *in vitro* bacterial translation system in the presence of lecithin liposomes (Fig. 4D).

277 The complex was resistant to 2M urea treatment, which would remove nonspecific
278 protein aggregates, and was still pelleted with the liposomes, suggesting that it was
279 inserted into the liposomes. These results showed that the mitochondrial GspD, despite
280 being significantly truncated when compared to its bacterial homologues, has retained the
281 major characteristics of bacterial secretins[43], including the capacity to form oligomers and
282 insert into membranes.
283 To test if the mitochondrial GspD forms *bona fide* pores in the membrane, we
284 aimed to produce His-tagged *Ng*GspD in *E. coli* and purify the membrane complexes for
285 the electrophysiology analysis. Inducing the production of this protein in the bacterial
286 cytoplasm was highly toxic (Fig. 5A). A similar phenomenon has been reported for
287 bacterial secretins, which form pores in the inner bacterial membrane[43]. While this
288 hampered protein production and purification, directing the GspD export to the periplasm
289 by fusing it to the N-terminal signal peptide of *E. coli* DsbA protein alleviated the
290 toxicity upon autoinduction. The His-*Ng*GspD variant was affinity-purified on a Ni-
291 column followed by size exclusion chromatography, during which two peaks of about
292 230 kDa and 125 kDa could be observed (Fig. 5B, Supplementary Fig.15). The pore-
293 forming activity of His-*Ng*GspD purified from both protein peaks was demonstrated by
294 conductivity measurements in black lipid membranes composed of an *E. coli* polar lipid
295 extract. The channel recordings illustrated a very high stability of the inserted membrane
296 pores (Fig. 5C). The amplitude histograms (Fig. 5D) suggested that the mitochondrial
297 GspD can form variable arrangements resulting in stable membrane pores of different
298 sizes.
299 The secretion mechanism of the T2SS relies on the assembly of pseudopilus made
300 up of GspG subunits[44]. Comparison of the mitochondrial GspG with bacterial
301 homologues revealed important similarities as well as differences. These proteins share a
302 complete pseudopilin domain preceded by a transmembrane domain, but the
303 mitochondrial proteins are substantially longer due to extensions at both N- and C-termini
304 (Fig. 6A). The N-terminal extension likely serves as a MTS, but the origin and function
305 of the C-terminal extension (amounting to ~100 amino acid residues) is unclear, as it is
306 well conserved among the mitochondrial GspG but lacks discernible homologues even
307 when analysed by highly sensitive homology-detection methods (HMM-HMM
308 comparisons with HHpred[45] and protein modelling using the Phyre2 server[46]). Structural
309 modelling of the pseudopilin domain into the recently obtained cryo-EM reconstruction
310 of the PulG (=GspG) complex from *Klebsiella oxytoca*[47] revealed the presence of key
311 structural features in the mitochondrial GspG from *G. okellyi* (*Go*GspG1), supporting
312 possible formation of a pseudopilus (Fig. 6B).
313 To test this directly, we purified a recombinant pseudopilin domain of *Ng*GspG1
314 under native conditions, which showed a uniform size of 25 kDa corresponding to the
315 monomer (Fig. 7A). Possible protein-protein interaction of the purified pseudopilin
316 domain was tested by thermophoresis of an NT-647-labelled protein. The measurements
317 revealed specific self-interaction and plotting of the change in thermophoresis yielded a
318 K_d of 216 (±15.1) nM (Fig. 7B). Additionally, the interaction properties of GspG were
319 followed by the bacterial two-hybrid assay (BACTH). When produced in bacteria, the
320 mitochondrial *Go*GspG1 showed specific oligomerisation, typical of bacterial major
321 pseudopilins[48], supporting its *in vivo* propensity to form a pseudopilus (Fig. 7C). In
322 addition, *Go*GspG1 showed positive interaction with *Go*GspF, consistent with the

323 analogous interaction of bacterial GspG with the IM-embedded GspF[48]. Moreover, the
324 mitochondrial *Go*GspF and *Go*GspE each formed dimers in the BACTH assay (Fig. 7C).
325 These interactions are consistent with the hypothesized role of both proteins as
326 mitochondrial T2SS components, as GspF forms dimers within the IM complex and
327 GspE assembles into an active hexameric ATPase in the bacterial T2SS system. Indeed,
328 bacterial GspG and GspF also interact in the BACTH assay[48]. Tests of all other possible
329 interactions of *G. okellyi* Gsp proteins were negative.
330 The *in silico* analyses and experiments described above are consistent with the
331 hypothesized existence of a functional mitochondrial secretion machinery derived from
332 the bacterial T2SS. However, the mitochondrial subunits identified would assemble only
333 a minimalist version of the secretion system, reduced to the functional core of the four
334 subcomplexes of the bacterial T2SS, i.e. the luminal ATPase (GspE), the IM pseudopilus
335 assembly platform (GspF), the intermembrane space pseudopilus (GspG), and the OM
336 pore (truncated GspD). Despite using sensitive HMM-based searches, we did not detect
337 homologues of other conserved T2SS subunits in any of the eukaryotes possessing GspD
338 to GspG proteins. One of the missing subunits is GspC, which connects the assembly
339 platform with the N0 domain of GspD pore[22,23]. Thus, the absence of GspC in eukaryotes
340 correlates with the lack of the N0 domain in the eukaryotic GspD. Analogously, the
341 absence of the C-terminal S-domain in the mitochondrial GspD (Fig. 4A), known to be
342 missing also from some bacterial GspD proteins, rationalizes the lack of a eukaryotic
343 homologue of the bacterial OM component GspS that binds to GspD via the S-domain
344 during the pore assembly[35]. The apparent lack of a eukaryotic homologue of GspL, which
345 interacts via its cytosolic domain with the N1E domain of the bacterial GspE [49], may
346 similarly be explained by the fact that the eukaryotic GspE protein seems to be
347 homologous only to the C-terminal (CTE) domain of its bacterial counterpart and lacks
348 an equivalent of the N1E domain (Supplementary Fig. 4E). The mitochondrial system
349 also apparently lacks a homologue of GspO, a bifunctional enzyme that is essential for
350 GspG maturation. Despite this absence, eukaryotic GspG homologues have conserved all
351 the characteristic sequence features required for GspG maturation (the polar anchor and
352 the trans-membrane domain with a conserved glutamate residue at position +5 relative to
353 the processing site) (Fig. 6A, Supplementary Fig. 4J). Notably, all the *Ng*GspG1 and
354 *Ng*GspG2-derived peptides detected in our proteomic analysis come from the region of
355 the protein downstream of the conserved processing site (Fig. 6C), suggesting that
356 analogous maturation of the pseudopilin also occurs in mitochondria.
357
358 ***Additional putative components of the mitochondrial T2SS-based functional pathway***
359 ***identified by phylogenetic profiling***
360 Since none of the eukaryotes with the Gsp homologues is currently amenable to
361 functional studies, we tried to further illuminate the role of the mitochondrial T2SS
362 system using a comparative genomic approach. Specifically, we reasoned that possible
363 additional components of the machinery, as well as its actual substrate(s), might show the
364 same phylogenetic distribution as the originally identified four subunits. Using a
365 combination of an automated identification of candidate protein families and subsequent
366 manual scrutiny by exhaustive searches of available eukaryote sequence data (for details
367 of the procedure see Methods), we identified 23 proteins (more precisely, groups of
368 orthologues) that proved to exhibit the same phylogenetic distribution in eukaryotes as

369 the four core T2SS components. Specifically, all 23 proteins were represented in each of
370 the heterolobosean, jakobid, and malawimonad species possessing all four core Gsp
371 proteins, whereas only 17 proteins were identified in the incomplete transcriptomic data
372 of hemimastigotes and seven of them were found in the transcriptomic data from the
373 *Percolomonas* lineage that possesses only GspD (Fig. 1B; Supplementary Data 1). Except
374 for two presumably Gsp-positive jakobids represented by incomplete EST surveys and a
375 case of a likely contamination (Supplementary Data 3), no orthologues of any of these
376 proteins were found in any other eukaryote (including the Gsp-lacking members of
377 heteroloboseans, jakobids and malawimonads). The sequences of these 23 proteins were
378 analysed by various *in silico* approaches, including HHpred and Phyre2 to assess their
379 possible function (Fig. 8A).
380 These analyses revealed that seven of the families have a direct link to the T2SS
381 suggested by discerned homology to known T2SS components. One of them represents
382 an additional, more divergent homologue of the C-terminal part of the bacterial GspD.
383 Hence, the protein has been marked as GspDL (GspD-like). Three other families, referred
384 to as GspDN1 to GspDN3, proved to be homologous to the Secretin_N domain (Pfam
385 family PF03958), present in the bacterial GspD protein as domains N1, N2 and N3 (Fig.
386 4A). The N1-N3 array protrudes into the periplasm, where it oligomerizes to form three
387 stacked rings[20]. As mentioned above, the initially identified eukaryotic GspD homologues
388 lack the N-terminal region, suggesting that the gene was split into multiple parts in
389 eukaryotes. While GspDN1 corresponds to a full single N-domain, GspDN2 and GspDN3
390 relate only to its C-terminal and N-terminal halves, respectively (Fig. 4A; Supplementary
391 Fig. 4B-D). Unfortunately, high sequence divergence makes it impossible to identify
392 potential specific correspondence between the N1 to N3 domains of the bacterial GspD
393 and the eukaryotic GspDN1 to GspDN3 proteins. Importantly, a Y2H assay indicated that
394 the two separate polypeptides GspD and GspDN1 of *N. gruberi* may interact *in vivo* (Fig.
395 4E), perhaps forming a larger mitochondrial complex. In addition, we identified most of
396 the discovered GspD-related proteins (GspDL/N) in the *N. gruberi* mitochondrial
397 proteome (the exception being GspDN1, which was not detected in a sufficient number of
398 replicates to be included in the downstream analysis; Supplementary Data 2).
399 The final three proteins linked to the T2SS based on their sequence features seem
400 to be evolutionarily derived from GspE. One, denoted GspEL (GspE-like) represents a
401 divergent homologue of the C1E (i.e. nucleotide-binding) domain of GspE, although with
402 some of the characteristic motifs (Walker A, Asp box, Walker B) abrogated
403 (Supplementary Fig. 4F), indicating the loss of the ATPase function. The other two
404 proteins, which we denote GspEN2A and GspEN2B, are suggested by HHpred to be
405 related to just the N2E domain of the bacterial GspE (Supplementary Fig.4G and 4H).
406 GspEN2A and GspEN2B were identified among *N. gruberi* mitochondrial proteins in the
407 proteomic analysis, whereas GspEL was found in the cluster of putative peroxisomal
408 proteins. Importantly, a polyclonal antibody raised against *Ng*GspEN2A confirmed the
409 mitochondrial localization of the protein (Fig. 3C, D; Supplementary Fig. 6, 7).
410 The remaining sixteen proteins co-occurring with the core eukaryotic T2SS
411 subunits, hereafter referred to as Gcp (Gsp-co-occurring proteins), were divided into three
412 categories. The first comprises four proteins that constitute paralogues within broader
413 common eukaryotic (super)families. Three of them (Gcp1 to Gcp3) belong to the WD40
414 superfamily and seem to be most closely related to the peroxisomal protein import co-

415 receptor Pex7 (Supplementary Fig. 9). None of these proteins has any putative N-terminal
416 targeting sequence, but interestingly, the peroxisomal targeting signal 1 (PTS1) could be
417 predicted on most Gcp1 and some Gcp2 proteins (Supplementary Data 1). However,
418 these predictions are not fully consistent with the results of our proteomic analysis:
419 *Ng*Gcp1 was found among the mitochondrial proteins and *Ng*Gcp2 in the cluster of
420 putative peroxisomal proteins (Supplementary Data 2), but PTS1 is predicted to be
421 present in the *Ng*Gcp1 protein (Supplementary Data 1). The fourth Gcp protein (Gcp4) is
422 a paralogue of the ubiquitin-like superfamily, distinctly different from the previously
423 characterized members including ubiquitin, SUMO, NEDD8 and others (Supplementary
424 Fig. 10).
425 The second Gcp category comprises eleven proteins (Gcp5 to Gcp15) well
426 conserved at the sequence level among the Gsp-containing eukaryotes, yet lacking any
427 discernible homologues in other eukaryotes or in prokaryotes. Two of these proteins
428 (Gcp8, Gcp15) were not identified in the proteomic analysis of *N. gruberi*
429 (Supplementary Data 1 and 2). Of those identified, several (Gcp5, Gcp6, Gcp13) were
430 found among the mitochondrial proteins, whereas some others (Gcp9, Gcp10, Gcp11)
431 clustered with peroxisomal markers. Specific localization of the three remaining proteins
432 (Gcp7, Gcp12, and Gcp14) could not be determined due to their presence at the
433 boundaries of the mitochondrial or peroxisomal clusters. No homology to other proteins
434 or domains could be discerned for the Gsp5 to Gsp15 proteins even when sensitive
435 homology-detection algorithms were employed. However, four of them are predicted as
436 single-pass membrane proteins, with the transmembrane segment in the N- (Gcp7,
437 Gcp11, Gcp15) or C-terminal (Gcp5) regions (Fig. 8A; Supplementary Fig. 11).
438 Interestingly, Gcp6 and Gcp12 proteins contain multiple absolutely conserved cysteine or
439 histidine residues (Fig. 8A, B; Supplementary Fig. 12). We were not able to determine
440 their localization in *N. gruberi* by microscopy, but we tested the localization of *Go*Gcp12
441 upon expression in *T. brucei*, where it co-localized with the mitochondrial tubules (Fig.
442 8C).
443 Gcp16 constitutes a category of its own, as it typifies a newly described protein
444 family present also in bacteria of the PVC superphylum (Supplementary Fig. 13).
445 Phylogenetic analysis confirmed that the eukaryotic members of the family are of the
446 same origin (Supplementary Fig. 14). Gcp16 proteins are predicted to harbour two
447 transmembrane domains (Supplementary Fig. 13). Furthermore, HHpred searches
448 suggested possible homology of a region of the Gcp16 protein (upstream of the
449 transmembrane domains) to various metallopeptidases, although with inconclusive
450 statistical support. However, inspection of the HHpred alignments revealed that Gcp16
451 shares with these hits an absolutely conserved motif HEXXH (Supplementary Fig. 13),
452 which is the catalytic, metal-binding motif of the zincin tribe of metallopeptidases[50].
453 Interestingly, close to the HEXXH motif, Gcp16 possesses an absolutely conserved
454 EXXA motif, which is diagnostic of a zincin subgroup called gluzincins[50], further
455 supporting the notion that Gcp16 may function as a membrane-embedded peptidase.
456 Most eukaryotic Gcp16 proteins exhibit an N-terminal extension compared to the
457 bacterial homologues (Supplementary Fig. 13), but only some of these extensions are
458 recognized as putative MTSs and the *N. gruberi* Gcp16 was not identified either in
459 putative mitochondrial or peroxisomal proteome.
460

461 **Discussion**
462 Our analyses revealed that a subset of species belonging to four eukaryotic lineages share
463 a set of at least 27 proteins (or families of orthologues) absent from other eukaryotes for
464 which genomic or transcriptomic data are currently available (Fig. 1B). At least eleven of
465 these proteins (the Gsp proteins) are evolutionarily related to components of the bacterial
466 T2SS, although seven of them are so divergent that their evolutionary connection to the
467 T2SS could be recognized only retrospectively after their identification based on their
468 characteristic phylogenetic profile. For the sixteen remaining proteins (Gcp1 to Gcp16)
469 no other obvious evolutionary or functional link to the T2SS is evident apart from the
470 same phyletic pattern as exhibited by the T2SS subunit homologues. Nevertheless,
471 similar phylogenetic profiles are generally a strong indication for proteins being parts of
472 the same functional system or pathway, and have enabled identification of additional
473 components of different cellular structures or pathways (e.g. refs [51,52]). Is it, therefore,
474 possible that the 27 Gsp/Gcp proteins similarly belong to a single functional pathway?
475 The phylogenetic profile shared by the eukaryotic Gsp and Gcp proteins is not
476 trivial, as it implies independent gene losses in a specific set of multiple eukaryotic
477 branches (Fig. 1B). The likelihood of a chance emergence of the same taxonomic
478 distribution of these proteins is thus low. Nevertheless, false positives cannot be
479 completely excluded among the Gcp proteins and their list may be revised when a more
480 comprehensive sampling of eukaryote genomes or transcriptomes becomes available. It is
481 also possible that the currently inferred phylogenetic profile of some of the Gsp/Gcp
482 proteins is incomplete due to limited sampling of the actual gene repertoire of species
483 represented by transcriptome assemblies only. The inherently incomplete nature of
484 single-cell transcriptome assemblies available for hemimastigotes potentially explains our
485 failure to identify homologues of some Gsp and Gcp proteins in this group (Fig. 1B;
486 Supplementary Data 1). An incomplete set is evident also in the heterolobosean
487 *Percolomonas* lineage, as transcriptomic data from three different members revealed only
488 the presence of GspD, GspDL, the three GspDN proteins, and four Gcp proteins (Fig. 1B,
489 Supplementary Data 1). The relatively coherent pattern of Gsp/Gcp protein occurrence in
490 the three independently sequenced transcriptomes and the fact that in other Gsp/Gcp-
491 containing eukaryotes (except for hemimastigotes) all 27 families are always represented
492 in the respective transcriptome assembly (Supplementary Data 1) suggest that the
493 *Percolomonas* lineage has indeed preserved only a subset of Gsp/Gcp families. Genome
494 sequencing is required to test this possibility.
495 All uncertainties notwithstanding, our data favour the idea that a hitherto
496 unrecognized complex functional pathway exists in some eukaryotic cells, underpinned
497 by most, if not all, of the 27 Gsp/Gcp proteins and possibly others yet to be discovered.
498 Direct biochemical and cell biological investigations are required for testing its existence
499 and the actual cellular role. Nevertheless, we have integrated the experimental data
500 gathered so far with the insights from bioinformatic analyses to propose a hypothetical
501 working model (Fig. 9).
502 Our main proposition is that the eukaryotic homologues of the bacterial Gsp
503 proteins assemble a functional transport system, here denoted miT2SS, that spans the
504 mitochondrial OM and mediates the export of specific substrate proteins from the
505 mitochondrion. Although the actual architecture of the miT2SS needs to be determined,
506 the available data suggest that it departs in detail from the canonical bacterial T2SS

507 organization, as homologues of some of the important bacterial T2SS components are
508 apparently missing. Most notable is the absence of GspC, presumably related to the
509 modified structure of its interacting partner GspD, which in eukaryotes is split into
510 multiple polypeptides and seems to completely lack the N0 domain involved in GspC
511 binding. It thus remains unclear whether and how the IM assembly platform and the OM
512 pore interact in mitochondria. One possible explanation is that GspC has been replaced
513 by an unrelated protein. It is notable that three Gcp proteins (Gcp7, Gcp11, and Gcp15)
514 have the same general architecture as GspC: they possess a transmembrane segment at
515 the N-terminus and a (predicted) globular domain at the C-terminus (Fig. 9A,
516 Supplementary Fig. 11). Testing possible interactions between these proteins and T2SS
517 core subunits (particularly GspG, GspF and GspDN) using BACTH or Y2H assays will
518 be of future interest.
519 Further investigations also must address the question of whether the
520 mitochondrial GspG is processed analogously to the bacterial homologues and how such
521 processing occurs in the absence of discernible homologues of GspO. The mitochondrial
522 GspG is presumably inserted into the IM by the Tim22 or Tim23 complex, resulting in a
523 GspG precursor with the N-terminus, including the MTS, protruding into the matrix. It is
524 possible that N-terminal cleavage by matrix processing peptidase serves not only to
525 remove the transit peptide, but at the same time to generate the mature N-terminus of the
526 processed GspG form, ready for recruitment into the pseudopilus. A different hypothesis
527 is offered by the discovery of the Gcp16 protein with sequence features suggesting that it
528 is a membrane-embedded metallopeptidase (certainly non-homologous to GspO, which is
529 an aspartic acid peptidase[53]). Although the subcellular localization of Gcp16 needs to be
530 established, we speculate that it might be a mitochondrial IM protein serving as an
531 alternative prepilin peptidase.
532 In parallel with its apparent simplification, the miT2SS may have been
533 specifically elaborated compared to the ancestral bacterial machinery. This possibility is
534 suggested by the existence of two pairs of proteins corresponding to different parts of the
535 bacterial GspE subunit. We propose that the eukaryotic GspE makes a heterodimer with
536 the GspEN2A to reconstitute a unit equivalent to most of the bacterial GspE protein
537 (lacking the GspL-interacting N1E domain), whereas GspEN2B, which is much more
538 divergent from the standard N2E domain than GspEN2A, may pair specifically with
539 GspEL to make an enzymatically inactive GspE-like version. We can only speculate as to
540 the function of these proteins, but the fact that the bacterial GspE assembles into a
541 homohexamer raises the possibility that in eukaryotes catalytically active and inactive
542 versions of GspE are mixed together in a manner analogous to the presence of
543 catalytically active and inactive paralogous subunits in some well-known protein
544 complexes, such as the proton-pumping ATPase (e.g. refs [54,55]). The co-occurrence of two
545 different paralogues of the GspD C-domain, one (GspDL) being particularly divergent,
546 suggests a eukaryote-specific elaboration of the putative pore in the mitochondrial OM.
547 Moreover, the electrophysiology measurements of the pores built of mitochondrial GspD
548 indicated stable complexes of variable sizes; a property not observed for bacterial
549 proteins. Finally, the C-terminal extension of the mitochondrial GspG representing a
550 conserved domain without discernible homology to other proteins suggests a eukaryote-
551 specific modification of the pseudopilus functioning.

552 An unanswered key question is what is the actual substrate (or substrates)
553 possibly exported from the mitochondrion by the miT2SS. No bioinformatic tool for
554 T2SS substrate prediction is available due to the enigmatic nature of the mechanism of
555 substrate recognition by the pathway[14], so at the moment we can only speculate. It is
556 notable that no protein encoded by the mitochondrial genomes of jakobids,
557 heteroloboseans and malawimonads stands out as an obvious candidate for the miT2SS
558 substrate, since they either have well-established roles in the mitochondrion or are
559 hypothetical proteins with a restricted (genus-specific) distribution. Therefore, we
560 hypothesize that the substrate could be encoded by the nuclear genome and imported into
561 the mitochondrion to undergo a specific processing/maturation step. This may include
562 addition of a prosthetic group – a scenario modelled on the process of cytochrome *c* or
563 Rieske protein maturation[56,57]. Interestingly, the proteins Gcp6 and Gcp12, each
564 exhibiting an array of absolutely conserved cysteine and histidine residues
565 (Supplementary Fig. 12), are good candidates for proteins to which a specific prosthetic
566 group might be attached, so any of them could be the sought-after miT2SS substrate.
567 Some of the other Gcp proteins may then represent components of the hypothetical
568 machinery responsible for the substrate modification. The putative functionalization step
569 may occur either in the mitochondrial matrix or in the intermembrane space (IMS), but
570 we note that the former localization would necessitate a mechanism of protein
571 translocation across the mitochondrial IM in the matrix-to-IMS direction, which has not
572 been demonstrated yet. Regardless, the T2SS system would eventually translocate the
573 modified protein across the mitochondrial OM to the cytoplasm.
574 However, this may not be the end of the journey, since there are hints of a link
575 between the miT2SS-associated pathway and peroxisomes. First, three Gcp proteins,
576 namely Gcp1 to Gcp3, are specifically related to Pex7, a protein mediating import of
577 peroxisomal proteins characterized by the peroxisomal targeting signal 2 (PTS2)[58].
578 Second, some of the Gcp proteins (especially Gcp1 and Gcp13) have at the C-terminus a
579 predicted PTS1 signal (at least in some species; Supplementary Data 1). Third, several
580 Gcp proteins (Gcp2, Gcp9, Gcp10, and Gcp11) and GspEL were assigned to the putative
581 peroxisomal proteome in our proteomic analysis (Supplementary Data 2). We note the
582 discrepancy between the PTS1 signal predictions and the actual set of experimentally
583 defined peroxisomal proteins, which might be due to an incomplete separation of
584 peroxisome and mitochondria by our purification procedure, but may also reflect protein
585 shuttling between the two organelles. We thus hypothesize that upon its export from the
586 mitochondrion, the miT2SS substrate might be eventually delivered to the peroxisome.
587 This is possibly mediated by the Gcp1/2/3 trio, but other Gcp proteins might participate
588 as well. One such protein might be the ubiquitin-related protein Gcp4. Ubiquitination and
589 deubiquitination of several components of the peroxisome protein import machinery are a
590 critical part of the import mechanism[58] and Gcp4 might serve as an analogous peptide
591 modifier in the hypothetical peroxisome import pathway functionally linked to the
592 miT2SS.
593 Altogether, our data suggest the existence of an elaborate functional pathway
594 combining components of bacterial origin with newly evolved eukaryote-specific
595 proteins. The extant phylogenetic distribution of the pathway is sparse, but our current
596 understanding of eukaryote phylogeny suggests that it was ancestrally present in
597 eukaryotes and for some reason dispensed with multiple times during evolution.

598 Although we could not define a specific bacterial group as the actual source of the
599 eukaryotic Gsp genes, it is tempting to speculate that the T2SS was introduced into
600 eukaryotes by the bacterial progenitor of mitochondria and that it was involved in
601 delivering specific proteins from the endosymbiont into the host cell, as is known in the
602 case of current intracellular bacteria[59]. Elucidating the actual role of this communication
603 route in establishing the endosymbiont as a fully integrated organelle requires
604 understanding the cellular function of the modern miT2SS-associated pathways, which is
605 a challenge for future research.
606
607 **Methods**
608 **Sequence data and homology searches**
609 Homologues of relevant genes/proteins were searched in sequence databases accessible
610 via the National Center for Biotechnology Information BLAST server
611 (https://blast.ncbi.nlm.nih.gov/Blast.cgi), including the nucleotide and protein non-
612 redundant (nr) databases, whole-genome shotgun assemblies (WGAs), expressed
613 sequence tags (ESTs), and transcriptome shotgun assemblies (TSAs). Additional public
614 databases searched included the data provided by the Marine Microbial Eukaryote
615 Transcriptome Sequencing Project (MMETSP[60]) comprising TSAs from hundreds of
616 diverse protists (https://www.imicrobe.us/#/projects/104), the OneKP project[61]
617 (https://sites.google.com/a/ualberta.ca/onekp/) comprising TSAs from hundreds of plants
618 and algae, and individual WGAs and TSAs deposited at various on-line repositories
619 (Supplementary Data 4). To further improve the sampling, draft genome assemblies were
620 generated in this study for the heterolobosean *Neovahlkampfia damariscottae* and the
621 malawimonad informally called "*Malawimonas californiana*". Details on the sequencing
622 and assembly are provided in Supplementary Methods. Finally, the analyses also included
623 sequence data from genome and/or transcriptome projects for several protists that are
624 underway in our laboratories and will be published in full elsewhere upon completion
625 (Supplementary Data 4). Relevant sequences were extracted from these unpublished
626 datasets and either deposited in GenBank or included in Supplementary Data 1.
627 Similarity searches were done using BLAST[62] (blastp or tblastn, depending on the
628 database queried) and HMMER[63] using profile HMMs built from sequence alignments of
629 proteins of interest. Hits were evaluated by BLAST (blastp or blastx) searches against the
630 nr protein dataset at NCBI to distinguish orthologues of Gsp and Gcp proteins from
631 paralogous proteins or non-specific matches. This was facilitated by a high degree of
632 conservation of individual eukaryotic Gsp/Gcp proteins among different species (see also
633 Supplementary Figs 4 and 11-13) and in most cases by the lack of other close
634 homologues in eukaryotic genomes (the exceptions being members of broader protein
635 families, including the ATPase GspE, the WD40 superfamily proteins Gcp1 to Gcp3, and
636 the ubiquitin-related protein Gcp4). All identified eukaryotic Gsp and Gcp sequences
637 were carefully manually curated to ensure maximal accuracy and completeness of the
638 data, which included correction of existing gene models, extension of truncated
639 sequences by manual analysis of raw sequencing reads, and correction of assembly errors
640 (for details see Supplementary Methods). All newly predicted or curated Gsp and Gcp
641 sequences are provided in Supplementary Data 1; additional Gsp and Gcp sequences
642 from non-target species are listed in Supplementary Data 4. The nomenclature of the Gsp

643 and Gcp genes proposed in this study was also reflected in the annotation of the *A. godoyi*
644 genome, recently published as part of a separate study[5].
645
646 **Phylogenetic profiling**
647 In order to identify genes with the same phylogenetic distribution as the eukaryotic
648 homologues of the four core T2SS components, we carried out two partially overlapping
649 analyses based on defining groups of putative orthologous genes in select Gsp-positive
650 species and phylogenetically diverse Gsp-negative eukaryotic species. The list of taxa
651 included is provided in Supplementary Data 5. The first analysis was based on 18 species,
652 including three Gsp-positive ones (*N. gruberi*, *A. godoyi* and *M. jakobiformis*), for the
653 second analysis the set was expanded by adding one additional Gsp-positive species (*G.*
654 *okellyi*) and one Gsp-negative species (*Monocercomonoides exilis*). Briefly, the protein
655 sequences of a given species were compared to those of all other species using blastp
656 followed by fast phylogenetic analyses, and orthologous relationships between proteins
657 were then inferred from this set of phylogenetic trees using a reference-species-tree-
658 independent approach. This procedure was repeated for each species and all resulting sets
659 of orthologous relationships, also known as phylomes[64], were combined in a dense
660 network of orthologous relationships. This network was finally trimmed in several
661 successive steps to remove weak or spurious connections and to account for (genuine or
662 artificial) gene fusions, with the first analysis being less restrictive than the second.
663 Details of this pipeline are provided in Supplementary Methods. For each of the two
664 analyses, the final set of defined groups of orthologs (orthogroups) was parsed to identify
665 those comprising genes from at least two Gsp-positive species yet lacking genes from any
666 Gsp-negative species. The orthogroups passing this criterion were further analysed
667 manually by blastp and tblastn searches against various public and private sequence
668 repositories (see the section "Sequence data and homology searches") to exclude those
669 orthogroups with obvious orthologs in Gsp-negative species. *Percolomonas* lineage
670 exhibiting only GspD and jakobids represented by incomplete EST surveys (these species
671 likely possess the miT2SS system) were not considered Gsp-negative. The orthogroups
672 that remained were then evaluated for their conservation in Gsp-positive species and
673 those that proved to have a representative in all these species (*N. gruberi*, *N. fowleri*, *N.*
674 *damariscottae*, *P. kirbyi*, *A. godoyi*, *R. americana*, *M. jakobiformis*, *G. okellyi*) were
675 considered as bona fide Gcp (Gsp-co-occurring protein) candidates. It is of note that
676 some of these proteins are short and were missed by the automated annotation of some of
677 the genomes, so using relaxed criteria for the initial consideration of candidate
678 orthogroups (i.e. allowing for their absence from some of the Gsp-positive species)
679 proved critical for decreasing the number of false-negative identifications.
680
681 **Sequence analyses and phylogenetic inference**
682 Subcellular targeting of Gsp and Gcp proteins was evaluated using TargetP-1.1 (ref. [65];
683 http://www.cbs.dtu.dk/services/TargetP-1.1/index.php), TargetP-2.0 (ref. [66];
684 http://www.cbs.dtu.dk/services/TargetP/), Mitoprot II (ref. [67];
685 https://ihg.gsf.de/ihg/mitoprot.html), MitoFates[68] (http://mitf.cbrc.jp/MitoFates/cgi-
686 bin/top.cgi)[68], WoLF PSORT (https://wolfpsort.hgc.jp/), and PTS1 predictor[69]
687 (http://mendel.imp.ac.at/pts1/). Transmembrane domains were predicted using
688 TMHMM[70] (http://www.cbs.dtu.dk/services/TMHMM/). Homology of Gsp and Gcp

689 protein families to other proteins was evaluated by searches against Pfam v. 31 (ref. [71];
690 http://pfam.xfam.org/) and Superfamily 1.75 database[72]
691 (http://supfam.org/SUPERFAMILY/index.html), and by using HHpred[45]
692 (https://toolkit.tuebingen.mpg.de/#/tools/hhpred) and the Phyre2 server[46]
693 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index). The relative position of the
694 Gcp4 family among ubiquitin-like proteins was analysed by a cluster analysis using
695 CLANS[73] (https://www.eb.tuebingen.mpg.de/protein-evolution/software/clans/); for the
696 analysis the Gcp4 family was combined with all 59 defined families included in the clan
697 Ubiquitin (CL0072) as defined in the Pfam database (each family was represented by
698 sequences from the respective seed alignments stored in the Pfam database). For further
699 details on the procedure see the legend of Supplementary Fig. 10A. Multiple sequence
700 alignments used for presentation of the conservation and specific sequence features of
701 Gsp and Gcp families were built using MUSCLE[74] and shaded using BioEdit
702 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html).
703 In order to obtain datasets for the phylogenetic analyses of eukaryotic GspD to
704 GspG proteins, the protein sequences were aligned using MAFFT[75] and trimmed
705 manually. Profile hidden Markov models (HMMs) built on the basis of the respective
706 alignments were used as queries to search the UniProt database using HMMER. All
707 recovered sequences were assigned to components of the T4P superfamily machineries
708 using HMMER searches against a collection of profile HMMs reported by Abby et al.
709 (ref. [76]). For each GspD to GspG proteins, a series of alignments was built by
710 progressively expanding the sequence set by including more distant homologues (as
711 retrieved by the HMMER searches). Specifically, the different sets of sequences were
712 defined by the HMMER score based on the formula $score_{cutoff} = c*score_{best\ prokaryotic\ hit}$,
713 with the coefficient c decreasing from 0.99 to 0.70 incrementally by 0.01. The sequences
714 were then aligned using MAFFT, trimmed with BMGE[77] and the phylogenies were
715 computed with IQ-TREE[78] using the best-fit model (selected by the program from
716 standard protein evolution models and the mixture models[79] offered). The topologies
717 were tested using 10,000 ultra-fast bootstraps. The resulting trees were systematically
718 analyzed for support of the monophyly of eukaryotic sequences and for the taxonomic
719 assignment of the parental prokaryotic node of the eukaryotic subtree. The assignment
720 was done using the following procedure. The tree was artificially rooted between the
721 eukaryotic and prokaryotic sequences. From sub-leaf nodes to the deepest node of the
722 prokaryotic subtree, the taxonomic affiliation of each node was assigned by
723 proportionally considering the known or inferred taxonomic affiliations (at the phylum or
724 class level) of the descending nodes. See the legend to Supplementary Fig. 3 for further
725 details.
726 The phylogenetic analysis of the WD40 superfamily including Gcp1 to Gcp3
727 proteins was performed as follows. The starting dataset was prepared by a combination of
728 two different approaches: 1) each identified sequence of Gcp1 to Gcp3 proteins was used
729 as a query in a blastp search against the non-redundant (nr) NCBI protein database and
730 the 500 best hits for each sequence were kept; 2) protein sequences of each the Gcp1 to
731 Gcp3 family were aligned using MAFFT and the multiple alignment was used as a query
732 in a HMMER3 search (https://toolkit.tuebingen.mpg.de/#/tools/hmmer) against the
733 UniProt database. Best hits (E-value cutoff 1e-50) from the two searches were pooled and
734 de-duplicated, and the resulting sequence set (including Gcp1 to Gcp3 sequences) was

735 aligned using MAFFT and trimmed manually to remove poorly conserved regions.
736 Because WD40 proteins are very diversified, sequences that were too divergent were
737 eliminated from the starting dataset during three subsequent rounds of sequence removal,
738 based on a manual inspection of the alignment and phylogenetic trees computed by IQ-
739 TREE (using the best-fit model as described above). The final dataset was enriched by
740 adding PEX7 and WDR24 orthologues from eukaryotes known to possess miT2SS
741 components. The final phylogenetic tree was computed using IQ-TEE as described in the
742 legend to Supplementary Fig. 9. IQ-TREE was used also for inferring trees of the
743 heterolobosean 18S rRNA gene sequences (Supplementary Fig. 1), ubiquitin-related
744 proteins (Supplementary Fig. 10B) and the Gcp16 family (Supplementary Fig. 14);
745 details on the analyses are provided in legends to the respective figures.
746
747 **Structural homology modelling**
748 The PDB database was searched by the SWISS-MODEL server[80] for structural
749 homologues of *Go*GspD and *Go*GspG1. *V. cholerae* GspD[20] (PDB entry 5WQ9) and *K.*
750 *oxytoca* PulG[47] pseudopilus (PDB entry 5WDA) were selected as the top matches,
751 respectively. Models were built based on the target-template alignment using ProMod3
752 (ref. [80]). Coordinates that were conserved between the target and the template were
753 copied from the template to the model. Insertions and deletions were remodelled using a
754 fragment library, followed by rebuilding side chains. Finally, the geometry of the
755 resulting model was regularized by using a force field. In the case of loop modelling with
756 ProMod3 fails, an alternative model was built with PROMOD-II (ref. [81]). The quaternary
757 structure annotation of the template was used to model the target sequence in its
758 oligomeric form[82].
759
760 **Cultivation and fractionation of *N. gruberi* and proteomic analysis**
761 *Naegleria gruberi* str. NEG-M was axenically cultured in M7 medium with PenStrep
762 (100 U/mL of penicillin and 100 μg/mL of streptomycin) at 27°C in vented tissue culture
763 flasks. Mitochondria of *N. gruberi* were isolated in seven independent experiments and
764 were analyzed individually (see below). Each time ~1×10^9 *N. gruberi* cells were
765 resuspended in 2 mL of SM buffer (250 mM sucrose, 20 mM MOPS, pH 7.4)
766 supplemented with DNase I (40 μg/mL) and Roche cOmplete™ EDTA-free Protease
767 Inhibitor Cocktail and homogenized by eight passages through a 33-gauge hypodermic
768 needle (Sigma Aldrich). The resulting cell homogenate was then cleaned of cellular
769 debris using differential centrifugation and separated by a 2-hr centrifugation in a
770 discontinuous density OptiPrep gradient (10%, 15%, 20%, 30% and 50%) as described
771 previously[83]. Three visually identifiable fractions corresponding to 10-15% (OPT-1015),
772 15-20% (OPT-1520) and 20-30% (OPT-2023) OptiPrep densities were collected (each in
773 five biological replicates) and washed with SM buffer.
774 Proteins extracted from these samples were then digested with trypsin and
775 peptides were separated by nanoflow liquid chromatography and analyzed by tandem
776 mass spectrometry (nLC-MS2) on a Thermo Orbitrap Fusion (q-OT-IT) instrument as
777 described elsewhere[84]. The quantification of mass spectrometry data in the MaxQuant
778 software[85] provided normalized intensity values for 4,198 proteins in all samples and all
779 three fractions. These values were further processed using the Perseus software[86]. Data
780 were filtered and only proteins with at least two valid values in one fraction were kept.

781 Imputation of missing values, which represent low-abundance measurements, was
782 performed with random distribution around the value of instrument sensitivity using
783 default settings of Perseus software[86].
784 The data were analyzed by principle component analysis (PCA). The first two
785 loadings of the PCA were used to plot a two-dimensional graph. Based on a set of marker
786 proteins (376 mitochondrial and 26 peroxisomal, Supplementary Data 2), clusters of
787 proteins co-fractionating with mitochondria and peroxisomes were defined and the
788 proteins within the clusters were further analyzed. This workflow was set up on the basis
789 of the LOPIT protocol[87]. As a result, out of the 4,198 proteins detected, 946 putative
790 mitochondrial and 78 putative peroxisomal proteins were defined. All proteins were
791 subjected to *in silico* predictions concerning their function (BLAST, HHpred[45]) and
792 subcellular localization (Psort II, https://psort.hgc.jp/form2.html; TargetP,
793 http://www.cbs.dtu.dk/services/TargetP/; MultiLoc2, https://abi.inf.uni-
794 tuebingen.de/Services/MultiLoc2).
795
796 **Fluorescence *in situ* hybridization (FISH)**
797 The PCR products of the *NgGspE* and *NgGspF* genes were labelled by alkali-stable
798 digoxigenin-11-dUTP (Roche) using DecaLabel DNA Labeling Kit (Thermo Scientific).
799 Labelled probes were purified on columns of QIAquick Gel Extraction Kit (Qiagen,
800 28704) in a final volume of 50 µL. Labelling efficiencies were tested by dot blotting with
801 anti-digoxigenin alkaline phosphatase conjugate and CSPD chemiluminescence substrate
802 for alkaline phosphatase from DIG High Prime DNA Labelling and Detection Starter Kit
803 II (Roche) according to the manufacturer`s protocol. FISH with digoxigenin-labelled
804 probes was performed essentially according to the procedure described in Zubacova at
805 al.[88] with some modifications. *N. gruberi* cells were pelleted by centrifugation for 10 min
806 at 2,000 × g at 4°C. Cells were placed in hypotonic solution, fixed twice with a freshly
807 prepared mixture of methanol and acetic acid (3:1) and dropped on superfrost microscope
808 slides (ThermoScientific). Preparations for hybridizations were treated with RNase A, 20
809 µg in 100 µL 2 × SSC, for 1 hr at 37°C, washed twice in 2 × SSC for 5 min, dehydrated
810 in a methanol series and air-dried. Slides were treated with 50% acetic acid followed by
811 pepsin treatment and postfixation with 2% paraformaldehyde. Endogenous peroxidase
812 activity of the cell remnants (undesirable for tyramide signal amplification) was
813 inactivated by incubation in 1% hydrogen peroxide, followed by dehydration in a graded
814 methanol series. All slides were denatured together with 2 µL (25 ng) of the probe in 50
815 µL of hybridization mixture containing 50% deionised formamide (Sigma) in 2 × SSC for
816 5 min at 82°C. Hybridizations were carried out overnight. Slides were incubated with
817 tyramide reagent for 7 min. Preparations were counterstained with DAPI in VectaShield
818 and observed under an Olympus IX81 microscope equipped with a Hamamatsu Orca-AG
819 digital camera using the Cell^R imaging software.
820
821 **Heterologous gene expression, preparation of antibodies**
822 The selected Gsp genes from *G. okellyi* and *N. gruberi* were amplified from
823 commercially synthesized templates (Genscript; for primers used for PCR amplification
824 of the coding sequences see Supplementary Data 6) and cloned into the pUG35 vector.
825 The constructs were introduced into *S. cerevisiae* strain YPH499 by the lithium
826 acetate/PEG method. The positive colonies grown on SD-URA plates were incubated

827 with MitoTracker Red CMXRos (Thermo Fisher Scientific) and observed for GFP and
828 MitoTracker fluorescence (using the same equipment as used for FISH, see above). For
829 the expression in *T. brucei,* sequences encoding full-length *Go*GspD, *Go*GspG2,
830 *Ng*GspG1 as well as the first 160 amino acid residues from *Ng*GspG1 were amplified
831 from the commercially synthesized templates and cloned into the pT7 plasmid, encoding
832 either three C-terminal V5 tags (full-length genes) or C-terminal mNeonGreen followed
833 by three V5 tags (*Ng*GspG1 targeting sequence). *T. brucei* cell line SMOX 927 (ref. [89])
834 were grown in SDM79 media[90] supplemented with 10% fetal bovine serum (Gibco).
835 *NotI*-linearized plasmids (50 µg) were nucleofected into procyclic *T. brucei* cells using an
836 Amaxa nucleofector (Lonza) as described before[91]. Expression of the genes was induced
837 by an overnight incubation with doxycycline (1 µg/ml). For bacterial expression, genes
838 encoding *Ng*GspG1 and *Ng*GspEN2A were amplified from commercially synthesized
839 templates and cloned into the pET42b vector (for primers used for PCR amplification of
840 the coding sequences, see Supplementary Data 6). The constructs were introduced into
841 the chemically-competent *E. coli* strain BL21(DE3) and their expression induced by 1
842 mM IPTG. The recombinant proteins were purified under denaturing conditions on Ni-
843 NTA agarose (Qiagen). The purified proteins were used for rat immunization in an in-
844 house animal facility at the Charles University.
845
846 **Immunofluorescence microscopy**
847 Yeast, procyclic *T. brucei* or *N. gruberi* cells were pre-treated by incubation with
848 MitoTracker CMX Ros (1:000 dilution) for 20 min to stain mitochondria, washed twice
849 in PBS and placed on coverslips. After a 5-min incubation, the cells were fixed with 4%
850 PFA in PBS for 15 min. The solution was replaced by 0.1% Triton X-100 in PBS and the
851 slides were incubated for 15 min. The slides were then treated with blocking buffer (1%
852 BSA and 0.033% Triton X-100 in PBS) for 1 hr at room temperature. After blocking, the
853 samples were stained overnight at 4°C with a blocking solution supplemented with the
854 primary antibody (in-house-produced rat anti-*Ng*GspG1 and anti-*Ng*GspEN2A
855 antibodies, dilutions 1:100, rat anti-V5 antibody Abcam, dilution 1:1000 dilution). The
856 slides were washed three times for 10 min with 0.033% Triton X-100 and incubated with
857 an anti-rat antibody conjugated with Alexa Fluor® 488 (1:1000 dilution, Thermo Fisher
858 Scientific) in blocking buffer for 1 hr at room temperature. Slides were washed twice in
859 PBS supplemented with 0.033% Triton X-100 for 10 min followed by a single 10-min
860 wash with PBS only. The slides were mounted in Vectashield containing DAPI (Vector
861 laboratories). Static images were acquired on a Leica SP8 FLIM inverted confocal
862 microscope equipped with 405 nm and white light (470-670 nm) lasers and a FOV SP8
863 scanner using an HC PL APO CS2 63x/1.4 NA oil-immersion objective. Laser
864 wavelengths and intensities were controlled by a combination of AOTF and AOBS
865 separately for each channel. Emitting fluorescence was captured by internal spectrally-
866 tunable HyD detectors. Imaging was controlled by the Leica LAS-X software. Images
867 were deconvolved using the SVI Huygens Professional software (Scientific Volume
868 Imaging) with the CMLE algorithm. Maximum intensity projections and
869 brightness/contrast corrections were performed in the FIJI ImageJ software.
870
871 **Purification of native *Ng*GspD and *Ng*GspG**

His-tagged *Ng*GspD carrying the signal peptide of *E. coli* DsbA was produced in the *E. coli* strain BL21 (DE3) in auto-induction media (50 mM Na_2HPO_4, 50 mM KH_2PO_4, 2% Tryptone, 0.5% Yeast extract, 85 mM NaCl, 0.5% glycerol, 0.05% glucose and 0.2% lactose) as described in ref. [92]. The cells were grown at 37°C for 16 hr, centrifuged at 6,000 × g for 15 min at 4°C and resuspended in 20 mM Tris pH 8, 5 mM EDTA. Bacteria were incubated for 30 min on ice in the presence of lysozyme (1 mg/ml) and DNAse, and lysed in a French press. The cell lysate was centrifuged at 6,000 × g for 15 min at 4°C to pellet cell debris. The cleared lysate was then centrifuged at 100,000 × g for 2 h at 4°C and membrane pellet was washed twice in 20 mM Tris pH 8, with 1-hr centrifugation steps (100,000 × g at 4°C). The membranes were resuspended to the final protein concentration of 1 mg/ml in 50 mM Tris pH 8, 250 mM NaCl, 1% Zwittergent 3-14, and solubilized for 2 hr at 4°C. The sample was then centrifuged at 100,000 × g for 1 hr at 4°C. His-tagged proteins from the supernatant were incubated overnight with Ni-NTA agarose (Qiagen) at 4°C. The next day, the agarose was collected on a column and washed with 50 ml of 50 mM Tris pH 8, 250 mM NaCl, 20 mM Imidazole, 0.5% Zwittergent 3-14. Bound proteins were eluted by 5 × 0.5 ml of 50 mM Tris pH 8, 250 mM NaCl, 250 mM imidazole, Zwittergent 3-14. Collected fractions were analyzed by SDS PAGE and western blotting. Selected samples were then pooled together, re-buffered into 50 mM Tris pH 8, 250 mM NaCl, 0.5% Zwittergent 3-14, and analyzed by size exclusion chromatography.

For *Ng*GspG1, the BL21 cells expressing the pseudopilin domain lacking the N-terminal hydrophobic part were collected after 4 hr of IPTG induction at 37°C. The cells were collected, washed with PBS, and then resuspended in 35 ml of 50 mM Tris, 100 mM NaCl, pH 8 with added inhibitors (cOMPLETE™ tablets, EDTA-free, Roche), DNAse (1 mg/ml), 5 mM $MgCl_2$ and lysozyme (1 mg/ml). The suspension was incubated on ice for 30 min. After lysis via French press the resulting suspension was spun down for 20 min at 100,000 × g, 4 °C. Ni-NTA agarose beads (1 ml; Qiagen), washed and resuspended in loading buffer (50 mM Tris, 100 mM NaCl, 10 mM imidazole, pH 8), were added to the supernatant. The supernatant was incubated with the beads for 1 hr on a tube rotator at 4°C. The suspension was then applied to a column. The beads on the column were then washed with 8 ml of wash buffer (50mM Tris, 100 mM NaCl, 20 mM imidazole, pH 8). The protein was eluted by 4 ml elution buffer I (50 mM Tris, 100 mM NaCl, 100 mM imidazole, pH 8) and 4 ml of elution buffer II (50 mM Tris, 100 mM NaCl, 150 mM imidazole, pH 8). All elutions were then rebuffered to 50 mM Tris pH 8, 100 mM NaCl, using Amicon Ultra-4 10k centrifugal filter tubes. The protein binding was measured by the microscale thermophoresis technique on a Monolith NT.115 instrument (Nanotemper). Protein (10 nM) in 50 mM HEPES buffer with 50 mM NaCl was labelled with Red-NHS dye NT-647. Maximum concentration of titrated protein was 50 µM.

Black lipid membrane measurements

Planar lipid membrane experiments were performed as described previously[93]. The electrolyte solution contained 1 M KCl and 10 mM Tris-HCl (pH 7.4). 5 µl of purified protein (80 ng/ml) was mixed with 1500 µl KCl and was added to the *cis* compartment with a positive electrode, whereas the *trans* compartment was grounded. Planar lipid membrane was formed across a 0.5-mm aperture by painting the *E. coli* polar lipids

918 extract (3% wt/vol, Avanti Polar Lipids, Alabaster, USA) dissolved in *n*-decan and
919 butanol (9:1). The membrane current was registered using Ag/AgCl electrodes with salt
920 bridges connected to an LCA-200-10G amplifier (Femto, Germany) and digitized with a
921 KPCI-3108 16-Bit A/D card (Keithley Instruments, USA) with a 1-kHz sampling rate.
922 Single-pore recordings were processed in the program QuB[94]. The histogram of single-
923 pore conductance (n=100) was constructed by kernel density estimation (with the
924 Gaussian kernel of 100-pS width) to overcome bin edge effects.
925
926 ***In vitro* protein translation and mitochondrial protein import**
927 The *Go*GspD and *Ng*GspD genes were amplified from commercially synthesized
928 templates (for primers used for PCR amplification of the coding sequences, see
929 Supplementary Data 6) and cloned into pDHFR vector provided in the PURExpress *In*
930 *Vitro* Protein Synthesis Kit (NEB). The translation into liposomes was done as described
931 previously[95] and the output was analyzed by Blue Native PAGE using 2% digitonin and
932 NativePAGE Novex 4-16% Bis-Tris Protein Gel (Thermo Fisher Scientific). For the *in*
933 *vitro* mitochondrial protein import, the mitochondria were isolated from *S. cerevisiae*
934 YPH499 according to the method described in ref. [96]. The *in vitro*-translated *Ng*GspD and
935 Su9-DHFR chimeric construct[41] were incubated with mitochondria as described in ref. [97].
936 The import reactions were incubated with 50 µg/ml of trypsin for 30 min on ice to
937 remove unimported protein precursor.
938
939 **Testing protein interactions using two-hybrid systems**
940 Bacterial two-hybrid system (BACTH) analysis was performed as described before[98].
941 Gsp genes were amplified with specific primers (listed in Supplementary Data 6) and
942 cloned into pKT25 and pUT18c plasmids. *E. coli* strain DHT1 competent cells were co-
943 transformed with two plasmids with different combinations of Gsp genes. Co-
944 transformants were selected on LB plates with ampicillin (Ap) (100 µg/mL) and
945 kanamycin (Km) (25 µg/mL). Colonies were grown at 30°C for 48 to 96 hr. From each
946 plate three colonies were picked, transferred to 1 mL of LB medium with Ap and Km,
947 and grown overnight at 30°C with shaking. Next day, precultures (0.25 mL) were
948 inoculated to 5 mL of LB medium with Ap, Km and 1 mM IPTG. Cultures were grown
949 with shaking at 30°C to OD_{600} of about 1-1.5. Bacteria (0.5 mL) were mixed with 0.5 mL
950 of Z buffer and β-galactosidase activity was measured[99].
951 The yeast two-hybrid system (Y2H) was employed as described in ref. [100]. Cells
952 of *S. cerevisiae* strain AH109 were co-transformed with two plasmids (pGADT7,
953 pGBKT7) with different combinations of Gsp genes. Co-transformants were selected on
954 double-dropout SD-Leu/-Trp and triple-dropout SD-Leu/-Trp/-His plates. The colonies
955 were grown for a few days. Positive colonies from the triple dropout were grown
956 overnight at 30°C with shaking and then the serial dilution test was performed on double-
957 and triple-dropout plates.
958
959 **Data availability**
960 All sequences of Gsp and Gcp proteins analysed in the study are provided in
961 Supplementary Data 1. Gsp and Gcp genes extracted from an unpublished *Malawimonas*
962 *jakobiformis* genome assembly have been deposited at GenBank with accession numbers
963 MT460910-MT460938 (https://www.ncbi.nlm.nih.gov/nuccore/MT460910.1/ etc.). Raw

964 genome sequencing reads from "*Malawimonas californiana*" and *Neovahlkampfia*
965 *damariscottae* are available from NCBI under the BioProject PRJNA549687
966 (https://www.ncbi.nlm.nih.gov/bioproject/?term=prjna549687). The genome assembly of
967 *N. damariscottae* has been deposited at GenBank with the accession number
968 JABLTG000000000 (https://www.ncbi.nlm.nih.gov/nuccore/JABLTG000000000.1/).
969 The transcriptome assembly of *Gefionella okellyi*, the genome assembly and predicted
970 proteins of "*Malawimonas californiana*", and partial genome assemblies of
971 *Reclinomonas americana* are available from
972 https://megasun.bch.umontreal.ca/papers/T2SS-2020/. The mass spectrometry proteomics
973 data have been deposited in the ProteomeXchange Consortium via the PRIDE[101] partner
974 repository with the dataset identifier PXD007764
975 (https://www.ebi.ac.uk/pride/archive/projects/PXD007764). Other relevant data (e.g.
976 multiple sequence alignments used for phylogenetic analyses) are available from the
977 authors upon request.
978
979 **References**
980

981 1. Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. The origin and diversification
982 of mitochondria. *Curr. Biol.* **27**, R1177–R1192 (2017).
983 2. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial
984 origin outside the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).
985 3. Leger, M. M. *et al.* An ancestral bacterial division system is widespread in
986 eukaryotic mitochondria. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 10239–46 (2015).
987 4. Beech, P. L. Mitochondrial FtsZ in a chromophyte alga. *Science* **287**, 1276–1279
988 (2000).
989 5. Gray, M. W. *et al.* The draft nuclear genome sequence and predicted mitochondrial
990 proteome of Andalucia godoyi, a protist with the most gene-rich and bacteria-like
991 mitochondrial genome. *BMC Biol.* **18**, (2020).
992 6. Natale, P., Brüser, T. & Driessen, A. J. M. Sec- and Tat-mediated protein secretion
993 across the bacterial cytoplasmic membrane—Distinct translocases and
994 mechanisms. *Biochim. Biophys. Acta - Biomembr.* **1778**, 1735–1756 (2008).
995 7. Costa, T. R. D. *et al.* Secretion systems in Gram-negative bacteria: structural and
996 mechanistic insights. *Nat. Rev. Microbiol.* **13**, 343–359 (2015).
997 8. Dolezal, P., Likic, V., Tachezy, J. & Lithgow, T. Evolution of the molecular
998 machines for protein import into mitochondria. *Science* **313**, 314–8 (2006).
999 9. Petru, M. *et al.* Evolution of mitochondrial TAT translocases illustrates the loss of
1000 bacterial protein transport machines in mitochondria. *BMC Biol.* **16**, (2018).
1001 10. Schäfer, K., Künzler, P., Klingl, A., Eubel, H. & Carrie, C. The plant
1002 mitochondrial TAT pathway is essential for complex III biogenesis. *Curr. Biol.* **30**,
1003 840-853.e5 (2020).
1004 11. Lang, B. F. *et al.* An ancestral mitochondrial DNA resembling a eubacterial
1005 genome in miniature. *Nature* **387**, 493–7 (1997).
1006 12. Burger, G., Gray, M. W., Forget, L. & Lang, B. F. Strikingly bacteria-like and
1007 gene-rich mitochondrial genomes throughout jakobid protists. *Genome Biol. Evol.*
1008 **5**, 418–38 (2013).
1009 13. Tong, J. *et al.* Ancestral and derived protein import pathways in the mitochondrion

1010 of *Reclinomonas america*. *Mol. Biol. Evol.* **28**, 1581–91 (2011).
1011 14. Korotkov, K. V., Sandkvist, M. & Hol, W. G. J. The type II secretion system:
1012 biogenesis, molecular architecture and mechanism. *Nat. Rev. Microbiol.* **10**, 336–
1013 51 (2012).
1014 15. Thomassin, J.-L., Santos Moreno, J., Guilvout, I., Tran Van Nhieu, G. & Francetic,
1015 O. The trans-envelope architecture and function of the type 2 secretion system:
1016 new insights raising new questions. *Mol. Microbiol.* **105**, 211–226 (2017).
1017 16. Berry, J.-L. & Pelicic, V. Exceptionally widespread nanomachines composed of
1018 type IV pilins: the prokaryotic Swiss Army knives. *FEMS Microbiol. Rev.* **39**,
1019 134–54 (2015).
1020 17. Nivaskumar, M. & Francetic, O. Type II secretion system: a magic beanstalk or a
1021 protein escalator. *Biochim. Biophys. Acta* **1843**, 1568–77 (2014).
1022 18. Denise, R., Abby, S. S. & Rocha, E. P. C. Diversification of the type IV filament
1023 superfamily into machines for adhesion, protein secretion, DNA uptake, and
1024 motility. *PLoS Biol.* **17**, e3000390 (2019).
1025 19. Guilvout, I. *et al.* In vitro multimerization and membrane insertion of bacterial
1026 outer membrane secretin PulD. *J. Mol. Biol.* **382**, 13–23 (2008).
1027 20. Yan, Z., Yin, M., Xu, D., Zhu, Y. & Li, X. Structural insights into the secretin
1028 translocation channel in the type II secretion system. *Nat. Struct. Mol. Biol.* **24**,
1029 177–183 (2017).
1030 21. Py, B., Loiseau, L. & Barras, F. An inner membrane platform in the type II
1031 secretion machinery of Gram-negative bacteria. *EMBO Rep.* **2**, 244–248 (2001).
1032 22. Wang, X. *et al.* Cysteine scanning mutagenesis and disulfide mapping analysis of
1033 arrangement of GspC and GspD protomers within the type 2 secretion system. *J.*
1034 *Biol. Chem.* **287**, 19082–93 (2012).
1035 23. Korotkov, K. V. *et al.* Structural and functional studies on the interaction of GspC
1036 and GspD in the type II secretion system. *PLoS Pathog.* **7**, e1002228 (2011).
1037 24. Korotkov, K. V & Hol, W. G. J. Structure of the GspK-GspI-GspJ complex from
1038 the enterotoxigenic Escherichia coli type 2 secretion system. *Nat. Struct. Mol. Biol.*
1039 **15**, 462–8 (2008).
1040 25. Peabody, C. R. *et al.* Type II protein secretion and its relationship to bacterial type
1041 IV pili and archaeal flagella. *Microbiology* **149**, 3051–72 (2003).
1042 26. Lu, C. *et al.* Hexamers of the type II secretion ATPase GspE from Vibrio cholerae
1043 with Increased ATPase activity. *Structure* **21**, 1707–1717 (2013).
1044 27. Lax, G. *et al.* Hemimastigophora is a novel supra-kingdom-level lineage of
1045 eukaryotes. *Nature* **564**, 410–414 (2018).
1046 28. Adl, S. M. *et al.* The revised classification of eukaryotes. *J. Eukaryot. Microbiol.*
1047 **59**, 429–514 (2012).
1048 29. Derelle, R. *et al.* Bacterial proteins pinpoint a single eukaryotic root. *Proc. Natl.*
1049 *Acad. Sci. U. S. A.* **112**, E693-9 (2015).
1050 30. Karnkowska, A. *et al.* A eukaryote without a mitochondrial organelle. *Curr. Biol.*
1051 **26**, 1274–84 (2016).
1052 31. Heiss, A. A. *et al.* Combined morphological and phylogenomic re-examination of
1053 malawimonads, a critical taxon for inferring the evolutionary history of
1054 eukaryotes. *R. Soc. Open Sci.* **5**, 171707 (2018).
1055 32. Brown, M. W. *et al.* Phylogenomics places orphan protistan lineages in a novel

1056 eukaryotic super-group. *Genome Biol. Evol.* **10**, 427–433 (2018).
1057 33. Mach, J. *et al.* Iron economy in *Naegleria gruberi* reflects its metabolic flexibility.
1058 *Int. J. Parasitol.* **48**, 719–727 (2018).
1059 34. Nouwen, N. *et al.* Secretin PulD: Association with pilot PulS, structure, and ion-
1060 conducting channel formation. *Proc. Natl. Acad. Sci.* **96**, 8173–8177 (1999).
1061 35. Hardie, K. R., Lory, S. & Pugsley, A. P. Insertion of an outer membrane protein in
1062 *Escherichia coli* requires a chaperone-like protein. *EMBO J.* **15**, 978–88 (1996).
1063 36. Dunstan, R. A. *et al.* Assembly of the secretion pores GspD, Wza and CsgG into
1064 bacterial outer membranes does not require the Omp85 proteins BamA or TamA.
1065 *Mol. Microbiol.* **97**, 616–629 (2015).
1066 37. Collin, S., Guilvout, I., Chami, M. & Pugsley, A. P. YaeT-independent
1067 multimerization and outer membrane association of secretin PulD. *Mol. Microbiol.*
1068 **64**, 1350–1357 (2007).
1069 38. Korotkov, K. V, Pardon, E., Steyaert, J. & Hol, W. G. J. Crystal structure of the N-
1070 terminal domain of the secretin GspD from ETEC determined with the assistance
1071 of a nanobody. *Structure* **17**, 255–265 (2009).
1072 39. Guilvout, I. *et al.* Prepore stability controls productive folding of the BAM
1073 independent multimeric outer membrane secretin PulD. *J. Biol. Chem.* **292**, 328–
1074 338 (2017).
1075 40. Chernyatina, A. A. & Low, H. H. Core architecture of a bacterial type II secretion
1076 system. *Nat. Commun.* **10**, 5437 (2019).
1077 41. Pfanner, N., Tropschug, M. & Neupert, W. Mitochondrial protein import:
1078 Nucleoside triphosphates are involved in conferring import-competence to
1079 precursors. *Cell* **49**, 815–823 (1987).
1080 42. Yin, M., Yan, Z. & Li, X. Structural insight into the assembly of the Type II
1081 secretion system pilotin-Secretin complex from enterotoxigenic *Escherichia coli*.
1082 *Nat Microbiol* (2018). doi:10.2210/PDB5ZDH/PDB
1083 43. Guilvout, I. *et al.* In Vitro Multimerization and Membrane Insertion of Bacterial
1084 Outer Membrane Secretin PulD. *J. Mol. Biol.* **382**, 13–23 (2008).
1085 44. Sauvonnet, N., Vignon, G., Pugsley, A. P. & Gounon, P. Pilus formation and
1086 protein secretion by the same machinery in Escherichia coli. *EMBO J.* **19**, 2221–
1087 2228 (2000).
1088 45. Alva, V., Nam, S.-Z., Söding, J. & Lupas, A. N. The MPI bioinformatics Toolkit
1089 as an integrative platform for advanced protein sequence and structure analysis.
1090 *Nucleic Acids Res.* **44**, W410–W415 (2016).
1091 46. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The
1092 Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**,
1093 845–858 (2015).
1094 47. López-Castilla, A. *et al.* Structure of the calcium-dependent type 2 secretion
1095 pseudopilus. *Nat. Microbiol.* **2**, 1686–1695 (2017).
1096 48. Nivaskumar, M. *et al.* Pseudopilin residue E5 is essential for recruitment by the
1097 type 2 secretion system assembly platform. *Mol. Microbiol.* **101**, 924–41 (2016).
1098 49. Lu, C., Korotkov, K. V. & Hol, W. G. J. Crystal structure of the full-length
1099 ATPase GspE from the Vibrio vulnificus type II secretion system in complex with
1100 the cytoplasmic domain of GspL. *J. Struct. Biol.* **187**, 223–235 (2014).
1101 50. Cerdà-Costa, N. & Xavier Gomis-Rüth, F. Architecture and function of

1102 metallopeptidase catalytic domains. *Protein Sci.* **23**, 123–144 (2014).
1103 51. Tabach, Y. *et al.* Identification of small RNA pathway genes using patterns of
1104 phylogenetic conservation and divergence. *Nature* **493**, 694–698 (2012).
1105 52. Nevers, Y. *et al.* Insights into ciliary genes and evolution from multi-level
1106 phylogenetic profiling. *Mol. Biol. Evol.* **34**, 2016–2034 (2017).
1107 53. McLaughlin, L. S., Haft, R. J. F. & Forest, K. T. Structural insights into the Type
1108 II secretion nanomachine. *Current Opinion in Structural Biology* **22**, 208–216
1109 (2012).
1110 54. Okuno, D., Iino, R. & Noji, H. Rotation and structure of FoF1-ATP synthase. *J.*
1111 *Biochem.* **149**, 655–664 (2011).
1112 55. Tomko, R. J. & Hochstrasser, M. Molecular architecture and assembly of the
1113 eukaryotic proteasome. *Annu. Rev. Biochem.* **82**, 415–445 (2013).
1114 56. Babbitt, S. E., Sutherland, M. C., San Francisco, B., Mendez, D. L. & Kranz, R. G.
1115 Mitochondrial cytochrome c biogenesis: no longer an enigma. *Trends Biochem.*
1116 *Sci.* **40**, 446–55 (2015).
1117 57. Hartl, F. U., Schmidt, B., Wachter, E., Weiss, H. & Neupert, W. Transport into
1118 mitochondria and intramitochondrial sorting of the Fe/S protein of ubiquinol-
1119 cytochrome c reductase. *Cell* **47**, 939–51 (1986).
1120 58. Francisco, T. *et al.* Protein transport into peroxisomes: Knowns and unknowns.
1121 *BioEssays* **39**, 1700047 (2017).
1122 59. Nguyen, B. D. & Valdivia, R. H. Virulence determinants in the obligate
1123 intracellular pathogen *Chlamydia trachomatis* revealed by forward genetic
1124 approaches. *Proc. Natl. Acad. Sci.* **109**, 1263–1268 (2012).
1125 60. Keeling, P. J. *et al.* The marine microbial eukaryote transcriptome sequencing
1126 project (MMETSP): illuminating the functional diversity of eukaryotic life in the
1127 oceans through transcriptome sequencing. *PLoS Biol.* **12**, e1001889 (2014).
1128 61. Matasci, N. *et al.* Data access for the 1,000 Plants (1KP) project. *Gigascience* **3**, 17
1129 (2014).
1130 62. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein
1131 database search programs. *Nucleic Acids Res.* **25**, 3389–402 (1997).
1132 63. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence
1133 similarity searching. *Nucleic Acids Res.* **39**, W29-37 (2011).
1134 64. Huerta-Cepas, J., Dopazo, H., Dopazo, J. & Gabaldón, T. The human phylome.
1135 *Genome Biol.* **8**, R109 (2007).
1136 65. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular
1137 localization of proteins based on their N-terminal amino acid sequence. *J. Mol.*
1138 *Biol.* **300**, 1005–16 (2000).
1139 66. Armenteros, J. J. A. *et al.* Detecting sequence signals in targeting peptides using
1140 deep learning. *Life Sci. Alliance* **2**, (2019).
1141 67. Claros, M. G. MitoProt, a Macintosh application for studying mitochondrial
1142 proteins. *Comput. Appl. Biosci.* **11**, 441–7 (1995).
1143 68. Fukasawa, Y. *et al.* MitoFates: improved prediction of mitochondrial targeting
1144 sequences and their cleavage sites. *Mol. Cell. Proteomics* **14**, 1113–26 (2015).
1145 69. Neuberger, G., Maurer-Stroh, S., Eisenhaber, B., Hartig, A. & Eisenhaber, F.
1146 Prediction of peroxisomal targeting signal 1 containing proteins from amino acid
1147 sequence. *J. Mol. Biol.* **328**, 581–92 (2003).

1148 70. Käll, L., Krogh, A. & Sonnhammer, E. L. L. Advantages of combined
1149 transmembrane topology and signal peptide prediction--the Phobius web server.
1150 *Nucleic Acids Res.* **35**, W429-32 (2007).
1151 71. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable
1152 future. *Nucleic Acids Res.* **44**, D279-85 (2016).
1153 72. de Lima Morais, D. A. *et al.* SUPERFAMILY 1.75 including a domain-centric
1154 gene ontology method. *Nucleic Acids Res.* **39**, D427-34 (2011).
1155 73. Frickey, T. & Lupas, A. CLANS: a Java application for visualizing protein
1156 families based on pairwise similarity. *Bioinformatics* **20**, 3702–3704 (2004).
1157 74. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high
1158 throughput. *Nucleic Acids Res.* **32**, 1792–7 (2004).
1159 75. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software
1160 version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–
1161 780 (2013).
1162 76. Abby, S. S. *et al.* Identification of protein secretion systems in bacterial genomes.
1163 *Sci. Rep.* **6**, 23080 (2016).
1164 77. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with
1165 Entropy): a new software for selection of phylogenetic informative regions from
1166 multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
1167 78. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast
1168 and effective stochastic algorithm for estimating maximum-likelihood
1169 phylogenies. *Mol. Biol. Evol.* **32**, 268–74 (2015).
1170 79. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site
1171 heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–
1172 109 (2004).
1173 80. Bienert, S. *et al.* The SWISS-MODEL Repository-new features and functionality.
1174 *Nucleic Acids Res.* **45**, D313–D319 (2017).
1175 81. Guex, N., Peitsch, M. C. & Schwede, T. Automated comparative protein structure
1176 modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective.
1177 *Electrophoresis* **30**, S162–S173 (2009).
1178 82. Biasini, M. *et al.* SWISS-MODEL: modelling protein tertiary and quaternary
1179 structure using evolutionary information. *Nucleic Acids Res.* **42**, W252-8 (2014).
1180 83. Jedelský, P. L. *et al.* The minimal proteome in the reduced mitochondrion of the
1181 parasitic protist *Giardia intestinalis*. *PLoS One* **6**, e17285 (2011).
1182 84. Černá, M., Kuntová, B., Talacko, P., Stopková, R. & Stopka, P. Differential
1183 regulation of vaginal lipocalins (OBP, MUP) during the estrous cycle of the house
1184 mouse. *Sci. Rep.* **7**, 11674 (2017).
1185 85. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed
1186 normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell.*
1187 *Proteomics* **13**, 2513–2526 (2014).
1188 86. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis
1189 of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
1190 87. Dunkley, T. P. J., Watson, R., Griffin, J. L., Dupree, P. & Lilley, K. S.
1191 Localization of organelle proteins by isotope tagging (LOPIT). *Mol. Cell.*
1192 *Proteomics* **3**, 1128–1134 (2004).
1193 88. Zubáčová, Z., Krylov, V. & Tachezy, J. Fluorescence in situ hybridization (FISH)

1194 mapping of single copy genes on *Trichomonas vaginalis* chromosomes. *Mol.*
1195 *Biochem. Parasitol.* **176**, 135–137 (2011).
1196 89. Poon, S. K., Peacock, L., Gibson, W., Gull, K. & Kelly, S. A modular and
1197 optimized single marker system for generating *Trypanosoma brucei* cell lines
1198 expressing T7 RNA polymerase and the tetracycline repressor. *Open Biol.* **2**,
1199 110037 (2012).
1200 90. Brun, R. & Schönenberger. Cultivation and *in vitro* cloning or procyclic culture
1201 forms of Trypanosoma brucei in a semi-defined medium. Short communication.
1202 *Acta Trop.* **36**, 289–92 (1979).
1203 91. Kaurov, I. *et al.* The diverged trypanosome MICOS complex as a hub for
1204 mitochondrial cristae shaping and protein import. *Curr. Biol.* **28**, 3393-3407.e5
1205 (2018).
1206 92. Studier, F. W. Protein production by auto-induction in high-density shaking
1207 cultures. (2005). doi:10.1016/j.pep.2005.01.016
1208 93. Seydlová, G. *et al.* Lipophosphonoxins II: design, synthesis, and properties of
1209 novel broad spectrum antibacterial agents. *J. Med. Chem.* **60**, 6098–6118 (2017).
1210 94. Nicolai, C. & Sachs, F. Solving ion channel kinetics with the QuB software.
1211 *Biophys. Rev. Lett.* **08**, 191–211 (2013).
1212 95. Pyrihová, E. *et al.* A Single Tim translocase in the mitosomes of *Giardia*
1213 *intestinalis* illustrates convergence of protein import machines in anaerobic
1214 eukaryotes. *Genome Biol. Evol.* **10**, 2813–2822 (2018).
1215 96. Daum, G., Böhni, P. C. & Schatz, G. Import of proteins into mitochondria.
1216 Cytochrome b2 and cytochrome c peroxidase are located in the intermembrane
1217 space of yeast mitochondria. *J. Biol. Chem.* **257**, 13028–13033 (1982).
1218 97. Dolezal, P. *et al.* Legionella pneumophila secretes a mitochondrial carrier protein
1219 during infection. *PLoS Pathog.* **8**, e1002459 (2012).
1220 98. Battesti, A. & Bouveret, E. The bacterial two-hybrid system based on adenylate
1221 cyclase reconstitution in *Escherichia coli*. *Methods* **58**, 325–334 (2012).
1222 99. Miller, J. H. *Experiments in molecular genetics*. (Cold Spring Harbor Laboratory,
1223 1972).
1224 100. Fields, S. & Song, O. A novel genetic system to detect protein-protein interactions.
1225 *Nature* **340**, 245–246 (1989).
1226 101. Vizcaíno, J. A. *et al.* 2016 update of the PRIDE database and its related tools.
1227 *Nucleic Acids Res.* **44**, 11033–11033 (2016).
1228 102. Adl, S. M. *et al.* Revisions to the classification, nomenclature, and diversity of
1229 eukaryotes. *J. Eukaryot. Microbiol.* **66**, 4–119 (2019).
1230 103. Strassert, J. F. H., Jamy, M., Mylnikov, A. P., Tikhonenkov, D. V & Burki, F.
1231 New Phylogenomic Analysis of the Enigmatic Phylum Telonemia Further
1232 Resolves the Eukaryote Tree of Life. *Mol. Biol. Evol.* **36**, 757–765 (2019).
1233
1234
1235 **Acknowledgements**
1236 We would like to thank Michelle Leger and Alastair Simpson for granting us access, prior
1237 to publication, to their transcriptomic data from *A. godoyi* and *G. okellyi*, respectively,
1238 data that were instrumental in annotating the Gsp and Gcp genes in our genome
1239 assemblies. This work was supported by Czech Science Foundation grants 13-29423S to

PD and 18-18699S to ME, and the KONTAKT II grant LH15253 provided by Ministry of Education, Youth and Sports of CR (MEYS) to PD. This work was also supported by MEYS within the National Sustainability Program II (Project BIOCEV-FAR, LQ1604) the project BIOCEV (CZ.1.05/1.1.00/02.0109), and the project "Centre for research of pathogenicity and virulence of parasites" (No. CZ.02.1.01/0.0/0.0/16_019/0000759) funded by European Regional Development Fund (ERDF) and MEYS and by Moore-Simons Project on the Origin of the Eukaryotic Cell to PD. The work in the OF laboratory was funded by the ANR-14-CE09-0004 grant. JP was supported by a grant from the Gordon and Betty Moore Foundation to ADT. This work was supported by The Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project "IT4Innovations National Supercomputing Center – LM2015070".

Author information

Affiliations

Department of Parasitology, Faculty of Science, Charles University, BIOCEV, Průmyslová 595, Vestec, 252 42, Czech Republic
L. Horváthová, V. Žárský, A. Motyčková, V. Klápšťová, M. Vinopalová, L. Marková, L. Voleman, 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

Department of Genetics and Microbiology, Faculty of Science, Charles University, Viničná 7, Prague 2, 128 44, Czech Republic
K. Hryzáková

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

Center for Cellular Imaging and NanoAnalytics, University of Basel, Mattenstrasse 26, CH-4058 Basel, Switzerland
M. Chami

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

Competing interests

None declared.

Corresponding author

Correspondence to M. Eliáš or P. Doležal.

Figure Legends

Fig. 1 Some eukaryotes harbour homologues of core components of the bacterial T2SS machinery. (A) Schematic representation of the complete bacterial T2SS; subunits having identified eukaryotic homologues are highlighted in colour. For simplicity, GspDN represents in the figure three different eukaryotic proteins (GspDN1 to GspDN3), together corresponding to two versions of a conserved domain present as a triplicate in the N-terminal region of the bacterial GspD protein. (B) Phylogenetic distribution of eukaryotic homologues of bacterial T2SS subunits (Gsp proteins) and co-occurring proteins (Gcp). Core T2SS components (cyan), eukaryote-specific T2SS components (dark blue), Gcp proteins carrying protein domains found in eukaryotes (magenta), and Gcp proteins without discernible homologues or with homologues only in prokaryotes (pink). Coloured sections indicate proteins found to be present in genome or transcriptome data; white sections, proteins absent from complete genome data; grey sections, proteins absent from transcriptome data. The asterisk indicates the presence of the particular protein in at least two of the three *Percolomonas* or at least one of the two Hemimastigophora species analyzed. The species name in parentheses has not yet been formally published. Sequence IDs and additional details on the eukaryotic Gsp and Gcp proteins are provided in Supplementary Data 1. The tree topology and taxon names reflect most recent phylogenomic studies of eukaryotes[32,102,103]; the root (LECA) is placed according to ref. [29].

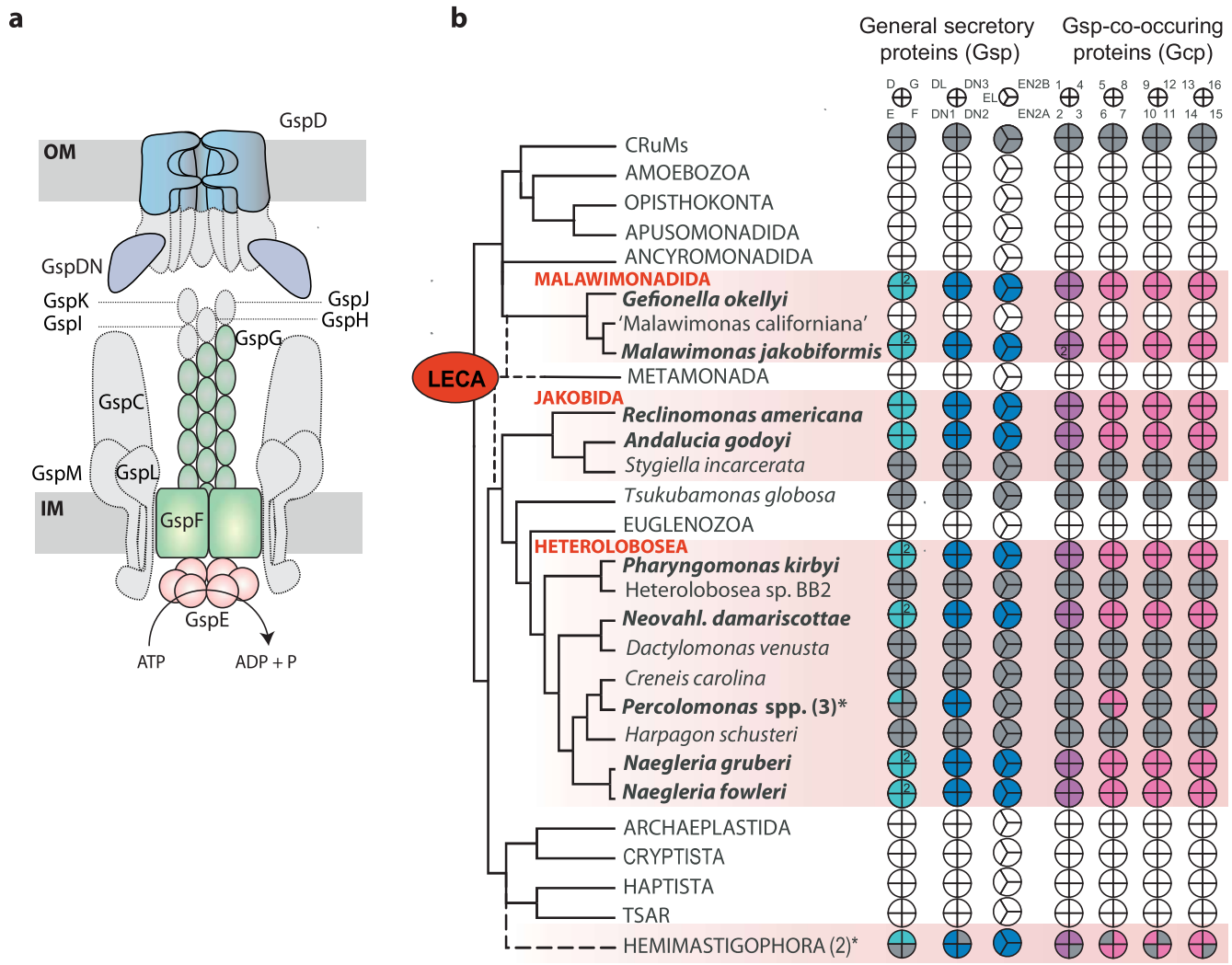
Fig. 2 Eukaryotic Gsp homologues are monophyletic. Maximum likelihood (ML) phylogenetic tree of eukaryotic and selected bacterial GspG proteins demonstrating the monophyletic origin of the eukaryotic GspG proteins and their separation from bacterial homologues by a long branch (the tree inferred using IQ-TREE). Branch support (bootstrap) was assessed by ML ultrafast bootstrapping and is shown only for branches where support is > 50.

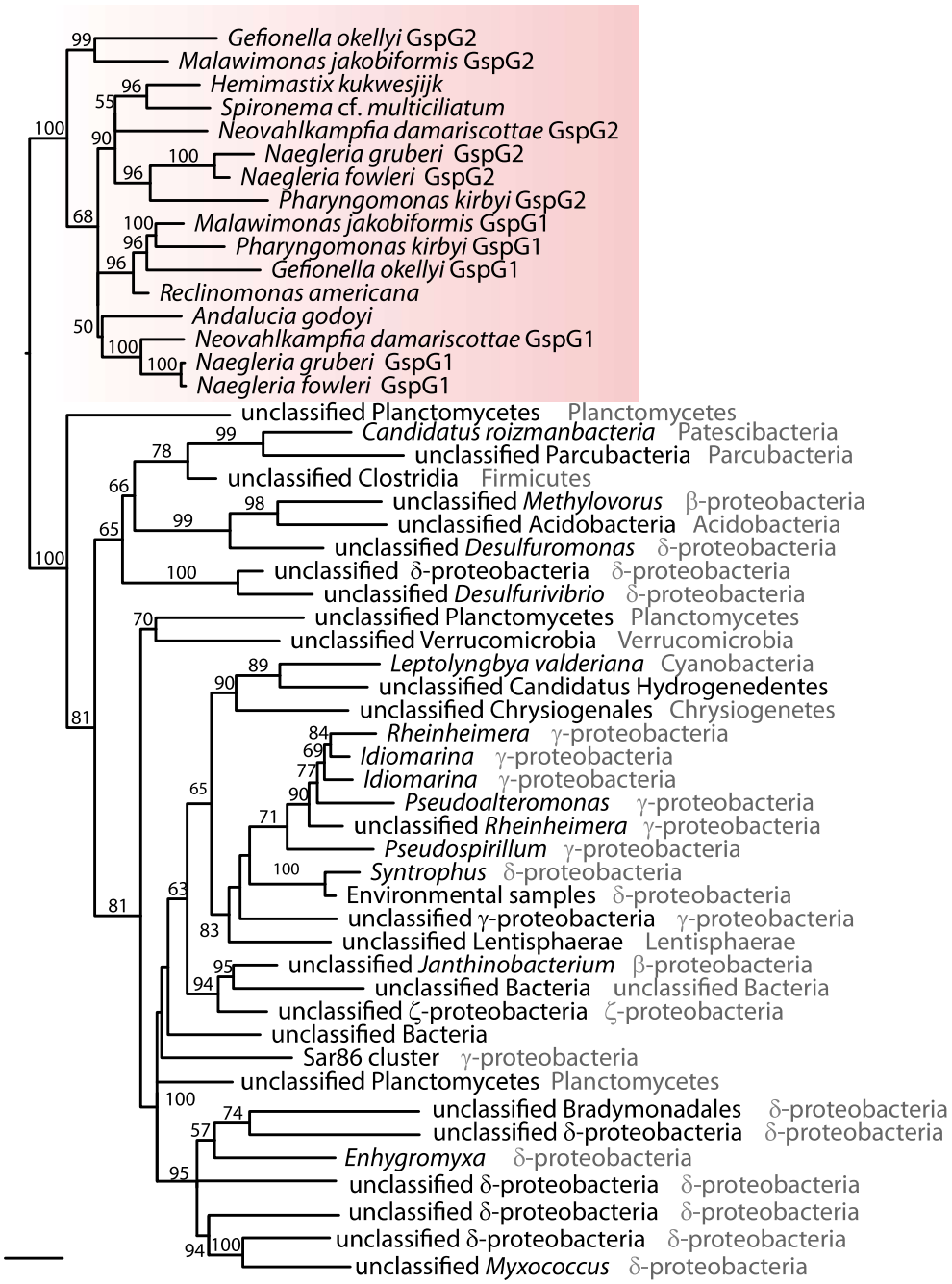
Fig. 3 Eukaryotic T2SS components are localized in mitochondria. (A) *S. cerevisiae* expressing *G. okellyi* and *N. gruberi* Gsp proteins as C-terminal GFP fusions stained with MitoTracker red CMX ROS to visualize mitochondria. (B) Expression of *G. okellyi* GspD and GspG2 with the C-terminal V5 Tag in *T. brucei* visualized by immunofluorescence; the cells are co-stained with MitoTracker red. (C) Cellular fractions of *N. gruberi* labelled by specific polyclonal antibodies raised against GspG1, GspEN2A and mitochondrial (alternative oxidase – AOX, TatC,) cytosolic (hemerythrin) marker proteins (D) Immunofluorescence microscopy of *N. gruberi* labelled with specific polyclonal antibodies raised against GspG1 and GspEN2A, and co-stained with MitoTracker red. Scale bar (parts A-D), 10 µm. (for A-D, representative images of multiple, at least three, experiments are shown), (E) PCA analysis of 4,198 proteins identified in the proteomic analysis of *N. gruberi* subcellular fractions differentially enriched in mitochondria. The cluster of mitochondrial proteins was defined on the basis of 376 mitochondrial markers. The boundaries of the cluster of co-purified peroxisomal proteins were defined by 26 peroxisomal markers.

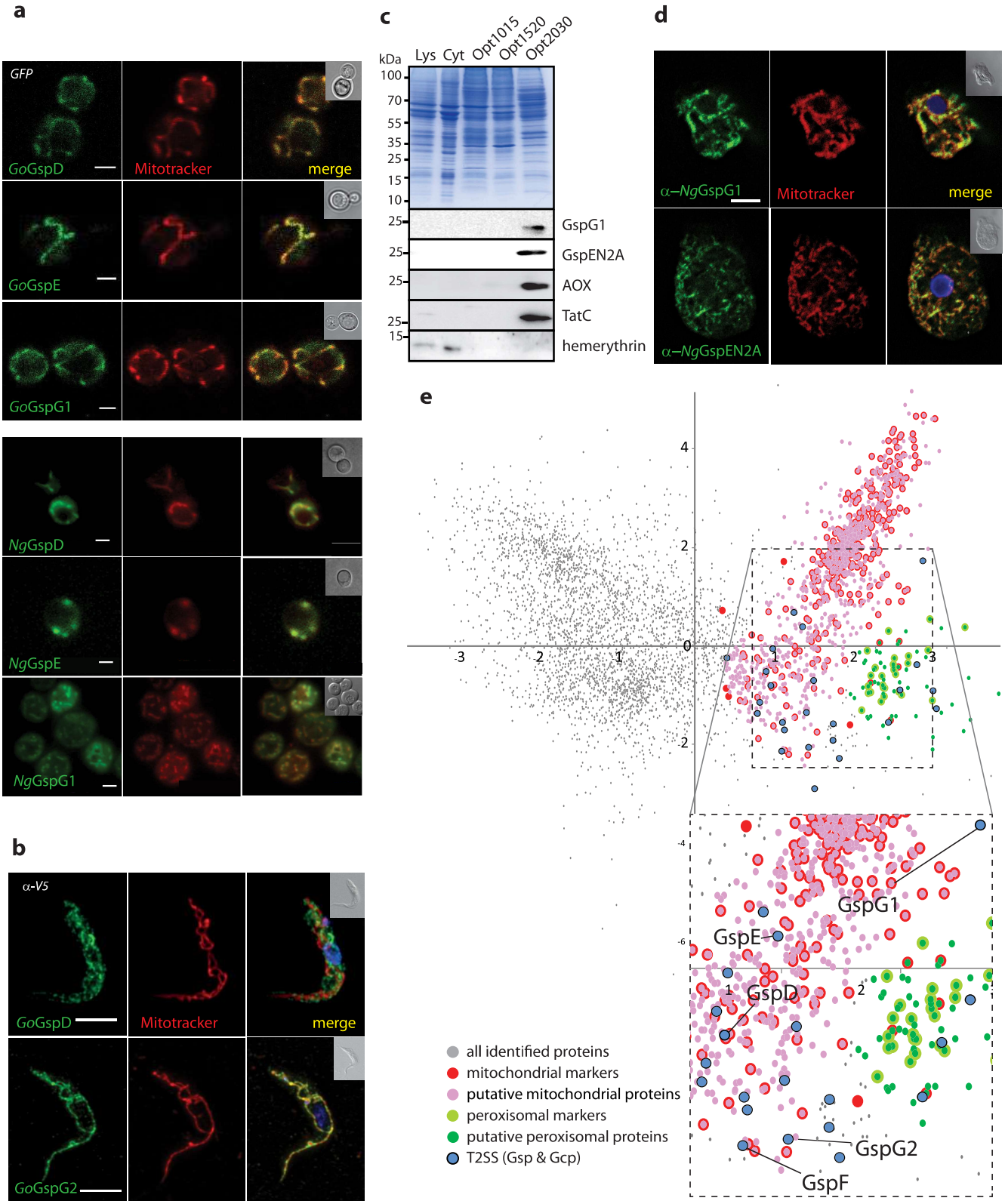
1377
1378 **Fig. 4 Mitochondrial GspD oligomerizes towards the formation of membrane pores.**
1379 (A) Domain architecture of the canonical bacterial GspD and the eukaryotic proteins
1380 homologous to its different parts (short N-terminal region of mitochondrial GspD of
1381 unidentified homology shown in grey). Below, protein sequence alignment of the secretin
1382 C-domain of bacterial and mitochondrial orthologues (mitochondrial GspD or GspDL
1383 and the respective molecular complex of bacterial secretins is depicted in grey, the
1384 numbers on the right depict the position of the amino acid in the particular sequence). (B)
1385 *In vitro* import of *Ng*GspD into isolated yeast mitochondria over a period of 16 min.
1386 Dissipation of the membrane potential (ΔΨ) by AVO mix abolished the import of matrix
1387 reporter protein (Su9-DHFR) bud did not affect the mitochondrial GspD; p – precursor of
1388 Su9-DHFR, m – mature form of the protein upon cleavage of the mitochondrial targeting
1389 sequence (C) Structural model of *Go*GspD built by ProMod3 on the *Vibrio cholerae*
1390 GspD template. Top and side view of a cartoon and a transparent surface representation
1391 of the *Go*GspD pentadecamer model is shown in blue. The amphipathic helical loop
1392 (AHL), a signature of the secretin family, is highlighted and coloured according to the
1393 secondary structure with strands in magenta, helices in cyan and loops in light brown.
1394 The C-terminal GpsD residues are highlighted as spheres. The detailed view of the AHL
1395 region shows the essential residues V162 and F166 pointing towards the membrane
1396 surface. (D) *In vitro* translation and assembly of mitochondrial *Go*GspD into a high-
1397 molecular-weight complex; lipo – liposomes added, urea – liposome fraction after 2M
1398 urea treatment. (E) Y2H assay suggests the self- and mutual interaction of *Ng*GspD and
1399 *Ng*GspDN1. (for B - D, representative images of three experiments are shown)
1400
1401 **Fig. 5 Membrane pore formation of mitochondrial GspD.** (A) Expression of
1402 mitochondrial *Ng*GspD in bacteria quickly causes cell death upon induction. (B) Size
1403 exclusion chromatography of *Ng*GspD shows two peaks of putative GspD complexes
1404 (bottom) and SDS-PAGE of the corresponding fractions (top) . (C) Single channel
1405 recording of the *Ng*GspD complex. The applied membrane potential was 15 mV. (D)
1406 Histogram of different recorded amplitudes during channel opening indicates multiple
1407 types of *Ng*GspD-formed pores in the membrane (For A-D, representative images of
1408 three experiments are shown)
1409
1410 **Fig. 6 Structure and maturation of mitochondrial GspG**. (A) Domain architecture of
1411 the bacterial and the mitochondrial pseudopilin GspG (top). The arrow indicates the
1412 processing site of the bacterial GspG during protein maturation. MTS – mitochondria
1413 targeting sequence, + – polar anchor, TMD – transmembrane domain, ND – novel
1414 domain. (Bottom) Protein sequence alignment of the pseudopilin domains of
1415 mitochondrial and bacterial GspG proteins (in case two paralogues are present, only
1416 GspG1 is shown) (B) Homology modelling of *Go*GspG1. Top: pairwise alignment of
1417 protein sequences of *Go*GspG1 and *Klebisella oxytoca* PulG (the template used in model
1418 building). Regions of high sequence similarity are highlighted, including the hydrophobic
1419 segment (cyan), the α-β loop (magenta) and GspG signature loop with conserved Pro
1420 residues (green). Bottom: side and top views of the cartoon and transparent surface
1421 representation of *Go*GspG pilus model based on the pseudopilus cryo-EM
1422 reconstruction[47]. Proteins are coloured based on the secondary structure, with helix

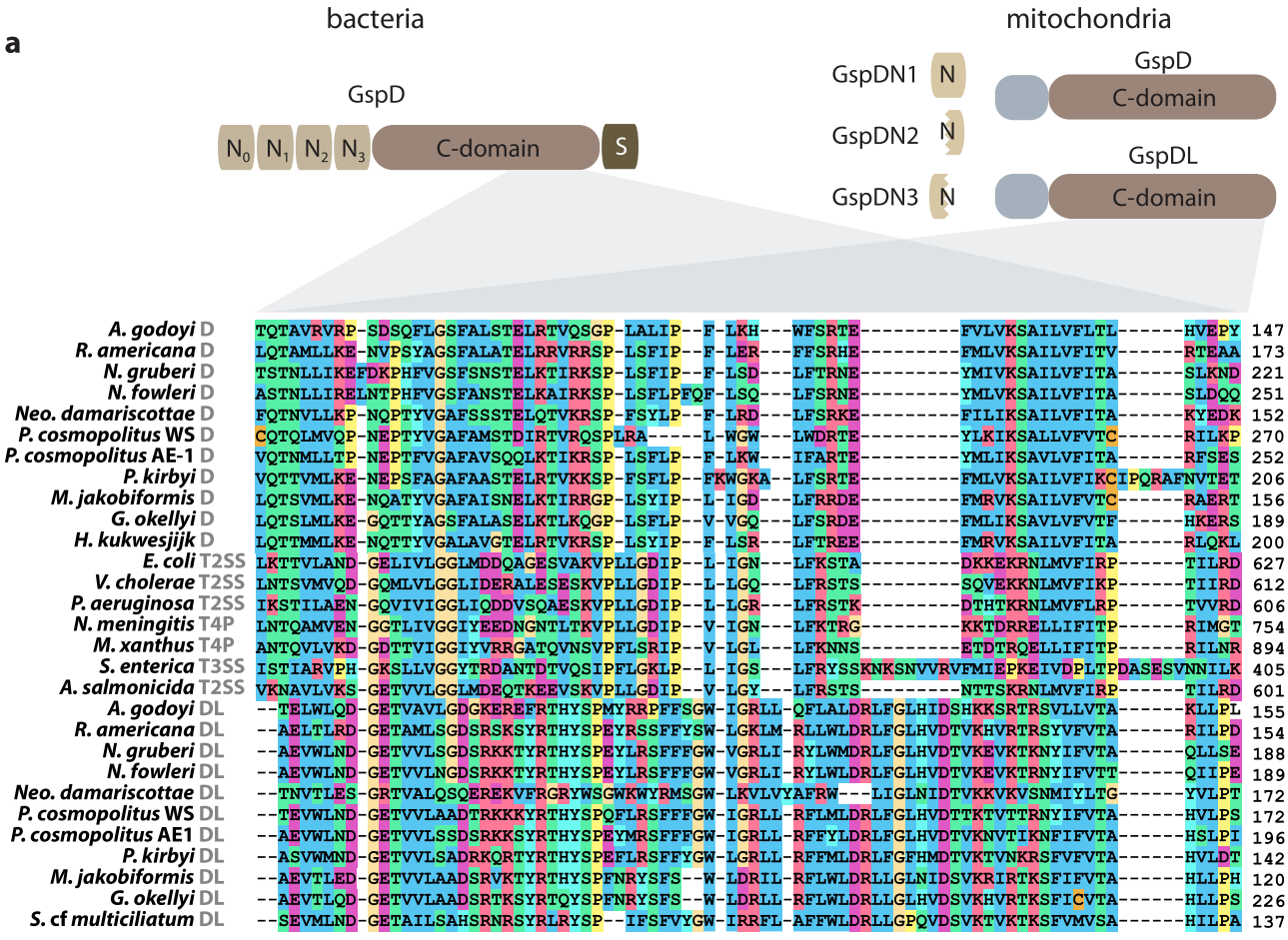
1423 regions in cyan and loops in light brown. The regions of GspG1 sharing high similarity
1424 with PulG are highlighted with the same colour code as in the sequence alignment, with
1425 side chains shown in magenta and green. Inset: Detail of the structurally conserved loop
1426 regions. The novel domain (ND) specific to mitochondrial GspG proteins was omitted
1427 from the modelling. (C) Peptides specific to *Ng*GspG1 retrieved from the proteomic
1428 analysis of *N. gruberi* mitochondria. The Leu residue highlighted in blue indicates
1429 residue +1 following the processing site of bacterial GspG proteins.
1430
1431 **Fig. 7 The protein interactions of mitochondrial GspG.** (A) The soluble part of the
1432 pseudopilin domain was purified to homogeneity on and ran through a gel filtration
1433 column, which demonstrated the stable monomeric state of the protein (representative
1434 image of three experiments is shown). (B) Thermophoresis of NT-647-labelled protein
1435 showed specific self-interaction of the domain (top); plotting of the change in
1436 thermophoresis yielded a K_d of 216 ± 15.1 nm, concentration of the labelled GspG was
1437 10 nM (bottom), The error bars depict standard deviation; n = 3. (C) Positive interaction
1438 between the mitochondrial GspG protein and other mitochondrial T2SS subunits as
1439 determined by the BACTH assay, n=6.
1440
1441 **Fig. 8 Proteins with the same phylogenetic profile as the originally identified**
1442 **mitochondrial Gsp homologues.** (A) Schematic domain representation of 23 proteins
1443 occurring in heteroloboseans, jakobids and malawimonads with the core T2SS subunits
1444 but not in other eukaryotes analyzed. Proteins with a functional link to the T2SS
1445 suggested by sequence homology are shown in blue, proteins representing novel
1446 paralogues within broader (super)families are shown in violet, and proteins without
1447 discernible homologues or with homologues only in prokaryotes are shown in pink. The
1448 presence of conserved protein domains or characteristic structural motifs is shown if
1449 detected in the given protein. Grey block – predicted transmembrane domain (see also
1450 Supplementary Fig. 11); "C H C H" – the presence of absolutely conserved cysteine and
1451 histidine residues (see also Supplementary Fig. 12) that may mediate binding of a
1452 prosthetic group; "HEXXH" and "EXXA" in Gcp16 indicate absolutely conserved motifs
1453 suggesting that the protein is a metallopeptidase of the gluzincin group (see text). The
1454 length of the rectangles corresponds to the relative size of the proteins. (B) Protein
1455 sequence alignment of Gcp12 proteins with highlighted conserved cysteine (black circles)
1456 and histidine (red circles) residues. (C) The expression of *Go*Gcp12 in *T. brucei* with the
1457 C-terminal V5 tag (green) showed partial co-localization with the mitochondrion (red)
1458 (representative image of three experiments is shown). Scale bar 10 μm.
1459
1460 **Fig. 9 The mitochondrial T2SS (miT2SS) as part of a hypothetical eukaryotic**
1461 **functional pathway connecting the mitochondrion and the peroxisome.** The scheme
1462 presents the most reasonable interpretation of the findings reported in this study, but
1463 further work is needed to test details of the working model. According to the model, a
1464 nucleus-encoded protein (green), possibly one of the newly identified Gcp proteins, is
1465 imported via the TOM complex into the mitochondrial inner membrane space, where it is
1466 modified by addition of a specific prosthetic group. After folding it becomes a substrate
1467 of the miT2SS machinery, is exported from the mitochondrion and finally reaches the
1468 peroxisome. The loading of the prosthetic group, the delivery to the peroxisome and

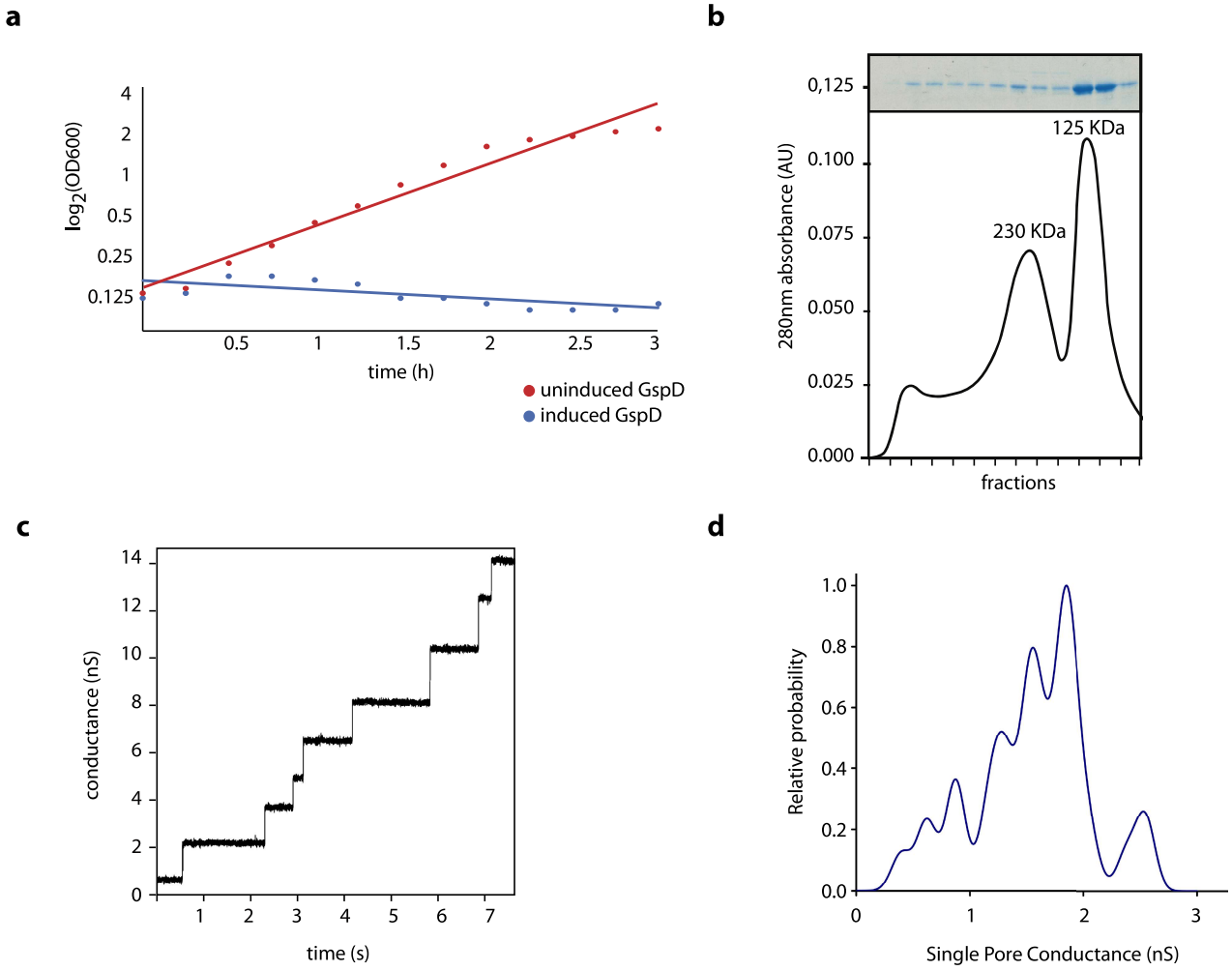
possibly also the actual function of the protein in the peroxisome is assisted by specific subsets of other Gcp proteins. The hypothetical presence of Gcp proteins in specific (sub)compartments is depicted as a group of orange ovals. OMM – outer mitochondrial membrane, IMS – intermembrane space, IMM – inner mitochondrial membrane, MM – mitochondrial matrix


a
OM
GspD
GspDN
GspK
GspI
GspJ
GspH
GspG
GspC
GspM
GspL
IM
GspF
GspE
ATP
ADP + P
b
General secretory proteins (Gsp)
Gsp-co-occuring proteins (Gcp)
D G DL DN3 EN2B 1 4 5 8 9 12 13 16
E F DN1 DN2 EL EN2A 2 3 6 7 10 11 14 15
LECA
CRuMs
AMOEBOZOA
OPISTHOKONTA
APUSOMONADIDA
ANCYROMONADIDA
MALAWIMONADIDA
Gefionella okellyi
'Malawimonas californiana'
Malawimonas jakobiformis
METAMONADA
JAKOBIDA
Reclinomonas americana
Andalucia godoyi
Stygiella incarcerata
Tsukubamonas globosa
EUGLENOZOA
HETEROLOBOSEA
Pharyngomonas kirbyi
Heterolobosea sp. BB2
Neovahl. damariscottae
Dactylomonas venusta
Creneis carolina
Percolomonas spp. (3)*
Harpagon schusteri
Naegleria gruberi
Naegleria fowleri
ARCHAEPLASTIDA
CRYPTISTA
HAPTISTA
TSAR
HEMIMASTIGOPHORA (2)*

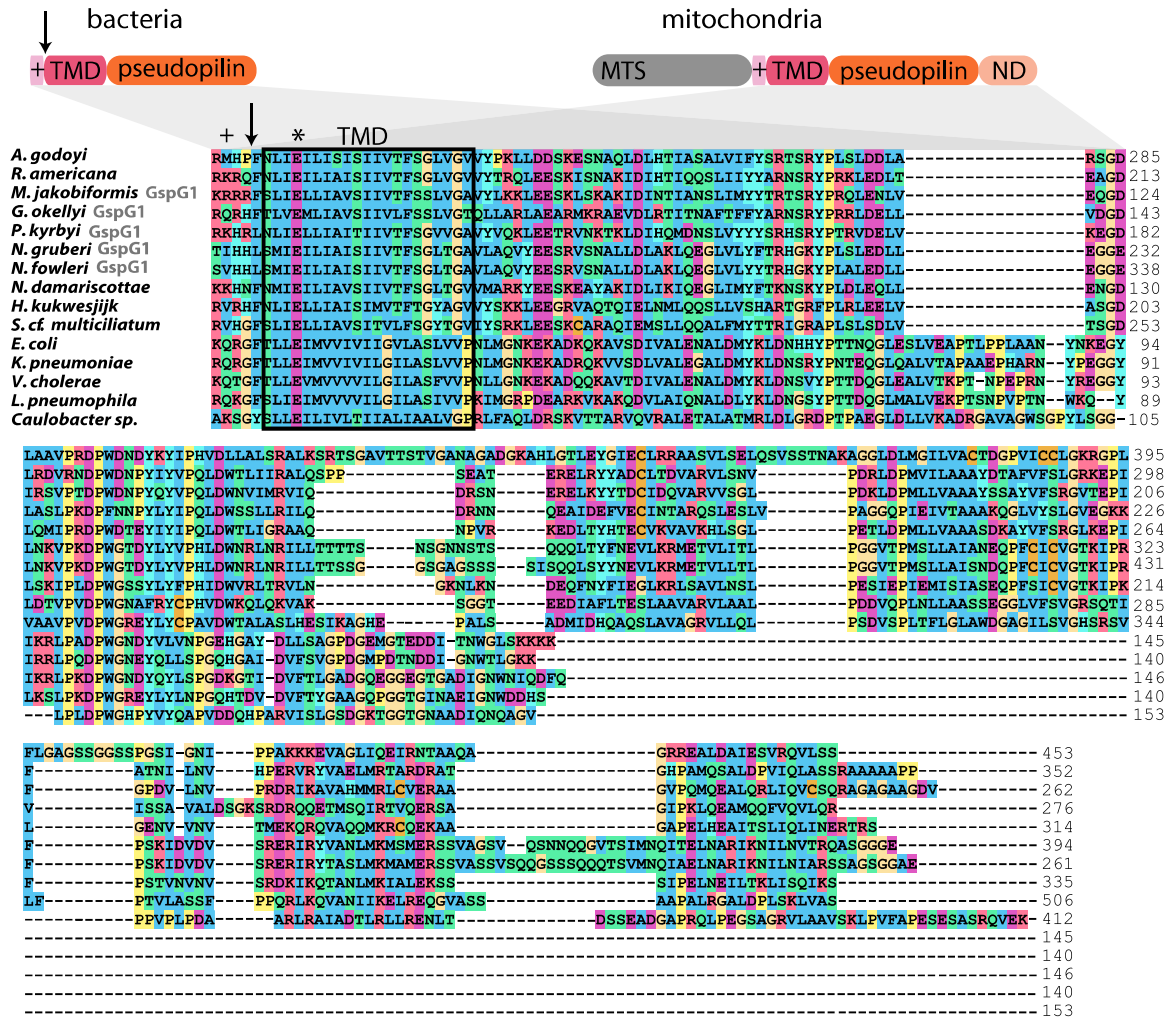

Gefionella okellyi GspG2
Malawimonas jakobiformis GspG2
Hemimastix kukwesijik
Spironema cf. multiciliatum
Neovahlkampfia damariscottae GspG2
Naegleria gruberi GspG2
Naegleria fowleri GspG2
Pharyngomonas kirbyi GspG2
Malawimonas jakobiformis GspG1
Pharyngomonas kirbyi GspG1
Gefionella okellyi GspG1
Reclinomonas americana
Andalucia godoyi
Neovahlkampfia damariscottae GspG1
Naegleria gruberi GspG1
Naegleria fowleri GspG1
unclassified Planctomycetes Planctomycetes
Candidatus roizmanbacteria Patescibacteria
unclassified Parcubacteria Parcubacteria
unclassified Clostridia Firmicutes
unclassified Methylovorus β-proteobacteria
unclassified Acidobacteria Acidobacteria
unclassified Desulfuromonas δ-proteobacteria
unclassified δ-proteobacteria δ-proteobacteria
unclassified Desulfurivibrio δ-proteobacteria
unclassified Planctomycetes Planctomycetes
unclassified Verrucomicrobia Verrucomicrobia
Leptolyngbya valderiana Cyanobacteria
unclassified Candidatus Hydrogenedentes
unclassified Chrysiogenales Chrysiogenetes
Rheinheimera γ-proteobacteria
Idiomarina γ-proteobacteria
Idiomarina γ-proteobacteria
Pseudoalteromonas γ-proteobacteria
unclassified Rheinheimera γ-proteobacteria
Pseudospirillum γ-proteobacteria
Syntrophus δ-proteobacteria
Environmental samples δ-proteobacteria
unclassified γ-proteobacteria γ-proteobacteria
unclassified Lentisphaerae Lentisphaerae
unclassified Janthinobacterium β-proteobacteria
unclassified Bacteria unclassified Bacteria
unclassified ζ-proteobacteria ζ-proteobacteria
unclassified Bacteria
Sar86 cluster γ-proteobacteria
unclassified Planctomycetes Planctomycetes
unclassified Bradymonadales δ-proteobacteria
unclassified δ-proteobacteria δ-proteobacteria
Enhygromyxa δ-proteobacteria
unclassified δ-proteobacteria δ-proteobacteria
unclassified δ-proteobacteria δ-proteobacteria
unclassified δ-proteobacteria δ-proteobacteria
unclassified Myxococcus δ-proteobacteria


a
GFP
GoGspD
Mitotracker
merge
GoGspE
GoGspG1
NgGspD
NgGspE
NgGspG1
b
α-V5
GoGspD
Mitotracker
merge
GoGspG2
c
kDa
Lys
Cyt
Opt1015
Opt1520
Opt2030
100
70
55
35
25
15
10
25
25
25
25
15
GspG1
GspEN2A
AOX
TatC
hemerythrin
d
α–NgGspG1
Mitotracker
merge
α–NgGspEN2A
e
all identified proteins
mitochondrial markers
putative mitochondrial proteins
peroxisomal markers
putative peroxisomal proteins
T2SS (Gsp & Gcp)
GspG1
GspE
GspD
GspG2
GspF

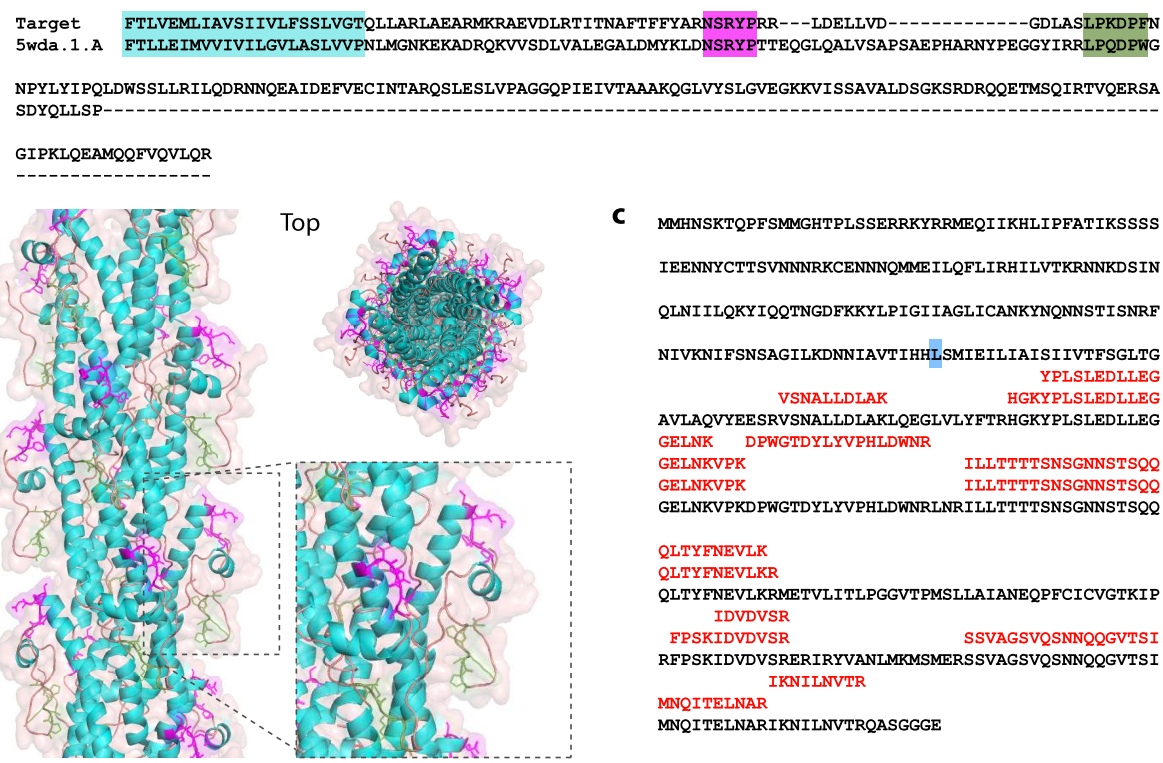

a
bacteria
mitochondria
GspD
N0 N1 N2 N3 C-domain S
GspDN1 N
GspDN2 N
GspDN3 N
GspD
C-domain
GspDL
C-domain
A. godoyi D TQTAVRVRP-SDSQFLGSFALSTELRTVQSGP-LALIP--F-LKH---WFSRTE---------FVLVKSAILVFLTL------HVEPY 147
R. americana D LQTAMLLKE-NVPSYAGSFALATELRRVRRSP-LSFIP--F-LER---FFSRHE---------FMLVKSAILVFITV------RTEAA 173
N. gruberi D TSTNLLIKEFDKPHFVGSFSNSTELKTIRKSP-LSFIP--F-LSD---LFTRNE---------YMIVKSAILVFITA------SLKND 221
N. fowleri D ASTNLLIRELNTPHFVGSFANSTELKAIRKSP-LSFLPFQF-LSQ---LFSRNE---------YMLVKSAILVFITA------SLDQQ 251
Neo. damariscottae D FQTNVLLKP-NQPTYVGAFSSSTELQTVKRSP-FSYLP--F-LRD---LFSRKE---------FILIKSAILVFITA------KYEDK 152
P. cosmopolitus WS D CQTQLMVQP-NEPTYVGAFAMSTDIRTVRQSPLRA-----L-WGW---LWDRTE---------YLKIKSALLVFVTC------RILKP 270
P. cosmopolitus AE-1 D VQTNMLLTP-NEPTFVGAFAVSQQLKTIKRSP-LSFLP--F-LKW---IFARTE---------YMLIKSAVLVFITA------RFSES 252
P. kirbyi D VQTTVMLKE-NEPSFAGAFAASTELRTVKKSP-FSFLP--FKWGKA--LFSRTE---------FMLVKSAILVFIKCIPQRAFNVTET 206
M. jakobiformis D LQTSVMLKE-NQATYVGAFAISNELKTIRRGP-LSYIP--L-IGD---LFRRDE---------FMRVKSAILVFVTC------RAERT 156
G. okellyi D LQTSLMLKE-GQTTYAGSFALASELKTLKQGP-LSFLP--V-VGQ---LFSRDE---------FMLIKSAVLVFVTF------HKERS 189
H. kukwesjijk D LQTTMMLKE-NQTTYVGALAVGTELRTVKRSP-LSYIP--F-LSR---LFTREE---------FMRVKSAILVFITA------RLQKL 200
E. coli T2SS LKTTVLAND-GELIVLGGLMDDQAGESVAKVPLLGDIP--L-IGN---LFKSTA---------DKKEKRNLMVFIRP------TIIRD 627
V. cholerae T2SS LNTSVMVQD-GQMLVLGGLIDERALESESKVPLLGDIP--L-LGQ---LFRSTS---------SQVEKKNLMVFIKP------TIIRD 612
P. aeruginosa T2SS IKSTILAEN-GQVIVIGGLIQDDVSQAESKVPLLGDIP--L-LGR---LFRSTK---------DTHTKRNLMVFLRP------TVVRD 606
N. meningitis T4P LNTQAMVEN-GGTLIVGGIYEEDNGNTLTKVPLLGDIP--V-IGN---LFKTRG---------KKTDRRELLIFITP------RIMGT 754
M. xanthus T4P ANTQVLVKD-GDTTVIGGIYVRRGATQVNSVPFLSRIP--V-LGL---LFKNNS---------ETDTRQELLIFITP------RILNR 894
S. enterica T3SS ISTIARVPH-GKSLLVGGYTRDANTDTVQSIPFLGKLP--L-IGS---LFRYSSKNKSNVVRVFMIEPKEIVDPLTPDASESVNNILK 405
A. salmonicida T2SS VKNAVLVKS-GETVVLGGLMDEQTKEEVSKVPLLGDIP--V-LGY---LFRSTS---------NTTSKRNLMVFIRP------TILRD 601
A. godoyi DL --TELWLQD-GETVAVLGDGKEREFRTHYSPMYRRPFFSGW-IGRLL-QFLALDRLFGLHIDSHKKSRTRSVLLVTA------KLLPL 155
R. americana DL --AELTLRD-GETAMLSGDSRSKSYRTHYSPEYRSSFFYSW-LGKLM-RLLWLDRLFGLHVDTVKHVRTRSYVFVTA------RILPD 154
N. gruberi DL --AEVWLND-GETVVLSGDSRKKTYRTHYSPEYLRSFFFGW-VGRLI-RYLWMDRLFGLHVDTVKEVKTKNYIFVTA------QLLSE 188
N. fowleri DL --AEVWLND-GETVVLNGDSRKKTYRTHYSPEYLRSFFFGW-VGRLI-RYLWLDRLFGLHVDTVKEVKTRNYIFVTT------QIIPE 189
Neo. damariscottae DL --TNVTLES-GRTVALQSQEREKVFRGRYWSGWKWYRMSGW-LKVLVYAFRW---LIGLNIDTVKKVKVSNMIYLTG------YVLPT 172
P. cosmopolitus WS DL --TEVWLND-GETVVLAADTRKKKYRTHYSPQFLRSFFFGW-IGRLL-RFLMLDRLFGLHVDTTKTVTTRNYIFVTA------HVLPS 172
P. cosmopolitus AE1 DL --AEVWLND-GETVVLSSDSRKKSYRTHYSPEYMRSFFFGW-IGRLL-RFFYLDRLFGLHVDTVKNVTIKNFIFVTA------HSLPI 196
P. kirbyi DL --ASVWMND-GETVVLSADRKQRTYRTHYSPEFLRSFFYGW-LGRLL-RFFMLDRLFGFHMDTVKTVNKRSFVFVTA------HVLDT 142
M. jakobiformis DL --AEVTLED-GETVVLAADSRVKTYRTHYSPFNRYSFS--W-LDRIL-RFLWLDRLLGLNIDSVKRIRTKSFIFVTA------HLLPH 120
G. okellyi DL --AEVTLQD-GETVVLAADSRTKSYRTQYSPFNRYSFS--W-LDRLL-RFLWLDRLLGLHVDSVKHVRTKSFICVTA------HLLPS 226
S. cf multiciliatum DL --SEVMLND-GETAILSAHSRNRSYRLRYSP--IFSFVYGW-IRRFL-AFFWLDRLLGPQVDSVKTVKTKSFVMVSA------HILPA 137
b
L
time (min)
0' 1' 2' 4' 8' 16' -Δ 16'
Su9-DHFR
NgGspD
p
m
c
Top
Side
V162
F166
AHL
d
lipo - + +
urea - - +
kDa
720
450
242
146
66
20
GoGspD oligomer
GoGspD monomer
e
neg x neg
NgGspD x NgGspD
NgGspD x NgGspDN1
NgGspDN1 x NgGspDN1


a
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
b
0,125
0.100
0.075
0.050
0.025
0.000
280nm absorbance (AU)
125 KDa
230 KDa
fractions
c
conductance (nS)
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7
time (s)
d
Relative probability
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
Single Pore Conductance (nS)

a

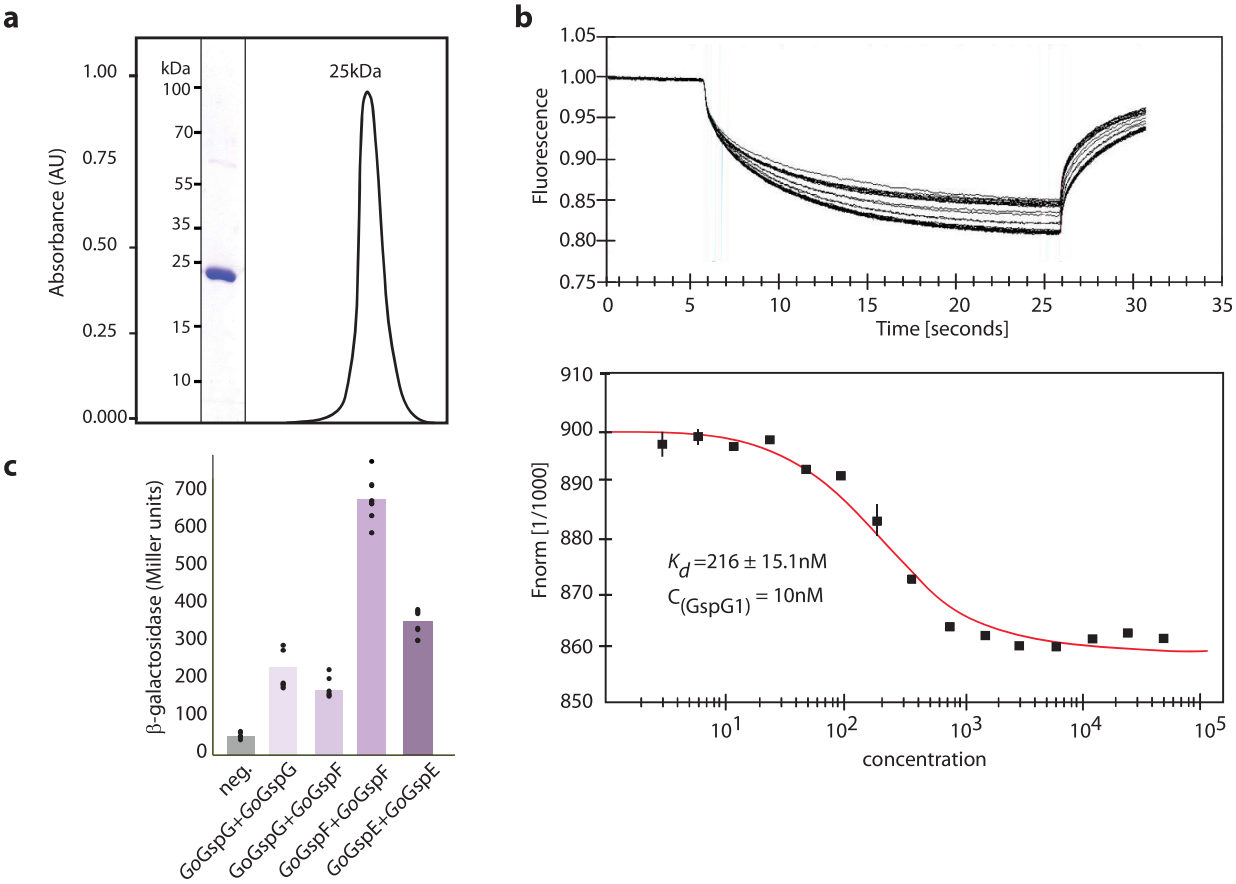

bacteria
mitochondria
+TMD pseudopilin
MTS +TMD pseudopilin ND
+ * TMD
A. godoyi
R. americana
M. jakobiformis GspG1
G. okellyi GspG1
P. kyrbyi GspG1
N. gruberi GspG1
N. fowleri GspG1
N. damariscottae
H. kukwesjijk
S. cf. multiciliatum
E. coli
K. pneumoniae
V. cholerae
L. pneumophila
Caulobacter sp.

b

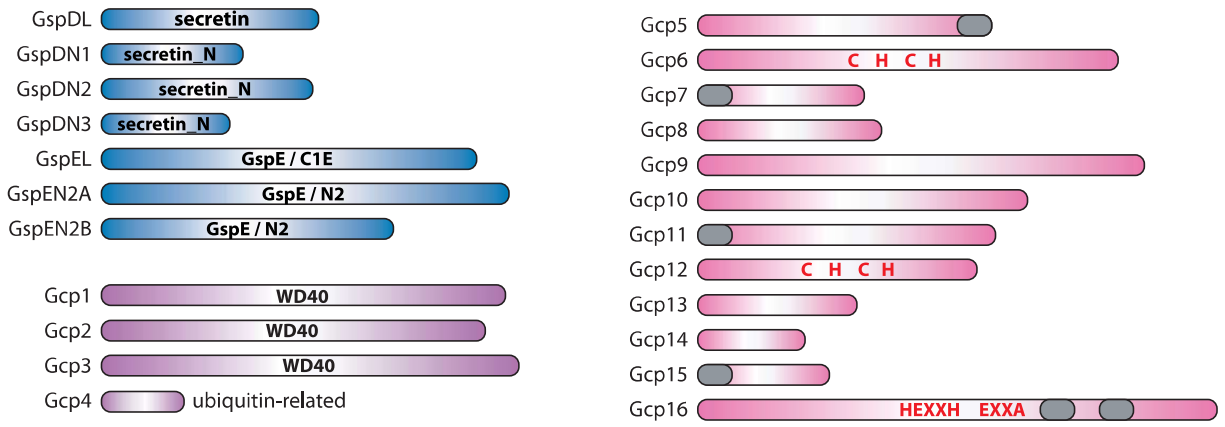

Target
5wda.1.A
Top
Side

c

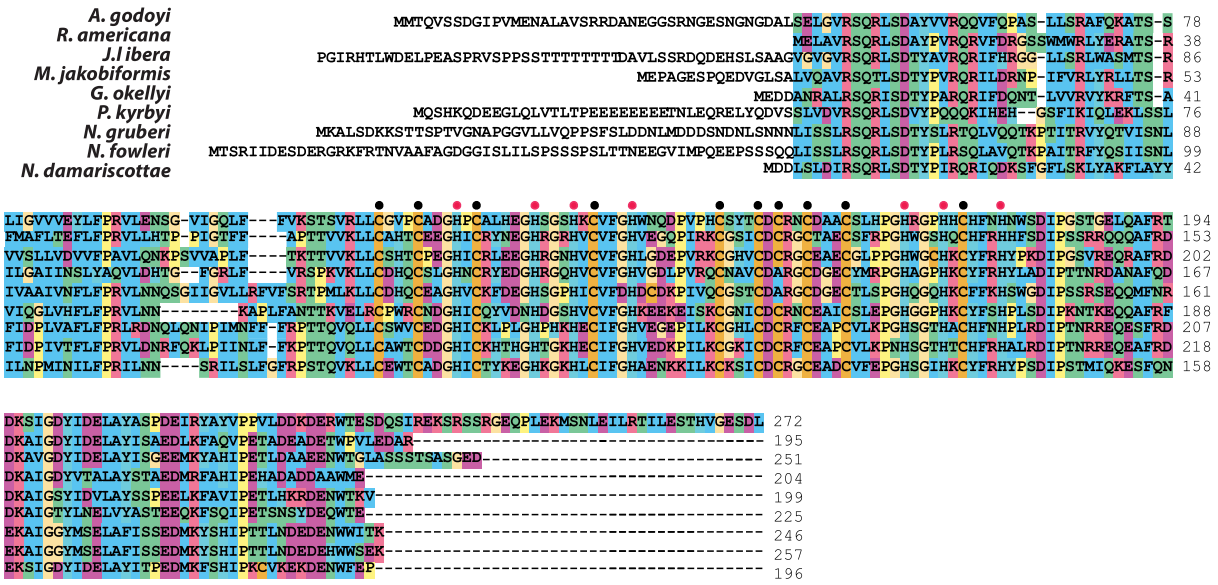
MMHNSKTQPFSMMGHTPLSSERRKYRRMEQIIKHLIPFATIKSSSS
IEENNYCTTSVNNNRKCENNNQMMEILQFLIRHILVTKRNNKDSIN
QLNIILQKYIQQTNGDFKKYLPIGIIAGLICANKYNQNNSTISNRF
NIVKNIFSNSAGILKDNNIAVTIHHLSMIEILIAISIIVTFSGLTG
YPLSLEDLLEG
VSNALLDLAK
HGKYPLSLEDLLEG
AVLAQVYEESRVSNALLDLAKLQEGLVLYFTRHGKYPLSLEDLLEG
GELNK
DPWGTDYLYVPHLDWNR
GELNKVPK
ILLTTTTSNSGNNSTSQQ
GELNKVPK
ILLTTTTSNSGNNSTSQQ
GELNKVPKDPWGTDYLYVPHLDWNRLNRILLTTTTSNSGNNSTSQQ
QLTYFNEVLK
QLTYFNEVLKR
QLTYFNEVLKRMETVLITLPGGVTPMSLLAIANEQPFCICVGTKIP
IDVDVSR
FPSKIDVDVSR
SSVAGSVQSNNQQGVTSI
RFPSKIDVDVSRERIRYVANLMKMSMERSSVAGSVQSNNQQGVTSI
IKNILNVTR
MNQITELNAR
MNQITELNARIKNILNVTRQASGGGE


a
kDa
100
70
55
35
25
15
10
25kDa
Absorbance (AU)
1.00
0.75
0.50
0.25
0.000
b
Fluorescence
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0
5
10
15
20
25
30
35
Time [seconds]
Fnorm [1/1000]
910
900
890
880
870
860
850
K_d=216 ± 15.1nM
$C_{(GspG1)}$ = 10nM
10^1
10^2
10^3
10^4
10^5
concentration
c
β-galactosidase (Miller units)
700
600
500
400
300
200
100
0
neg.
GoGspG+GoGspG
GoGspG+GoGspF
GoGspF+GoGspF
GoGspE+GoGspE

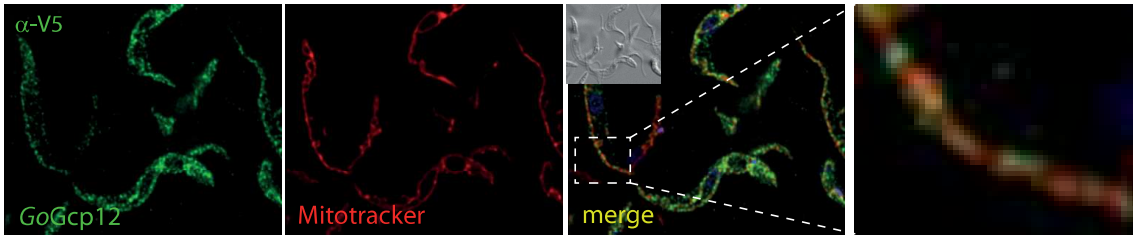
a

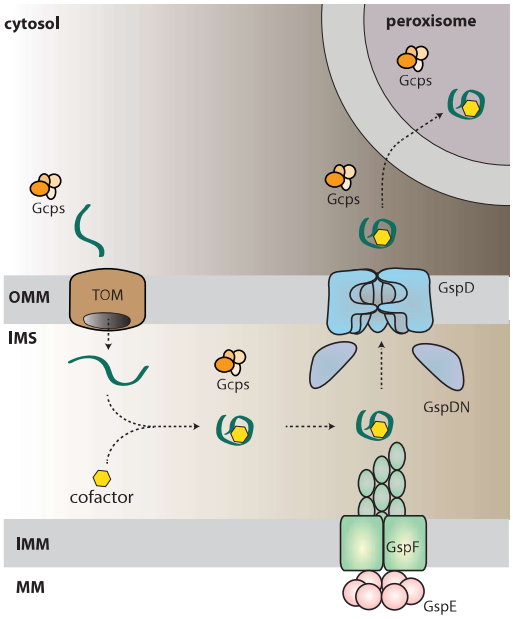

GspDL
secretin
GspDN1
secretin_N
GspDN2
secretin_N
GspDN3
secretin_N
GspEL
GspE / C1E
GspEN2A
GspE / N2
GspEN2B
GspE / N2
Gcp1
WD40
Gcp2
WD40
Gcp3
WD40
Gcp4
ubiquitin-related
Gcp5
Gcp6
C H C H
Gcp7
Gcp8
Gcp9
Gcp10
Gcp11
Gcp12
C H C H
Gcp13
Gcp14
Gcp15
Gcp16
HEXXH EXXA

b


A. godoyi
R. americana
J.l ibera
M. jakobiformis
G. okellyi
P. kyrbyi
N. gruberi
N. fowleri
N. damariscottae

c


α-V5
GoGcp12
Mitotracker
merge


cytosol
peroxisome
Gcps
Gcps
Gcps
OMM
TOM
GspD
IMS
Gcps
GspDN
cofactor
IMM
GspF
MM
GspE