

This is a repository copy of *Simulating the evolutionary trajectories of metabolic pathways for insect symbionts in the Sodalis genus*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/160085/>

Version: Accepted Version

Article:

Hall, Rebecca J, Thorpe, Stephen, Thomas, Gavin Hugh orcid.org/0000-0002-9763-1313 et al. (1 more author) (Accepted: 2020) Simulating the evolutionary trajectories of metabolic pathways for insect symbionts in the Sodalis genus. Microbial Genomics. ISSN 2057-5858 (In Press)

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

1 Simulating the evolutionary trajectories of
2 metabolic pathways for insect symbionts in the
3 *Sodalis* genus

4 Rebecca J. Hall*^{1,2}, Stephen Thorpe*^{1,3}, Gavin H. Thomas¹ and
5 A. Jamie Wood^{1,4}

6 **Author affiliations**

7 *These authors contributed equally to this work.

8 ¹Department of Biology, University of York, York, YO10 5NG, UK

9 ²School of Life Sciences, University of Nottingham, Nottingham, NG7 2TQ, UK

10 ³Department of Chemistry, University of Oxford, Oxford, OX1 3TA, UK

11 ⁴Department of Mathematics, University of York, York, YO10 5DD, UK

12 Corresponding author: A. Jamie Wood (jamie.wood@york.ac.uk)

13 **Abstract**

14 Insect-bacterial symbioses are ubiquitous, but there is still much to uncover about how
15 these relationships establish, persist and evolve. The tsetse endosymbiont *Sodalis*
16 *glossinidius* displays intriguing metabolic adaptations to its microenvironment, but the
17 process by which this relationship evolved remains to be elucidated. The recent chance
18 discovery of the free-living species of the *Sodalis* genus, *S. praecaptivus*, provides a
19 serendipitous starting point from which to investigate the evolution of this symbiosis.
20 Here, we present a flux balance model for *S. praecaptivus* and empirically verify its
21 predictions. Metabolic modelling is used in combination with a multi-objective
22 evolutionary algorithm to explore the trajectories that *S. glossinidius* may have
23 undertaken from this starting point after becoming internalised. The order in which key
24 genes are lost is shown to influence the evolved populations, providing possible targets
25 for future *in vitro* genetic manipulation. This method provides a detailed perspective on

26 possible evolutionary trajectories for *S. glossinidius* in this fundamental process of
27 evolutionary and ecological change.

28 Keywords

29 Symbiosis, evolution, *Sodalis*, flux balance analysis, evolutionary algorithm

30 Data Summary

31 The Python code for running the algorithm with an example data set is available on
32 GitHub at <https://github.com/St659/SodalisFBAEvolution>. The data generated by the
33 simulations are available on the York Research Database.

34 Data Statement

35 All supporting data, code and protocols have been provided within the article or through
36 supplementary data files. Supplementary material is available with the online version of
37 this article.

38 Impact Statement

39 Insect-microbe symbioses are challenging to study as the symbionts may not be
40 amenable to *in vitro* culture or traditional genetic manipulation techniques. The
41 establishment and tracking of symbiosis from initiation to infection also presents
42 technical challenges. A metabolic model of a free-living plausible starting organism is
43 presented and verified against empirical data. This work provides a computational
44 method to examine the potential evolutionary trajectories that symbionts may have
45 taken once becoming internalised by a host. It enables new questions to be asked about
46 genome reduction and niche adaptation by symbiotic bacteria. This technique has wider
47 implications beyond symbiosis, with potential applications in directed evolution for
48 industrial biotechnology.

49 Introduction

50 Symbioses are both fundamental and ubiquitous in nature. Understanding their
51 evolution poses an ongoing challenge, as well as an expanse of unresolved research
52 questions. Bacterial symbionts of insects provide a range of benefits including stress
53 tolerance [1, 2], protection from predation [1, 3, 4] and the provision of metabolites [5–
54 10]. The latter forms arguably the strongest link within the symbioses. Host and
55 symbiont frequently share metabolic substrates, as well as the products and
56 components of individual biosynthetic pathways [7,11–15]. These relationships typically
57 enable the host to survive on a nutritionally restricted diet, such as blood [8,16,17] or
58 plant sap [18–20].

59 Deciphering the evolutionary pressures that affect the organisms within a symbiosis is
60 an essential part of understanding the relationship. This includes establishing how the
61 symbioses develop over time and the way in which the metabolism of the individuals is
62 intertwined. It is, however, often hindered by biological difficulties. Symbiotic bacteria
63 undergo genomic streamlining, may not be cultivatable *in vitro*, may no longer express
64 stress response genes and might lack a sound outer membrane [4,5,21–25]. It is
65 therefore impossible in many cases to test hypotheses about host-symbiont interactions
66 in controlled experimental conditions. In these circumstances, computational
67 techniques offer a viable, and currently the only, alternative to investigating metabolic
68 potential and pseudogenisation in symbiotic bacteria.

69 Computational biology is now well established as a key tool of scientific discovery, now
70 that vast amounts of data are generated quickly and cheaply from advancements in
71 sequencing technology [26,27]. Genome scale metabolic modelling of microorganisms
72 enables predictions to be made about metabolite preferences, transporter use and the
73 functionality of biosynthetic pathways [26,28]. Microbial metabolism can be simulated
74 using flux balance analysis (FBA), a constraint-based quantitative approach that
75 reconstructs a metabolic network from a genome annotation [27,29]. FBA is a powerful
76 tool when based on a well annotated genome and with the provision of *in vitro*
77 experimental validation [29].

78 FBA is widely used for biotechnology applications, and this can be re-purposed to
79 examine symbiosis. There are several published examples of using FBA to analyse the
80 metabolism of symbiotic bacteria, including for *Buchnera aphidicola* [7,30,31], *Sodalis*
81 *glossinidius* [32,33], *Portiera aleyrodidarum* [34], *Hamiltonella defensa* [34] and strains
82 of *Blattabacterium* [35]. There are also models published for the *Synechocystis* species

83 used in the study of artificially induced symbiosis [36–40]. FBA is useful in this instance,
84 as experiments that would not be possible empirically, due to culturability issues, can
85 be performed *in silico*. Furthermore, the genomes of symbiotic bacteria are often
86 unusual, with large pathway deletions or widespread pseudogenisation [6,25,41,42].
87 Analysis of the resulting metabolic network via FBA can suggest which pathways are
88 being used when supplied with different media, and predict which external metabolites
89 might be required to support growth *in vitro*.

90 FBA has been applied to several microbiological problems. Boolean logical operators
91 have been incorporated into *Escherichia coli* metabolic models to investigate the impact
92 of gene regulation on a system [43–46]. Dynamic FBA (dFBA), where a rate of change in
93 flux constraints is included, has successfully modelled diauxic growth in *E. coli* [47]. FBA
94 has been used to compare strains of *Blattabacterium* from separate cockroach lineages
95 to assess their divergence [35], and to predict the evolution of metabolism from *E. coli*
96 experimental data sets [48]. The evolution of metabolic networks in isolation has also
97 been simulated with the aim of identifying key metabolites [49], and to investigate
98 pseudogenisations in specific metabolic pathways [50]. FBA has not yet been harnessed
99 to its full potential with regards to the investigation of symbiont evolution. This is
100 perhaps surprising given that several models of *E. coli* metabolism are available as an
101 evolutionary starting point [51–56]. The evolution of *B. aphidicola* and *Wigglesworthia*
102 *glossinidia* from an *E. coli* ancestor has been simulated using FBA [56,57]. This work,
103 whilst elegant, has a key limitation. Reactions that are lost at the start have no chance
104 of being reintroduced. This limits the evolutionary space that can be explored, as the
105 loss of a key reaction at the start will fundamentally affect which reactions can be lost
106 subsequently. A similar approach to that used by Pal *et al.* [56] was used with dFBA to
107 study the evolution of cooperation and cross-feeding in *E. coli* [58]. Using FBA in isolation
108 to remove reactions successively may not therefore be the optimal way to simulate the
109 evolution of symbiosis.

110 *In silico* evolution has been used increasingly in recent years to complement *in vivo*
111 experimental evolution [59]. *In silico* evolution benefits from being able to test widely
112 different ecological conditions whilst controlling key variables [60]. For example, it
113 allows the investigation of groups of mutations that lead to a specific phenotype, or
114 mutations that are difficult to induce *in vitro* [61]. This has enabled the study of many
115 aspects of evolution, including simulating the reduction of genome size in an individual
116 [60]. Multi-objective evolutionary algorithms (MOEA) have been used in many
117 disciplines for solving problems that have two or more conflicting objectives [62]. The
118 use of MOEA in combination with metabolic models has been implemented for the

119 design of minimal genomes [63] and for the production of industrially relevant
120 molecules [64,65]. It has however seen only limited use for *in silico* evolution [66]. When
121 viewed computationally, the evolution of symbiosis can be considered as a multi-
122 objective optimisation; symbiotic bacteria undergo genome reduction whilst trying to
123 maximise their individual growth.

124 A free-living organism within the *Sodalis* genus has been characterised and sequenced
125 only recently [67,68]. *Sodalis praecaptivus* was isolated from a human wound, caused
126 by an impalement on a crab apple tree branch, and it is assumed that the tree was the
127 likely source of the *S. praecaptivus* infection. *S. praecaptivus* is a prototroph, capable of
128 growth in minimal media and at 37°C [68]. The annotated genome sequence for *S.*
129 *praecaptivus* is also available [67]. It is of particular interest given its close relation *S.*
130 *glossinidius*, secondary symbiont of the tsetse [25]. The tsetse, genus *Glossina* is
131 medically important as the vector for *Trypanosoma brucei*, causative agent of human
132 African trypanosomiasis [69]. *S. praecaptivus* therefore provides a rich set of data from
133 which to begin investigations into the origin of, and adaptations within, the tsetse-*S.*
134 *glossinidius* symbiosis.

135 Here, we present a flux balance model for *S. praecaptivus*, *iRH830*. This model, and a
136 previously presented model of *S. glossinidius* metabolism, *iLF517* [33], both represent
137 adaptations of the organisms to their contrasting environments. The *Sodalis* system is
138 therefore an excellent candidate for assessing the ability of FBA to describe the
139 evolution of symbioses. A MOEA has been used to evolve *iRH830* under various
140 biological conditions. The aim was to investigate computationally the route that *S.*
141 *glossinidius* may have taken in its transition to symbiosis. It is not known whether the
142 solutions found by *S. glossinidius*, described in *iLF517*, are the only possible outcomes
143 given the metabolic constraints of the microenvironment, or whether the symbiont's
144 unusual metabolic network evolved by chance. The application of the MOEA to *iRH830*
145 enabled us to ask in which order of the evolutionary sequence key pseudogenisations
146 may have occurred. The effect of exposing the ancestral *Sodalis* to contrasting diets was
147 modelled to mirror the different trajectories that this genus has taken within blood- and
148 sap-feeding insects. The techniques used here could be applied to other symbiotic
149 systems to drive forward the discovery of novel relationship criteria.

150 Results

151 A model of *S. praecaptivus* metabolism, *iRH830*

152 In order to investigate computationally the path that *S. glossinidius* has taken to
153 symbiosis, a metabolic model describing its close, free-living relative *S. praecaptivus* was
154 constructed (Fig. 1). Full details are given in Supplementary Data 1. *iRH830* contains 830
155 genes, 891 metabolites and 1246 reactions (excluding pseudoreactions), and is a
156 prototroph for all essential amino acids. An iterative process of gap filling was
157 undertaken by comparing the draft *S. praecaptivus* model to *iLF517* (*S. glossinidius*) and
158 *iJO1366*, a model of *E. coli* metabolism [54]. *iRH830* is supplied with an oxygen uptake
159 value of 20 mmol gr DW⁻¹ hr⁻¹, reflecting the highly aerated conditions the organism is
160 grown in and to retain consistency with models of *E. coli* metabolism [53,54].

161

162 Fig. 1. The construction process for *iRH830*. The *S. praecaptivus* genome was mined for
163 orthologues to metabolic genes in *E. coli* and *S. glossinidius*, before compiling into a draft
164 model. An iterative process of testing and gap filling was then performed, using
165 information provided in various databases (see key).

166 A series of biochemical screens were conducted using Biolog phenotypic microplates to
167 strengthen the model. In total, 190 metabolites were tested for their ability to act as the
168 sole carbon source for *S. praecaptivus*. Experiments were conducted in triplicate with
169 full results detailed in Supplementary Data 2. Through this phenotypic screen, it was
170 found that *S. praecaptivus* was able to use 19 of the metabolites tested as a sole source
171 of carbon (Table S1). When these metabolites were tested *in silico* by the exogenous
172 addition to *iRH830*, it was found that all but two mirrored the *in vitro* data; *N*-acetyl D-
173 galactosamine (GalNAc) and xylitol. This was then confirmed quantitatively in a 96-well
174 microplate with xylitol or GalNAc supplemented into M9 minimal medium (Fig. S1a, Fig.
175 S1b). Neither models of *S. glossinidius* (*iLF517*, [33]) or *E. coli* (*iJO1366*, [54]) were able
176 to produce a positive biomass output with xylitol or GalNAc as sole sources of carbon
177 (Table S1).

178 Comparison of the *S. praecaptivus* genome to other known D-xylitol consumers such as
179 *Morganella morganii* subsp. *morganii* revealed a highly conserved catabolic operon
180 containing the distinguishing D-xylitol dehydrogenase (79.7% identity between the *S.*
181 *praecaptivus* orthologue, AFW03778/Sant_3108, and the *Morganella morganii* subsp.

182 *morganii* protein, Uniprot ID: Q59545). The cluster contains a xylulose reductase as the
183 second enzyme required to convert D-xylitol to the central metabolite D-xylulose-5-
184 phosphate (Fig. S1c) and a complete ABC transporter that is likely specific for D-xylitol.
185 Interestingly, the pathway is also complete in the reduced *S. glossinidius* genome (Table
186 S2), suggesting that this is a conserved metabolic trait of the *Sodalis* genus. The cluster
187 is not present in *E. coli* with only weakly matching homologues are found fragmented
188 over the genome (Table S2). The proposed pathway for GalNAc metabolism was also
189 constructed using known pathways (Fig. S1d).

190 Robustness analysis of the *S. praecaptivus* metabolic network

191 Robustness analysis was used to examine reaction essentiality and therefore
192 redundancy in the *iRH830* network. *iRH830* was run on a simple, tsetse-specific nutrient
193 limited medium ("famine") and a blood medium simulating the internal tsetse
194 environment and informed by *S. glossinidius* requirements [33] ("blood", Table S3). All
195 media are detailed in Supplementary Data 1. Reactions were removed individually and
196 the resulting effect on biomass output noted. The same analysis was also run on *iLF517*
197 in blood as a comparison.

198 There are 282 essential reactions in *iRH830* when the medium (famine) is nutritionally
199 limited, and 228 in the tsetse-specific blood medium (Fig. 2a). The overall pattern for
200 the two conditions is very similar. The subsystem most represented in either condition
201 is for cofactor and prosthetic group biosynthesis, with 88 and 87 essential reactions for
202 the famine and blood media, respectively. The main difference at the subsystem level
203 can be attributed to amino acid metabolism; 15.8% of the total number of essential
204 reactions in blood and 29.8% in the famine medium that does not contain amino acids
205 are involved in these pathways. Of these, the essential reactions involved in L-arginine,
206 L-proline, L-threonine and L-lysine metabolism are highly prevalent in both media types
207 (Fig. 2b).

208

209 Fig. 2. Robustness analysis of *iRH830*. (a) Essential reactions in famine (top) and blood
210 (bottom) media. Essential reactions are categorised by subsystem. (b) Essential
211 reactions involved in amino acid metabolism in *iRH830* in famine (top) and blood
212 (bottom) media.

213 Media provisioning affects evolutionary trajectories

214 NSGA-II is a heuristic multi-objective optimisation algorithm used to evaluate multi-
215 objective problems without giving weight to any specific outcome. Evolution within a
216 constrained environment, such as the tsetse microenvironment, can be considered a
217 multi-objective optimisation problem of trying to reduce the genome size to increase
218 replication speed, while still retaining sufficient capacity to grow [70]. The MOEA was
219 used to explore the potential evolutionary trajectories of *S. praecaptivus* when exposed
220 to similar environmental conditions to *S. glossinidius*. A graphical description of the
221 MOEA is provided in Fig. 7. A key feature of this is the option of reactions that have been
222 removed being re-introduced later in the simulation. This helps to prevent the model
223 from consistently finding the same solutions and instead allows a greater evolutionary
224 space to be explored. The conditions under which *i*RH830 and *i*LF517 were evolved are
225 detailed in Table 1. In Scenario i, *i*RH830 was evolved in blood and famine growth media,
226 as well as a medium that mimics plant sap (Supplementary Data 1), to examine the effect
227 of metabolite availability. In Scenario ii, three key reactions, ASPTA, PDH and PPC, were
228 removed from *i*RH830 prior to evolution to compare the trajectories that arise as a result
229 of pseudogenisations, and to investigate if these were possible adaptations prior to
230 symbiont establishment. The gene encoding PPC is thought to be pseudogenised in *S.*
231 *glossinidius*, whereas the PDH and ASPTA reactions are predicted to be functional [33].
232 In Scenario iii, the MOEA is applied to *i*LF517 to investigate the possible future of *S.*
233 *glossinidius* as a symbiont.

234 Table 1. *In silico* evolution conditions. Conditions under which *i*RH830 and *i*LF517 were
235 evolved, including wild-type (WT) or reaction knockouts and media type.

Scenario	Test	Model	Media
i	Effect of growth media	<i>i</i> RH830 (WT)	Blood, sap, famine
ii	Effect of gene loss	<i>i</i> RH830 (Δ ASPTA, Δ PDH, Δ PPC)	Blood
iii	Future of <i>S. glossinidius</i>	<i>i</i> RH830 (WT), <i>i</i> LF517 (WT)	Blood

236 Species of the *Sodalis* genus have been found in insects that feed on a variety of
237 contrasting diets, including blood (e.g. tsetse [25] and ticks [71–73]) and plant tissue
238 (e.g. weevils [74]). To replicate *Sodalis* evolution in different environments, the MOEA
239 was applied to *i*RH830 that was supplied with the tsetse-specific blood medium, the
240 famine medium, and a medium that mimics plant sap (Table 1, Scenario i). Sap was
241 chosen as a comparison medium as *Sodalis*-allied symbionts have been identified in a
242 range of phytophagic insects [75–80]. The algorithm underwent ten runs of 3000
243 generations and the resulting solutions were collated.

244 In all conditions, the models evolved to completion, demonstrated by the convergence
245 of solutions to the left of the plots (Fig. 3a). The number of reactions decreases over

246 evolutionary time, with the majority of solutions clustering at the maximum biomass
247 output. This is an indication that sub-optimal solutions are being removed successfully.
248 After 3000 generations, there are a range of solution sizes at the maximum biomass
249 output found in sap, whereas in blood and famine all solutions at this time point cluster
250 at the minimum number of reactions. The two complex media, blood and sap (Fig. 3a),
251 produce a lot of metabolic flexibility, with a complete range of possible biomass outputs
252 produced by the smallest models. When grown in the simple famine medium, there is
253 significantly less flexibility in terms of possible solutions found (Fig. 3a). Here, the
254 majority of the solutions cluster at the minimum reactions/maximum biomass output.
255 This is as expected, given the fitness function of the MOEA. In blood and sap, the biomass
256 outputs reach near zero, made possible by the variety of available substrates. In famine,
257 the options for streamlining are limited, resulting in few solutions that are able to
258 deviate away from what is selected by the fitness function.

259

260 Fig. 3. *i*RH830 evolved under different starting conditions. (a) Evolution of *i*RH830 in a
261 tsetse-specific blood medium (left), a nutritionally-limited famine medium (centre), and
262 a medium mimicking plant sap (right). (b) *i*RH830 evolution in a blood medium with the
263 reactions ASPTA (left), PDH (centre) and PPC (right) removed at the start. The MOEA was
264 run for 3000 generations, with the plot depicting new populations every 50 generations
265 (blue to green). Black boxes indicate individual solutions selected for further analysis.

266 A number of individual solutions that were representative of the biomass output of
267 *i*LF517 [33] were then selected from each of these simulations (Fig. 3, black boxes). The
268 raw, binary data were translated back into reaction names and this was subsequently
269 processed to produce a list of "core non-essential reactions". These reactions are found
270 in all individuals selected, and do not produce a lethal phenotype when removed. A full
271 list of all core non-essential reactions described here can be found in Table S4. There are
272 14 core non-essentials reactions found in all 1194 of the individuals examined when
273 *i*RH830 was supplied with blood; AGDC, ARGabc, ASnt2r, G6PDA, H2Ot, HlSt2r, lLEt2r,
274 NH4t, RPE, TKT1, TKT2, TMK, TRPt2r and TYRt2r. In sap, only one non-essential reaction
275 is found in all 1888 individuals; the L-arginine ABC transporter reaction ARGabc. As
276 anticipated, when grown in the limited famine medium there are a higher number of
277 core nonessential reactions (22 found in each of the 2989 individuals tested); ATPS4r,
278 CO2t,
279 ENO, FORT, GAPD, GHMT2r, GLUDy, ORNDC, PAPSr, PGCD, PGK, PGM, PPPGO3, PSERT,
280 PSP_L, RPE, TALA, THRAi, TKT1, TPI, TRDR and TRPS1.

281 The rare core non-essential reactions were then calculated. In famine, there are 73
282 unique reactions that occur in less than 0.1% of the 2989 evolved models. This is
283 significantly more than for sap (13 in less than 0.1% of 1888 models) and blood (five in
284 less than 0.1% of 1194 models).

285 These core non-essential reactions were then analysed by subsystem to assess themes
286 across the different conditions. In blood, over half (eight of 14) of these are secondary
287 transporter reactions (Fig. 4a). This reflects what is observed in *S. glossinidius*, which has
288 retained, for example, secondary amino acid transporters, as well as losing metabolic
289 pathways whilst maintaining functional transporters in order to scavenge free
290 metabolites [33, 81]. As mentioned previously, the only core nonessential reaction in
291 sap is a transport reaction. In contrast, the set of core nonessentials are more varied
292 when metabolites are limited (famine), with a particular emphasis towards central
293 metabolism and amino acid metabolism.

294 Fig. 4. Core non-essential reactions in evolved *iRH830* populations. (a) The proportion of
295 core non-essential reactions per conditions by subsystem when the ancestral *iRH830* is
296 exposed to blood (left), famine (centre), or sap (right) media. (b) Core non-essential
297 reactions in Δ ASPTA (left), Δ PDH (centre), and Δ PPC (right) *iRH830* models in a blood
298 medium, grouped by subsystem.

299 The order of gene loss can be estimated

300 A characteristic of *S. glossinidius* and other symbiotic bacteria is their propensity to
301 accumulate pseudogenes [41]. It is not known whether certain genes are lost early in
302 the tsetse-*Sodalis* symbiosis in order to facilitate the establishment of the relationship,
303 or whether their loss is an inevitable consequence of genomic streamlining. To
304 investigate the effect that pseudogenising key genes early in evolutionary time has on
305 the trajectory of a symbiont, the MOEA was run on *iRH830* with one of three reactions
306 involved in the TCA cycle removed at the start, with the resulting solutions compared to
307 wild-type (WT) (Table 1, Scenario ii). An assumption is made in these simulations that
308 pseudogenes are non-functional. The first reaction selected was PPC
309 (phosphoenolpyruvate carboxylase) (Fig. 5). The gene encoding this reaction, *ppc*, is
310 pseudogenised in *S. glossinidius* [33] and it was thought that the loss of this gene would
311 have had a significant impact on the resulting evolution of the symbiont. The two other
312 reactions selected were PDH (pyruvate dehydrogenase) and ASPTA (aspartate
313 transaminase). These reactions are both encoded by genes predicted to be functional in
314 *S. glossinidius* (Fig. 5). It was hypothesised that the loss of PPC may result in a different

315 evolutionary trajectory compared to the loss of PDH or ASPTA, with the former
316 potentially producing solutions that were more similar to *S. glossinidius* metabolism.
317

318 Fig. 5. TCA cycle reactions examined by the MOEA. Three reactions were removed from
319 *iRH830* to investigate the resulting trajectories following application of the MOEA;
320 ASPTA, PDH and PPC. Blue arrows show reactions functional in *S. praecaptivus* and *S.*
321 *glossinidius*, white arrows show reaction not functional in *S. glossinidius*. Gene
322 associations in *S. praecaptivus* and *S. glossinidius* are given in blue and black text,
323 respectively. Adapted from Hall *et al.* [33].

324 When considering the population plots, there is minimal qualitative difference between
325 Δ PDH and Δ PPC (Fig. 3b). Δ ASPTA, in contrast, produces solutions with a much lower
326 biomass output and with fewer individuals that deviate away from the optimum as
327 defined by the fitness function. A selection of individuals were then examined and the
328 number of core non-essential reactions in the evolved models analysed as described
329 previously (Fig. 3b, black boxes). There are one, eleven and nine core non-essential
330 reactions in the WT, Δ PDH and Δ PPC solutions, respectively, whereas there are 61 in
331 Δ ASPTA. These 61 reactions function in a variety of subsystems, particularly transport,
332 central metabolism, amino acid metabolism and nucleotide salvage pathways. There are
333 minimal differences between the core non-essential reactions at the subsystem level
334 between Δ PDH and Δ PPC (Fig. 4b). The main difference of note is the presence of
335 reactions involved in amino acid metabolism in the Δ ASPTA, but not the Δ PDH or Δ PPC,
336 solutions. This could be of relevance given the amino acid-rich hematophagous tsetse diet.
337 The order in which key pseudogenisations occurred can therefore be estimated, using
338 the resulting evolutionary trajectories as a guide. The gene encoding PPC could have
339 been lost early by *S. glossinidius* in the sequence of pseudogenisations with minimal
340 impact on its fitness.

341 A prediction of the evolutionary future of *S. glossinidius* as a symbiont

342 *S. glossinidius* is a secondary symbiont. Both bacterium and insect can survive
343 independently of one another, and the former is likely a more recent acquisition [25]. It
344 is however unclear how recently *S. glossinidius* was captured, or, given the
345 pseudogenisations already present, how much more streamlined its genome can
346 become. The algorithm was therefore applied to *iLF517* in a blood medium with the aim

347 of evaluating potential future evolutionary trajectories within the bounds of its
348 relationship with host and primary symbiont (Table 1, Scenario iii). There are a spread
349 of biomass outputs found at the end of the simulation (Fig. 6a), as observed when
350 *iRH830* was evolved in blood. The smallest solutions contain approximately 300
351 reactions. After 3000 generations the *iLF517* model retained $51 \pm 1.1\%$ of the starting
352 606 reactions compared to *iRH830* which was reduced to $27 \pm 1.8\%$ of the 1247 starting
353 reactions. The *S. glossinidius* genome can therefore reduce its potential coding capacity
354 for metabolic genes to approximately half the size that it is currently.

355 The evolved solutions were then compared to the evolved *iRH830* models to assess their
356 similarity. *iLF517* converges on a minimum after approximately 1000 generations,
357 compared to the 2500 generations taken by the evolved *iRH830* model to find the
358 minimum number of reactions (Fig. S2). The greater standard deviation of the *iRH830*
359 solutions compared to *iLF517* is likely due to the larger starting point of the free-living
360 model. The evolved *iRH830* and *iLF517* solutions ultimately converge at a similar point.
361 To investigate this further, ten evolved models for both *iRH830* and *iLF517* were
362 analysed. All exchange reactions and those that carried zero flux were removed from
363 further analysis. Full evolved models with fluxes can be found in Supplementary Data 3.
364 Of the 383 unique reactions that carry flux in the evolved *iRH830* models, 289 (75.5%)
365 are found in all ten. For the *iLF517* solutions, 301 of the 316 (95.3%) unique reactions
366 that carry flux are found in all ten. This suggests that the smaller *S. glossinidius* model
367 has fewer viable trajectories compared to the larger *S. praecaptivus* model. Of the 441
368 unique reactions across the 20 evolved models, 225 (51%) were found in all of the
369 *iRH830* and *iLF517* evolved solutions; they are core across the two species. The biomass
370 outputs for the evolved *iRH830* and *iLF517* solutions range from 0.064 to 0.281 (gr DW
371 $(\text{mmol glucose})^{-1} \text{hr}^{-1}$), and 0.075 to 0.331 (gr DW $(\text{mmol glucose})^{-1} \text{hr}^{-1}$), respectively (Fig.
372 6b). Given the differences between the solutions from the two simulations, and the
373 lower proportion of core conserved reactions, it is possible that *S. praecaptivus* may not
374 be the free-living species of *Sodalis* most closely related to *S. glossinidius*, and that there
375 may be others yet to be discovered, or may now be extinct or unrecognisable from the
376 *S. glossinidius* progenitor.

377

378

379 Fig. 6. Evolution of *iLF517*. (a) Evolution of *iLF517* in a blood medium. MOEA was run for
380 3000 generations (blue to green). (b) Biomass output (gr DW $(\text{mmol glucose})^{-1} \text{hr}^{-1}$) and
381 the number of reactions carrying flux in evolved *iRH830* (blue triangle) and evolved
382 *iRH830* (yellow circle) models. The evolved solutions produce comparable biomass
383 outputs. Ten evolved solutions are given for each, some duplicates are present.

384 Discussion

385 Classical studies of microbial evolution, whilst useful, are ultimately limited by their
386 inherent inability to replicate adaptations over large evolutionary time scales. Here, we
387 present a computational approach by combining a MOEA with FBA, with the *Sodalis*
388 system as a model for this. The *Sodalis* genus is ideal for the study of the evolution of
389 symbiosis in this way, as within the genus are a free-living and a host restricted species,
390 both well defined with complete genome sequences and existing protocols for culture.

391 Here, we present a model for *S. praecaptivus* metabolism, *iRH830*, accompanied by
392 experimental verification, that has been used in subsequent *in silico* evolution
393 experiments. Supplying the ancestral *iRH830* with contrasting growth media
394 demonstrates the effect that nutrient provisioning may have on evolutionary
395 trajectories of symbiotic bacteria. Exposure to the famine medium reflects what might
396 be expected in a nutrient-limited environment *in vivo*, in which evolutionary pressures
397 result in the retention of pathways to synthesise key, essential metabolites. Here, this
398 has shown to be particularly evident in the pathways retained for
399 glycolysis/gluconeogenesis, the pentose phosphate pathway, and amino acid
400 metabolism. This indicates that key pathways in central metabolism are being retained
401 when the external environment is nutrient-limited. The evolved solutions therefore
402 reflect what is observed in symbiotic bacteria; the retention or loss of pathways can be
403 used to inform about the microenvironment it resides within. Some reactions in these
404 pathways are also likely to be retained as they produce essential components of the
405 biomass reaction. It is expected that the biomass reaction for symbiotic bacteria will
406 change over evolutionary time, and therefore this work is limited by maintaining a
407 consistent biomass reaction throughout the simulation. Incorporating a biologically
408 accurate, variable biomass reaction, for example one which will at certain time points
409 lose components that the host can synthesise, would be an interesting progression to
410 this work.

411 It has been shown in the simulations presented here that the evolved famine solutions
412 contain a much greater number of core non-essential reactions that are present in a
413 small percentage of the solutions. This suggests a lack of flexibility in the evolved
414 network; either the reaction is found repeatedly, or not at all. This is not observed in the
415 solutions provided with complex media (blood or sap), where a greater degree of
416 flexibility is demonstrated by more reactions being included repeatedly across the
417 evolutionary space. This implies that, *in vivo*, there are many possible trajectories for an
418 early symbiont if there are sufficient nutrients in the microenvironment.

419 The work here demonstrates the power of evolutionary algorithms in the study of
420 symbiont evolution. A strength of this system is that removal of a reaction from the
421 model is not irreversible; it is possible for a reaction to be added back into an individual
422 at any point. This reduces the likelihood of repeatedly encountering the same
423 evolutionary solutions as a result of losing the same key reaction, or reactions, early in
424 the simulation. Although according to Muller's ratchet [82] the reduction of symbiont
425 genomes should be irreversible [70], *S. praecaptivus* is as yet not obligately intracellular.
426 A simple model to allow the (re-)acquisition of reactions is therefore appropriate for this
427 system. Whilst there is no evidence currently for horizontal gene transfer (HGT) within
428 species of the *Sodalis* genus, the NSGA-II algorithm is only intended to be used as a tool
429 to explore the possible evolutionary space rather than as a biologically accurate model
430 of genome reduction. Previous examples of evolving minimal metabolic networks do not
431 allow for full exploration of the possible evolutionary space [50,56,57]. Decisions that
432 are made at the start of process persist, which, whilst biologically relevant, does not
433 allow the full complement of evolutionary routes to be examined. Expanding this
434 algorithm to include the possibility of the model acquiring reactions that are not
435 currently encoded by *S. praecaptivus*, therefore more closely reflecting HGT, may be an
436 interesting avenue of future research.

437 This tool can produce biologically relevant simulations that accurately reflect the
438 metabolic pressures that symbionts are exposed to. An example of this was
439 demonstrated by the investigation of key knockouts in *S. glossinidius*. The symbiont has
440 a pseudogenisation in *ppc*, a key gene in central metabolism [33]. It is not possible to
441 deduce when this loss occurred from the genome annotation alone. Reactions are
442 related to genes in multiple ways through the gene-protein-reaction relations which
443 may or may not be 1:1; here we evolve reactions to focus on the phenotypic effects. As
444 the MOEA enables a flexible search methodology this will cause minimal difference in
445 outcomes to a gene-centred approach. By removing the PPC reaction from *S.*
446 *praecaptivus* at the start of the simulation, the resulting trajectories can be analysed and
447 compared to WT. The loss of PPC appeared to have minimal effect on the evolved
448 populations compared to WT, in contrast to what was observed when the ASPTA
449 reaction was removed at the start (Fig. 3). This would indicate that, *in vivo*, the loss of
450 the gene encoding the ASPTA reaction would have a greater impact on a bacterial
451 symbiont if it was lost early in the relationship in comparison to the lower burden that
452 the loss of the genes encoding PDH or PPC would have. This result can then be used to
453 infer the possible sequence of gene loss in the tsetse-*S. glossinidius* symbiosis. *S.*
454 *glossinidius* has lost the *ppc* gene, whereas it has retained the genes encoding the PDH

455 (SG0467-9) and ASPTA (SG1006) reactions [25,33]. As the profile of Δ PPC evolution is
456 similar to that of WT, it could be suggested that the *ppc* gene could have been lost early
457 in evolutionary time without heavily bottlenecking *S. glossinidius* evolution
458 subsequently. The gene encoding the ASPTA reaction may have been retained by *S.*
459 *glossinidius* because of the detrimental impact that its loss may have caused. This is
460 therefore a useful tool for making general predictions about when key
461 pseudogenisations in insect-bacterial symbioses may have occurred.

462 It has been shown here that it is possible for *S. glossinidius* to reduce its metabolic
463 network to approximately half of the size that it is currently. This provides support for
464 the published hypothesis that *S. glossinidius* is a recent acquisition by the tsetse [25].
465 The number of reactions remains slightly higher in evolved *i*RH830 models compared to
466 evolved *i*LF517 solutions, possibly due to difficulties in finding the minima from a larger
467 starting point. *i*RH830 can however be reduced down to look phenotypically similar to
468 *i*LF517 at the level of biomass output, but with differences at the individual reaction level
469 (Supplementary Data 3). These results suggest therefore that the route that *S.*
470 *glossinidius* has taken within the tsetse is perhaps just one of several possible routes.
471 The differences also indicate that *S. praecaptivus* may not be the ancestor that initiated
472 the tsetse-*S. glossinidius* symbiosis. The unusual ability of *S. praecaptivus* to metabolise
473 xylitol may be related to the frequent presence of the *Sodalis* genus as a symbiont
474 amongst sap-feeding insects, by hinting that it may naturally subsist on this important
475 plant-derived sugar.

476 A possible, if challenging, extension to this work could be to incorporate the influence
477 of the host and other members of the microbiome on the evolution of *Sodalis*. We
478 acknowledge that this modelling method does not account for changes in host fitness
479 that may arise from the evolution of the symbiont. The host could, for example,
480 constrain the population of symbiont when it increases beyond a certain density, as
481 demonstrated by the weevil *Sitophilus oryzae* that produce antimicrobial peptides to
482 constrict the symbiont population size [83]. Alternatively, the host may benefit from
483 reduced costs of symbiont maintenance [84], or an increased fitness or efficiency of the
484 symbiont via the provision of metabolites. It may also suffer if the bacterial population
485 becomes less fit. The latter is less likely to be an issue here, given that it is not yet known
486 for certain whether *S. glossinidius* provides a benefit to the tsetse. This level of nuance
487 is not captured by FBA as it focuses entirely on the fitness of an individual, with the only
488 reference here to a population being the selection of the next generation. This tool is
489 therefore most useful as a technique to examine broad changes that may occur during
490 microbial evolution.

491 Previous uses of metabolic models to simulate evolution have focused on *E. coli* and *B.*
492 *aphidicola* as a proof of concept [56,57]. The availability of both a genome sequence and
493 a culturable organism for a free-living and symbiont of the same genus makes the *Sodalis*
494 system a candidate model system to investigate the evolution of symbiosis. The work
495 described here has augmented knowledge about the loss of key genes in *S. glossinidius*
496 central metabolism. Combining FBA with a MOEA in this way could be used for any
497 organism for which a well-annotated genome is available. It could be applied not only
498 to the evolution of symbiosis but to the directed evolution of, for example, industrially
499 relevant microorganisms or to the study of rapid genome evolution in pathogenic
500 bacteria.

501 Materials and Methods

502 Bacterial strains, growth conditions and reagents

503 *S. praecaptivus* was obtained from DMSZ (Brunswick, Germany). Working stocks were
504 established by incubating starter cultures on LB (Merck, Darmstadt, Germany) agar
505 plates overnight at 37°C. A single colony was then sub-cultured on to a fresh LB plate
506 and incubated overnight at 37°C. A single colony was selected with a sterile pipette tip
507 and used for downstream experimentation as per Biolog, Inc (Hayward, CA, USA)
508 manufacturer protocol. Briefly, the colony was vortexed in IF-0 media before a redox
509 dye was added (Biolog). Phenotypic microplates were used to screen for the ability of *S.*
510 *praecaptivus* to grow on a range of carbon sources, using PM1 and PM2A microplates
511 (Biolog). A 100 μ L bacterial suspension in the relevant media was added per well. Optical
512 density was measured at 590 nm and 730 nm in a microplate reader (Epoch, BioTek,
513 Winooski, VT, USA), and incubated with double orbital shaking at 37°C for 24 hours.

514 Discrepancies between *in silico* and *in vitro* Biolog results were re-examined by
515 establishing individual cultures of *S. praecaptivus* in M9 salts in 96-well microplates, and
516 supplemented with the metabolite of interest at a range of concentrations from 25 mM
517 to 50 μ M. Cultures were incubated in a microplate reader with double orbital shaking at
518 37°C for 36 hours.

519 Construction of the *S. praecaptivus* metabolic network

520 The annotated *S. praecaptivus* genome sequence, CP006569.1, was downloaded from
521 NCBI in GenBank format. Genes in *S. praecaptivus* with the same annotation as genes in
522 the *E. coli* str. K-12 substr. MG1655 genome (ASM584v2) were highlighted, and the
523 reactions encoded by these genes extracted from the BiGG Models database [85]. These
524 processes were automated using custom scripts written in Python.

525 FBA models of *S. glossinidius* (*iLF517* [33]) and *E. coli* (*iJO1366* [54,55], *iJR904* [53],
526 *iAF1260* [52]) were then used to aid the identification of missing reactions. The reactions
527 and corresponding gene assignments in these published models were compared to the
528 draft *S. praecaptivus* model. These gene assignments were then used to guide translated
529 nucleotide and protein BLAST searches of the *S. praecaptivus* genome. KEGG [86,87] and
530 EcoCyc [88] databases were used to confirm the identity of the *E. coli* genes encoding
531 each reaction. *S. glossinidius* gene assignments were taken from *iLF517* [33]. These
532 orthologues in *E. coli* and *S. glossinidius*, with sequences taken from UniProt [89], were
533 used as BLAST search queries.

534 KEGG, BiGG Models, and MetaCyc [90] were used to assign reaction stoichiometry.
535 Candidate pseudogenes were aligned with known functional orthologues using ClustalX
536 2.1 [91]. Those with sequences missing or mutations in key residues were not included
537 in the model. FBA and literature searches were used to identify and fill gaps in metabolic
538 pathways appropriately [92]. The xylitol pathway components in *Morganella morganii*
539 subsp. *morganii* were identified using KEGG, with candidate protein sequences
540 extracted from UniProt and used in a protein BLAST search against *S. praecaptivus*. KEGG
541 was also used to identify known GalNAc degradation pathways.

542 Flux balance analysis

543 FBA solutions were generated using the GNU linear programming kit (GLPK) integrated
544 with custom software in Java. Oxygen uptake was constrained to 20 mmol gr DW⁻¹ hr⁻¹,
545 comparable to other models of free-living Gram negative bacteria. The uptake of
546 ammonia, water, phosphate, sulphate, potassium, sodium, calcium, carbon dioxide,
547 protons and essential transition metals was unconstrained for all media conditions.
548 Cofactor constraints were implemented by introducing these metabolites to the
549 biomass functions at small fluxes (1 x 10⁻⁵ mmol gr DW⁻¹ hr⁻¹) [7]. *iRH830* was supplied
550 with either 6 mmol gr DW⁻¹ hr⁻¹ GlcNAc and 1 mmol gr DW⁻¹ hr⁻¹ thiamine ("famine"), a

551 tsetse-specific medium ("blood", Table S3), or a sap-inspired medium (from [92], "sap").
552 Full recipes are provided in Supplementary Data 1. The phenotype was considered viable
553 if the biomass production rate was greater than 1×10^{-4} gr DW (mmol glucose)⁻¹ hr⁻¹.
554 Futile cycles, closed loops of a number of reactions, were detected by the presence of
555 unsustainably large fluxes. Futile cycles often occur when several reversible reactions
556 are present in which the product of one becomes the substrate of another. These
557 reactions were examined individually, and solved by adjusting the reversability with
558 guidance from EcoCyc and BiGG Models.

559 To investigate the concordance between the *in vitro* screen and the *in silico* outputs,
560 *i*RH830 was, where possible, supplemented with the carbon sources analysed at an
561 exogenous concentration of 6 mmol gr DW⁻¹ hr⁻¹. A qualitative presence/absence of a
562 positive biomass output was noted. Full description of the model is provided in
563 Supplementary Data 1.

564 Robustness analysis

565 Robustness analysis of the *i*RH830 network was executed using COBRApy [94] to conduct
566 single reaction deletions. *i*RH830 was supplied with either famine or blood media under
567 aerated conditions. The flux through reactions was set to zero individually and the
568 resulting effect on biomass output measured. Reactions were categorised as essential if
569 the resulting biomass output was less than 1×10^{-3} gr DW (mmol glucose)⁻¹ hr⁻¹.

570 Implementation of multi-objective evolutionary algorithm

571 A MOEA was used to explore possible evolutionary trajectories in the *Sodalis* genus. An
572 overview of the process is provided in Fig. 7. The non-dominated sorted genetic
573 algorithm (NSGA-II) [95] from the Distributed Evolutionary Algorithms in Python (DEAP)
574 [96] package was used in combination with the COBRApy package [94] for FBA
575 evaluation. Equal weight was placed on reducing the number of reactions used in the
576 model whilst maximising the biomass output. The Python code for running the algorithm
577 with an example data set is available on GitHub
578 (<https://github.com/St659/SodalisFBAEvolution>). The full datasets generated by the
579 simulations are available on the York Research Database.

580

581 Fig. 7. Process of the MOEA. A starting population of individuals is initialised, and the
582 fitness calculated by solving the FBA model to calculate biomass output and the number
583 of active reactions. For each generation the population is allowed to mutate and then
584 the fitness of each individual is evaluated from the biomass output and the sum of the
585 active reactions. A new population is then selected using nondominated sorting,
586 generating a Pareto front of biomass output to active reactions. The process of mutation
587 and selection is repeated for 3000 generations resulting in a final population. Green
588 boxes represent the start and final populations, pink boxes represent the iterative
589 process of mutation and selection.

590 Population initiation

591 Prior to starting an evolutionary run, reactions essential to growth were identified using
592 a single reaction knockout. Essential reactions were defined as those producing a
593 biomass output of less than 1×10^{-3} gr DW (mmol glucose)⁻¹ hr⁻¹. Reactions that were
594 identified as essential were not included in the subsequent mutation strategy, therefore
595 reducing the solution space and computational time taken to run the MOEA. The
596 essential reactions were added back to the evolved populations for downstream
597 analysis.

598 At the start of the algorithm an initial population of 100 genotype copies was created,
599 with all non-essential reactions being active (Fig. 7). Each genotype consisted of a binary
600 number, where a 1 or 0 corresponded to the reaction being active or inactive,
601 respectively. This is a proxy for gene loss, where a one-to-one gene-protein reaction
602 mapping is assumed. Reactions, rather than genes, were used to reduce the potential
603 search space whilst maintaining the key phenotypic effect. All post-evolution analysis
604 focused on the reactions lost or retained.

605 Mutation

606 Mutation was performed on each genotype by flipping the value of each reaction with a
607 probability of 0.005 (Fig. 7). The fitness of each individual is evaluated by solving the FBA
608 model to calculate both its biomass output and the sum of number of active reactions.

609 Fitness evaluation and selection

610 The population was first evaluated for non-dominated individuals. This gave a
611 population of individuals that has the highest biomass output for their current number

612 of active reactions (Fig. 7). From the non-dominated population, the Euclidean distance
613 between each individual was calculated. A greater priority was given to selecting
614 individuals with a larger Euclidean distance. This prevented the clustering of similar
615 potential solutions, thereby reducing the likelihood of becoming trapped in sub-optimal
616 local minima within the search space. The resulting population maximised the
617 convergence on the highest biomass output, lowest number of reactions, and the
618 distribution of those solutions. There will be a set of solutions whereby the number of
619 reactions cannot be minimised further without also reducing the corresponding biomass
620 output. This set of solutions is known as a Pareto front. The algorithm was repeated for
621 3000 generations to produce genotypes that converged. This indicated that minimal
622 new solutions were being found. The biomass output from the slim optimisation
623 COBRAPy function and the summation of the number of active reactions was used to
624 evaluate the fitness.

625 MOEA variations

626 The MOEA was run under several conditions in order to investigate aspects of symbiont
627 evolution. Full details are provided in Table 1. Scenario i investigated the trajectories
628 taken when the *S. praecaptivus* model was provided with blood, sap, and famine growth
629 media. In Scenario ii, gene knockouts were simulated by removing individual reactions
630 from the *S. praecaptivus* model prior to commencing the evolution. The reactions
631 chosen were ASPTA, PDH, and PPC. In Scenario iii, the MOEA was applied to a model of
632 *S. glossinidius* metabolism, *iLF517* [33]. Here, *iLF517* was supplied with the blood
633 medium for 3000 generations.

634 Analysis of evolved populations

635 For all conditions the algorithm was independently run ten times, giving a total of 1000
636 final solutions. All of the solutions were pooled together for analysis. To identify key
637 reactions in the evolved populations, individuals were selected from each condition and
638 the remaining non-essential reactions extracted. The subset of reactions that were
639 present in every individual selected were designated as "core nonessentials", and are
640 referred to hereafter as such. When examining the similarity between evolved models,
641 exchange reactions and reactions carrying zero flux were discounted.

642 Acknowledgements

643 The authors would like to thank Greg Hurst for insightful comments on this work, and
644 the two anonymous reviewers for their constructive and thorough feedback.

645 Conflicts of interest

646 The authors declare no conflicts of interest.

647 Funding information

648 RJH was funded by the BBSRC White Rose DTP (BB/M011151/1) and ST by the Wellcome
649 Trust CIDCATS programme (WT095024MA).

650 References

651 [1] Wilcox JL, Dunbar HE, Wolfinger RD, Moran NA. Consequences of reductive
652 evolution for gene expression in an obligate endosymbiont. *Molecular*
653 *Microbiology*. 2003 5;48(6):1491 – 1500.

654 [2] Dunbar HE, Wilson ACC, Ferguson NR, Moran NA. Aphid thermal tolerance is
655 governed by a point mutation in bacterial symbionts. *PLoS Biology*. 2007 4;5(5):e96.

656 [3] Oliver KM, Russell JA, Moran NA, Hunter MS. Facultative bacterial symbionts in
657 aphids confer resistance to parasitic wasps. *Proceedings of the National Academy*
658 *of Sciences*. 2003;100(4):1803–1807.

659 [4] Nakabachi A, Ueoka R, Oshima K, Teta R, Mangoni A, Gurgui M, et al. Defensive
660 bacteriome symbiont with a drastically reduced genome. *Current Biology*. 2013
661 8;23(15):1478–1484.

662 [5] Aksoy S. *Wigglesworthia* gen. nov. and *Wigglesworthia glossinidia* sp. nov., taxa
663 consisting of the mycetocyte-associated, primary endosymbionts of tsetse flies.
664 *International Journal of Systematic and Evolutionary Microbiology*.
665 1995;45(4):848–851.

- 666 [6] Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. Genome sequence of
667 the endocellular bacterial symbiont of aphids *Buchnera* sp. *APS. Nature*.
668 2000;407:81–86.
- 669 [7] Thomas GH, Zucker J, Macdonald SJ, Sorokin A, Goryanin I, Douglas AE. A fragile
670 metabolic network adapted for cooperation in the symbiotic bacterium *Buchnera*
671 *aphidicola*. *BMC Systems Biology*. 2009;3:24.
- 672 [8] Snyder AK, Rio RVM. *Wigglesworthia morsitans* folate (Vitamin B9) biosynthesis
673 contributes to tsetse host fitness. *Applied and Environmental Microbiology*.
674 2015;81(16):5375 – 5386.
- 675 [9] Hrusa G, Farmer W, Weiss BL, Applebaum T, Roma JS, Szeto L, et al. TonBdependent
676 heme iron acquisition in the tsetse fly symbiont *Sodalis glossinidius*.
677 *Applied and Environmental Microbiology*. 2015;81(February):04166–14.
- 678 [10] Manzano-Marín A, Ocegüera-Figueroa A, Latorre A, Jiménez-García LF, Moya A.
679 Solving a bloody mess: B-vitamin independent metabolic convergence among
680 gammaproteobacterial obligate endosymbionts from blood-feeding arthropods
681 and the leech *Haementeria officinalis*. *Genome Biology and Evolution*.
682 2015;7(10):2871–2884.
- 683 [11] McCutcheon JP, Moran NA. Parallel genomic evolution and metabolic
684 interdependence in an ancient symbiosis. *Proceedings of the National Academy of*
685 *Sciences*. 2007 12;104(49):19392–19397.
- 686 [12] McCutcheon JP, McDonald BR, Moran NA. Convergent evolution of metabolic roles
687 in bacterial co-symbionts of insects. *Proceedings of the National Academy of*
688 *Sciences*. 2009 9;106(36):15394–9.
- 689 [13] McCutcheon JP, McDonald BR, Moran NA. Origin of an alternative genetic code in
690 the extremely small and GC-rich genome of a bacterial symbiont. *PLoS Genetics*.
691 2009 7;5(7):e1000565.
- 692 [14] Wilson ACC, Ashton PD, Caleviro F, Charles H, Colella S, Febvay G, et al. Genomic
693 insight into the amino acid relations of the pea aphid, *Acyrtosiphon pisum*, with
694 its symbiotic bacterium *Buchnera aphidicola*. *Insect Molecular Biology*. 2010
695 2;19:249–258.
- 696 [15] McCutcheon JP, von Dohlen CD. An interdependent metabolic patchwork in the
697 nested symbiosis of mealybugs. *Current Biology*. 2011 8;21(16):1366– 1372.

- 698 [16] Rio RVM, Lefevre C, Heddi A, Aksoy S. Comparative genomics of insect symbiotic
699 bacteria: influence of host environment on microbial genome composition. *Applied*
700 *and Environmental Microbiology*. 2003;69(11):6825–6832.
- 701 [17] Michalkova V, Benoit JB, Weiss BL, Attardo GM, Aksoy S. Vitamin B6 generated by
702 obligate symbionts is critical for maintaining proline homeostasis and fecundity in
703 tsetse flies. *Applied and Environmental Microbiology*. 2014;80(18):5844–5853.
- 704 [18] Baumann P, Baumann L, Lai CY, Rouhbakhsh D, Moran NA, Clark MA.
705 Genetics, physiology, and evolutionary relationships of the genus *Buchnera*:
706 Intracellular symbionts of aphids. *Annual Review of Microbiology*. 1995;49:55– 94.
- 707 [19] Akman Gündüz E, Douglas AE. Symbiotic bacteria enable insect to use a nutritionally
708 inadequate diet. *Proceedings of the Royal Society B*. 2009;276:987– 991.
- 709 [20] Richards S, Gibbs RA, Gerardo NM, Moran N, Nakabachi A, Richards S, et al. Genome
710 Sequence of the Pea Aphid *Acyrtosiphon pisum*. *PLoS Biology*. 2010
711 2;8(2):e1000313.
- 712 [21] Wu M, Sun LV, Vamathevan J, Riegler M, Deboy R, Brownlie JC, et al. Phylogenomics
713 of the reproductive parasite *Wolbachia pipientis* wMel: A streamlined genome
714 overrun by mobile genetic elements. *PLoS Biology*. 2004;2(3):e69.
- 715 [22] Pérez-Brocal V, Gil R, Ramos S, Lamelas A, Postigo M, Michelena JM, et al. A small
716 microbial genome: The end of a long symbiotic relationship? *Science*. 2006
717 10;314(5797):312–313.
- 718 [23] Moran NA, Mccutcheon JP, Nakabachi A. Genomics and evolution of heritable
719 bacterial symbionts. *Annual Review of Genetics*. 2008;42:165–190.
- 720 [24] Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, Hattori M, et al. Genome
721 sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia*
722 *glossinidia*. *Nature Genetics*. 2002;32(3):402–407.
- 723 [25] Dale C, Maudlin I. *Sodalis* gen. nov. and *Sodalis glossinidius* sp. nov., a
724 microaerophilic secondary endosymbiont of the tsetse fly *Glossina morsitans*
725 *morsitans*. *International Journal of Systematic Bacteriology*. 1999;49:267–275.
- 726 [26] Edwards JS, Covert M, Palsson B. Metabolic modelling of microbes: the fluxbalance
727 approach. *Environmental Microbiology*. 2002;4(3):133–140.
- 728 [27] Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. *Current*
729 *Opinion in Biotechnology*. 2003 10;14(5):491–496.

- 730 [28] Lewis NE, Nagarajan H, Palsson BO. Constraining the metabolic genotype–
731 phenotype relationship using a phylogeny of in silico methods. *Nature*
732 *Reviews Microbiology*. 2012 4;10(4):291–305.
- 733 [29] Orth JD, Thiele I, Palsson B. What is flux balance analysis? *Nature Biotechnology*.
734 2010;28(3):245–248.
- 735 [30] Macdonald SJ, Lin GG, Russell CW, Thomas GH, Douglas AE. The central role of the
736 host cell in symbiotic nitrogen metabolism. *Proceedings of the Royal Society B*.
737 2012 8;279:2965–2973.
- 738 [31] Macdonald SJ, Thomas GH, Douglas AE. Genetic and metabolic determinants of
739 nutritional phenotype in an insect-bacterial symbiosis. *Molecular Ecology*. 2011
740 5;20(10):2073–2084.
- 741 [32] Belda E, Silva FJ, Peretó J, Moya A. Metabolic networks of *Sodalis glossinidius*: A
742 systems biology approach to reductive evolution. *PLoS ONE*. 2012;7(1):e30652.
- 743 [33] Hall RJ, Flanagan LA, Bottery MJ, Springthorpe V, Thorpe S, Darby AC, et al. A tale
744 of three species: Adaptation of *Sodalis glossinidius* to tsetse biology,
745 *Wigglesworthia* metabolism, and host diet. *mBio*. 2019 1;10(1):02106– 18.
- 746 [34] Ankrah NYD, Luan J, Douglas AE. Cooperative metabolism in a three-partner insect-
747 bacterial symbiosis revealed by metabolic modeling. *Journal of Bacteriology*. 2017
748 8;199(15):00872–16.
- 749 [35] González-Domenech C, Belda E, Patiño-Navarrete R, Moya A, Peretó J, Latorre A.
750 Metabolic stasis in an ancient symbiosis: Genome-scale metabolic networks from
751 two *Blattabacterium cuenoti* strains, primary endosymbionts of cockroaches. *BMC*
752 *Microbiology*. 2012 1;12(Suppl 1):S5.
- 753 [36] Sørensen MES, Cameron DD, Brockhurst MA, Wood AJ. Metabolic constraints for a
754 novel symbiosis. *Royal Society Open Science*. 2016;3(3):150708.
- 755 [37] Shastri AA, Morgan JA. Flux balance analysis of photoautotrophic metabolism.
756 *Biotechnology Progress*. 2005 12;21(6):1617–1626.
- 757 [38] Nogales J, Gudmundsson S, Knight EM, Palsson BO, Thiele I. Detailing the optimality
758 of photosynthesis in cyanobacteria through systems biology analysis. *Proceedings*
759 *of the National Academy of Sciences*. 2012 2;109(7):2678– 83.
- 760 [39] Knoop H, Zilliges Y, Lockau W, Steuer R. The metabolic network of *Synechocystis* sp.
761 PCC 6803: Systemic properties of autotrophic growth. *Plant Physiology*. 2010
762 9;154(1):410–422.

- 763 [40] Knoop H, Gründel M, Zilliges Y, Lehmann R, Hoffmann S, Lockau W, et al. Flux
764 balance analysis of cyanobacterial metabolism: The metabolic network of
765 *Synechocystis* sp. PCC 6803. *PLoS Computational Biology*. 2013 6;9(6):e1003081.
766 [pcbi.1003081](https://doi.org/10.1371/journal.pcbi.1003081).
- 767 [41] Toh H, Weiss BL, Perkin SAH, Yamashita A, Oshima K, Hattori M, et al. Massive
768 genome erosion and functional adaptations provide insights into the symbiotic
769 lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Research*.
770 2006;16(2):149–156.
- 771 [42] Goodhead I, Blow F, Brownridge P, Hughes M, Kenny J, Krishna R, et al.
772 Large scale and significant expression from pseudogenes in *Sodalis glossinidius* - a
773 facultative bacterial endosymbiont. *bioRxiv*. 2018 4;p. 124388.
- 774 [43] Covert MW, Palsson BO. Transcriptional regulation in constraints-based metabolic
775 models of *Escherichia coli*. *The Journal of Biological Chemistry*. 2002
776 8;277(31):28058–28064.
- 777 [44] Lee JM, Gianchandani EP, Papin JA. Flux balance analysis in the era of
778 metabolomics. *Briefings in Bioinformatics*. 2006;7(2):140–150.
- 779 [45] Covert MW, Palsson BO. Constraints-based models: regulation of gene expression
780 reduces the steady-state solution space. *Journal of Theoretical Biology*.
781 2003;221:309–325.
- 782 [46] Covert MW, Schilling CH, Palsson B. Regulation of gene expression in flux balance
783 models of metabolism. *Journal of Theoretical Biology*. 2001 11;213(1):73– 88.
- 784 [47] Mahadevan R, Edwards JS, Doyle FJ. Dynamic flux balance analysis of diauxic growth
785 in *Escherichia coli*. *Biophysical Journal*. 2002 9;83(3):1331– 1340.
- 786 [48] Harcombe WR, Delaney NF, Leiby N, Klitgord N, Marx CJ. The ability of flux balance
787 analysis to predict evolution of central metabolism scales with the initial distance
788 to the optimum. *PLoS Computational Biology*. 2013 6;9(6):e1003091.
- 789 [49] Pfeiffer T, Soyer OS, Bonhoeffer S. The evolution of connectivity in metabolic
790 networks. *PLoS Biology*. 2005;3(7):e228.
- 791 [50] Ponce-de Leon M, Tamarit D, Calle-Espinosa J, Mori M, Latorre A, Montero F, et al.
792 Determinism and contingency shape metabolic complementation in an
793 endosymbiotic consortium. *Frontiers in Microbiology*. 2017;8:2290.

- 794 [51] Edwards JS, Palsson BO. Metabolic flux balance analysis and the in silico analysis of
795 Escherichia coli K-12 gene deletions. BMC Bioinformatics. 2000 7;1(1):1.
- 796 [52] Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, et al. A genome-
797 scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for
798 1260 ORFs and thermodynamic information. Molecular Systems Biology. 2007
799 1;3(1):121.
- 800 [53] Reed JL, Vo TD, Schilling CH, Palsson BO. An expanded genome-scale model of
801 Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biology. 2003;4(9):R54.
- 802 [54] Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, et al. A comprehensive
803 genome-scale reconstruction of Escherichia coli metabolism–2011. Molecular
804 Systems Biology. 2011 10;7(1):535.
- 805 [55] Orth JD, Palsson BO. Gap-filling analysis of the iJO1366 Escherichia coli metabolic
806 network reconstruction for discovery of metabolic functions. BMC Systems Biology.
807 2012 5;6(1):30.
- 808 [56] Pál C, Papp B, Lercher MJ, Csermely P, Oliver SG, Hurst LD. Chance and necessity in
809 the evolution of minimal metabolic networks. Nature. 2006 3;440(7084):667–670.
- 810 [57] Yizhak K, Tuller T, Papp B, Ruppin E. Metabolic modeling of endosymbiont genome
811 reduction on a temporal scale. Molecular Systems Biology. 2011 1;7(1):479.
812 McNally CP, Borenstein E. Metabolic model-based analysis of the emergence of
813 bacterial cross-feeding through extensive gene loss. bioRxiv. 2017;p.
814 10.1101/180208.
- 815 [58] McNally CP, Borenstein E. Metabolic model-based analysis of the emergence of
816 bacterial cross-feeding through extensive gene loss. bioRxiv. 2017.
817 p.10.1101/18020.
- 818 [59] Hindré T, Knibbe C, Beslon G, Schneider D. New insights into bacterial adaptation
819 through in vivo and in silico experimental evolution. Nature Reviews Microbiology.
820 2012 5;10(5):352–365.
- 821 [60] Batut B, Parsons DP, Fischer S, Beslon G, Knibbe C. In silico experimental evolution:
822 a tool to test evolutionary scenarios. BMC Bioinformatics. 2013 10;14(Suppl
823 15):S11.

- 824 [61] François P, Hakim V. Design of genetic networks with specified functions by
825 evolution in silico. Proceedings of the National Academy of Sciences.
826 2004;101(2):580–585.
- 827 [62] Budinich M, Bourdon J, Larhlimi A, Eveillard D. A multi-objective constraintbased
828 approach for modeling genome-scale microbial ecosystems. PLoS ONE. 2017
829 2;12(2).
- 830 [63] Wang L, Maranas CD. MinGenome: An n silico top-down approach for the synthesis
831 of minimized genomes. ACS Synthetic Biology. 2018;7:462–473.
- 832 [64] Fong SS, Burgard AP, Herring CD, Knight EM, Blattner FR, Maranas CD, et al. In silico
833 design and adaptive evolution of Escherichia coli for production of lactic acid.
834 Biotechnology and Bioengineering. 2005 9;91(5):643–648.
- 835 [65] Garcia S, Trinh C. Comparison of multi-objective evolutionary algorithms to solve
836 the modular cell design problem for novel biocatalysis. bioRxiv. 2019;p.
837 10.1101/616078.
- 838 [66] Machado D, Herrgård MJ. Co-evolution of strain design methods based on flux
839 balance and elementary mode analysis. Metabolic Engineering Communications.
840 2015 12;2:85–92.
- 841 [67] Clayton AL, Oakeson KF, Gutin M, Pontes A, Dunn DM, von Niederhausern AC, et al.
842 A novel human-infection-derived bacterium provides insights into the evolutionary
843 origins of mutualistic insect–bacterial symbioses. PLoS Genetics.
844 2012;8(11):e1002990.
- 845 [68] Chari A, Oakeson KF, Enomoto S, Jackson DG, Fisher MA, Dale C. Phenotypic
846 characterisation of *Sodalis praecaptivus* sp. nov., a close non-insect associated
847 member of the *Sodalis*-allied lineage of insect endosymbionts. International Journal
848 of Systematic and Evolutionary Microbiology. 2015;65:1400–1405.
- 849 [69] WHO. Control and surveillance of African trypanosomiasis. WHO Technical Report
850 Series. 1998;881:1-113.
- 851 [70] Moran NA. Accelerated evolution and Muller’s ratchet in endosymbiotic bacteria.
852 Proceedings of the National Academy of Sciences. 1996;93:2873–2878.
- 853 [71] Novakova E, Hypsa V. A new *Sodalis* lineage from bloodsucking fly *Craterina melbae*
854 (Diptera, Hippoboscoidea) originated independently of the tsetse flies symbiont
855 *Sodalis glossinidius*. FEMS Microbiology Letters. 2007 4;269(1):131– 135.

- 856 [72] Chrudimský T, Husník F, Nováková E, Hypša V. Candidatus *Sodalis melophagi* sp.
857 nov.: Phylogenetically independent comparative model to the tsetse fly symbiont
858 *Sodalis glossinidius*. PLoS ONE. 2012 7;7(7):e40354.
- 859 [73] Boyd BM, Allen JM, Koga R, Fukatsu T, Sweet AD, Johnson KP, et al. Two bacterial
860 genera, *Sodalis* and *Rickettsia*, associated with the seal louse *Proechinophthirus*
861 *fluctus* (Phthiraptera: Anoplura).
- 862 [74] Oakeson KF, Gil R, Clayton AL, Dunn DM, von Niederhausern AC, Hamil C, et al.
863 Genome Degeneration and Adaptation in a Nascent Stage of Symbiosis. *Genome*
864 *Biology and Evolution*. 2014;6(1):76–93.
- 865 [75] Szklarzewicz T, Kalandyk-Kołodziejczyk M, Michalik K, Jankowska W, Michalik A.
866 Symbiotic microorganisms in *Puto superbus* (Leonardi, 1907) (Insecta, Hemiptera,
867 Coccothraupidae: Putoidea). *Protoplasma*. 2018 1;255(1):129–138.
- 868 [76] Koga R, Bennett GM, Cryan JR, Moran NA. Evolutionary replacement of obligate
869 symbionts in an ancient and diverse insect lineage. *Environmental*
870 *Microbiology*. 2013 7;15(7):2073–2081.
- 871 [77] Koga R, Moran NA. Swapping symbionts in spittlebugs: evolutionary replacement
872 of a reduced genome symbiont. *The ISME Journal*. 2014 1;8:1237.
- 873 [78] Michalik A, Jankowska W, Kot M, Gołas A, Szklarzewicz T. Symbiosis in the green
874 leafhopper, *Cicadella viridis* (Hemiptera, Cicadellidae). Association in statu
875 nascendi? *Arthropod Structure & Development*. 2014 11;43(6):579– 587.
- 876 [79] Kaiwa N, Hosokawa T, Kikuchi Y, Nikoh N, Meng XY, Kimura N, et al. Primary gut
877 symbiont and secondary, *Sodalis*-allied symbiont of the Scutellerid stinkbug *Cantao*
878 *ocellatus*. *Applied and Environmental Microbiology*. 2010 6;76(11):3486–94.
- 879 [80] Hosokawa T, Kaiwa N, Matsuura Y, Kikuchi Y, Fukatsu T. Infection prevalence of
880 *Sodalis* symbionts among stinkbugs. *Zoological Letters*. 2015 12;1(1):5.
- 881 [81] Snyder AK, Deberry JW, Runyen-Janecky L, Rio RVM. Nutrient provisioning
882 facilitates homeostasis between tsetse fly (Diptera: Glossinidae) symbionts.
883 *Proceedings of the Royal Society B: Biological Sciences*. 2010;277(1692):2389–
884 2397.
- 885 [82] Muller HJ. The relation of recombination to mutational advance. *Mutation*
886 *Research/Fundamental and Molecular Mechanisms of Mutagenesis*. 1964 5;1(1):2–
887 9.
- 888 [83] Login FH, Balmand S, Vallier A, Vincent-Monégat C, Vigneron A, WeissGayet M, et
889 al. Antimicrobial peptides keep insect endosymbionts under control. *Science*. 2011
890 10;334(6054):362 – 365.
- 891 [84] Ankrah NYD, Chouaia B, Douglas AE. The cost of metabolic interactions in symbioses
892 between insects and bacteria with reduced genomes. *mBio*. 2018 9:e01433-18.

893 [85] King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, et al. BiGG Models: A
894 platform for integrating, standardizing and sharing genome-scale models. *Nucleic
895 Acids Research*. 2016 1;44(D1):D515–D522.

896 [86] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic
897 Acids Research*. 2000 1;28(1):27–30.

898 [87] Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for
899 understanding genome variations in KEGG. *Nucleic Acids Research*. 2019
900 1;47(D1):D590–D595.

901 [88] Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martínez C, Caspi
902 R, et al. The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-
903 12. *Nucleic Acids Research*. 2017 1;45(D1):D543–D550.

904 [89] Bateman A, Martin MJ, Orchard S, Magrane M, Alpi E, Bely B, et al. UniProt: a
905 worldwide hub of protein knowledge. *Nucleic Acids Research*. 2019 1;47(D1):D506–
906 D515.

907 [90] Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, et al.
908 The MetaCyc database of metabolic pathways and enzymes and the BioCyc
909 collection of pathway/genome databases. *Nucleic Acids Research*. 2007
910 12;36(Database):D623–D631.

911 [91] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et
912 al. Clustal W and Clustal X version 2.0. *Bioinformatics*. 2007 11;23(21):2947–2948.

913 [92] Thiele I, Palsson BO. A protocol for generating a high-quality genome-scale
914 metabolic reconstruction. *Nature Protocols*. 2010;5(1):93–121.

915 [93] Krishnan HB, Natarajan SS, Bennett JO, Sicher RC. Protein and metabolite
916 composition of xylem sap from field-grown soybeans (*Glycine max*). *Planta*.
917 2011;233:921–931.

918 [94] Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: COstraintsBased
919 Reconstruction and Analysis for Python. *BMC Systems Biology*. 2013 8;7(1):74.

920 [95] Deb K, Pratap A, Agarwal S, Meyarivan T. A fast and elitist multiobjective genetic
921 algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*.
922 2002;6(2):182–197.

923 [96] De Rainville FM, Fortin FA, Gardner MA, Parizeau M, Gagné C. DEAP: A Python
924 framework for evolutionary algorithms. In: *Proceedings of the 14th International
925 Conference on Genetic and Evolutionary Computation*; 2012. p.
926 85–92.