Titles: Robustness encoded across essential and accessory replicons in an ecologically versatile bacterium

Running title: Genome effects on gene essentiality

Authors: George C diCenzo[1]*, Alex B Benedict[2], Marco Fondi[1], Graham C Walker[3], Turlough M Finan[4], Alessio Mengoni[1], Joel S Griffitts[2]

Affiliations: [1] Department of Biology, University of Florence, Sesto Fiorentino, FI, 50019, Italy.
[2] Department of Microbiology and Molecular Biology, Brigham Young University, Provo, UT, 84602, USA.
[3] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA.
[4] Department of Biology, McMaster University, Hamilton, ON, L8S 4K1, Canada.

*** Corresponding author:** George C diCenzo
Department of Biology
University of Florence
Sesto Fiorentino, FI, 50019
Italy
Email: georgecolin.dicenzo@unifi.it

Keywords: genetic interactions / synthetic lethality / systems biology / Tn-seq / constraint based modelling

ABSTRACT

Bacterial genome evolution is characterized by gains, losses, and rearrangements of functional genetic segments. The extent to which genotype-phenotype relationships are influenced by large-scale genomic alterations has not been investigated in a high-throughput manner. In the symbiotic soil bacterium *Sinorhizobium meliloti*, the genome is composed of a chromosome and two large extrachromosomal replicons (pSymA and pSymB, which together constitute 45% of the genome). Massively parallel transposon insertion sequencing (Tn-seq) was employed to evaluate contributions of chromosomal genes to fitness in both the presence and absence of these extrachromosomal replicons. Ten percent of chromosomal genes from diverse functional categories are shown to genetically interact with pSymA and pSymB. These results demonstrate the pervasive robustness provided by the extrachromosomal replicons, which is further supported by constraint-based metabolic modelling. A comprehensive picture of core *S. meliloti* metabolism was generated through a Tn-seq-guided *in silico* metabolic network reconstruction, producing a core network encompassing 726 genes. This integrated approach facilitated functional assignments for previously uncharacterized genes, while also revealing that Tn-seq alone misses over a quarter of wild type metabolism. This work highlights the strong functional dependencies and epistatic relationships that may arise between bacterial replicons and across a genome, while also demonstrating how Tn-seq and metabolic modelling can be used together to yield insights not obtainable by either method alone.

INTRODUCTION

The prediction of genotype-phenotype relationships is a fundamental goal of genetic, biomedical, and eco-evolutionary research, and this problem underpins the design of synthetic microbial systems for biotechnological applications [1]. The last decades have witnessed a shift away from the functional characterization of single genes towards whole-genome, systems-level analyses [for recent reviews, see [2,3]]. Such studies have been facilitated by the development of methods that allow for the direct interrogation of a genome to determine all genetic elements required for adaptation to a specified environment. Two primary methods are *in silico* metabolic modelling [4,5], and massively parallel sequencing of transposon insertions in bacterial mutant libraries (Tn-seq) [6,7].

The process of *in silico* genome-scale metabolic modelling consists of two stages. First, a reconstruction of all cellular metabolism is built that contains all reactions expected to be present, as well as which genes encode the enzymes performing each reaction, thereby linking genetics to metabolism [8]. Next, mathematical models such as flux balance analysis (FBA) are used to simulate the flux distribution through the reconstructed metabolic network [9], which can be used to predict how environmental perturbations or gene disruptions influence growth phenotypes. This approach allows for phenotypic predictions of all possible single, double, or higher-order gene deletion mutations within a matter of days [10,11], something that is infeasible using a direct experimental approach. However, the quality of the predictions is highly dependent on the accuracy of the metabolic reconstruction. Outside of a few model species like *Escherichia coli*, experimental genetic and biochemical data are not available at the resolution necessary to provide accurate assignment of all metabolic gene functions.

66 The Tn-seq approach involves the generation of a library of hundreds of thousands of
67 mutant clones, each containing a single transposon insertion at a random genomic location [12].
68 The library of pooled clones is then cultured in the presence of a defined environmental
69 challenge. Insertions resulting in altered fitness in the environment under investigation become
70 under- or over-represented in the population, and this is monitored by deep sequencing to
71 identify the genomic location and frequency of all transposon insertions. This approach is
72 imperfect, as important biochemical functions may be encoded redundantly in the genome [13-
73 15], and the loss of some essential genes can be compensated for by evolution of alternative
74 cellular processes [16]. Moreover, fitness changes brought about by mutation in one gene may be
75 dependent on mutation of a second gene bearing no resemblance to the first—a phenomenon
76 known as a genetic interaction [17,18]. Such genetic interactions may cause the apparent
77 functions of some genes to be strictly dependent on their genomic environment [19]. In other
78 words, a gene may be essential for growth in one organism, but its orthologous counterpart in
79 another organism may be non-essential. This significantly complicates efforts to generalize
80 genotype-phenotype relationships [20].

81 Resolving the problem of genome-conditioned gene function is of broad significance in
82 the areas of functional genomics, population genetics, and synthetic biology. For example, the
83 ability to design and build optimized minimal cell factories on the basis of single-mutant fitness
84 data is expected to present numerous complications [21], as evidenced by the recent effort to
85 rationally build a functional minimal genome [22]. Tn-seq studies have suggested there is as
86 little as 50% to 25% overlap in the essential genome of any two species [23-25]. As a striking
87 example, 210 of the Tn-seq determined essential genes of *Pseudomonas aeruginosa* PA14 are
88 not even present in the genome of *P. aeruginosa* PAO1 [26]. Comparison of Tn-seq data for

Shigella flexneri with the deletion analysis data for closely related *E. coli* suggested only a small number of genes were specifically essential in one species. Mutation of about 100 genes, however, appeared to result in a growth rate decrease specifically in *E. coli* [27]. Similarly, comparison of Tn-seq datasets from two *Salmonella* species revealed that mutation of nearly 40 genes had a stronger growth phenotype in one of the two species [28]. Overall, these studies suggest that the genomic environment (here defined as the genomic components that may vary from organism to organism) influences the fitness contributions of a significant proportion of an organism's genes. However, no large-scale analysis has been performed that directly illustrates how the phenotypes of individual genes are impacted when a small or large part of the genome is modified.

Here, we provide a quantitative, genome-scale evaluation of how large-scale genomic variance influences genotype-phenotype relationships. We have accomplished this in a way that minimizes the effects of laboratory-to-laboratory variation, and removes the effects of complex genome evolution. The model system used is *Sinorhizobium meliloti*, an α-proteobacterium whose 6.7-Mb genome consists of a chromosome and two additional replicons, the pSymA megaplasmid and the pSymB chromid. The pSymA and pSymB replicons constitute 45% of the *S. meliloti* genome (~2,900 genes); yet, by simply transferring only two essential genes from pSymB to the chromosome, both pSymA and pSymB can be completely removed from the genome, yielding a viable single-replicon organism [29]. We report a comparison of gene essentiality (via Tn-seq) for wild-type *S. meliloti* and the single-replicon derivative. This analysis was supplemented by an *in silico* double gene deletion analysis of a *S. meliloti* genome-scale metabolic network reconstruction. We further examine how integration of Tn-seq data with *in silico* metabolic modelling, through a Tn-seq-guided reconstruction process, overcomes the

112 limitations of using either of these approaches in isolation to develop a consolidated view of the
113 core metabolism of the organism. This process produced a fully referenced *core S. meliloti*
114 metabolic reconstruction.
115
116 **RESULTS**
117 **Development and validation of the Tn*5*-based transposon Tn*5*-714.**
118 In order to interrogate the *S. meliloti* genome using a Tn-seq based approach, we first
119 developed a new construct based on the Tn*5* transposon as described in the Materials and
120 Methods. The resulting transposon (Figure S1) contains constitutive promoters reading out from
121 both ends of the transposon to ensure the production of non-polar mutations. Analysis of the
122 insertion site locations validated that the transposon performed largely as expected. Gene
123 disruptions caused by transposon insertions were confirmed to be non-polar as illustrated by the
124 case reported in Figure 1, and there was no strong bias in the distribution of insertions around the
125 chromosome (Figures 2A, S2). However, there did appear to be somewhat of a bias for
126 integration of the transposon in GC rich regions (Figure S3). Given the high GC content (62.7%)
127 of the *S. meliloti* chromosome, it is unlikely that this moderate bias had a discernable influence
128 on the results of this study.
129 **Overview of the Tn-seq output.**
130 The Tn-seq experiments reported here were undertaken with two primary aims: i) to
131 identify the core set of genes contributing to *S. meliloti* growth in laboratory conditions, and ii) to
132 determine the extent to which the phenotypic consequence of a gene deletion is influenced by the
133 genomic environment (i.e. presence/absence of the secondary replicons). To accomplish this, Tn-
134 seq libraries of two *S. meliloti* strains were prepared: a wild type strain (designated RmP3499)

135 containing the entire genome, and a strain with both the pSymA and pSymB replicons removed
136 (designated RmP3496 or ΔpSymAB; strains described previously in [30]). Transposon library
137 sizes were skewed to compensate for the difference in genome sizes, resulting in nearly identical
138 insertion site density for each library (Table S1). Both libraries were passed through selective
139 growth regimens in either complex BRM broth (rich medium) or minimal VMM broth (defined
140 medium) in duplicates. Following approximately nine generations of growth, the location of the
141 transposon insertions in the population was determined, a gene essentiality index (GEI) was
142 calculated for all chromosomal genes, and each gene was classified into one of five fitness
143 categories (Table 1) using the procedure described in the Materials and Methods. Four genes
144 (*pdxJ*, *fumC*, *smc01011*, *smc03995*), including two of unknown function, were independently
145 mutated in the wild-type background, and in all cases, the mutations yielded the expected no-
146 growth phenotype (Figure S4), supporting the accuracy of the Tn-seq output. All Tn-seq data is
147 available as Data Set S1.

148 A strong correlation was observed between the number of insertions per gene in each set
149 of duplicates (Figure S5), indicating that there was high reproducibility of the results and that
150 differences between conditions were unlikely to reflect random fluctuations in the output. On
151 average, insertions were found in 190,000 unique chromosomal positions with a median of 39
152 unique insertion positions per gene (Table S1). The similarity in the number of unique insertion
153 positions between samples suggested that differences in the Tn-seq outputs were also unlikely to
154 be an artefact of the quality of the libraries.

155 **Elucidation of the core genetic components of *S. meliloti*.**

156 There were 307 genes classified as essential independently of growth medium or strain
157 (Figure S6). This set of 307 genes includes those encoding functions commonly understood to be

158 essential: the DNA replication apparatus, the four RNA polymerase subunits, the housekeeping
159 sigma factor, the general transcriptional termination factor Rho, 40 out of 55 of the annotated
160 ribosomal protein subunits, 18 out of 20 of the annotated aminoacyl-tRNA synthetases, and 6 out
161 of 10 of the annotated ATP synthase subunits. Considering genes classified as essential plus
162 those genes whose mutation resulted in a large growth defect (Groups I and II in Table 1), a core
163 growth promoting genome of 489 genes, representing ~ 15% of the chromosome, was identified
164 (Figure 2B). This expanded list includes 51 out of 55 of the annotated ribosomal protein
165 subunits, 19 out of 20 of the annotated aminoacyl-tRNA synthetases, and 9 out of 10 of the
166 annotated ATP synthase subunits These 489 genes appeared to be mostly dispersed around the
167 chromosome, although there was a bias for these genes to be found in the leading strand (Figure
168 2A). Based on published RNA-seq data for *S. meliloti* grown in a glucose minimal medium,
169 these 489 genes tend to be highly expressed, with a median expression level above the 90%
170 percentile (Figure S7). Compared to the entire chromosome (Fisher exact test, p-value < 0.05
171 following a Bonferroni correction for 18 tests), this set of 489 genes was enriched for genes
172 involved in translation (5.2-fold), lipid metabolism (2.7-fold), cofactor metabolism (3.3-fold),
173 and electron transport (2.1-fold), whereas genes involved in transport (2.1-fold),
174 motility/attachment (9.4-fold), and hypothetical genes (2.7-fold) were under-represented (Figure
175 2C). Additionally, cell wall (2.2-fold) and cell division (2.3-fold) were over-represented while
176 transcription (1.9-fold) was under-represented (Figure 2C), although these differences where not
177 considered statistically significant.

178 A clear influence of the growth medium on the fitness phenotypes of gene mutations was
179 observed. The degree to which mutant phenotypes were impacted by growth medium type is
180 reflected in the synthetic medium index (SMI) calculated as described in the Materials and

181 Methods. Focusing on the wild-type strain, a core of 519 genes were identified as contributing
182 equally to growth in both media (Figure 2D). Forty genes were identified as more important
183 during growth in rich medium than in defined medium, and these genes had a median SMI score
184 of 7 (values of 1 and -1 are neutral). Only translation functions (5.8-fold) displayed a statistically
185 significant enrichment in these genes, which may reflect the faster growth rate in the rich
186 medium (Figure S8), while there was also a non-statistically significant enrichment in signal
187 transduction (5.1-fold) (Figure 2C). The extent of specialization for growth in the defined
188 medium was more pronounced; 93 genes were more important during growth in the defined
189 medium with a median SMI score of -20. These genes were enriched (statistically significant) in
190 amino acid (9.0-fold) and nucleotide (6.7-fold) metabolism presumably due to the requirement of
191 their biosynthesis, and carbohydrate metabolism (3.6-fold) likely as the sole carbon source was a
192 carbohydrate (Figure 2C). The same overall pattern was observed between media for the
193 ΔpSymAB strain (Figure S9).

194 **Mutant fitness phenotypes are strongly influenced by their genomic environment.**

195 The Tn-seq data sets for the wild-type and the ΔpSymAB strains were compared to
196 evaluate the robustness of the observed fitness phenotypes in response to changes in the gene's
197 genomic environment. Similar results were observed for both growth media, suggesting that the
198 results were generalizable and not medium specific. Depending on the medium, either 484 or 488
199 genes had an equal contribution to growth in both strains, 81 or 89 genes led to stronger growth
200 impairment when mutated in wild-type cells, and either 250 or 251 genes led to stronger growth
201 impairment when mutated in ΔpSymAB cells (Figures 2E, 2F, and Table 2). Only minor
202 functional bias was observed in the genes that displayed larger fitness defects in the ΔpSymAB
203 background (Figure 2C); in both media, only electron transport (3-fold) and oxidoreductases

204 (9.5-fold) were over- and under-represented, respectively. Similarly, little functional bias was
205 detected in genes with larger fitness defects in the wild-type background (Figure 2C); in both
206 media, lipid metabolism (4.5-fold) and hypothetical genes (2-fold) were over- and under-
207 represented, respectively, while nucleotide metabolism (5.5-fold) was also enriched in the rich
208 medium. Overall, these results were consistent with pervasive effects of the genomic
209 environment on the genotype-phenotype relationship that was largely independent of the
210 biological role of the gene products.

211 Approximately half (9 of 16) of the genes that were independently mutated in both strains
212 yielded the expected phenotypes on rich agar plates (Figures S10). Of the other seven genes,
213 which were expected to be essential specifically in the ΔpSymAB strain, at least three were non-
214 lethal but displayed obvious growth rate defects or extended lag phases during liquid culture
215 experiments (Table S2 and Figure S11). The remaining three genes may represent false positives
216 from the Tn-seq screen, or may reflect differences in the growth conditions, namely, competitive
217 growth versus isogenic growth. Nevertheless, the observation that at least 75% of the selected
218 genes were confirmed to have a genome content-dependent fitness phenotype validates that the
219 large majority of the strain specific phenotypes observed in the Tn-seq screen represent true
220 differences.

221 **Level of genetic and phenotypic conservation of the essential *S. meliloti* genes.**

222 Several recent studies have used Tn-seq to study the essential genome of *Rhizobium*
223 *leguminosarum* [31-33]. We compared our Tn-seq datasets with those reported in by Perry *et al*
224 [32] to examine the conservation of the essential genome of these two closely related N_2-fixing
225 species. Putative orthologs for ~ 75% of all *S. meliloti* chromosomal genes were identified in *R.*
226 *leguminosarum* via a Blast Bidirectional Best Hit (Blast-BBH) approach (Data Set S2). Much

227 higher conservation of the growth promoting genome was observed; 97% of the 489 core growth
228 promoting genes and 99% of the 307 core essential genes had a putative ortholog in *R.*
229 *leguminosarum*. However, conservation of the gene did not necessarily correspond to
230 conservation of the phenotype. Considering only the 303 conserved core essential *S. meliloti*
231 genes (as these were the least likely to have been falsely identified as essential), 8% (25 of 303)
232 of their orthologous genes were classified as having little contribution to growth on defined
233 medium in *R. leguminosarum* (Figure 3A). An additional 34 genes were considered to be non-
234 essential but growth defective when mutated (Figure 3A). Independent mutation of two genes
235 (*fumC*, *pdxJ*) identified as specifically essential in *S. meliloti* confirmed their essentiality (Figure
236 S4), supporting the Tn-seq data. A similar pattern is observed starting with the *R. leguminosarum*
237 genes classified as essential in both minimal and complex medium by Perry *et al*. [32]. Of the
238 241 core essential *R. leguminosarum* genes with an ortholog on the *S. meliloti* chromosome, 21
239 (9%) of the orthologs were classified as non-essential in *S. meliloti* for growth in defined
240 medium, while an additional 8 were considered to have a moderate growth defect (Figure 3B).

241 To further test the species specificity of the above-mentioned genes, the experiment was
242 replicated *in silico*. Fifteen of the 25 orthologs specifically essential in *S. meliloti* were present
243 both in our existing *S. meliloti* genome-scale metabolic model [34] as well as in a draft *R.*
244 *leguminosarum* metabolic model (see Materials and Methods). Flux balance analysis was used to
245 examine the *in silico* effect of deleting these 15 pairs of orthologs on growth. Three pairs of
246 orthologs were classified as essential in both models, five were classified as non-essential in both
247 models, and seven were classified as essential specifically in the *S. meliloti* model. Thus, at least
248 half of the gene essentiality differences observed in the Tn-seq data are corroborated by the *in*
249 *silico* metabolic simulation, despite the preliminary nature of the draft *R. leguminosarum* model.

250 An *in silico* analysis of the genes identified as specifically essential in *R. leguminosarum* on the
251 basis of the Tn-seq data was not performed as only two of these genes were present in the *R.*
252 *leguminosarum* model.

253 ***In silico* analyses support a high potential for genetic redundancy in the *S. meliloti* genome.**

254 The results of the previous two sections are consistent with a strong genomic
255 environment effect on the phenotypic consequences of gene mutations. One possible explanation
256 is the presence of widespread genetic redundancy, at the gene and/or pathway level. In support of
257 this, ~ 14% of chromosomal genes had a Blast-BBH hit when the chromosomal proteome was
258 compared against the combined pSymA/pSymB proteome (Data Set S3). Therefore, this
259 phenomenon was further explored using a constraint-based metabolic modelling approach.

260 We first tested the *in silico* effect of chromosomal single gene deletions on growth rate in
261 the presence and absence of pSymA/pSymB (Figure 4A). This analysis identified 67 genes (~
262 7% of all chromosomal model genes) as having a more severely impaired growth phenotype
263 when deleted in the absence of pSymA/pSymB genes, 38 of which were lethal. This appeared to
264 be due to a combination of direct functional redundancy of the gene products as well as through
265 metabolic bypasses, as deletion of 50 reactions dependent on chromosomal genes had a more
266 severe phenotype in the absence of pSymA/pSymB, 42 of which were lethal (Figure S12).

267 Next, a double gene deletion analysis was performed to examine the effect on growth rate
268 of deleting every possible pair of model genes. This analysis suggested that 49 chromosomal
269 genes had a more significant impact on growth than expected when simultaneously deleted with
270 a single pSymA or pSymB gene (Figure 4B). Additionally, synthetic negative phenotypes were
271 observed for 97 chromosomal genes when simultaneously deleted with another chromosomal
272 gene (Figure 4C). Overall, 14% of chromosomal genes were predicted to have a synthetic

273 negative phenotype when co-deleted with a second gene, consistent with a high potential for
274 metabolic robustness being encoded by the *S. meliloti* genome, and with a significant influence
275 of the genomic environment on the fitness phenotype of gene mutations.

276 **A consolidated view of core *S. meliloti* metabolism through Tn-seq-guided *in silico***
277 **metabolic reconstruction.**

278 The results described in the previous sections made it evident that a Tn-seq approach
279 alone is insufficient to elucidate all processes contributing to growth in a particular environment.
280 This is especially true if also considering non-essential metabolism that is nevertheless actively
281 present in wild type cells, such as exopolysaccharide production. Moreover, it is difficult to fully
282 comprehend the core functions of a cell by simply examining a list of essential genes and their
283 predicted functions. We therefore attempted to overcome these limitations by using the Tn-seq
284 data to guide a manual *in silico* reconstruction of the core metabolic processes of *S. meliloti*. A
285 detailed description of this process is provided in the Materials and Methods. In brief, the
286 existing metabolic model iGD1575 was treated as a database of reactions and gene-reaction
287 associations. Each pathway involved in central carbon metabolism or the production of essential
288 or non-essential biomass components (Table S3) were then rebuilt in a new (initially empty)
289 reconstruction drawing from the reactions present in iGD1575. At the same time, the genes
290 associated with each reaction were compared to the Tn-seq data and published literature to
291 confirm the linkage of the correct gene(s) to each reaction.

292 The resulting model, termed iGD726 and included as in SBML format in File S2, is
293 summarized in Figure 5 and Table 3, and the entire model including genes, reaction formulas,
294 and references is provided as an easy to read Excel table in Data Set S4. The process of
295 integrating the Tn-seq data with *in silico* metabolic reconstruction resulted in a major refinement

296 of the core metabolism compared to the existing genome-scale model: 228 new reactions were
297 added, 115 new genes were added, and the genes associated with 135 of the 432 reactions
298 common to both reconstructions were updated. In addition to improving the metabolic
299 reconstruction, this process significantly expanded the view of core *S. meliloti* metabolism
300 compared to that gained solely through the application of Tn-seq. The genes associated with
301 approximately one third of the iGD726 reactions were not detected as growth promoting in the
302 Tn-seq datasets (Figure 5, Table 3). While many of the additional reactions present in iGD726
303 are due to the inclusion of non-essential biomass components, which are part of the wild type
304 cell but are nonetheless dispensable for growth, others are from essential metabolic pathways
305 (Figures 5, S13). Overall, the combined approach of integrating Tn-seq data and *in silico*
306 metabolic modelling allowed for the development of a high-quality representation of core *S.*
307 *meliloti* metabolism in a way that neither approach alone was capable of accomplishing.

308 **Tn-seq-guided *in silico* metabolic reconstruction facilitates novel gene annotation.**

309 Over 20 of the reactions of the core metabolic reconstruction initially had no gene
310 attributed with producing the enzyme responsible for its catalysis. Similarly, many genes with no
311 clear biological function were found to be essential in the Tn-seq screen. By attempting to fill the
312 gaps in the *in silico* model with the uncharacterized essential genes, we were able to assign
313 putative functions to eight previously uncharacterized genes (Table S4). Two of these genes were
314 chosen for further characterization, *smc01361* and *smc04042*. The *smc01361* gene was annotated
315 as encoding a dihydroorotase, and mutation of *smc01361* resulted in pyrimidine auxotrophy
316 (Figure S14). Given its location next to *pyrB*, and the presence of an essential PyrC
317 dihydroorotoase encoded elsewhere in the genome (Data Set S1), we propose that *smc01361*
318 encodes an inactive dihydroorotase (PyrX) required for PyrB activity as has been observed in

319 some other species including *Pseudomonas putida* [35,36]. The essential *smc04042* gene was
320 annotated as an inositol-1-monphosphatase family protein. It was previously observed that
321 rhizobia lack a gene encoding a classical L-histidinol-phosphate phosphohydrolase, and it was
322 suggested an inositol monophosphatase family protein may fulfill this function instead [37].
323 Mutation of *smc04042* resulted in histidine auxotrophy (Figure S14), consistent with this enzyme
324 fulfilling the role of a L-histidinol-phosphate phosphohydrolase. It is likely that this is true for
325 most rhizobia, as putative orthologs of this gene were identified in all 10 of the examined
326 Rhizobiales genomes (Data Set S4). These examples illustrate the power of the combined Tn-seq
327 and metabolic reconstruction process in the functional annotation of bacterial genomes.
328

329 **DISCUSSION**

330 In this study, we developed a new variant of the Tn*5* transposon for construction of non-
331 polar insertion mutations that should be readily adaptable for use with other α-proteobacteria.
332 The Tn*5* transposon was chosen as it was expected to have low insertion site specificity in *S.*
333 *meliloti* [38]. However, we observed a moderate sequence insertion bias for GC rich regions
334 consistent with previous studies of the Tn*5* transposon [39-41]. The consensus sequence of ~
335 190,000 unique insertion locations was largely consistent, but not identical, with that previously
336 reported [41]; however, the specificity appeared to extend past the 9 base pair region that is
337 duplicated during Tn*5* insertion (Figure S3). While this bias is unlikely to have a significant
338 influence on the results in species with high GC content genomes, such as *S. meliloti*, accounting
339 for this bias may be important when applying Tn*5* mutagenesis to species with low GC content
340 genomes.

341 Greater than 10% of species with a sequenced genome contain a genomic architecture
342 similar to *S. meliloti*, that is, with at least two large DNA replicons [42,43]. Several studies have
343 revealed that, in many ways, each replicon acts as a functionally and evolutionarily distinct entity
344 (for a review, refer to [43]); yet, there can also be regulatory cross-talk [44], as well as the
345 exchange of genetic material between the replicons [45]. The Tn-seq analyses reported here
346 provide new insights into the functional integration of secondary replicons into the host
347 organism. The pSymB replicon of *S. meliloti* is known to have two essential genes (which were
348 transferred to the chromosome for this study) [45], while pSymA has no essential genes [46].
349 However, the large number of chromosomal genes—across many functional groups (Figure 2)—
350 that became conditionally essential following the removal of pSymA and pSymB indicate the
351 presence of many genes whose products can perform essential metabolic capabilities but that
352 remain cryptic due to inter-replicon epistatic interactions. It was also interesting to note that the
353 strength of the correlation between duplicates (Figure S5), as determined by the size of the
354 absolute residuals, was higher for the ΔpSymAB strain than for the wild type strain in both
355 media (p-value $< 2.2 \times 10^{-16}$ for both media, as determined with Welch two-sample t-tests). This
356 may be reflective of the genetic robustness encoded by the secondary replicons and the stochastic
357 activation of these processes in the mutant population. Potentially, the high level of inter-
358 replicon redundancy may reduce the level of purifying selection on the chromosomal copies of
359 the genes, facilitating more rapid diversification of gene functionality and increased rates of
360 chromosomal gene evolution. Overall, the results of these analyses suggest that secondary
361 replicons may influence the evolution of the chromosome and play a vital role in the biology of
362 the organism, even if these activities remain cryptic due to inter-replicon epistatic interactions.

363 More generally, the Tn-seq data reported here provide a unique perspective of how a
364 gene's genomic environment influences its genotype-phenotype relationship. Previous studies
365 have illustrated that the fitness phenotypes of orthologous genes of both distant and closely
366 related species may differ [21,23-28,47], and even how intercellular effects within microbial
367 communities can modify the essential genome of a species [48]. The data reported here more
368 directly addressed the influence of the genomic environment by comparing the fitness
369 phenotypes of mutating the exact same set of ~ 3,500 genes in two very different genomic
370 environments. It was found that the non-essential genome had a remarkable influence on what
371 was classified as a growth-promoting gene, with 10% of *S. meliloti* chromosomal genes
372 exhibiting fitness-based genetic interactions with the non-essential component of the genome
373 (Figure 2). This observation was not growth medium-dependent, was not unique to a specific
374 gene functional class, and was not simply due to an overall reduced fitness of the ΔpSymAB
375 strain as the findings could be largely replicated *in silico* (Figure 4).
376 The majority of the genes whose fitness phenotype was dependent on the genomic
377 environment became more important for fitness following the genome reduction. In many cases,
378 this is expected to reflect a loss of functional redundancy; the increased importance of the
379 chromosomal cytochrome genes likely reflects a compensation for the loss of the pSymA/pSymB
380 encoded cytochrome complexes (Figure 6). In other cases, it may reflect newly activated
381 pathways that must compensate for the loss of a normal housekeeping pathway. The specific
382 essentiality of proline biosynthesis, and the second half of histidine biosynthesis, in the
383 ΔpSymAB strain during growth in rich medium presumably reflects the inability of these strains
384 to transport these compounds and must therefore synthesize them *de novo* (Figure 6). Indeed,
385 previous metabolomics work is consistent with the ΔpSymAB strain being unable to transport

386 many amino acids, including proline and histidine [49]. Similarly, glycolysis appeared
387 specifically essential in the ΔpSymAB strain in rich medium (Figure 6), likely as the reduced
388 metabolic capacity of this strain [29] led to a greater reliance on catabolism of the abundant
389 sucrose for energy and biosynthetic precursors. Specific gene essentiality in the ΔpSymAB
390 background may also occur as a result of synthetic negative interactions that are not associated
391 with metabolic redundancy, for example, synthetic effects of disrupting two independent aspects
392 of the cell envelope. This may be reflected in the specific essentiality of the *feuNPQ* and *ndvAB*
393 genes involved in production of periplasmic cyclic β-glucans (Figure 6) [50-53]. The cell
394 envelope of the ΔpSymAB strain is altered compared to the wild-type, due to the loss of
395 succinoglycan production [54] and the *bacA* gene [55], and the membrane lipid composition
396 contains signs of increased stress [49]. The fitness of disrupting periplasmic cyclic β-glucans
397 biosynthesis in this background, further altering the cell envelope, may therefore represent a
398 synthetic negative interaction.

399 Somewhat surprisingly, approximately a quarter of the genes with a genomic
400 environment effect had a greater fitness defect in the wild type strain. In some cases this may
401 have been due to the reduced nutrient demand of the ΔpSymAB strain as a result of the smaller
402 genome content. For example, mutations of genes for arginine biosynthesis and the biosynthesis
403 of AICAR and UMP, common precursors in the synthesis of purines and pyrimidines,
404 respectively, had fitness defects in rich medium specifically in the wild-type (Figure 6). This
405 may reflect that in this environment, the uptake of these nutrients is growth limiting to the wild-
406 type in the absence of their *de novo* synthesis, whereas this is not the case in the ΔpSymAB
407 strain due to the reduced genome size, and thus lower nutrient requirement, and the already
408 reduced growth rate (Figure S8). Another possibility is that removal of pSymAB evokes

409 phenotypes that are epistatic to many of those brought about by chromosomal mutations. For
410 example, the removal of pSymB is expected to have resulted in alterations of the cell membrane
411 [49,54,55]; our observation that many mutations causing greater relative fitness defects in wild-
412 type cells are associated with lipid metabolism, such as biosynthesis of the lipopolysaccharide
413 core oligosaccharide (Figure 6) may be a result of those mutations being phenotypically masked
414 in the absence of pSymB.

415 Our work in integrating the Tn-seq data with *in silico* metabolic modelling made it
416 evident that Tn-seq alone is insufficient to identify the entire core metabolism of an organism;
417 almost a third of the reactions present in the core metabolic reconstruction were not supported by
418 Tn-seq data (Figure 5 and Table 3). Similarly, the large number of changes made in the gene-
419 reaction relationships when producing the core model illustrated the limitations in the quality of
420 metabolic reconstructions when high-throughput mutagenesis data are lacking. In some cases,
421 the gaps in the Tn-seq data were due to genomic environment effects, such as genetic
422 redundancy, in other cases it was due to the inclusion of reactions that are non-essential but that
423 are nonetheless required for production of 'wild type' cells, and sometimes the gene associated
424 with a reaction is simply unknown. A fourth possibility is phenotypic complementation through
425 cross-feeding. Given that Tn-seq involves growth of a population of mutants, a mutant unable to
426 produce an essential metabolite may still grow if the metabolite is excreted and transferred to the
427 mutant from the rest of the population.

428 Regardless of the reasons why Tn-seq may have missed so many central metabolic
429 reactions, this limitation can have a significant practical impact in the modern era of synthetic
430 biology. The results of Tn-seq studies may be used to guide engineering of designer microbial
431 factories with specific properties [56], or for the identification of putative new therapeutic targets

432 [25,57]. While Tn-seq studies undoubtedly give invaluable information to be used towards these
433 goals, basing engineered cells solely on Tn-seq studies is insufficient, as evidenced in the recent
434 monumental efforts to design and synthesize a minimal bacterial genome [22] . Importantly, this
435 limitation can be overcome by combining Tn-seq with metabolic modelling. We are aware of
436 only a few other studies making use of both Tn-seq data and metabolic reconstruction [58-62];
437 however, these studies almost always focus on using the Tn-seq data to refine the metabolic
438 reconstruction. As illustrated here, combining an experimental Tn-seq approach with a ground-
439 up *in silico* metabolic reconstruction strategy can improve not only the reconstruction but also
440 overcome the limitations of the Tn-seq approach. A Tn-seq-guided reconstruction process forces
441 the identification of missing essential reactions, while ensuring correct gene-reaction
442 associations, and the integrated approach can facilitate functional annotation of genes without
443 clear biological roles. This process allows one to obtain a very high-quality representation of the
444 metabolism, and the underlying genetics, of the organism in the given environment. The
445 resulting model can serve as a blueprint to simply understand the workings of the cell, or as a
446 basis for developing new cell factories.
447

448 **MATERIALS AND METHODS**

449 **Bacterial strains, media, and growth conditions.**

450 The wild type and ΔpSymAB strains used throughout this work are the RmP3499 and
451 RmP3496 strains, respectively, whose construction was described previously [30]. All *E. coli* or
452 *S. meliloti* strains used in this study are described in Table S5 and were grown at 37°C or 30°C,
453 respectively. BRM medium was used as the rich medium for growth of the *S. meliloti* strains,
454 and it consisted of 5 g/L Bacto Tryptone, 5 g/L Bacto Yeast Extract, 50 mM NaCl, 2 mM

455 $MgSO_4$, 2 μM $CoCl_2$, 0.5% (w/v) sucrose, and supplemented with the following antibiotics, as
456 appropriate: streptomycin (Sm, 200 μg/ml), neomycin (Nm, 100 μg/ml), gentamycin (Gm, 15
457 μg/ml). The defined medium for growth of *S. meliloti* contained 50 mM NaCl, 10 mM KH_2PO_4,
458 10 mM NH_4Cl, 2 mM $MgSO_4$, 0.2 mM $CaCl_2$, 0.5% (w/v) sucrose, 2.5 μM thiamine, 2 μM
459 biotin, 10 μM EDTA, 10 μM $FeSO_4$, 3 μM $MnSO_4$, 2 μM $ZnSO_4$, 2 μM H_3BO_3, 1 μM $CoCl_2$,
460 0.2 μM Na_2MoO_4, 0.3 μM $CuSO_4$, 50 μg/ml streptomycin, and 30 μg/ml neomycin. *E. coli*
461 strains were grown on Luria-Bertani (LB) supplemented with the following antibiotics as
462 appropriate: chloramphenicol (30 mg/ml), kanamycin (Km, 30 μg/ml), gentamycin (Gm, 3
463 μg/ml).

464 **Growth curves.**

465 Overnight cultures grown in rich media with the appropriate antibiotics were pelleted,
466 washed with a phosphate buffer (20 mM KH_2PO_4 and 100 mM NaCl), and resuspended to an
467 OD_{600} of 0.25. Twelve μl of each cell suspension was mixed with 288 μl of growth medium,
468 without antibiotics, in wells of a 100-well Honeycomb microplate. Plates were incubated in a
469 Bioscreen C analyzer at 30°C with shaking, and OD_{600} recorded every hour for at least 48 hours.

470 ***S. meliloti* mutant construction for Tn-seq validation.**

471 Single gene knockout mutants were generated through single cross-over plasmid
472 integration of the suicide plasmid pJG194 [63]. Approximately 400-bp fragments homologous to
473 the central portion of the target genes were PCR amplified using the primers listed in Table S6.
474 PCR products as well as the pJG194 and pJG796 vectors were digested with the restriction
475 enzymes *Eco*RI/*Hind*III, *Bam*HI/*Xba*I, or *Sal*I/*Xho*I, and each PCR fragment was ligated into the
476 appropriately digested pJG194 or pJG796 vector using standard molecular biology techniques
477 [64], and all recombinant plasmids verified. Recombinant plasmids were mobilized from *E. coli*

478 to *S. meliloti* via tri-parental matings as described before [52], and transconjugants isolated on
479 BRM Sm Nm agar plates. All *S. meliloti* mutants were verified by PCR.

480 Transduction of the integrated plasmids into the *S. meliloti* wild type and ΔpSymAB
481 strains was performed using phage N3 as described elsewhere [65], with transductants recovered
482 on BRM medium containing the appropriate antibiotics.

483 **Construction of the transposon delivery vector pJG714**

484 The plasmid pJG714 is a variant of the previously reported mini-Tn*5* delivery plasmid,
485 pJG110 [63], with the primary modifications being removal of the *bla* gene and pUC origin of
486 replication, and introduction of the *pir*-dependent R6K replication origin. A map of pJG714 is
487 given in Figure S1A, and the complete sequence of the transposable region is provided in Figure
488 S1B. This delivery plasmid is maintained in *E. coli* strain MFD*pir* [66], which possesses
489 chromosomal copies of R6K *pir* and RK2 transfer functions. MFD*pir* is unable to synthesize
490 diaminopimelic acid (DAP), thus disabling growth on rich or defined medium lacking
491 supplemental DAP. The MFD*pir*/pJG714 strain is cultured on rich medium containing
492 kanamycin and 12.5 μg/ml DAP.

493 **Tn-seq experimental setup.**

494 Transposon mutagenesis was accomplished in the wild-type and ΔpSymAB strains in
495 parallel. Flask cultures of MFD*pir*/pJG714 and the two *S. meliloti* strains were grown overnight
496 to saturation, and pellets were washed and suspended in BRM to a final OD_{600} value of
497 approximately 40. Equal volumes of each suspension were mixed as bi-parental matings, to
498 accomplish mobilization of the transposon delivery vector into the *S. meliloti* recipient strains.
499 These cell mixtures were plated on BRM supplemented with 50 μg/ml DAP and incubated at
500 30°C for 6 h. Mating mixtures were collected in BRM with 10% glycerol, and cell clumps were

501 broken up by shaking the suspended material for 30 min at 225 rpm. Aliquots were stored at -
502 80°C. For selection of transposants, mating mixes were thawed and plated at a density of 15,000
503 cfu/plate (150-mm plates) on BRM supplemented with Sm and Nm. To accomplish equivalent
504 coverage of each genome with transposon insertions, 675,000 and 360,000 colonies were
505 selected for the wild-type and ΔpSymAB strains, respectively. For each recipient, transposon
506 mutant colonies were collected and cell clumps were broken up as described above. The selected
507 clone libraries were aliquoted and stored at -80°C.

508 For whole-population selection and massively parallel sequencing of transposon ends,
509 $1x10^9$ cells from each of the two clone libraries were transferred into 500 ml of either BRM or
510 defined medium, allowing approximately 8-10 generations of growth at 30°C before reaching
511 saturation. At this stage, cells were pelleted, DNA was extracted using the MoBio microbial
512 DNA isolation kit (#12255-50), and the resulting DNA was fragmented with NEB fragmentase
513 (#M0348S) to an average molecular weight of 1000 bp. After clean-up (Qiagen #27106), the
514 resulting DNA fragments were appended with short 3' homopolymer (oligo-dCTP) tails using
515 terminal deoxynucleotidyl transferase (NEB #M0315S), and this sample was used as the
516 template for a two-round PCR that gave rise to the final Illumina-ready libraries. In the first
517 round, a transposon end-specific primer (1TN) and oligo-G primer (1GG) were used (all primer
518 sequences can be found in Table S6). After clean-up, a portion of the first-round product was
519 used as the template for the second-round reaction employing a nested transposon-specific
520 primer (2TNA-C) and a reverse index-incorporating primer (2BAR01-08). The series of three
521 2TN primers (A-C) were designed to incorporate base diversity in the opening cycles of Illumina
522 sequencing, and the series of eight 2BAR primers were designed to uniquely identify each
523 experimental condition in a single multiplexed sequencing sample. After PCR amplification of

524 transposon-flanking sequences with concomitant incorporation of Illumina adapters and
525 barcodes, the samples were size-selected for 200-600-bp fragments, and sequenced on an
526 Illumina Hi-Seq instrument as 50-bp single-end reads. Raw reads were used as input into a
527 custom-built Tn-seq analytical pipeline, which was recently described [57].

528 **Calculation of gene and synthetic indexes.**

529 For calculation of Gene Essentiality Index (GEI) scores, a pseudo count of one was first
530 added to all gene read counts for each replicate. GEI were then calculated by summing the
531 number of reads that mapped to the gene in both replicates, and dividing this number by the
532 nucleotide length of the gene. GEI scores were calculated for each gene separately in each
533 medium and in each strain. All GEI values are available in Data Set S1.

534 Synthetic Media Index (SMI) scores were calculated to represent the difference in GEI
535 scores between the two media for the same strain. Raw SMI scores were determined by dividing
536 the GEI of the gene in defined medium by the GEI of the gene in rich medium. Processed SMI
537 scores, those shown throughout the manuscript, were determined as follows. If the raw value was
538 above one, the processed SMI and the raw SMI are the same. Raw SMI scores that were below
539 one were converted to processed SMI scores through the transformation, "1 / raw SMI score",
540 and presenting the value as a negative number.

541 Raw and processed Synthetic Rich Index (SRI) and Synthetic Defined Index (SDI) scores
542 were calculated to represent the difference in the GEI scores of a gene between the wild-type and
543 ΔpSymAB strains when grown in rich or defined medium, respectively. SRI and SDI indexes
544 were calculated using the same procedure as described for the SMI scores above. All synthetic
545 index scores are provided in Data Set S1.

546

547 **Statistical analysis of the Tn-seq output.**

548 The output of the Tn-seq analysis pipeline was used in the fitness classification of genes
549 as follows. First, all genes with no observed insertions were classified as essential. Next, GEI
550 scores were imported into R version 3.2.3 and log transformed. Initial clustering of the log
551 transformed GEI scores into fitness categories was performed using the *Mclust* function of the
552 *Mclust* package in R [67]. In short, this function attempts to explain the distribution of GEI
553 values by fitting a series of overlapping Guassian distributions, with the number and shape of the
554 distributions determined by *Mclust*. The data are then assigned to different categories based on
555 the probability of the data point arising from each of the distributions. As high uncertainty in the
556 classification of genes at the borders of groupings exists, the clusters were refined through the
557 use of affinity propagation implemented by the *apcluster* function of the *apcluster* package of R
558 [68]. All genes belonging to an *apcluster* grouping that contained an essential gene, as
559 determined in any of the previous steps, were re-annotated as essential. Additionally, all genes
560 belonging to an *apcluster* grouping that spanned the border of two *Mclust* goups were transferred
561 to the same classification, based on which cluster the genes had a higher median probability of
562 being derived from in the *Mclust* analysis. Finally, genes that were classified as 'essential' in one
563 medium and 'large growth impairment' in the second medium, but that were identified as having
564 no medium specificity based on their SMI scores, were considered as essential in both media.

565 Genes with GEI scores significantly different between conditions were determined as
566 follows. The synthetic indexes (SMI, SDI, SRI) scores were imported into R and log
567 transformed, and the following clustering performed independently for each index. The log
568 transformed synthetic scores were clustered using *Mclust* and *apcluster* in R as described above
569 for the GEI scores. In the case of the SMI scores, three clusters were produced 'Little to no

570 difference', 'Moderate difference', and 'Large difference'; only genes with a SMI scores
571 classified as 'Large difference' were considered to display a medium specificity. In the case of
572 SDI and SRI scores, only two clusters were produced: 'Little to no difference' and 'Difference
573 between strains'.

574 **Gene functional enrichments.**

575 Assignment of chromosomal genes into specific functional categories was performed
576 largely based on the annotations provided on the *S. meliloti* Rm1021 online genome database
577 (https://iant.toulouse.inra.fr/bacteria/annotation/cgi/rhime.cgi). This website pulls annotations
578 from several databases including PubMed, Swissprot, trEMBL, and Interpro. Additionally, it
579 provides enzyme codes, PubMed IDs, functional classifications, and suggested Gene Ontology
580 (GO) terms for most genes. The numerous classifications were simplified to 18 functional
581 categories, designed to adequately cover all core cellular processes. Occasionally, ambiguous or
582 conflicting annotations were observed. In these cases, protein BLASTp searches through the
583 NCBI server were performed against the non-redundant protein database. If putative domains
584 were detected within the amino acid sequence, a combination of the best hit (lowest E-value) and
585 consensus among domain annotations were used to categorize the gene in question. If no putative
586 domains were detected, the functional annotation was based on the best scoring protein hits in
587 the database. The functional annotations of all chromosomal genes are provided in Data Set S5.

588 **Data visualization.**

589 Tn-seq results were visualized using the Integrative Genomics Viewer v2.3.97 [69].
590 Scatter plots, functional enrichment plots, box plots, and line plots were generated in R using the
591 *ggplot2* package [70]. Venn diagrams were produced in R using the *VennDiagram* package [71].
592 The genome map was prepared using the circos v0.67-7 software [72]; the sliding window

593 insertion density was calculated with the *geom_histogram* function of *ggplot2*, and the GC skew
594 was calculated using the analysis of sequence heterogeneity sliding window plots online
595 webserver [73]. The metabolic model was visualized using the iPath v2.0 webserver [74]. The
596 logo of the transposon insertion site specificity was generated by first extracting the nucleotides
597 surrounding all unique insertion sites in one replicate of the wild-type grown in rich medium
598 using Perl v5.18.2, followed by generation of a hidden Markov model with the *hmmbuild*
599 function of HMMER v3.1b2 [75] and visualization with the Skylign webserver [76].

600 **Blast Bidirectional Best Hit (Blast-BBH) strategy.**

601 Putative orthologous proteins between species were identified with a Blast-BBH
602 approach, implemented using a modified version of our in-house Shell and Perl pipeline [77].
603 This pipeline involved GNU bash v4.3.48(1), Perl v5.22.1, Python v2.7.12 and the Blast v2.6.0+
604 software [78]. Proteomes were downloaded from the National Center for Biotechnology
605 Information repository, and the Genbank annotations were used. As a threshold to limit false
606 positives, Blast-BBH pairs were only maintained if they displayed a minimum of 30% amino
607 acid identify over at least 60% of the protein. To identify putative duplicate proteins in *S.*
608 *meliloti*, the same Blast-BBH approach was employed to compare the *S. meliloti* chromosomal
609 proteome with the proteins encoded by pSymA and pSymB.

610 ***In silico* metabolic modeling procedures.**

611 All simulations were performed in Matlab 2017a (Mathworks) with scripts from the
612 Cobra Toolbox (downloaded May 12, 2017 from the openCOBRA repository) [79], and using
613 the Gurobi 7.0.2 solver (www.gurobi.com), the SBMLToolbox 4.1.0 [80], and libSBML 5.15.0
614 [81]. Boundary conditions for simulation of the defined medium are given in Table S7. *In silico*
615 analysis of redundancy in the *S. meliloti* genome was performed using the iGD1575b metabolic

616 reconstruction, whose development is described in the following section. Single and double gene
617 deletion analyses were performed using the *singleGeneDeletion* and *doubleGeneDeletion*
618 functions, respectively, using the Minimization of Metabolic Adjustment (MOMA) method. All
619 Matlab scripts used in this work are provided as File S3.

620 For all deletion mutants, the growth rate ratio (grRatio) was calculated (growth rate of
621 mutant / growth rate of wild type). Single gene deletion mutants were considered to have a
622 growth defect if the grRatio was < 0.9. For the double gene deletion analysis, if the grRatio of
623 the double mutant was less than 0.9 the expected grRatio (based on multiplying the grRatio of
624 the two corresponding single mutants), the double deletion was said to have a synthetic negative
625 phenotype.

626 **Development of iGD1575b.**

627 For *in silico* analysis of redundancy in the *S. meliloti* genome, the previously published *S.*
628 *meliloti* genome-scale metabolic model iGD1575 [34] was modified slightly. As indicated in
629 Table S8, the biomass composition was updated to include 31 additional compounds at trace
630 concentrations, including vitamins, coenzymes, and ions, in order to ensure the corresponding
631 transport or biosynthetic pathways were essential. However, the original model iGD1575 was
632 unable to produce vitamin B12 and holo-carboxylate. To rectify this, the reversibility of
633 rxn00792_c0 was changed from 'false' to 'true', and the reactions rxn01609, rxn06864, and
634 rxnBluB were added to the model. However, no new genes were included in the model. This
635 updated model was termed iGD1575b and is available in SBML and Matlab format in File S2.

636 **Simulating the removal of pSymA and pSymB *in silico*.**

637 Several modifications to iGD1575b were required in order to produce a viable model
638 following the deletion of all pSymA and pSymB genes. As described previously [34],

639 succinoglycan was removed from the biomass composition, 'gapfill' GPRs (gene-protein-
640 reaction relationships) were added to the reactions 'rxn01675_c0', 'rxn01997_c0',
641 'rxn02000_c0', and 'rxn02003_c0' in order to allow the continued production of the full LPS
642 molecule, as well as to 'rxn00416_c0' to allow asparagine biosynthesis. Additionally, 'gapfill'
643 GPRs were added to the reactions 'rxn03975_c0' and 'rxn03393_c0' so that removal of pSymA
644 and pSymB did not prevent biosynthesis of vitamin B12 and ubiquinone-8, respectively. Finally,
645 a glycerol export reaction via diffusion (rxnBLTPcpd00100b) was added to remove the glycerol
646 build-up resulting from cardiolipin biosynthesis. The modified version of the model was termed
647 iGD1575c, and is available in in SBML and Matlab format in File S2. For simulating the
648 removal of pSymA and pSymB in Matlab, all pSymA and pSymB genes were deleted from the
649 iGD1575b model using the *deleteModelGenes* function, followed by the removal of all
650 constrained reactions using the *removeRxns* function.

651 **Building the draft *R. leguminosarum* metabolic model.**

652 A draft, fully automated model containing no manual curation for *R. leguminosarum* bv.
653 *viciae* 3841 was built using the KBase webserver (www.kbase.us). The Genbank file
654 (GCA_000009265.1_ASM926v1_genomic.gbff) of the *R. leguminosarum* genome [82] was
655 uploaded to KBase and re-annotated using the 'annotate microbial genome' function,
656 maintaining the original locus tags. An automated metabolic model was then built using the
657 'build metabolic model' function, with gap-filling. This model included 1537 genes, 1647
658 reaction, and 1731 metabolites, and is available in in SBML and Matlab format in File S2. The
659 biomass composition was not modified from the default Gram negative biomass of Kbase. All
660 essential model genes were determined using the Cobra Toolbox in Matlab with the

661 *singleGeneDeletion* function and the MOMA protocol, with exchange reaction bounds set as
662 provided in Table S7.

663 **Building the *S. meliloti* core metabolic reconstruction, iGD726.**

664 The iGD726 model was built from the ground-up using the existing iGD1575 model as a
665 reaction and GPR database, and with the Tn-seq data as a guide. Each metabolic pathway
666 included in iGD726 was rebuilt in a new file by adding individual reactions to the file. These
667 reactions were taken from iGD1575, or were taken from other sources, primarily the Kyoto
668 Encyclopedia of Genes and Genomes (KEGG) database [83], if an appropriate reaction was
669 missing in iGD1575. Following the transfer of each reaction, the genes associated with the
670 reaction were checked against the Tn-seq data, and a literature search for each associated gene
671 was performed. The gene associations were then modified as necessary to ensure the model
672 accurately captured the experimental data. For example, if gene was experimentally determined
673 to be essential, but the corresponding reaction for the gene was associated with multiple
674 alternative genes, all but the essential gene were removed from the reaction. Similarly, if a non-
675 essential gene was associated with an essential reaction, a second gene or an Unknown was
676 added to reflect the apparent redundancy in the genome. Where possible, unknowns in the gene
677 associations were replaced with genes whose gene product may catalyze the reaction.

678 During the construction of the core model, the biomass composition was updated. This
679 included modifying the membrane lipid composition to include lipids with different sized fatty
680 acids based on the ratio experimentally determined [84]; the original iGD1575 model contained
681 only one representative per each membrane lipid class. Additionally, essential vitamins,
682 cofactors, and ions were added to the biomass composition at trace concentrations to ensure that

683 their biosynthesis or transport was essential. The complete biomass composition is provided in
684 Table S3.

685 The necessary metabolic and transport reactions to allow the model to growth with
686 sucrose, glucose, or succinate were included in the reconstruction. Once the model was capable
687 of producing all biomass components using any of the three carbon sources, the list of model
688 genes was compared with the list of 489 core growth promoting genes to identify genes not
689 included in the model but experimentally determined to contribute to growth. When possible,
690 missing genes and their corresponding reactions were added to the core model. The final model
691 contained 726 genes, 681 reactions, and 703 metabolites, and is provided in SBML and Matlab
692 format in File S2, and as an Excel file in Data Set S4. The Excel file contains all necessary
693 information for use as a *S. meliloti* metabolic resource, including the reaction name, the reaction
694 equation using the real metabolite names, the associated genes/proteins, and references.
695 Additionally, for each reaction, the putative orthologs of the associated genes in 10 related
696 Rhizobiales species are included, allowing the model to provide useful information for each of
697 these organisms.

698

699 **FUNDING INFORMATION**

700 This work was funded by National Science Foundation grant IOS-1054980 to Joel S
701 Griffitts. George C diCenzo was supported by the Natural Sciences and Engineering Research
702 Council of Canada (NSERC) through a PDF fellowship. Work in the laboratory of Turlough M
703 Finan was funded by NSERC. The funders had no role in the study design, data collection and
704 interpretation, or the decision to submit the work for publication.

705

706 **REFERENCES**

707 1. Orgogozo V, Morizot B, Martin A. The differential view of genotype-phenotype
708 relationships. Front Genet. 2015;6: 179. doi:10.3389/fgene.2015.00179
709 2. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to
710 uncover genotype-phenotype interactions. Nat Rev Genet. 2015;16: 85–97.
711 doi:10.1038/nrg3868
712 3. Palsson B. Metabolic systems biology. FEBS Lett. 2009;583: 3900–3904.
713 doi:10.1016/j.febslet.2009.09.031
714 4. Durot M, Bourguignon P-Y, Schachter V. Genome-scale models of bacterial metabolism:
715 reconstruction and applications. FEMS Microbiol Rev. 2009;33: 164–190.
716 doi:10.1111/j.1574-6976.2008.00146.x
717 5. O'Brien EJ, Monk JM, Palsson BØ. Using genome-scale models to predict biological
718 capabilities. Cell. 2015;161: 971–987.
719 6. van Opijnen T, Bodi KL, Camilli A. Tn-seq: high-throughput parallel sequencing for
720 fitness and genetic interaction studies in microorganisms. Nat Methods. 2009;6: 767–772.
721 doi:10.1038/nmeth.1377
722 7. van Opijnen T, Camilli A. Transposon insertion sequencing: a new tool for systems-level
723 analysis of microorganisms. Nature Rev Microbiol. 2013;11: 435–442.
724 doi:10.1038/nrmicro3033
725 8. Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic
726 reconstruction. Nat Protoc. 2010;5: 93–121. doi:10.1038/nprot.2009.203
727 9. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. 2010;28:
728 245–248. doi:10.1038/nbt.1614

729 10. Fowler ZL, Gikandi WW, Koffas MAG. Increased malonyl coenzyme A biosynthesis by
730 tuning the *Escherichia coli* metabolic network and its application to flavanone production.
731 Appl Environ Microbiol. 2009;75: 5831–5839. doi:10.1128/AEM.00270-09
732 11. Pratapa A, Balachandran S, Raman K. Fast-SL: an efficient algorithm to identify synthetic
733 lethal sets in metabolic networks. Bioinformatics. 2015;31: 3299–3305.
734 doi:10.1093/bioinformatics/btv352
735 12. Chao MC, Abel S, Davis BM, Waldor MK. The design and analysis of transposon
736 insertion sequencing experiments. Nature Rev Microbiol. 2016;14: 119–128.
737 doi:10.1038/nrmicro.2015.7
738 13. Thomaides HB, Davison EJ, Burston L, Johnson H, Brown DR, Hunt AC, et al. Essential
739 bacterial functions encoded by gene pairs. J Bacteriol. 2007;189: 591–602.
740 doi:10.1128/JB.01381-06
741 14. diCenzo GC, Finan TM. Genetic redundancy is prevalent within the 6.7 Mb *Sinorhizobium*
742 *meliloti* genome. Mol Genet Genomics. 2015;290: 1345–1356. doi:10.1007/s00438-015-
743 0998-6
744 15. Bergmiller T, Ackermann M, Silander OK. Patterns of evolutionary conservation of
745 essential genes correlate with their compensability. PLOS Genet. 2012;8: e1002803.
746 doi:10.1371/journal.pgen.1002803
747 16. Liu G, Yong MYJ, Yurieva M, Srinivasan KG, Liu J, Lim JSY, et al. Gene essentiality is a
748 quantitative property linked to cellular evolvability. Cell. 2015;163: 1388–1399.
749 17. Butland G, Babu M, Díaz-Mejía JJ, Bohdana F, Phanse S, Gold B, et al. eSGA: *E. coli*
750 synthetic genetic array analysis. Nat Methods. 2008;5: 789–795. doi:10.1038/nmeth.1239
751 18. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, et al. The genetic

752 landscape of a cell. Science. 2010;327: 425–431. doi:10.1126/science.1180823
753 19. Phillips PC. Epistasis—the essential role of gene interactions in the structure and evolution
754 of genetic systems. Nat Rev Genet. 2008;9: 855–867. doi:10.1038/nrg2452
755 20. Juhas M. On the road to synthetic life: the minimal cell and genome-scale engineering. Crit
756 Rev Biotechnol. 2016;36: 416–423. doi:10.3109/07388551.2014.989423
757 21. Juhas M, Reuß DR, Zhu B, Commichau FM. *Bacillus subtilis* and *Escherichia coli*
758 essential genes and minimal cell factories after one decade of genome engineering.
759 Microbiology. 2014;160: 2341–2351. doi:10.1099/mic.0.079376-0
760 22. Hutchison CA, Chuang R-Y, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, et
761 al. Design and synthesis of a minimal bacterial genome. Science. 2016;351: aad6253–
762 aad6253. doi:10.1126/science.aad6253
763 23. Curtis PD, Brun YV. Identification of essential alphaproteobacterial genes reveals
764 operational variability in conserved developmental and cell cycle systems. Mol Microbiol.
765 2014;93: 713–735. doi:10.1111/mmi.12686
766 24. Pechter KB, Gallagher L, Pyles H, Manoil CS, Harwood CS. Essential genome of the
767 metabolically yersatile alphaproteobacterium *Rhodopseudomonas palustris*. J Bacteriol.
768 2016;198: 867–876. doi:10.1128/JB.00771-15
769 25. Lee SA, Gallagher LA, Thongdee M, Staudinger BJ, Lippman S, Singh PK, et al. General
770 and condition-specific essential functions of *Pseudomonas aeruginosa*. Proc Natl Acad Sci
771 USA. 2015;112: 5189–5194. doi:10.1073/pnas.1422186112
772 26. Skurnik D, Roux D, Aschard H, Cattoir V, Yoder-Himes D, Lory S, et al. A
773 comprehensive analysis of *in vitro* and *in vivo* genetic fitness of *Pseudomonas aeruginosa*
774 using high-throughput sequencing of transposon libraries. PLOS Pathog. 2013;9:

775 e1003582. doi:10.1371/journal.ppat.1003582

776 27. Freed NE, Bumann D, Silander OK. Combining *Shigella* Tn-seq data with gold-standard *E.*
777 *coli* gene deletion data suggests rare transitions between essential and non-essential gene
778 functionality. BMC Microbiol. 2016;16: 203. doi:10.1186/s12866-016-0818-0

779 28. Canals R, Xia X-Q, Fronick C, Clifton SW, Ahmer BM, Andrews-Polymenis HL, et al.
780 High-throughput comparison of gene fitness among related bacteria. BMC Genomics.
781 2012;13: 212. doi:10.1186/1471-2164-13-212

782 29. diCenzo GC, MacLean AM, Milunovic B, Golding GB, Finan TM. Examination of
783 prokaryotic multipartite genome evolution through experimental genome reduction. PLOS
784 Genet. 2014;10: e1004742. doi:10.1371/journal.pgen.1004742

785 30. diCenzo GC, Zamani M, Milunovic B, Finan TM. Genomic resources for identification of
786 the minimal N_2-fixing symbiotic genome. Environ Microbiol. 2016;18: 2534–2547.
787 doi:10.1111/1462-2920.13221

788 31. Perry BJ, Yost CK. Construction of a *mariner*-based transposon vector for use in insertion
789 sequence mutagenesis in selected members of the *Rhizobiaceae*. BMC Microbiol. 2014;14:
790 298. doi:10.1186/s12866-014-0298-z

791 32. Perry BJ, Akter MS, Yost CK. The use of transposon insertion sequencing to interrogate
792 the core functional genome of the legume symbiont *Rhizobium leguminosarum*. Front
793 Microbiol. 2016;7: 1873. doi:10.3389/fmicb.2016.01873

794 33. Wheatley RM, Ramachandran VK, Geddes BA, Perry BJ, Yost CK, Poole PS. The role of
795 O2 in the growth of *Rhizobium leguminosarum* bv. *viciae* 3841 on glucose and succinate. J
796 Bacteriol. 2016;199: e00572–16. doi:10.1128/JB.00572-16

797 34. diCenzo GC, Checcucci A, Bazzicalupo M, Mengoni A, Viti C, Dziewit L, et al. Metabolic

798 modelling reveals the specialization of secondary replicons for niche adaptation in
799 *Sinorhizobium meliloti*. Nat Commun. 2016;7: 12219. doi:10.1038/ncomms12219
800 35. Schurr MJ, Vickrey JF, Kumar AP, Campbell AL, Cunin R, Benjamin RC, et al. Aspartate
801 transcarbamoylase genes of *Pseudomonas putida*: requirement for an inactive
802 dihydroorotase for assembly into the dodecameric holoenzyme. J Bacteriol. 1995;177:
803 1751–1759.
804 36. Labedan B, Xu Y, Naumoff DG, Glansdorff N. Using quaternary structures to assess the
805 evolutionary history of proteins: the case of the aspartate carbamoyltransferase. Mol Biol
806 Evol. 2004;21: 364–373. doi:10.1093/molbev/msh024
807 37. Dunn MF. Key roles of microsymbiont amino acid metabolism in rhizobia-legume
808 interactions. Critical Reviews in Microbiology. 2015;41: 411–451.
809 doi:10.3109/1040841X.2013.856854
810 38. De Bruijn FJ, Lupski JR. The use of transposon Tn*5* mutagenesis in the rapid generation of
811 correlated physical and genetic maps of DNA segments cloned into multicopy plasmids--a
812 review. Gene. 1984;27: 131–149.
813 39. Berg DE, Schmandt MA, Lowe JB. Specificity of transposon Tn*5* insertion. Genetics.
814 1983;105: 813–828.
815 40. Goryshin IY, Miller JA, Kil YV, Lanzov VA, Reznikoff WS. Tn*5*/IS*50* target recognition.
816 Proc Natl Acad Sci USA. 1998;95: 10716–10721.
817 41. Green B, Bouchier C, Fairhead C, Craig NL, Cormack BP. Insertion site preference of Mu,
818 Tn*5*, and Tn*7* transposons. Mob DNA. 2012;3: 3. doi:10.1186/1759-8753-3-3
819 42. Harrison PW, Lower RPJ, Kim NKD, Young JPW. Introducing the bacterial "chromid":
820 not a chromosome, not a plasmid. Trends Microbiol. 2010;18: 141–148.

821 doi:10.1016/j.tim.2009.12.010

822 43. diCenzo GC, Finan TM. The divided bacterial genome: structure, function, and evolution.
823 Microbiol Mol Biol Rev. 2017;81: e00019–17. doi:10.1128/MMBR.00019-17

824 44. Galardini M, Brilli M, Spini G, Rossi M, Roncaglia B, Bani A, et al. Evolution of intra-
825 specific regulatory networks in a multipartite bacterial genome. PLOS Comput Biol.
826 2015;11: e1004478. doi:10.1371/journal.pcbi.1004478

827 45. diCenzo G, Milunovic B, Cheng J, Finan TM. The $tRNA^{arg}$ gene and *engA* are essential
828 genes on the 1.7-mb pSymB megaplasmid of *Sinorhizobium meliloti* and were translocated
829 together from the chromosome in an ancestral strain. J Bacteriol. 2013;195: 202–212.
830 doi:10.1128/JB.01758-12

831 46. Oresnik IJ, Liu SL, Yost CK, Hynes MF. Megaplasmid pRme2011a of *Sinorhizobium*
832 *meliloti* is not required for viability. J Bacteriol. 2000;182: 3582–3586.

833 47. Koo B-M, Kritikos G, Farelli JD, Todor H, Tong K, Kimsey H, et al. Construction and
834 analysis of two genome-scale deletion libraries for *Bacillus subtilis*. Cell Systems. 2017;4:
835 291–305.e7.

836 48. Armbruster CE, Forsyth-DeOrnellas V, Johnson AO, Smith SN, Zhao L, Wu W, et al.
837 Genome-wide transposon mutagenesis of *Proteus mirabilis*: Essential genes, fitness factors
838 for catheter-associated urinary tract infection, and the impact of polymicrobial infection on
839 fitness requirements. PLOS Pathog. 2017;13: e1006434. doi:10.1371/journal.ppat.1006434

840 49. Fei F, diCenzo GC, Bowdish DME, McCarry BE, Finan TM. Effects of synthetic large-
841 scale genome reduction on metabolism and metabolic preferences in a nutritionally
842 complex environment. Metabolomics. 2016;12: 23. doi:10.1007/s11306-015-0928-y

843 50. Stanfield SW, Ielpi L, O'brochta D, Helinski DR, Ditta GS. The *ndvA* gene product of

844 *Rhizobium meliloti* is required for beta-(1----2)glucan production and has homology to the
845 ATP-binding export protein HlyB. J Bacteriol. 1988;170: 3523–3530.
846 doi:10.1128/jb.170.8.3523-3530.1988
847 51. Ielpi L, Dylan T, Ditta GS, Helinski DR, Stanfield SW. The *ndvB* locus of *Rhizobium*
848 *meliloti* encodes a 319-kDa protein involved in the production of beta-(1----2)-glucan. J
849 Biol Chem. 1990;265: 2843–2851.
850 52. Griffitts JS, Carlyon RE, Erickson JH, Moulton JL, Barnett MJ, Toman CJ, et al. A
851 *Sinorhizobium meliloti* osmosensory two-component system required for cyclic glucan
852 export and symbiosis. Mol Microbiol. 2008;69: 479–490.
853 53. Carlyon RE, Ryther JL, VanYperen RD, Griffitts JS. FeuN, a novel modulator of two-
854 component signalling identified in *Sinorhizobium meliloti*. Mol Microbiol. 2010;77: 170–
855 182. doi:10.1111/j.1365-2958.2010.07198.x
856 54. Finan TM, Kunkel B, De Vos GF, Signer ER. Second symbiotic megaplasmid in
857 *Rhizobium meliloti* carrying exopolysaccharide and thiamine synthesis genes. J Bacteriol.
858 1986;167: 66–72.
859 55. Ferguson GP, Roop RM, Walker GC. Deficiency of a *Sinorhizobium meliloti bacA* mutant
860 in alfalfa symbiosis correlates with alteration of the cell envelope. J Bacteriol. 2002;184:
861 5625–5632. doi:10.1128/JB.184.20.5625-5632.2002
862 56. Chan CH, Levar CE, Jiménez-Otero F, Bond DR. Genome scale mutational analysis of
863 *Geobacter sulfurreducens* reveals distinct molecular mechanisms for respiration and
864 sensing of poised electrodes vs. Fe(III) oxides. J Bacteriol. 2017;: JB.00340–17.
865 doi:10.1128/JB.00340-17
866 57. Arnold MFF, Shabab M, Penterman J, Boehme KL, Griffitts JS, Walker GC. Genome-

867 wide sensitivity analysis of the microsymbiont *Sinorhizobium meliloti* to symbiotically
868 important, defensin-like host peptides. mBio. 2017;8: e01060–17.
869 doi:10.1128/mBio.01060-17
870 58. Yang H, Krumholz EW, Brutinel ED, Palani NP, Sadowsky MJ, Odlyzko AM, et al.
871 Genome-scale metabolic network validation of *Shewanella oneidensis* using transposon
872 insertion frequency analysis. PLOS Comput Biol. 2014;10: e1003848.
873 doi:10.1371/journal.pcbi.1003848
874 59. Broddrick JT, Rubin BE, Welkie DG, Du N, Mih N, Diamond S, et al. Unique attributes of
875 cyanobacterial metabolism revealed by improved genome-scale metabolic modeling and
876 essential gene analysis. Proc Natl Acad Sci USA. 2016;113: E8344–E8353.
877 doi:10.1073/pnas.1613446113
878 60. Bartell JA, Blazier AS, Yen P, Thøgersen JC, Jelsbak L, Goldberg JB, et al. Reconstruction
879 of the metabolic network of *Pseudomonas aeruginosa* to interrogate virulence factor
880 synthesis. Nat Commun. 2017;8: 14631. doi:10.1038/ncomms14631
881 61. Senior NJ, Sasidharan K, Saint RJ, Scott AE, Sarkar-Tyson M, Ireland PM, et al. An
882 integrated computational-experimental approach reveals *Yersinia pestis* genes essential
883 across a narrow or a broad range of environmental conditions. BMC Microbiol. 2017;17:
884 163. doi:10.1186/s12866-017-1073-8
885 62. Burger BT, Imam S, Scarborough MJ, Noguera DR, Donohue TJ. Combining genome-
886 scale experimental and computational methods to identify essential genes in *Rhodobacter*
887 *sphaeroides*. mSystems. 2017;2: e00015–17. doi:10.1128/mSystems.00015-17
888 63. Griffitts JS, Long SR. A symbiotic mutant of *Sinorhizobium meliloti* reveals a novel
889 genetic pathway involving succinoglycan biosynthetic functions. Mol Microbiol. 2008;67:

890 1292–1306. doi:10.1111/j.1365-2958.2008.06123.x

891 64. Sambrook J, Fritsch EF, Maniatis T. Molecular cloning: a laboratory manual. New York:
892 Cold Spring Harbor Laboratory Press; 1989.

893 65. Martin MO, Long SR. Generalized transduction in *Rhizobium meliloti*. J Bacteriol.
894 1984;159: 125–129.

895 66. Ferrières L, Hémery G, Nham T, Guérout A-M, Mazel D, Beloin C, et al. Silent mischief:
896 bacteriophage Mu insertions contaminate products of *Escherichia coli* random mutagenesis
897 performed using suicidal transposon delivery plasmids mobilized by broad-host-range RP4
898 conjugative machinery. J Bacteriol. 2010;192: 6418–6427. doi:10.1128/JB.00621-10

899 67. Fraley C, Raftery AE, Murphy TB, Scrucca L. mclust Version 4 for R: Normal mixture
900 modeling for model-based clustering, classification, and density estimation. Washington,
901 USA: Department of Statistics, University of Washington; 2012.

902 68. Bodenhofer U, Kothmeier A, Hochreiter S. APCluster: an R package for affinity
903 propagation clustering. Bioinformatics. 2011;27: 2463–2464.
904 doi:10.1093/bioinformatics/btr406

905 69. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al.
906 Integrative genomics viewer. Nat Biotechnol. 2011;29: 24–26. doi:10.1038/nbt.1754

907 70. Wickham H. ggplot2: elegant graphics for data analysis.[Internet]. 2009. New York, USA:
908 Springer-Verlag; 2009.

909 71. Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable
910 Venn and Euler diagrams in R. BMC Bioinform. 2011;12: 35. doi:10.1186/1471-2105-12-
911 35

912 72. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an

913 information aesthetic for comparative genomics. Genome Res. 2009;19: 1639–1645.
914 doi:10.1101/gr.092759.109
915 73. Mrázek J, Karlin S. Strand compositional asymmetry in bacterial and large viral genomes.
916 Proc Natl Acad Sci USA. 1998;95: 3720–3725.
917 74. Yamada T, Letunic I, Okuda S, Kanehisa M, Bork P. iPath2.0: interactive pathway
918 explorer. Nucleic Acids Res. 2011;39: W412–5. doi:10.1093/nar/gkr313
919 75. Eddy SR. A new generation of homology search tools based on probabilistic inference.
920 Genome Inform. 2009;23: 205–211.
921 76. Wheeler TJ, Clements J, Finn RD. Skylign: a tool for creating informative, interactive
922 logos representing sequence alignments and profile hidden Markov models. BMC
923 Bioinform. 2014;15: 1. doi:10.1186/1471-2105-15-7
924 77. diCenzo GC, Zamani M, Ludwig HN, Finan TM. Heterologous complementation reveals a
925 specialized activity for BacA in the *Medicago-Sinorhizobium meliloti* symbiosis. Mol Plant
926 Microbe Interact. 2017;30: 312-324. doi:10.1094/MPMI-02-17-0030-R
927 78. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+:
928 architecture and applications. BMC Bioinform. 2009;10: 421. doi:10.1186/1471-2105-10-
929 421
930 79. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, et al. Quantitative
931 prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0.
932 Nat Protoc. 2011;6: 1290–1307. doi:10.1038/nprot.2011.308
933 80. Keating SM, Bornstein BJ, Finney A, Hucka M. SBMLToolbox: an SBML toolbox for
934 MATLAB users. Bioinformatics. 2006;22: 1275–1277. doi:10.1093/bioinformatics/btl111
935 81. Bornstein BJ, Keating SM, Jouraku A, Hucka M. LibSBML: an API library for SBML.

936 Bioinformatics. 2008;24: 880–881. doi:10.1093/bioinformatics/btn051

937 82. Young JPW, Crossman LC, Johnston AW, Thomson NR, Ghazoui ZF, Hull KH, et al. The
938 genome of *Rhizobium leguminosarum* has recognizable core and accessory components.
939 Genome Biol. 2006;7: R34. doi:10.1186/gb-2006-7-4-r34

940 83. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference
941 resource for gene and protein annotation. Nucleic Acids Res. 2016;44: D457–D462.
942 doi:10.1093/nar/gkv1070

943 84. Basconcillo LS, Zaheer R, Finan TM, McCarry BE. A shotgun lipidomics study of a
944 putative lysophosphatidic acid acyl transferase (PlsC) in *Sinorhizobium meliloti*. J
945 Chromatogr B. 2009;877: 2873–2882. doi:10.1016/j.jchromb.2009.05.014

946 85. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-
947 performance genomics data visualization and exploration. Brief Bioinformatics. 2013;14:
948 178–192. doi:10.1093/bib/bbs017

949 86. Yuan Z-C, Zaheer R, Finan TM. Regulation and properties of PstSCAB, a high-affinity,
950 high-velocity phosphate transport system of *Sinorhizobium meliloti*. J Bacteriol. 2006;188:
951 1089–1102. doi:10.1128/JB.188.3.1089-1102.2006

952 87. diCenzo GC, Sharthiya H, Nanda A, Zamani M, Finan TM. PhoU allows rapid adaptation
953 to high phosphate concentrations by modulating PstSCAB transport rate in *Sinorhizobium*
954 *meliloti*. J Bacteriol. 2017;: JB.00143–17. doi:10.1128/JB.00143-17

955 88. diCenzo GC, Muhammed Z, Østerås M, O'Brien SAP, Finan TM. A key regulator of the
956 glycolytic and gluconeogenic central metabolic pathways in *Sinorhizobium meliloti*.
957 Genetics. 2017;: [epub ahead of print]. doi:10.1534/genetics.117.300212

958

959 **Table 1. Fitness classification of chromosomal genes.** Genes were ranked from lowest to highest GEI, with the lowest GEI being at

960 the 0 percentile and the highest GEI being at the 100th percentile. The approximate break points for the groupings, determined as

961 described in the Materials and Methods, are shown for each condition.

		GEI percentile range			
Group	Description	Wild type, rich medium	ΔpSymAB, rich medium	Wild type, defined medium	ΔpSymAB, defined medium
I	Essential	0-12	0-14	0-12	0-14
II	Strong growth defect	12-17	14-23	12-18	14-24
III	Moderate growth defect	17-36	23-49	18-28	24-47
IV	Little to no growth impact	36-100	49-96	28-99	47-99
V	Growth improvement	N/A	96-100	99-100	99-100

962 **Table 2. Sample genes showing strain specific phenotypes.** The top ten genes from each of the indicated groupings, as determined

963 based on the ratio of GEI scores of the two strains, are shown. GEI (Gene Essentiality Index) scores are shown first for the wild type

964 (WT) followed by the scores for the ΔpSymAB (dAB) strain.

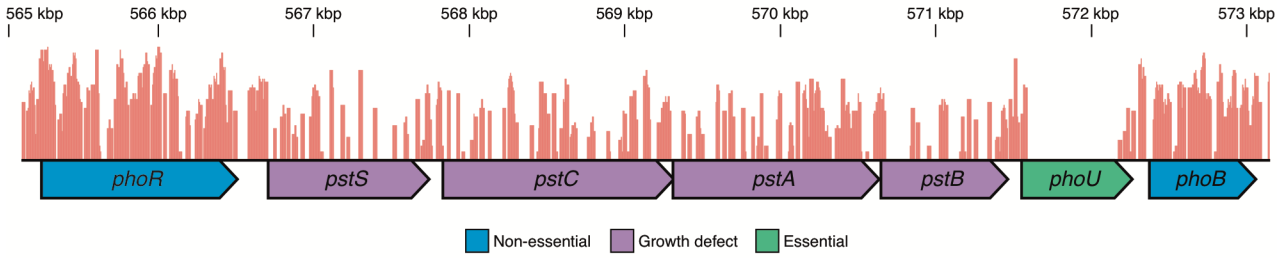
Gene	Function	GEI (WT ... dAB)	Gene	Function	GEI (WT ... dAB)
More Important ΔpSymAB - Rich Medium			More Important ΔpSymAB - Defined Medium		
kpsF3	capsule expression protein	5.951 ... 0.002	*smc03782*	signal peptide protein	9.670 ... 0.001
groEL	chaperonin GroEL	2.926 ... 0.001	*amiC*	N-acetylmuramoyl-L-alanine amidase	12.273 ... 0.003
aidB	oxidoreductase	2.658 ... 0.001	*groEL*	chaperonin GroEL	2.755 ... 0.001
proA	γ-glutamyl phosphate reductase	4.505 ... 0.001	*amn*	AMP nucleosidase	2.200 ... 0.001
amiC	N-acetylmuramoyl-L-alanine amidase	2.016 ... 0.002	*ndvA*	cyclic beta-1,2-glucan ABc transporter	1.434 ... 0.001
smc03782	signal peptide protein	1.847 ... 0.001	*smc02495*	translaldolase	3.933 ... 0.003
etfA1	electron transfer flavoprotein	2.447 ... 0.002	*glnA*	glutamine synthetase	2.535 ... 0.002
smc02495	translaldolase	3.127 ... 0.003	*glmS*	glucosamine--fructose-6P aminotransferase	1.918 ... 0.002
exoN2	UTP--glucose-1P uridylyltransferase	2.292 ... 0.002	*smc00717*	ABC transporter ATP-binding protein	6.624 ... 0.006
glnA	glutamine synthetase	2.153 ... 0.002	*etfA1*	electron transfer flavoprotein	2.156 ... 0.002
More Important Wild Type - Rich Medium			More Important Wild Type - Defined Medium		
carB	carbamoyl phosphate synthase	0.001 ... 1.930	*folD2*	5,10-methylene-THF dehydrogenase	0.003 ... 1.026
argG	argininosuccinate synthase	0.002 ... 1.281	*nuoK1*	NADH dehydrogenase subunit K	0.006 ... 1.476
carA	carbamoyl phosphate synthase	0.007 ... 3.928	*prfC*	peptide chain release factor RF-3 protein	0.001 ... 0.232
purB	adenylosuccinate lyase	0.002 ... 0.799	*smc00714*	1-acyl-SN-glycerol-3P acyltransferase	0.005 ... 0.522
hrm	histone-like protein	0.011 ... 2.586	*fpr*	ferredoxin--NADP reductase	0.002 ... 0.248
nuoK1	NADH dehydrogenase subunit K	0.006 ... 1.330	*smc00532*	hypothetical protein	0.002 ... 0.203
folD2	5,10-methylene-THF dehydrogenase	0.018 ... 3.110	*ubiE*	ubiquinone biosynthesis methyltransferase	0.002 ... 0.209
coaA	pantothenate kinase	0.004 ... 0.595	*asd*	aspartate-semialdehyde dehydrogenase	0.002 ... 0.157
argF1	ornithine carbamoyltransferase	0.012 ... 1.776	*secE*	preprotein translocase subunit SecE	0.010 ... 0.796
smc00914	oxidoreductase	0.002 ... 0.240	*smc01038*	hypothetical protein	0.040 ... 2.940

965 **Table 3. Summary of iGD726.** The last column indicates reactions whose genes associations are supported by the Tn-seq data of this

966 study. Percentage of all reactions in that category are indicated in brackets.

Pathways	Genes	Reactions	Reactions supported by Tn-seq
Overall	726	681	444 (65%)
Carbon metabolism, oxidative phosphorylation	105	54	37 (69%)
Amino acid metabolism	116	93	72 (77%)
Nucleotide metabolism	34	40	39 (98%)
Fatty acid, lipid metabolism	42	227	143 (63%)
Peptidoglycan, lipopolysaccharide, exopolysaccharide metabolism	47	43	27 (63%)
Nucleotide sugar metabolism	25	17	6 (35%)
Vitamin, cofactor, coenzyme metabolism	109	121	83 (69%)
Miscellaneous metabolism	23	15	7 (47%)
Transcription, translation, DNA replication, cell division	153	29	28 (97%)
Transport reactions	75	21	3 (14%)
Exchange reactions	0	21	N/A

967



968

969 **Figure 1. Visualization of the location of transposon insertion sites.** An image of the *pst* locus of *S. meliloti*
970 generated using the Integrative Genomics Viewer [85]. Chromosomal nucleotide positions are indicated along the
971 top of the image, and the location of transposon insertions are indicated by the red bars. Non-essential genes contain
972 a high density of transposon insertions, whereas essential genes have few to no transposon insertions. Genes are
973 color coded based on their fitness classification, and transcripts are indicated by the arrows below the genes. The
974 *pstS*, *pstC*, *pstA*, *pstB*, *phoU*, and *phoB* genes are co-transcribed as a single operon [86], and previous work
975 demonstrated that polar *phoU* mutations are lethal in *S. meliloti*, whereas non-polar mutations are not lethal [87].
976 The lack of insertions within the *phoU* coding region is therefore consistent with the non-polar nature of the
977 transposon.

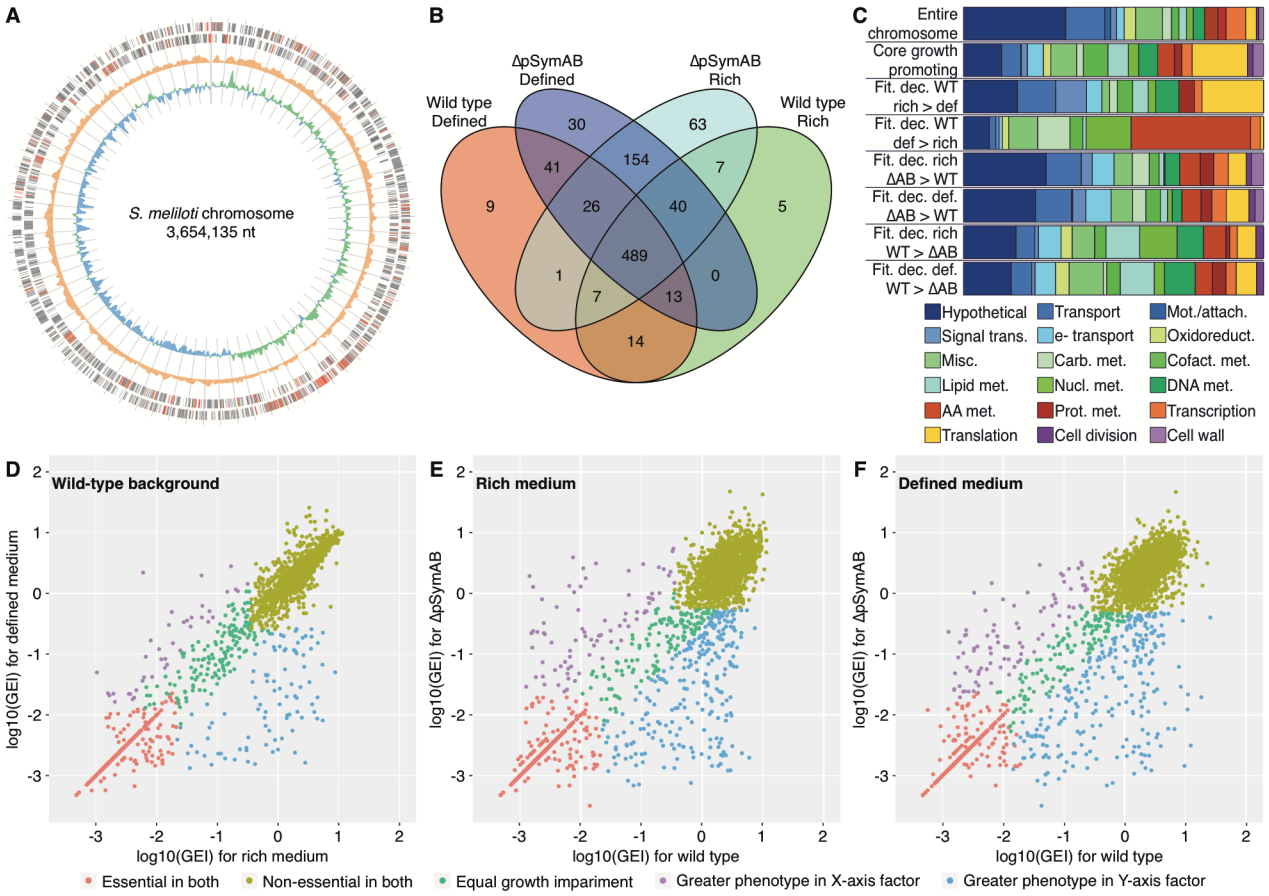


Figure 2. Characteristics of the core genetic components of *S. meliloti*. (**A**) A plot of the *S. meliloti* chromosome is shown. From the outside to inside: positive strand coding regions, negative strand coding regions, total insertion density, and GC skew. For the positive and negative strands, red lines indicate the core 489 growth promoting genes. The insertion density displays the total transposon insertions across all experiments over a 10,000-bp window. The GC skew was calculated over a 10,000-bp window, with green showing a positive skew and blue showing a negative skew. Tick marks are every 50,000 bp. (**B**) A comparison of the overlap between the growth promoting genome (Group I and II genes) of each Tn-seq data set. Each data set is labelled with the strain (wild type or ΔpSymAB) and the growth medium (defined medium or rich medium). (**C**) Functional enrichment plots for the indicated gene sets. Name abbreviations: Fit – fitness; Dec – decrease; WT – wild type; ΔAB - ΔpSymAB; Def – defined medium; Rich – rich medium. For example, 'Fit. dec. WT def > rich' means the genes with a greater fitness decrease in wild type grown in defined medium compared to rich medium. Legend abbreviations: AA – amino acid; Attach – attachment; Carb – carbohydrate; Cofact – cofactor; e- – electron; Met – metabolism; Misc – miscellaneous; Mot – motility; Nucl – nucleotide; Oxidoreduct – oxidoreductase activity; Prot – protein; Trans – transduction. (**D-F**) Scatter plots comparing the fitness phenotypes, shown as the $\log_{10}$ of the GEI scores (Gene Essentiality Index scores; i.e., number of insertions within the gene divided by gene length in nucleotides) of (**D**) wild type grown in rich medium versus wild type grown in defined medium, (**E**) wild type grown in rich medium versus ΔpSymAB grown in rich medium, and (**F**) wild type grown in defined medium versus ΔpSymAB grown in defined medium.

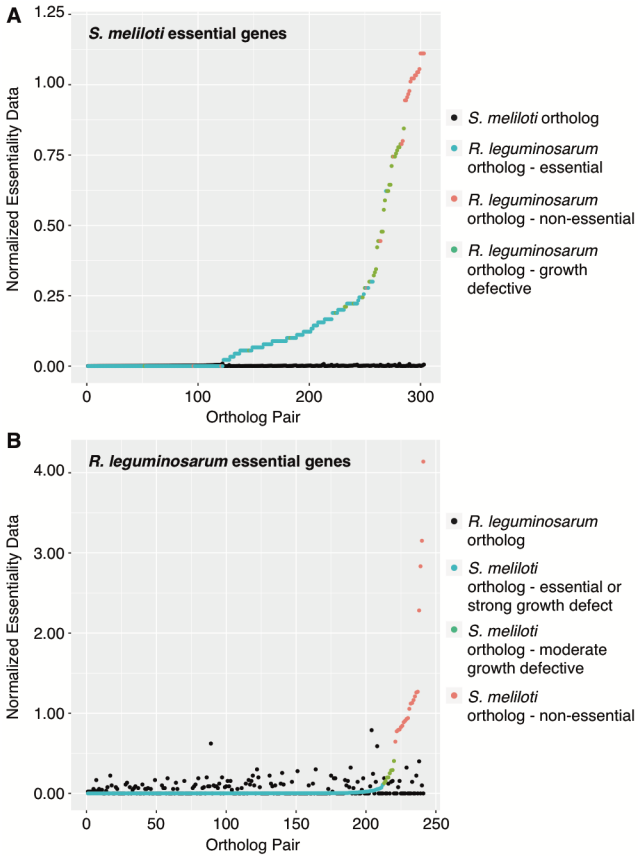


Figure 3. Comparison of *S. meliloti* and *R. leguminosarum* Tn-seq data. (**A**) The fitness phenotypes of essential *S. meliloti* genes, as determined in this study, is compared to the fitness phenotypes of the orthologous *R. leguminosarum* genes, as determined by Perry *et al*. [32]. *S. meliloti* orthologs are shown in black, while the *R. leguminosarum* orthologs are colored according to their classification by Perry *et al*. [32]. (**B**) The fitness phenotypes of essential *R. leguminosarum* genes is compared to the fitness phenotypes of the orthologous *S. meliloti* genes. *R. leguminosarum* orthologs are shown in black, while the *S. meliloti* orthologs are colored according to their classification in this study. (**A**,**B**) Normalized fitness values are used to facilitate direct comparison between the studies as different output statistics were calculated. For *S. meliloti*, the GEI score of each gene for wild type grown in minimal medium broth was divided by the median GEI for all genes under the same conditions. For *R. leguminosarum*, the insertion density of each gene during growth on minimal medium plates was divided by the median insertion density of all strains.

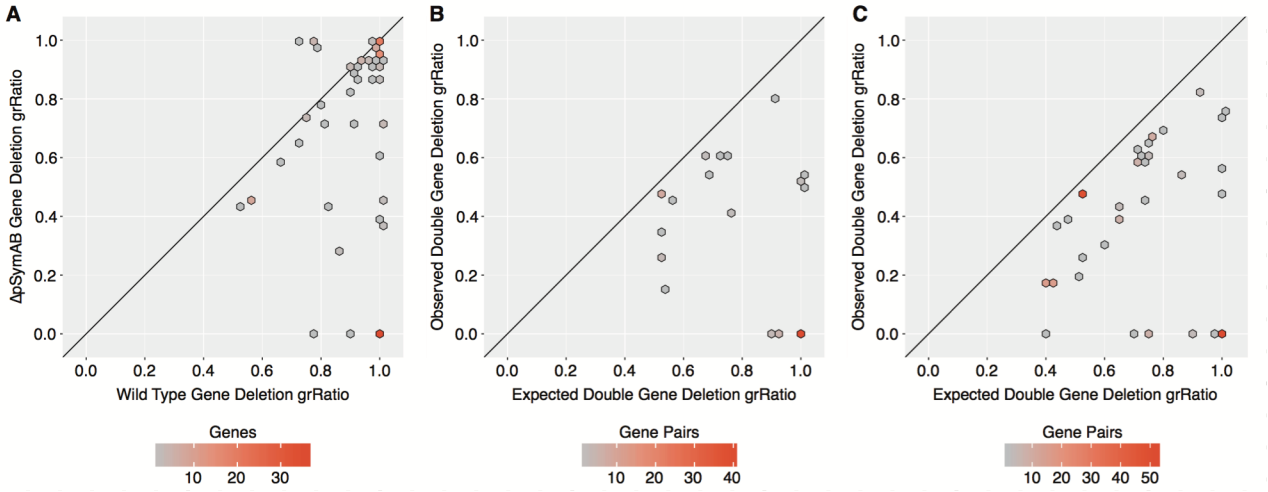


Figure 4. *In silico* analysis of genetic redundancy in *S. meliloti*. The effects of single or double gene deletion mutants were predicted *in silico* with the genome-scale *S. meliloti* metabolic model. (**A**) A scatter plot comparing the grRatio (growth rate of mutant / growth rate of non-mutant) for gene deletion mutations in the presence (wild type) versus absence (ΔpSymAB) of the pSymA/pSymB model genes. Genes whose deletion had either no effect or were lethal in both cases are not shown. (**B**) A scatter plot comparing the grRatio for each double gene deletion pair (where one gene was on the chromosome and the other on pSymA or pSymB) observed *in silico* versus the predicted grRatio based on the grRatio of the single deletions (grRatio1 * grRatio2). Only gene pairs with an observed grRatio at least 10% less than the expected are shown. (**C**) A scatter plot comparing the grRatio for each double gene deletion pair (both genes on the chromosome) observed *in silico* versus the predicted grRatio. Only gene pairs with an observed grRatio at least 10% less than the expected are shown. (**A-C**) The color of each hexagon is representative of the number of reactions plotted at that location, as illustrated by the density bar below each panel. The diagonal line serves as a reference line.

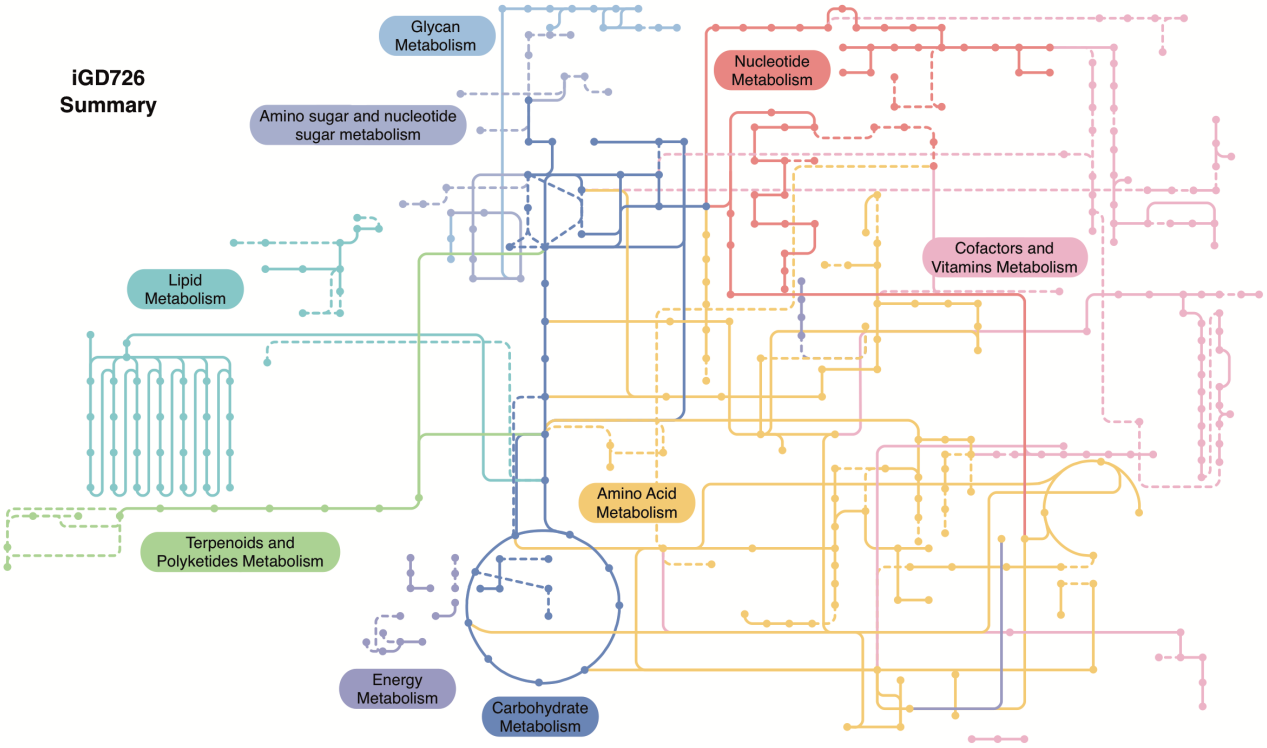


Figure 5. Summary schematic of core *S. meliloti* metabolism. The iGD726 core metabolic model was visualized using the iPath v2.0 webserver [74], which maps the reactions of the metabolic model to KEGG metabolic pathways; it therefore does not capture metabolism not present in the KEGG pathways included in iPath. Reactions and metabolites are colour coded according to their biological role, as indicated. Reactions whose associated genes were not identified as growth promoting in this study are in dashed lines.

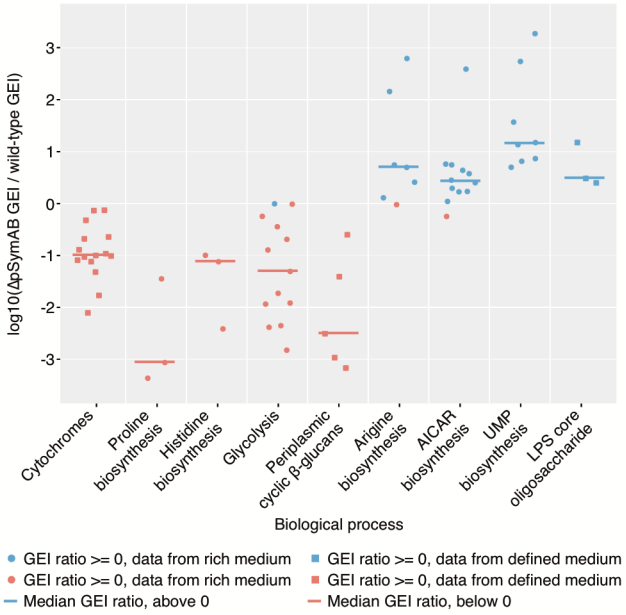


Figure 6. Gene essentiality index (GEI) changes for genes of selected biological pathways. Each circle or square represents an individual gene, and shows the $\log_{10}$ of the ratio of the GEI for that gene in the ΔpSymAB background compared to the wild-type background. Lines indicate the median value of all genes included from the biological process. The underlying data is given in Table S9. Genes included in each process are as follows: Cytochrome C oxidase related genes – *ctaB*, *ctaC*, *ctaD*, *ctaE*, *ctaG*, *ccsA*, *cycH*, *cycJ*, *cycK*, *cycL*, *ccmA*, *ccmB*, *ccmC*, *ccmD*, *ccmG*; Proline biosynthesis – *proA*, *proB1*, *proC*; Histidine biosynthesis – *hisB*, *hisD*, *smc04042*; Glycolysis and related genes – *glk*, *frk*, *pgi*, *zwf*, *pgl*, *edd*, *eda2*, *gap*, *pgk*, *gpmA*, *eno*, *pykA*, *pyc*; Periplamic cyclic β-glucan biosynthesis – *feuN*, *feuP*, *feuQ*, *ndvA*, *ndvB*; Arginine biosynthesis – *argB*, *argC*, *argD*, *argF1*, *argG*, *argH1*, *argJ*; AICAR biosynthesis – *purB*, *purC*, *purD*, *purE*, *purF*, *purH*, *purK*, *purL*, *purM*, *purN*, *purQ*, *smc00494*; UMP biosynthesis – *carA*, *carB*, *pyrB*, *pyrC*, *pyrD*, *pyrE*, *pyrF*, *smc01361*; LPS core oligosaccharide biosynthesis – *lpsC*, *lpsD*, *lpsE*.