Genetic manipulation of candidate phyla radiation bacteria provides functional insights into microbial dark matter

Yaxi Wang[1,†], Larry A. Gallagher[1,†], Pia A. Andrade[1], Andi Liu[1], Ian R. Humphreys[2,3], Serdar Turkarslan[4], Kevin J. Cutler[5], Mario L. Arrieta-Ortiz[4], Yaqiao Li[1,4], Matthew C. Radey[1], Jeffrey S. McLean[1,6], Qian Cong[7,8,9], David Baker[2,3,10], Nitin S. Baliga[4], S. Brook Peterson[1], and Joseph D. Mougous[1,10,11,*]

[1]Department of Microbiology, University of Washington, Seattle, WA 98109, USA
[2]Department of Biochemistry, University of Washington, Seattle, WA 98109, USA
[3]Institute for Protein Design
[4]Institute for Systems Biology, Seattle, WA 98109, USA
[5]Department of Physics, University of Washington, Seattle, WA 98195, USA
[6]Department of Periodontics, University of Washington, Seattle, WA 98195, USA
[7]Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, USA.
[8]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, USA.
[9]Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX, USA.
[10]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA
[11]Microbial Interactions and Microbiome Center, University of Washington, Seattle, WA 98109, USA

† Authors contributed equally.

* To whom correspondence should be addressed:

Email – mougous@uw.edu

Summary

The study of bacteria has yielded fundamental insights into cellular biology and physiology, biotechnological advances and many therapeutics. Yet due to a lack of suitable tools, the significant portion of bacterial diversity held within the candidate phyla radiation (CPR) remains inaccessible to such pursuits. Here we show that CPR bacteria belonging to the phylum Saccharibacteria exhibit natural competence. We exploit this property to develop methods for their genetic manipulation, including the insertion of heterologous sequences and the construction of targeted gene deletions. Imaging of fluorescent protein-labeled Saccharibacteria provides high spatiotemporal resolution of phenomena accompanying epibiotic growth and a transposon insertion sequencing genome-wide screen reveals the contribution of enigmatic Saccharibacterial genes to growth on their Actinobacteria hosts. Finally, we leverage metagenomic data to provide cutting-edge protein structure-based bioinformatic resources that support the strain *Southlakia epibionticum* and its corresponding host, *Actinomyces israelii*, as a model system for unlocking the molecular underpinnings of the epibiotic lifestyle.

48 **Introduction**

49 The vast majority of metagenomic DNA sequences obtained from microbial species-rich
50 environmental sources is derived from Bacteria and Archaea that have not been cultivated.
51 Conservative estimates suggest that these sequences, often referred to as microbial dark matter,
52 represent organisms constituting approximately half of phylum level diversity within these
53 domains[1]. Microbial dark matter holds great interest as a reservoir of biosynthetic pathways and
54 enzymes with potential for biotechnological application[2]. In addition, understanding the
55 functions of these genes is paramount to defining the molecular processes supporting a given
56 ecosystem and for unraveling the physiology and cell biology of the organisms within[3].

57 The candidate phyla radiation (CPR) of bacteria are reported to contain well over 50
58 phyla, only four of which have cultivated representatives[4]. This large monophyletic lineage of
59 phyla and superphyla, including the group currently known as Patescibacteria, is thought to
60 represent 15-50% of total bacterial diversity and contributes disproportionately to microbial dark
61 matter[5]. Common features of CPR bacteria are their small size (as little as 100-200 nm in width),
62 reduced genomes (typically < 1 megabase), and limited metabolic capability[6]. This has led to the
63 hypothesis that these organisms broadly share a requirement for host organisms to support their
64 growth. Indeed, experiments show that most cultivated CPR bacteria attach to and proliferate on
65 the surface of other bacteria – living as obligate epibionts[4 7-10]. Genomic analyses have yielded
66 speculation regarding the molecular functions that support the epibiotic lifestyles of CPR
67 bacteria[5,11]. However, owing to the phylogenetic distance separating CPR and well characterized
68 bacteria, the function of much of their proteome cannot be predicted and a lack of genetic tools
69 for these organisms has heretofore precluded the experimental investigation of genotype–
70 phenotype relationships.

71 Among the CPR, members of the phylum Saccharibacteria, originally named Tm7, were
72 the first to be cultivated in the laboratory[8]. Saccharibacteria are found in a multitude of terrestrial
73 and marine environments, yet early interest in them stemmed from their widespread occurrence
74 in human oral microbiomes[12]. Archaeological findings show this association dates to before the
75 mesolithic period and recent work links Saccharibacteria to human oral health[12-14]. The growth of
76 Saccharibacteria relies on the co-cultivation of host bacteria belonging to the class
77 Actinomycetia with the phylum Actinomycetota, for which they exhibit strain level specificity[15-
78 17]. Employing panels of Actinomycetia strains for Saccharibacterial enrichment has facilitated
79 the isolation and sequencing dozens of strains[16,18,19]. Despite this progress, the phylum remains
80 poorly sampled, with many divergent clades uncultivated, and the extent of genetic diversity
81 unresolved.
82 The genome sequences of Saccharibacteria reveal common features that provide insight
83 into molecular functions underlying their cellular physiology and lifestyle. Akin to other CPR
84 bacteria, Saccharibacteria generally lack a respiratory chain, and pathways for the *de novo*
85 generation of amino acids, nucleotides, and fatty acids[6]. On the contrary, Saccharibacteria
86 universally possess a relative wealth of specialized secretory mechanisms including the type II
87 and IV secretion systems (T2SS, T4SS)[11]. These diverge significantly from related systems in
88 bacterial pathogens that deliver toxins and effectors to eukaryotic host cells; however, it has been
89 speculated that they function in an analogous fashion to support bacterial host co-option by
90 Saccharibacteria[11]. Saccharibacteria also possess type IV pili (T4P), which were implicated in
91 twitching motility and host adhesion through the use of a small molecule inhibitor of pilus
92 extrusion[17]. Though such genomic analyses and experiments provide fertile ground for the
93 formulation of hypotheses, progress toward a mechanistic understanding of the unique biology of

94 Saccharibacteria and CPR bacteria as a whole has been stymied by a lack of genetic tools[20].
95 Here, we discover that natural competence can be harnessed for genetic manipulation of
96 Saccharibacteria. With this capability in hand, we go on to use fluorescent protein expression to
97 conduct time-lapse microscopic analysis of the Saccharibacterial lifecycle and we perform
98 transposon mutagenesis to identify genes important for epibiotic growth. Our findings offer an
99 initial mechanistic glimpse into the cellular functions encoded in microbial dark matter.
100

101 **Results**

102 **Isolation and characterization of Saccharibacteria strains**

103 We collected, pooled, homogenized, and filtered saliva and dental plaque from
104 volunteers, and enriched this material for Saccharibacteria using the method described by Bor *et*
105 *al*[18]. This led to the isolation of two strains, which exhibit distinct host specificity (Figure 1A).
106 Phylogenetic analysis using concatenated alignments of 49 core, conserved protein sequences
107 obtained through complete genome sequencing placed the two strains within human oral
108 subclades of the G1 clade of Saccharibacteria (Figure 1B-E)[21]. Based on these assignments, we
109 named our strains *Candidatus Nanosynbacter lyticus* ML1 (*Nl*) and *Ca. Southlakia epibionticum*
110 ML1 (*Se*). The genome sequences of the two strains additionally indicated they bear features
111 typical of Saccharibacteria, such as genes associated with specialized secretion systems, cell
112 surface appendages, and competence, coupled with a lack of genes required for a multitude of
113 biosynthetic pathways and an overall reduced genome size (Figure 1D,E).
114

115 **Genetic manipulation of Saccharibacteria via natural transformation**

116 We sought to develop methods for reverse genetic analyses within Saccharibacteria.
117 Genes encoding the core components of the *com* DNA uptake system are conserved across CPR
118 phyla, including the Saccharibacteria[5]. These consist of ComEC, the central membrane DNA
119 conduit, DprA, a catalyst of Rec-mediated recombination, and ComFC, which binds ssDNA and
120 links import to recombination[22-24]. Com proteins function in concert with type IV pili, which are
121 also widely distributed in Saccharibacteria, and CPR bacteria more generally[5,23]. Given the lack
122 of nucleotide biosynthetic capability in CPR bacteria, it has been proposed that these systems
123 facilitate nucleotide acquisition[5]. Although the presence of the Com system is largely not
124 predictive of DNA transformation in a laboratory setting[25], we sought to test whether exogenous
125 DNA could be exploited for genetically manipulating Saccharibacteria.
126 As a first step toward assessing feasibility of genetics in Saccharibacteria, we searched
127 for antibiotics with convenient resistance determinants that potently inhibit the growth of *Se*
128 without impacting that of its preferred host, *Actinomyces israelii* F0345 (*Ai*). These experiments
129 revealed that across a wide range of concentrations, the aminoglycoside hygromycin fulfills
130 these criteria (Figure S1). Next, we designed and generated a linear cassette containing the
131 hygromycin resistance gene (*hph*) codon optimized for *Se* flanked by the promoter and
132 terminator regions of the TM7x *tuf* gene (elongation factor Tu), an open reading frame (ORF)
133 predicted to be highly expressed (Figure 2A). For the insertion of this cassette, we selected an
134 intergenic region located between two convergently transcribed ORFs, SEML1_0215 and
135 SEML1_0216, hereafter referred to as neutral site 1 (NS1). To promote homologous
136 recombination, approximately 1000 bp on either side of the insertion site were added to the 5′
137 and 3′ ends of our cassette.

138 To transform *Se*, we incubated *Se–Ai* co-cultures with 2.0 ng/μL of our linear cassette for
139 six hours before initiating selection with hygromycin (Figure 2B). Naïve host was added
140 concomitantly in order to permit the outgrowth of successfully transformed *Se* cells. Reasoning
141 that transformation may be inefficient, we passaged these co-cultures twice, at 48 hour intervals,
142 with continued hygromycin selection and the addition of naïve host. Cultures were then diluted
143 and plated to obtain colonies, where were selected and propagated with naïve host without
144 selection before genotyping (Figure 2B). The latter step was included in order to bottleneck the
145 *Se* population and facilitate the isolation of clonal populations. Remarkably, each *Se*-infected
146 culture we tested – accounting for the majority of colonies selected – contained our synthetic
147 cassette inserted at the expected location (Figure 2C). Whole genome sequencing confirmed
148 these PCR results and it further showed that cassette integration occurred without introducing
149 off-target mutations.
150 Quantitative PCR (qPCR) analysis of total *Se*, transformed *Se*, and *Ai* at regular intervals
151 during our transformation procedure demonstrated that approximately 0.2% of Se contain the
152 integrated cassette by the conclusion of the initial incubation period (Figure 2D). Though *Se*
153 levels remain low through the second passage under selection with hygromycin, all surviving *Se*
154 cells bear the cassette at this time point. In the final passage, the population of *Se* continues to
155 maintain the cassette and expands markedly, far surpassing levels of the host (Figure S2A). In
156 the absence of hygromycin, similar quantities of initially transformed *Se* are observed; however,
157 this small proportion fails to expand despite overall robust growth of *Se*. We observed similar
158 transformation behavior using lengths of DNA with homology to the insertion site flanking
159 regions as short as 221 bp (Figure S2A) and with as little as 0.02 ng/μL (Figure S2B). Finally, to
160 probe the generality of our methods, we identified a predicted neutral site within *Nl* and

161 subjected this second Saccharibacterial strain to an analogous transformation protocol.
162 Genotyping of transformed populations indicated cassette insertion at the desired location also
163 occurred within this strain (Figure S2C,D).

164 The ability to introduce heterologous DNA into *Se* has numerous foreseeable
165 applications, one of which is the expression of reporter genes that allow *Se* to be distinguished
166 and studied within the context of co-culture with their hosts. To explore this possibility, we
167 designed and generated NS1 insertion cassettes containing genes encoding nanoluciferase,
168 mCherry and GFP under the control of the *tuf* promoter and upstream of *hph* driven by a second
169 predicted strong promoter of TM7x, that of *rpsJ* (Figure 2E,F). Using our transformation
170 protocol, we obtained clonal populations of hygromycin-resistant *Se* containing each of these
171 genes. Luminescence assays and fluorescence microscopy demonstrated robust activity of each
172 reporter gene (Figure 2E,F). We did not detect their activity in host cells, indicating the
173 feasibility of achieving specific manipulation of *Se* in the context of a co-culture.

174 The expression of fluorescent proteins within *Se* provided the opportunity to visualize
175 CPR bacterial growth with extended, time-lapse fluorescence imaging. Over the course of 20 hrs,
176 both *Se* and *Ai* populations – deposited from co-cultures at low *Se*:*Ai* ratio onto an agar substrate
177 containing growth media – showed clear evidence of expansion (Videos S1-11). Apparent T4P-
178 mediated motility of *Se* was also observed, as reported by Xie *et al*[17]. Our time-lapse imaging
179 further captured features of the CPR lifecycle at unprecedented spatiotemporal resolution. For
180 instance, we could distinguish productive (*Se* growth supporting) versus non-productive (*Se*
181 adhered without concomitant growth) interactions, and directly measure their respective impact
182 on individual host cells (Figure 3). Additionally, we observed productively adhered mother cells
183 producing small swarmer cell progeny via repeated polar budding, and the differentiation of a

184 subset of these progeny into mother cells (Videos S1-11). Altogether, these findings show that
185 natural transformation can be exploited to genetically manipulate CPR bacteria in a directed
186 manner and open a window into the distinctive biology of this largely unexplored group of
187 organisms.
188
189 **Tn-seq of *Se* provides insights into the epibiotic lifestyle**
190 The ability to genetically manipulate Saccharibacteria enables myriad avenues of
191 investigation. As a first step toward genetic dissection of the *Se* epibiotic relationship with *Ai*, we
192 conducted transposon-insertion sequencing (Tn-seq) within *Se* during growth on *Ai*. To this end,
193 we performed *in vitro* Tn5-based transposition on purified *Se* genomic DNA, repaired gaps as
194 described by Manoil and colleagues, and used sequencing to confirm high frequency,
195 homogenous insertion across the genome. This DNA was then used to transform *Se*, with slight
196 modifications from our basic transformation protocol (see Methods). Most notably, we elected to
197 increase the scale of the experiment to account for our previously measured transformation
198 efficiency and thereby avoid population bottlenecking following the onset of selection. We
199 collected four samples for Tn-seq analysis, an initial sample following the transformation and
200 recovery period (T0), and three additional samples representing the population after 48 hr serial
201 passages with the addition of naïve host (T1-3). Measurements of *Se* and *Ai* levels at each of
202 these timepoints revealed a precipitous drop in *Se* levels at T1 that is not observed at a similar
203 timepoint in transformations targeting NS1, suggesting that the large majority of transposon
204 insertions in *Se* are deleterious (Figure S3). Our measurements also showed that, as expected, *Ai*
205 levels drop relative to those of *Se* at later passages, such that by T3, *Se* outnumbers *Ai* by
206 approximately 50-fold.

207 Preliminary analysis of our Tn-seq data indicated that the T0 population of *Se* contained a
208 large number of unique insertions (9,996) relative to subsequent timepoints (T1, 1709; T2, 1887;
209 T3, 1699) despite similar sequencing depth (Table S1, Figure 4A). Furthermore, application of
210 the TRANSIT algorithm to the T0 data identified a small number of essential genes relative to
211 our expectations (905 *Se* genes; 56 essential by Gumbel or HMM)[26]. Taken together with our
212 measurement of *Se* and *Ai* population levels, we interpret these findings as evidence that little
213 selection occurred prior to T0 sampling. Therefore, we compared subsequent timepoints to T0
214 using TRANSIT resampling in order to assess gene level fitness contributions throughout the
215 experiment. In total, this approach identified 214 genes that demonstrate statistically significant
216 (multiple-comparison adjusted P value < 0.05) fitness contributions across all timepoints, with
217 selection yielding greater numbers at each subsequent sampling (T1, 222; T2, 252; T3, 275)
218 (Table S1).
219 The transposon insertion density we observed in our T1-3 samples is below the value
220 needed to obtain optimal results in TRANSIT. Given this potential caveat, we elected to further
221 process our data with ALDEx2, a program widely used for differential abundance analysis of
222 compositional data[27]. Overlaying genes with significantly differential insertion abundance
223 determined by ALDEx2 with those we obtained using TRANSIT revealed a group of 79 genes
224 that we deem unambiguously critical for *Se* growth on *Ai* in the context of our experimental
225 conditions (Table S1 and Figure 4A). This group includes genes encoding proteins that comprise
226 core cellular machinery such as DNA polymerase, RNA polymerase and the ribosome.
227 Interestingly, the group also includes non-canonical essential genes that likely encode specialized
228 functions that support the epibiotic lifestyle of CPR bacteria; examples include genes within a
229 T4P operon and several genes lacking homologs in non-CPR bacteria (Figure 4B).

230 If our Tn-seq findings provide an accurate measure of gene-level fitness contributions,
231 then those identified as critical for *Se* fitness via TRANSIT or TRANSIT and ALDEx2 should be
232 difficult to inactivate by targeted mutagenesis. Accordingly, to validate these data, we selected
233 four genes identified as significant contributors to fitness in one or both analyses, along with four
234 control genes, and designed constructs to replace each with an *hph* expression cassette. The
235 constructs shared equivalent length flanking sequences to enable the direct comparison of their
236 behavior in our transformation protocol. We assessed the fitness impact of inactivating each gene
237 by quantifying *Se* populations following transformation and outgrowth under selection with
238 hygromycin. For the control genes not predicted to contribute to fitness, we achieved robust
239 levels of *Se* growth by this time point, comparable to that achieved when introducing the *hph*
240 expression cassette at NS1 (Figure 4C). In contrast, in transformations targeting each of the four
241 predicted essential genes examined, *Se* failed to proliferate, consistent with the inactivation of
242 these genes strongly impacting fitness. Together, these findings provide confirmation that our
243 Tn-seq analysis successfully identified the relative fitness contributions of *Se* genes during co-
244 culture with *Ai.*
245 We noted that genes encoding homologs of *com* system components ComEC, ComF and
246 DprA were not among those genes defined as contributing significantly to *Se* fitness in our Tn-
247 seq study. This finding suggests that this DNA competence machinery is not required for
248 acquiring nucleotides to support *Se* growth, as had been previously suggested[5]. To determine
249 whether the *com* system of *Se* instead functions to mediate natural transformation, we generated
250 an *Se* strain in which the *comEC* ORF is replaced by the *hph* expression cassette (*Se*
251 Δ*comEC*::*hph*). To assess whether this mutation affects *Se* transformation, we measured the
252 efficiency of inserting a second, unmarked cassette at NS1. At the conclusion of the

253 transformation protocol, insertion at the NS1 site was only detectable in the wild-type
254 background, supporting the hypothesis that the *com* system mediates natural transformation in
255 this species (Figure 4D,E).

256 Banfield and colleagues previously leveraged the large number of CPR genomes
257 available from metagenomic and traditional genome sequencing datasets (n=2,321) to define
258 CPR-enriched protein families[5]. To gain insight into which of these contribute to fitness during
259 *Se* growth on *Ai*, we identified *Se* proteins belonging to these families, and cross-referenced hits
260 against the gene lists obtained from our Tn-seq experiment following TRANSIT or TRANSIT
261 and ALDEx2 analysis (Figure S4). Interestingly, we found that the representation of CPR-
262 enriched protein families among proteins critical for *Se* survival on *Ai*, 22% and 19%,
263 respectively, is considerably higher than their proportion within the overall *Se* proteome (12%)
264 (Table S1 and Figure S4). As many of these protein families have no predicted function, nor
265 orthologs outside of CPR bacteria, their future study is likely to reveal unique features of these
266 organisms.

267

268 **Development of resources for dissecting Saccharibacteria–host cell interactions**

269 With available genetic tools and gene-level fitness data, *Se* could become a tractable
270 model for the study of Saccharibacteria. However, comprehensive genotype–phenotype
271 dissection of the interaction between *Se* and its host remains limited by lack of equivalent tools
272 and data for *Ai.* As a first step towards bridging this gap, we obtained a complete genome
273 sequence for the host strain we employed, *A. israelii* F0345 (Figure S5A). This strain was
274 isolated from the human oral cavity as part of the Human Microbiome Project, but a genome
275 sequence had not been previously deposited in a public database. Although genetic tools have

276 not been developed for this organism, gene inactivation via the electroporation-mediated
277 introduction of suicide plasmid or linear gene inactivation constructs has been achieved in
278 related species[28,29]. The availability of genetic information for this strain should facilitate
279 development of selectable markers and counter selection strategies.

280 Automated annotation of the *Se* genome sequence failed to predict functions for the
281 proteins encoded by many ORFs (278 of 855), including 14/79 belonging to the conservative list
282 of essential genes we identified by Tn-seq (Table S1). To improve functional predictions
283 associated with ORFs in the *Se* genome, we applied a battery of sequence and structure-based
284 computational tools. ProtNLM is a natural language model trained on UniProt to predict protein
285 names given a sequence[30]. Applying ProtNLM to the *Se* proteome yielded functional annotations
286 for 337 of 855 proteins (ProtNLM score ≥ 0.5, Table S2). This sequence-based annotation was
287 further improved by mapping each protein to Pfam domains using hmmscan, detecting Pfam
288 domains in an additional 270 proteins (see Methods for cutoffs), 245 of which could be assigned
289 functions[31,32] (Table S2).

290 To complement our sequence-based annotations, we took advantage of recent advances in
291 protein structure prediction to conduct genome-wide structure-based homology analyses of the
292 *Se* proteome[33,34]. Generation of structural models using AlphaFold (AF) relies on evolutionary
293 information extracted from multiple sequence alignments (MSAs)[34]. For nearly 25% of the *Se*
294 proteome (220 proteins), our initial MSAs based on HHblits searches of UniRef and BFD (the
295 default databases used by AF) were too shallow for high confidence structure prediction (<500
296 sequences post filtering, Figure 5A)[35-37]. To improve the MSA depth for these proteins, we
297 implemented a hidden Markov model (HMM)-based approach to identify and align additional
298 homologs for each protein from multiple metagenomic datasets. This resulted in considerably

299 deeper MSAs (>500 sequences) for an additional 9% of the proteome (Figure 5B). Using the
300 highest depth MSA obtained for each protein, we obtained AF models for >99% of the predicted
301 *Se* proteome. For comparison, we also computed AF models using the original 220 MSAs that
302 contained <500 sequences and compared the model confidence metric obtained using these and
303 the improved depth MSAs. Particularly for low confidence models obtained using MSAs
304 generated with the default databases (average pLDDT<50), we found that the use of the deeper
305 MSAs for structural prediction resulted in substantial model confidence improvement (Figure
306 5C). The structure model improvements enabled by using extensive metagenomic databases for
307 MSA generation were further underscored by structural homology search results obtained using
308 Foldseek (FS)[38]. For some proteins, the improved structural models led to the identification of
309 structural homologs for proteins that initially had no FS matches passing our cutoffs, while for
310 others, the changes in the overall predicted structure led to the identification of different and
311 more closely aligning top matches (Figure 5D,E, Figure S5B-D). In total, using FS we were able
312 to identify similar structures for 88% of modeled proteins (748/852, Table S2).
313 The incorporation of structural information resulted in functional predictions for 70 *Se*
314 proteins that had none assigned by sequence-based approaches (Table S2). One example relevant
315 to our interest in genetics within *Se* is the protein encoded by SEML1_0370, for which sequence-
316 based annotations assigned no function. A FS search performed using the AF-generated model
317 for this protein revealed it shares overall structural homology to and active site residues with the
318 class II restriction endonuclease *PvuII*[39]. We confirmed using the Restriction Enzyme Database
319 that SEML1_0370 does not share detectable primary sequence homology with characterized
320 restriction endonucleases. Characterization of the sequence recognized by this and other *Se*

321 restriction enzymes could lead to improvements in transformation efficiency, as DNA cassettes
322 could be designed to exclude their target sequence and thus subvert restriction[40].
323 Many of the AF models with similar structure identified for *Se* proteins are
324 uncharacterized (12%), providing little information regarding function. Moreover, unrelated
325 proteins may share similar structure due to convergence, and additional evidence is usually
326 needed to assess the homologous relationships. Domain Parser for AF Models (DPAM) is a
327 recently developed tool that parses structural domains from AF modeled structures, and
328 integrates both structure and sequence searches to map protein domains to Evolutionary
329 Classification of protein Domains (ECOD)[41,42]. Using DPAM, we detected homologous ECOD
330 domains for 80% of predicted *Se* proteins (Table S2). This included 39 proteins for which no
331 function was assigned by other methods. In supplementary material accompanying this report,
332 we provide alignments for each mapped *Se* protein domain, along with links to corresponding
333 Protein Data Bank identifiers (PDB IDs) and ECOD classifications
334 (https://conglab.swmed.edu/ECOD_Se/Se_ECOD.html). In total, these predictions provide a
335 foundation for future use of the *Se–Ai* system as a model for understanding CPR bacteria.
336
337 **Discussion**
338 Scientists have been aware of CPR bacteria in environmental samples for many years, yet
339 our understanding of this group of organisms has lagged[4]. One major challenge is their apparent
340 strict requirement for host bacteria, the identity of which cannot currently be determined a priori,
341 thus adding significant complexity to CPR isolation[7,18,43,44]. It is foreseeable that methods
342 bringing to bear both experimental and computational approaches could provide predictions of
343 host–epibiont associations in the future. Here we addressed a second impediment for studying

344 CPR bacteria – a lack of genetic tools for their manipulation. Indeed, despite streamlined
345 methods for isolating Saccharibacteria on host Actinobacteria that have been employed by
346 several laboratories, this latter challenge has to-date prohibited a molecular dissection of their
347 epibiotic lifestyle[16,17,44]. Our discovery that natural transformation can facilitate targeted
348 mutagenesis in Saccharibacteria opens the door to direct interrogation of genotype–phenotype
349 relationships within these CPR bacteria; however, challenges remain. The most fundamental
350 hurdle is that factors required for the Saccharibacteria–host cell interaction are also required for
351 viability. In analogous situations, researchers overcome this using assorted conditional
352 inactivation strategies (e.g. modulated expression, temperature sensitivity, inducible
353 degradation). While it is conceivable that such methods could be implemented in
354 Saccharibacteria, a powerful approach in this system may involve exploiting host genetics to
355 identify gene interactions that suppress otherwise lethal mutations in the epibiont.
356 The small genome of Saccharibacteria stands in contrast to the apparent complexity of
357 their lifecycle. Xie *et al.* proposed a model for the epibiotic growth of Saccharibacteria that
358 consists of four stages: T4P-mediated infection, growth, bud formation and asymmetric
359 division[17]. Our time-lapse microscopy generally supports this model; however, it further revealed
360 facets of the Saccharibacteria–host interaction not previously captured. While the
361 aforementioned model implies uniform progression, our data suggest the lifecycle is more
362 complex. We find that only a small subset of *Se* achieve productive infections, and that after an
363 extended enlargement period, these cells rapidly bud a large number of highly motile swarmer
364 cell progeny (>10 in a 20 hr period observed in many instances). Quantitative analyses of our
365 data additionally permitted us to link productive and non-productive *Se* infections with
366 corresponding host cell outcomes. We observed a striking inverse relationship between *Se*

367 growth and that of *Ai*; cells productively infected with as few as one *Se* not only failed to divide,
368 but diminished in size, whereas those infected unproductively by multiple *Se* readily proliferated.
369 Our microscopy observations suggests a potential division of labor within *Se* populations,
370 wherein one subpopulation is devoted to reproduction, and a second, motile subpopulation
371 searches for new compatible host cells. Operating under the assumption that the
372 Saccharibacteria–Actinobacteria relationship is chiefly one of parasitism, which we maintain is
373 not fully resolved in the literature, then the relevant prevailing evolutionary theory predicts that
374 *Se* fitness is dependent on both the rate at which it causes new infections and the duration of
375 those infections[45]. The reproductive strategy of *Se* appears consistent with this framework;
376 release of progeny cells destined to infect other host cells tempers host cell burden and prolongs
377 the infection, while concurrently maximizing *Se* reproduction. A similar outcome could be
378 achieved if *Se* employed active mechanisms to block superinfection, reminiscent of those utilized
379 by phage; however, swarmer cell production has the added advantage of increasing *Se* dispersal
380 in the face of low host cell densities[46]. Interestingly, most nascent swarmer cells we observed do
381 not themselves establish a productive infection within our 20 hr observation period, despite close
382 proximity and apparent adherence to host cells. It may be that the time required to establish a
383 productive interaction varies, and many of these cells would go on to become productive. It is
384 also conceivable that infected host cells defend against secondary infections, perhaps even
385 inducing a defensive state in neighboring kin cells. Future studies coupling quantitative
386 microscopy with genetics to further dissect the *Se–Ai* interaction will no doubt address these
387 questions and shed light on this fascinating interphyla relationship.
388 Pili are a basal feature of CPR bacteria and our Tn-seq results highlight the critical role
389 they play for these organisms[7,17,47]. Among the 15 genes that are both enriched in CPR and that

390 we conservatively deemed essential in *Se*, six are predicted T4P components (Table S1). A prior
391 study utilizing quercetin, a small molecule inhibitor of pilus retraction, reported that T4P are
392 important for host attachment and twitching motility in a second Saccharibacterial species[17]. In
393 naturally competent bacteria, T4P work in concert with the *com* system to bind and mediate the
394 uptake of extracellular DNA[48,49]. Our observation that *Se* cells lacking *comEC* are viable
395 suggests that the essentiality of T4P is unrelated to its role in competence. *Se* lacks *de novo*
396 nucleotide synthesis pathways, thus the organism presumably possesses an essential DNA or
397 nucleotide uptake pathway that operates independently of the *com* system. Whether T4P
398 participate in such an alternative DNA uptake mechanism remains unknown.

399 The arginine deiminase system (ADS) is an ATP generating catabolic pathway prevalent
400 in mammalian-adapted Saccharibacteria species[50]. A prior study demonstrated that arginine
401 supplementation to *N. lyticus* TM7x–*A. odontolyticus* co-cultures permits acid neutralization via
402 ammonia production, supporting the viability of both strains during acid stress[50]. The authors of
403 this study put forth a model in which hosts lacking ADS benefit from and are specifically
404 permissive to colonization by Saccharibacteria containing the pathway. In *Se*, arginine deiminase
405 (ArcA) is a fusion of two proteins, ArcA and ornithine carbamoyltransferase (ArcB); *arcA* is
406 among our conservative list of genes deemed essential for *Se* growth. Inactivation of genes
407 encoding the remaining ADS components, carbamate kinase ArcC and arginine/ornithine
408 antiporter ArcDE, also yielded significant *Se* fitness defects by one or more metric (Table S1).
409 Interestingly, in contrast to the proposed model for ADS function in Saccharibacterial–host cell
410 interactions, we find intact homologs of each ADS gene in *Ai*. Taken together with our Tn-seq
411 data, our results suggest that in some co-culture pairs, the organisms may compete for arginine.

412 However, the mechanism by which the pathway influences Saccharibacteria–host cell dynamics
413 *in vivo*, including specific niches in the oral cavity, remains an open question[50].
414 We anticipate that the methods and genetic tools presented here will facilitate molecular
415 level characterization of Saccharibacteria and CPR bacteria more broadly. In this regard, one key
416 question is the extent to which natural competence is active and able to be similarly exploited
417 across the CPR. The ability to manipulate CPR bacteria outside of Saccharibacteria, particularly
418 those with phylogenetically distinct hosts and inhabiting diverse niches, should aid in elucidating
419 the core requirements of the epibiotic lifestyle. Regardless of the precise methods utilized,
420 genetic manipulation of CPR bacteria will open the door to phenotypic studies of the rich
421 reserves of microbial dark matter these organisms contain, potentially revealing unprecedented
422 biological mechanisms.

423 **Acknowledgements**

424 The authors wish to thank Simon Dove and Joshua Woodward for insightful comments on the
425 manuscript, and members of the Mougous laboratory for helpful suggestions. This work was
426 supported by grants from the NIH (R01AI128215 and R01AI141953 to N.S.B. and
427 R01DE023810 to J.S.M.), the National Science Foundation (MCB-2105570 to N.S.B. and S.T.,
428 and IIBR-2042948 and IOS-2050550 to N.S.B.), the Department of Defense, Defense Threat
429 Reduction Agency (HDTRA1-21-1-0007 [DTRA J5B] to D.B.), the Bill and Melinda Gates
430 Foundation (OPP1156262 to D.B.), and the Welch Foundation (I-2095-20220331 to Q.C.). Q.C.
431 is a Southwestern Medical Foundation endowed scholar. J.D.M. and D.B. are HHMI
432 Investigators, and J.D.M. is supported by the Lynn M. and Michael D. Garvey Endowed Chair at
433 the University of Washington.

434

435 **Author contributions**

436 L.A.G, A.L., S.B.P and J.D.M conceived the study. Y.W., L.A.G., A.L., I.R.H, S.T., K.J.C.,
437 Q.C., N.S.B., S.B.P. and J.D.M. designed the study. Y.W., L.A.G., P.A.A., A.L., and K.J.C.
438 performed experiments. Y.W., L.A.G., I.R.H., S.T., K.J.C., M.L.A-O., Y.L., M.C.R., J.S.M. and
439 Q.C. processed, analyzed and visualized the data. Y.W., L.A.G., I.R.H., S.B.P. and J.D.M wrote
440 the manuscript. D.B., N.S.B., S.B.P. and J.D.M. provided supervision. Q.C., D.B., N.S.B. and
441 J.D.M funded the study. All authors contributed to manuscript editing and support the
442 conclusions.

443

444 **Declaration of interests**

445 The authors declare no competing interests.

446 **Figure Legends**

447 Figure 1. Phylogenetic placement and genome sequencing of newly isolated Saccharibacteria
448 strains *S. epibionticum* ML1 (*Se*) and *N. lyticus* ML1 (*Nl*). (A) Maximum growth (fold change)
449 achieved by *Se* and *N*l during co-culture with compatible host species *A. israelii* and
450 *Propionibacterium propionicum* respectively, and population change (growth or death) detected
451 at equivalent timepoints with an incompatible host. (B) Phylogeny of CPR and other bacteria
452 based on concatenated ribosomal proteins, with the placement of the Saccharibacteria phylum
453 indicated. Figure adapted from Castelle *et al.*[6] (C) Phylogeny constructed using 49 core,
454 universal genes indicating placement of *Se* and *N. l* (blue text) within Saccharibacteria. Family
455 names (as designated by the Genome Taxonomy Database) and groups previously designated by
456 McLean *et al.* (G1, etc.) are indicated for each clade[11]. HOT, human oral taxon. (D, E) Overview
457 of the genome sequences of *S. epibionticum* ML1 and *N. lyticus* ML1.

458

459 Figure 2. Harnessing natural transformation to generate mutant Saccharibacteria. (A) Schematic
460 depicting the intergenic neutral site (blue, NS1) targeted for insertion of a hygomycin resistance
461 cassette (yellow) in the *Se* genome and the linear DNA fragment employed in transformation
462 experiments. Primer binding sites used for genotyping are indicated (sites 1-3). (B) Overview of
463 the *Se* transformation protocol. After incubation with linear DNA, *Se* + *Ai* co-cultures are
464 enlarged concomitant with hygromycin addition and serially passage with addition of naïve host
465 at each dilution to promote *Se* growth (grey box). Clonal transformed *Se* populations were
466 obtained by plating to isolate single colonies of *Ai* with accompanying *Se* cells, followed by
467 growth in liquid culture, with additional *Ai,* to promote *Se* population expansion. (C) PCR-based
468 genotyping of *Se* clones obtained following transformation according to the protocol show in (B)

469 in the presence (right) or absence (left) of selection with hygromycin during the expansion and
470 passaging steps. Binding sites for primers targeting NS1 (1, 2) and *hph* (3) are shown in (A).
471 Positive control primers (*Se*) target a locus distant from NS1. (D) Growth (red) and percent of *Se*
472 transformed (grey) over the course of transformation protocol depicted in (B), in the presence
473 (squares) or absence (circles) of selection with hygromycin. (E) Luminescence production from
474 *Se–Ai* co-cultures (left) or co-culture filtrates (right) in which *Se* contains a nanoluciferase
475 expression cassette inserted at NS1 (shown at bottom). (F) Fluorescence and phase contrast
476 micrographs of *Se–Ai* co-cultures in which *Se* carries an *mcherry* (top) or *sfgfp* (bottom)
477 expression cassette inserted at NS1. See also Figure S2. Data in (E) represent mean ± s.d.
478 Asterisks indicate statistically significant differences (unpaired two-tailed student's t-test;
479 *$p<0.05$).
480
481 Figure 3. Fluorescent protein expression and quantitative microscopy enable tracking of the *S.*
482 *epibionticum* lifecycle. (A) and (D) Snapshots captured at the indicated time points from
483 timelapse fluorescence and phase contrast microscopy of GFP-expressing *Se* grown in co-culture
484 with *Ai.* Arrows indicate *Se* cells exhibiting productive (pink, purple) and non-productive (blue)
485 interactions with *Ai* cells. White outlines in the fluorescent channel indicate depict an *Ai* cell
486 affected by *Se* infection (*Ai* 1). (B) and (E) Omnipose-generated segmentation of *Se* and *Ai* cells
487 depicted in (A) and (D), at the start (left) and end (right) of the 20 or 22 hr growth period. (C)
488 and (F) Growth of individual *Ai* cells as impacted by productive (light grey) or non-productive
489 *Se* cells (black, dark grey). Colors correspond to cell masks shown in (B) and (E). For the full
490 time course captured in (A) and (D), see Videos S1 and S2, respectively. For additional examples
491 of tracked *Se–Ai* growth, see Videos S3-S11.

492

493 Figure 4. Identification of genes important for fitness of *S. epibionticum* during co-culture with
494 *Ai* identified by Tn-seq. (A) Overview of normalized transposon insertion frequency across the
495 *Se* genome detected in input DNA used for mutagenesis (dark grey), and from samples collected
496 immediately following transformation (T0, dark blue) and subsequent outgrowth time points
497 (T1-T3, shades of blue). Genes encoding proteins belonging to CPR-enriched protein families
498 (light green) and those found to be significantly important for fitness across all time points and
499 by two different metrics (dark green). The location of the arginine deiminase system (ADS)
500 genes and two loci containing T4P genes ($T4P_1$, $T4P_2$) are indicated in the outer circle. (B)
501 Schematic depicting $T4P_1$ and flanking genes (separated by vertical lines) and the relative fitness
502 contribution of each gene as determined by the TRANSIT resampling algorithm (adjusted p-
503 values) and ALDEx2 (*delta* scores). Asterisks indicated genes found to be significant by both
504 metrics across all time points. Gene annotations were derived from Foldseek queries using AF
505 models generated for each gene product (see methods). (C) *Se* population levels detected in *Se*–
506 *Ai* cocultures following transformation with constructs designed to replace the indicated genes
507 with *hph.* Candidate essential genes tested were considered significant for *Se* fitness by
508 TRANSIT and ALDEx2 at a minimum of 2 of 3 time points (see Table S1). (D, E) Total *Se*
509 population (D) and proportion transformed (E) following transformation with an unmarked
510 cassette targeted to NS1 in the indicated strains of *Se.* See also Figures S3 and S4, Table S1.
511 Data in (C-E) represent mean ± s.d. Asterisks indicate statistically significant differences (C,
512 one-way ANOVA followed by Dunnett's compared to no DNA control; E, unpaired two-tailed
513 student's t-test; *$p<0.05$, ns, not significant).

514

515 Figure 5. Inclusion of extensive metagenomic data in MSAs enables proteome-wide AF
516 modeling of *Se* protein structures. (A) Histograms depicting MSA depths obtained for *Se*
517 proteins using HHblits. (B) Maximum depths obtained for *Se* protein MSAs that initially
518 contained <500 sequences. Additional sequences were sourced from metagenomic sequence
519 databases and incorporated into MSAs using Jackhmmer or Phmmer (see Methods). (C)
520 Comparison of the AF confidence metric (pLDDT) determined using Hhblits or
521 Jackhmmer/Phmmer (Metagenome)-generated MSAs for *Se* proteins with initially shallow
522 MSAs (<500). *Se* proteins shown in (D) and (E) are highlighted in blue. (D and E) Example *Se*
523 protein structure models and associated predicted alignment matrices obtained using shallow
524 (right) or metagenomic sequence-improved (left) MSAs. *Se* proteins models (blue) are aligned to
525 models from top FoldSeek (FS) hits (light grey, A0A2H0BDY7 (D), A0A8B1YPH4,
526 metagenome and A0A3D0YBM7, shallow, (E)), when available. The annotation in (D) derives
527 from the best FS hit; in (E), the best FS hit is an uncharacterized protein. The function for this
528 protein was assigned using DPAM. Some structures are trimmed to highlight the alignment. See
529 also Figure S5, Table S2.
530

531 **Methods**

532 **Strains, media and growth conditions**

533 Saccharibacteria strains employed in this work include *Southlakia epibionticum* ML1 and
534 *Nanosynbacter lyticus* ML1, both isolated in this study. Host bacterial strains used include
535 *Actinomyces israelii* F0345, *A. odontolyticus* F0309*, A. urogenitalis* S6-C4*, Actinomyces sp.*
536 F0386 and *Propionibacterium propionicum* F0230 (*Pp*)[51-53]. Host bacteria mono-cultures, *Se-Ai*
537 co-cultures and *Nl-Pp* co-cultures were routinely grown statically under an atmosphere of
538 ambient air supplemented with 5% CO_2 at 37°C or anaerobically with shaking at 37°C using the
539 GasPak EZ Anaerobe Container System with Indicator (BD 260626 and 260001) in TSY media
540 (30 g/L tryptic soy and 5 g/L yeast extract) or TSYR media (TSY media supplemented with 10
541 mM arginine), or anaerobically on TSBY agar plates (30 g/L tryptic soy, 5 g/L yeast extract, 15
542 g/L agar, supplemented with 5% (v/v) horse blood). For selection of hygromycin-resistant
543 Saccharibacteria, hygromycin was used at 150 μg/mL. Saccharibacteria-host co-cultures were
544 stored at -80°C in TSY or TSYR supplemented with 10% (v/v) Dimethylsulfoxide (DMSO).
545

546 **Isolation of Saccharibacteria strains *Se* ML1 and *Nl* ML1**

547 Isolation of Saccharibacteria strains in co-culture with host bacteria was carried out as
548 essentially as previously described[44,54]. Anonymous volunteers aged over 18 years provided oral
549 samples for Saccharibacteria isolation. Supragingival plaque samples were collected with
550 toothpicks and dispersed in 1 mL of Maximum Recovery Diluent (MRD; Peptone 1.0 g/L,
551 Sodium Chloride 8.5 g/L, pH 7.0) buffer. 5 mL of saliva was collected by voluntary
552 expectoration into sterile 50 mL conical tubes. All saliva samples and plaque samples were then
553 pooled into 20 mL MRD. Pooled samples were then vigorously resuspended by vortexing and

554 filtered with a 0.22 μm filter, and the flowthrough was collected. Residual bacteria in the sample
555 tube and filter were collected by washing once with 10 mL MRD, filtered again, and combined
556 with the previous flowthrough. Saccharibacteria present in filtrates were pelleted by centrifuging
557 at 60,000 rcf for 1 hr at 4°C. Supernatant was removed, and pellets were resuspended in 1 mL
558 MRD. The presence of Saccharibacteria in these samples were confirmed by PCR, using phylum
559 specific primers[54] (Table S3). The collected Saccharibacteria were then added to mono-cultures
560 of a panel of five potential host species (*A. odontolyticus, A. urogenitalis, Actinomyces sp.* F0386,
561 *Ai,* and *Pp*) and the cultures were passaged every 24 or 48 hr. The presence of Saccharibacteria
562 cells in the final cultures was confirmed by qPCR with universal Saccharibacteria primers (Table
563 S3) and microscopic imaging. The cultures were then streaked on TSY or TSBY agar, isolated
564 colonies of host bacteria were tested for Saccharibacteria by PCR, and positive colonies were re-
565 cultured in liquid medium and stored as clonal co-cultures.
566 To create purified suspensions of Saccharibacteria from co-cultures, the co-culture was
567 centrifuged at 3,000 rcf for 5 min at room temperature, to pellet host cells. Supernatant
568 containing suspended Saccharibacteria was then collected and filtered twice using a pre-
569 sterilized 5 μm mixed cellulose esters (MCE) filter. Filtrate was centrifuged at 15,000 rcf for 30
570 min at room temperature to pellet Saccharibacteria. The resulting pellet was resuspended in a
571 small volume of TSYR supplemented with 10% (v/v) DMSO and stored at -80°C.
572
573 **Saccharibacteria host compatibility testing**
574 Purified populations of *Se* and *Nl* for host compatibility testing were generated by
575 growing 100 ml co-cultures of each with their respective host strains according to the standard
576 protocol described above. Co-cultures were passed through a 0.45 μm SFCA filter to remove

host cells then spun at 80,000 rcf for 20 min to pellet Saccharibacteria. Supernatant was removed and cell pellets were resuspended in 1 mL fresh medium. Purified Saccharibacteria cells were then add to OD_{600} = 0.2 cultures of *Ai* and *Pp* (compatible host for *Nl*) at an MOI of 0.2. Cultures were then incubated statically at 37°C under ambient air enriched with 5% CO_2 for up to 72 hr, and populations of *Se* and *Nl* were monitored over time using qPCR.

Whole genome sequencing

Genomic DNA of *Se* ML1, *Nl* ML1and *Ai* F0345 was isolated using the Wizard HMW DNA Extraction kit (Promega). Sequencing was performed on Illumina iSeq and MiSeq and Oxford Nanopore (ONT) MinION instruments after standard sequencing library preparation protocols (Illumina and Oxford Nanopore). De-novo assemblies were generated using the Trycyler pipeline[55]. Specifically, ONT long reads were filtered using Filtlong v0.2.1 (https://github.com/rrwick/Filtlong) with Illumina reference reads, using --keep_percent 95 to retain approximately 95% of the reads. Long reads were then subsampled into 12 bins and assembled into 12 assemblies using the Flye v2.9 (https://github.com/fenderglass/Flye/releases/tag/2.9. Accessed 4 October 2021), Raven v1.8.1[56] Miniasm v0.3r179[57] and Minipolish v0.1.3 (https://github.com/rrwick/Minipolish/blob/main/miniasm_and_minipolish.sh) assemblers. Assembly contigs were manually curated and then reconciled using Trycycler v0.5.2[55]. A consensus assembly for each bacterium was generated and then polished with Illumina short reads using Polypolish v0.5.0[55] and POLCA from MaSuRCA v4.0.9[58]. Initial annotations were generated using PROKKA v1.14.5[59].

599 To sequence clonal transformed *Se* (see Results and below), genomic DNA was isolated
600 from frozen pellets of purified *Se* by re-suspending in Buffer PB (Qiagen) to a total volume of
601 500 µL, adding 250 µL of 0.1mm zirconia/silica beads (BioSpec Catalog # 11079101z), 250 µL
602 of 20% SDS, and 550 µL of phenol:chloroform:IAA (25:24:1) (Invitrogen Catalog #15593-031),
603 and bead-beating in a Mini-BeadBeater-16 (Biospec Model 607) with settings 3450 RPM, 115V,
604 10A, and ½ HP, for four 30-second intervals, each followed by cooling on ice for 1 minute.
605 Purification of the DNA was performed by applying the aqueous phase directly to a DNeasy
606 Blood & Tissue Prep Kit (Qiagen) purification column and following the recommended protocol
607 for washing and elution. Sequencing was performed on an Illumina iSeq using standard library
608 preparation protocols (Illumina). Reads were mapped to the assembled *Se* ML1 genome using
609 minimap2 and variants were called using LoFreq v2.
610
611 **Phylogenetic analysis**
612 The *Se* and *Nl* ML1 genomes were phylogenetically placed using whole genome
613 information. A genome tree was generated from these newly isolated strain genomes with a
614 manually curated set of high-quality CPR genomes (complete and partial) and metagenome
615 assembled genomes (MAGS) to remove any contaminants deposited with the original assemblies
616 as described previously[11]. This species tree was constructed using a set of 49 core, universal
617 genes defined by COG (Clusters of Orthologous Groups) gene families with KBase[60]. It
618 combines the curated genome(s) and a set of closely related genomes selected from the public
619 KBase genomes imported from RefSeq. Genome(s) are inserted into curated multiple sequence
620 alignment (MSA) for each COG family. The curated alignments are trimmed using GBLOCKS

621 to remove poorly aligned sections of the MSA. Then, the MSAs are concatenated, and a
622 phylogenetic tree is reconstructed using FastTree2 with the -fastest setting[61].
623
624 **Measuring hygromycin sensitivities of *Ai* and *Se***
625 Hygromycin sensitivity of *Ai* and *Se* was measured in liquid cocultures. Duplicate co-
626 cultures were initiated by mixing purified *Se* with *Ai* at an OD_{600} of 0.2 in TSY and at an
627 approximate cellular ratio of 1:2 (*Se*:*Ai*). The co-cultures were divided into multiple aliquots in
628 96-well culture plates and hygromycin was added to the final concentrations shown in Figure S1.
629 The plates were covered with Breath-Easy sealing membrane (Sigma Z380059) and incubated
630 without lids at 37°C with 5% CO_2. At 1 day and 3 days, the cells within individual wells were
631 pelleted (>15 min at >15,000 rcf) and stored at -20°C for later genomic DNA isolation and
632 qPCR-based quantification of *Se* and *Ai* (see below).
633
634 **Design and generation of cassettes for heterologous gene expression in *Se* and *Nl***
635 Cassettes for heterologous gene expression in *Se* and *Nl* were designed by appending
636 promoter and terminator sequences from the TM7x genome[8] to the 5' and 3' ends, respectively,
637 of ORFs codon optimized for *Se*. The promoter and terminator elements were sourced from
638 TM7x rather than *Se* to reduce the likelihood of off-target integration at corresponding *Se* loci.
639 The TM7x promoters were chosen from genes expected to be highly and constitutively
640 expressed, *tuf* and *rpsJ*. *Se* codon usage was calculated using the Dynamic Codon Biaser
641 (DCB)[62]. Heterologous ORFs were optimized to match relative codon usage frequencies found in
642 *Se,* but omitting codons with less than 10% usage. Heterologous genes utilized included *hph*
643 (hygromycin B phosphotransferase) from *Streptomyces hygroscopicus*, superfolder GFP

644 (www.fpbase.org/protein/superfolder-gfp/), mCherry2 (www.fpbase.org/protein/mcherry2/) and
645 NanoLuc Luciferase[63,64]. Table S4 reports the composition and complete sequences of the
646 designed cassettes. Cassettes were obtained as gBlocks from Integrated DNA Technologies, Inc
647 (IDT).

648 Linear fragments used for transformations were generated by adding *Se* or *Nl* genomic
649 sequences corresponding to the targeted insertion or allelic replacement sites to the left and right
650 sides of a heterologous gene expression cassette or of two cassettes joined together. Overlap
651 extension PCR was used to join gBlock cassettes and genomic fragments, the latter of which
652 were individually amplified from *Se* or *Nl* genomic DNA. In some cases, complete fragments
653 including genomic sequences were obtained as gBlocks from IDT. Fragments used for flank-
654 length tests were generated by amplification using larger fragments as templates followed by gel-
655 purification. Table S4 reports the composition and complete sequences of the fragments utilized
656 and primers used are listed in Table S3.

657

658 **Genetic transformation of *Se* and *Nl* with targeted insertion and allelic replacement**
659 **constructs**

660 To prepare *Se-Ai* or *Nl-Pp* co-cultures for transformation, 1-mL aliquots of a previously
661 frozen co-culture (see methods on co-culturing) were thawed on ice and added to 9 mL of TSY
662 supplemented with freshly cultured *Ai* to a final OD_{600} of 0.2, incubated statically for 2 d at 37°C
663 in ambient air enriched with 5% CO_2, then enlarged by the addition of 10 volumes of TSY and
664 incubated anaerobically using the GasPak EZ Anaerobe Container System with Indicator (BD
665 260626 and 260001) at 37°C for 2 days with shaking at 160 rpm.

666 For transformation of *Se* or *Nl* with linear targeted insertion constructs, 0.22 - 0.3 mL
667 aliquots of prepared co-culture were incubated statically with transforming DNA for 6 h at 37°C
668 in ambient air enriched with 5% CO_2 in culture tubes. The mixtures were subsequently enlarged
669 by the addition of TSYR to approximately 5 mL and supplemented with recently passaged *Ai*
670 (for *Se*) or *Pp* (for *Nl*) to a final OD_{600} of approximately 0.06. This time point was designated
671 "TF" (time zero after transformation). After removal of a sample for later qPCR analysis, the
672 cultures were divided and one half was supplemented with hygromycin to 150 μg/ml. The
673 cultures were then incubated as above for an additional 2-3 days. This time point was designated
674 "P1" (end of initial passage). Cultures were subsequently serially passaged as many as four times
675 by five-fold dilution into fresh TSYR supplemented with *Ai* (for *Se*) or *Pp* (for *Nl*) to a final
676 OD_{600} of 0.06 and, for the hygromycin containing cultures, with additional hygromycin to 150
677 μg/ml. These passages (designated "P2", "P3", etc.) were each incubated for 2-3 days at 37°C in
678 ambient air enriched with 5% CO_2 and without agitation. Cultures were sampled at each passage
679 by removing and pelleting (>15 min at >15,000 rcf) of 0.1 to 1 ml, and the pellets were stored at
680 -20°C for later qPCR analysis.
681
682 **qPCR assays**
683 We employed qPCR for quantification of Saccharibacteria and host populations in
684 cocultures, as well as to monitor *Se* transformation. For these assays, genomic DNA was isolated
685 from frozen co-culture pellets using Instagene matrix (Bio-Rad) and quantitative PCR (qPCR)
686 was performed using a CFX Connect Real-Time PCR Detection System (Bio-Rad). To quantify
687 *Se* and *Ai*, amplification employed primers targeting *Se uvrB* or the 16S rRNA gene of *Ai* and
688 was performed in 20 μL reactions with 1X SsoAdvanced Universal SYBR Green Supermix (Bio-

689 Rad), 300 nM each primer and 4 µL template DNA (primer sequences provided in Table S3).
690 Thermocycling conditions were 95°C for 5 min followed by 35-40 cycles of 95°C for 20 s, 60°C
691 for 30 s and a read of fluorescence. To quantify transformed *Se* (insertions at NS1), amplification
692 employed primer pairs targeting the insertion sequence and sequence adjacent to NS1 but outside
693 of the linear transformation construct arms and was performed in 20 µL reactions with 1X
694 Phusion HF Buffer, 0.2 mM dNTPs, 250 nM each primer, 0.5X SYBR Green I (ThermoFisher),
695 4 µL template DNA and 0.4 units of Phusion High-Fidelity DNA Polymerase (NEB) (primer
696 sequences provided in Table S3). Thermocycling conditions were 98°C for 90 s followed by 35-
697 40 cycles of 98°C for 15 s, 70°C for 20 s, 72°C for 90 s and a read of fluorescence. Melt curves
698 were performed following each amplification to verify product homogeneity. Absolute target
699 sequence abundance was determined by comparison to duplicate standard curve reactions
700 performed in parallel with each assay. The standard curve templates were generated by serial
701 dilution of previously amplified and gel-purified products quantified by Qubit (ThermoFisher).
702

703 **Isolation of isogenic mutant co-cultures**

704 To obtain co-cultures with isogenic mutant (transformed) *Se*, transformation mixtures
705 grown for three passages under hygromycin selection were serially diluted, plated on TSBY agar
706 without antibiotic and incubated for 7 days anaerobically. Isolated colonies (representing *Ai*
707 potentially colonized with *Se*) were picked into 0.2 mL TSYR supplemented with *Ai* at a final
708 OD_{600} of 0.2 and incubated statically at 37C in ambient air enriched with 5% CO_2 for 2-4 days,
709 then screened for the presence of *Se* and for the mutant (genomic insertion) and wild-type (no
710 insertion) alleles by PCR (e.g., Figure 2A, primer sequences in Table S3). Cultures containing

711 pure mutant *Se* were frozen and/or further propagated for purification of the *Se* for WGS (see
712 above).
713
714 **Nanoluciferase assay**
715 Nanoluciferase assays were performed using isogenic wild-type *Se-Ai* co-cultures and *Se*
716 NS1::*nluc-hph2-Ai* co-cultures grown statically in ambient air supplemented with 5% CO_2 at
717 37°C in TSY media for three passages as described above. To separate *Se* from host cells, 20 mL
718 of each co-culture was passed through a 0.45 μM SCFM filter and the resulting filtrate was spun
719 at 80,000 rcf for 30 min to pellet *Se*, supernatant was removed, and the pelleted *Se* was
720 resuspended in 750 μL TSY. 100 μL of these purified *Se* cells or 100 μL of the corresponding
721 *Se-Ai* cocultures were mixed with 100 μL Nano-Glo Luciferase assay reagent (50:1 mixture of
722 substrate:buffer, Promega N1110) in a 96-well plate. Luminescence signal indicative of
723 nanoluciferase activity was detected using a Cytation 2 plate reader. Luminescence signal was
724 later normalized by *Se* abundance in co-cultures and filtrates as measured by qPCR using *Se*-
725 specific primers as described above.
726
727 **Microscopy**
728 Imaging was performed on a Nikon Eclipse Ti-E wide-field epi-fluorescence microscope,
729 equipped with a sCMOS camera (Hamamatsu) and X-cite LED for fluorescence imaging. We
730 imaged through a Nikon Plan Apo λ 60X 1.4 NA oil-immersion Ph3 objective. The microscope
731 was controlled by NIS-Elements 3.30.02. Isogenic *Se* NS1::*mcherry-hph2-Ai* co-cultures and *Se*
732 NS1::*sfgfp-hph2-Ai* co-cultures were grown statically in ambient air supplemented with 5% CO_2
733 at 37°C in TSYR media as described above. Cell samples were spotted on a 3% (w/v) agarose

734 pad made with TSYR media supplemented with 0.4% glucose placed on a microscope slide. The
735 microscope chamber was heated to 37°C for time-lapse experiments. Time-lapse images were
736 aligned and segmented with Omnipose using the phase contrast channel and the
737 published bact_phase_omni model[65]. Masks were manually linked and corrected for
738 segmentation errors in regions of *Ai* cell overlap using Napari. Regions corresponding to *Se* were
739 also removed to accurately track host *Ai* growth alone. For visualization purposes, fluorescence
740 intensity was gamma-corrected to normalize background levels on a frame-by-frame basis while
741 not distorting the *Se* signal. No bleaching correction was implemented. Figure 3A-F and Videos
742 S1-11 were generated using Python and annotated in Adobe Premiere.
743
744 **Mutagenesis of Sac1a genomic DNA and transposon mutant library generation**
745 The transposon used for *in vitro* transposition (here named T36) was generated by
746 amplification of the *hph1* insertion cassette using the 5'-phosphorylated primers T36-ampF and
747 T36-ampR, which add required 19-bp Tn*5* mosaic end sequences to each end of the amplicon,
748 followed by purification using the Qiagen PCR Purification Kit with elution in TE (primer
749 sequences provided in Table S3). *In vitro* transposition was performed in multiple 45-µl
750 reactions containing 1.8 µg *Se* genomic DNA, 290 ng purified T36, 1X EZ-Tn5 Reaction Buffer
751 and 1.2 U EZ-Tn*5* Transposase (Biosearch Technologies) with incubation for 2 hr at 37 °C. The
752 reactions were stopped, DNA was ethanol precipitated, and gaps were repaired as described[66].
753 Following gap repair, the DNA was purified using Qiagen PCR Purification Kit with elution in
754 water after the columns were washed twice with Buffer PE.
755 Transformation of *Se* to generate a transposon mutant library was performed similarly to
756 the procedure described above for transformation with targeted insertion fragments, but at a

757 larger scale and with modifications. Specifically, the initial 6 hr incubation contained 4.0 μg of
758 *in vitro* transposon-mutagenized genomic DNA and 15 mL of *Ai*-*Se* co-culture. This co- culture
759 was enlarged by adding 135 mL of TSYR supplemented with *Ai*, then incubated for two
760 additional hours before hygromycin addition. Concomitantly with hygromycin addition, $MgCl_2$
761 (to 1 mM) and benzonase (Sigma) (to 25 U/mL) were added to degrade extracellular DNA, and
762 the cultures were similarly supplemented with benzonase approximately every 24 h during
763 passaging. After the initial 2-day growth passage (ending at time point T0), the culture was
764 serially passaged four additional times by dilution of 30 mL culture into 120 mL in TSYR
765 supplemented with *Ai* at OD_{600} of 0.067. These passages were each incubated statically for 2
766 days at 37°C in ambient air enriched with 5% CO_2. Samples for Tn-seq analysis were taken at
767 the end of the initial passage (60 mL, T0) and at the end passages 1, 2 and 4 (80 mL, T1-T3).
768 These samples were treated with additional benzonase (50 U/mL) for 30 min, then EDTA was
769 added to 15 mM and pellets were collected by centrifugation for 30 min at 15,000 rcf and stored
770 at -80°C.
771
772 **Tn-seq library preparation and sequencing**
773 Genomic DNA was isolated from frozen co-culture pellets using the bead-beating method
774 described above (section on whole genome sequencing). Tn-seq libraries were prepared by the
775 C-Tailing method as described[67]. The transposon-specific primers utilized are listed in Table S3.
776 Sequencing was performed in multiplex as 50-bp single-end reads on an Illumina MiniSeq with
777 25-40% PhiX spike-in.
778 Custom scripts (https://github.com/lg9/Tn-seq[66,67]) were used to process the Illumina
779 reads. In brief, reads were first filtered for those displaying transposon end sequence as their

780 initial bases (the sequencing primer was designed to anneal six bases from the end of the
781 transposon). These reads were then mapped to the *Se* genome after removing the transposon end
782 sequences. Reads per unique mapping position and orientation were tallied and read counts per
783 gene were calculated by summing reads from all unique sites within a given gene.
784
785 ***Se* gene fitness and essentiality analysis**
786 We first checked the saturation level (i.e., number of insertions detected) of the Tn-seq
787 library sampled along four time points (T0-T3). Because the T0 sample had the highest number
788 of unique insertions (9,996), we initially focused on this time point to identify essential genes in
789 the *Se* genome. Gene essentiality was determined based on the number of insertions detected in
790 each gene (using the GUMBEL method, part of the TRANSIT tool), or alternatively by taking
791 into account the read counts per insertion(s) in each gene (using the Hidden Markov Model
792 method, HMM, also part of TRANSIT)[26]. Default parameters were used for both GUMBEL and
793 HMM analyses. To complement these analyses and to identify genes whose disruption was
794 deleterious for *Se* in later time points, we used a comparative approach with respect to T0. The
795 comparative approach was necessary due to the significant reduction in the number of unique
796 insertions detected at T1-T3 (the average number of insertions for these time points was 1,765),
797 which suggested a strong selection pressure over time. Thus, the distribution of insertions at T1,
798 T2 or T3 were independently compared with T0. We detected genes associated with
799 differentially abundant insertions at T1-T3 vs. T0 using TRANSIT and ALDEx2[26,27]. For the
800 TRANSIT-based analysis, we used the Resampling method to identify differentially abundant
801 genes between the two time points being compared. Genes with TRANSIT-estimated adjusted p-
802 value < 0.05 were considered as differentially abundant (this gene set would include under- and

803 over-represented genes at the late time points with respect to T0). ALDEx2 takes into account
804 the compositional nature of sequencing data generated during selection assays (i.e., competitive
805 growth mutant assays)[27]. Given the lack of replicates (expected in the standard ALDEx2
806 pipeline) in our data, we adapted the ALDEx2 workflow by using the aldex.clr function available
807 in the ALDEx2 R package to generate Dirichlet distributions of centered-log ratio (CLR)
808 transformed abundances for any pair of samples being compared (e.g., T1 vs. T0)[27]. We used
809 1,000 Monte Carlo instances to generate the Dirichlet distributions. Then, we computed the
810 median CLR value per gene for each sample. Finally, we computed a delta score similar to
811 ALDEx2 original "dif.btw" by computing gene wise median difference between the two samples
812 being compared (e.g., T1-T0). Genes with delta scores < -1.5 were considered under-represented
813 in the late time points (with respect to T0), suggesting that insertions that disrupt those genes
814 were deleterious for *Se*. To define a meaningful threshold for our delta score and confirm the
815 performance of our ALDEx2-based analysis, we leveraged barcode sequencing data we
816 previously generated for competition growth assays of an antibiotic-treated *Escherichia coli*
817 pooled single-gene deletion library[68]. This benchmarking dataset included four time points (T0-
818 T3), each one with four replicates. For the T1 vs. T0 and T2 vs. T0 comparisons we applied our
819 strategy for each individual replicate and then evaluated the overlap with the original results
820 obtained when considering all four replicates. We found that indeed the set of genes with
821 absolute delta score > 1.5 for single replicate analyses significantly overlap with the set of genes
822 in the analysis that included all replicates. Three out of the four single replicate analyses had
823 hypergeometric test p-values < 0.05 for T1 vs. T0 (two with p-values < 0.005). Moreover, all
824 single replicate analyses for T2 vs T0 had hypergeometric test p-values < 1E-202.
825

826 **Transformations to validate gene essentiality predictions**

827 To validate Tn-seq essentiality predictions, we designed constructs to replace target genes
828 with *hph.* Each construct contained 335-bp sequences targeting the desired site flanking the *hph*
829 expression cassette described above. 100 ng synthesize linear DNA or water were added to 300
830 μL wild-type *Se-Ai* co-cultures (4 biological replicates per condition) prepared according to our
831 standard protocol, and transformation and subsequent passaging were performed using our
832 general transformation protocol described above. At the end of passage 3, 1 mL of each
833 transformed co-culture was treated with 1 mM $MgCl_2$ and 25 U/mL benzonase for 30 min at
834 37°C to remove extracellular DNA. 15 mM EDTA was then added to stop the benzonase
835 reaction, samples were centrifuged at 21,000 rcf for 1 hr and the resulting pellets were stored at -
836 80°C. *Se* abundance was then measured by qPCR as described above.

837

838 **Generation and transformation efficiency of *Se* Δ*comEC::hph***

839 A clonal population of *Se* Δ*comEC::hph* was generated by replacing *comEC* with our
840 hygromycin resistance cassette following the transformation and clonal mutant isolation
841 protocols described above. To assess *Se* Δ*comEC::hph* transformability, 50 ng of synthesized
842 linear gene product for inserting unmarked *sfGFP* at NS1 (with 545 bp flanking regions) was
843 used to transform 150 μL isogenic wild-type or Δ*comEC Se-Ai* co-cultures (4 biological
844 replicates per condition). After 6.5 h of incubation between DNA and co-cultures at 37°C, the
845 entire transformation culture was treated with 1 mM $MgCl_2$ and 25 U/mL benzonase for 30 min
846 at 37°C to degrade extracellular DNA. 15 mM EDTA was added to stop the benzonase reaction
847 and the samples were centrifuged at 21,000 rcf for 1 hr and the resulting pellets were stored at -

848 80°C. We then used qPCR to measure the abundance of transformed and total *Se* using NS1-
849 sfGFP or *uvrB*- specific primers as described above.
850
851 **CPR-enriched protein families analysis**
852 To assess the protein family distribution of the *Se* and *Nl* genomes, we collected amino
853 acid sequences for 22,977 protein clusters (protein families) from Meheust *et al*.[5]. These protein
854 families were built through their occurrences across at least 5 distinct non-redundant or draft
855 CPR bacteria, non-CPR bacteria and Archaeal genomes. The presence or absence of any given
856 family in *Se* or *Nl* were determined through all-vs-all protein sequence search of every protein in
857 the genome against the database of protein family sequences. Protein sequence search was
858 performed by using the MMseqs2 (version: 14-7e284) algorithm with the following parameters:
859 module: easy-search; sensitivity (-s): 7.5; alignment coverage (-c): 0.5; greedy-best-hits: 1. The
860 resulting presence or absence matrix was combined with the core family matrix of 921 protein
861 families across 2890 genomes (collected from Meheust *et al*.[5]) for clustering analysis. A Jaccard
862 distance based on the presence or absence of core families across *Se*, *Nl* and 2890 other genomes
863 was calculated in R by using proxyC package (version: 0.3.3). Agglomerative hierarchical
864 clustering was performed by using cluster package (version: 2.1.4) in R with the "wand" method.
865 A hierarchical clustering heatmap was built using complexHeatmap package in R (version:
866 2.14.0) by plotting presence (black) or absence (white) of protein families (columns) in a given
867 genome (rows). Enrichment or depletion of core protein family clusters in *Se* and *Nl* was
868 performed using Fisher's exact test in R with input table of presence or absence of protein
869 families in the *Se* and *Nl* genomes. A threshold of Benjamini-Hochberg corrected p-values of <
870 5e-05 was used to determine the enrichment or depletion.

871

872 **Sequence-based gene annotation of the *Se* genome**

873 We used ProtNLM (https://github.com/google-research/google-
874 research/tree/master/protnlm) to predict the function annotations of proteins in the *Se* proteome
875 from sequence[30]. A prediction of score ≥0.5 corresponds to >70% accuracy for proteins without
876 close UniRef50 matches and ~95% accuracy for proteins with close matches. We thus
877 considered the 337 proteins with score ≥0.5 to have confident annotations (Table S2).
878 Additionally, we mapped each protein to Pfam domains (Nov, 2021 release v35.0[46]) using
879 hmmscan[31]. We filtered the hit Pfam domains by full sequence E-value ≤ 1e-3 and domain-based
880 CE-value ≤ 1e-5 or CI-value ≤ 1e-5, domain coverage ≥ 50% (coverage of aligned portion of
881 domain to full length domain) and selected the top hit for each aligned region.

882

883 **Generation of multiple sequence alignments**

884 To construct MSAs for each protein in the *Se* proteome, we initially used HHblits[35] to
885 search against UniRef (2022 release[37]) and BFD (2019 release[36]). For those MSA that contained
886 <500 sequences after this approach, we conducted extensive homology searches against
887 metagenome databases similar to Anishchenko *et al.*[69]. We first converted the initial HHblits (E-
888 value ≤ 1e-3) MSAs for each protein into a hidden Markov model (HMM) which we used as a
889 seed sequence profile to search against metagenomic and metatranscriptomic sequences from
890 JGI[70], MGnify[71], and UniRef with hmmsearch[31], respectively. Hits from JGI, MGnify, and
891 UniRef were gathered and aligned to the query protein by phmmer and jackhammer[31], to retain
892 sequences with E-value ≤ 1e-5 at each stage. We removed columns containing gaps in the query
893 sequence from the MSAs, which were then subjected to redundancy filtering at 95% sequence

894 identity and 50% sequence coverage. We selected the deepest MSA for each protein (from
895 HHblits against UniRef+BFD or phmmer/jackhmmer against metagenome) for downstream
896 analysis.
897
898 **Structure-based annotations**
899 Using MSAs described above, we computed AlphaFold (AF) models (model_1 without
900 structural template search, 10 recycles, and version 1.0 weights) for 852 of the 855 proteins (3
901 were excluded because of MSA depth or protein size)[34]. We then applied FoldSeek[38] against
902 AFdatabase50 (AF Protein Structure Database clustered at 50% sequence identity and 80%
903 coverage). We report the top hits in Table S2 based on E-value ≤ 0.00001, TM-alignment score
904 ≥ 0.5 (which correlates well to fold-family level similarity[51]), and query coverage and target
905 coverage ≥ 50%. To classify proteins into evolutionary contexts based on structure and sequence
906 similarity, we used Domain Parser for AlphaFold Models (DPAM)[42], and the full results are
907 available at https://conglab.swmed.edu/ECOD_Se/Se_ECOD.html.
908
909 **Nucleotide sequence accession numbers.**
910 The complete genome sequences of *Se* ML1, *Nl* ML1and *Ai* F0345 have been deposited
911 in GenBank under BioProject PRJNA957798 with accession numbers SAMN34266291,
912 SAMN34266292 and SAMN34266293, respectively. Tn-seq data generated in this study have
913 been deposited in the Sequence Reads Archive (SRA, BioProject PRJNA957798).

914 **Supplemental information**

915 Figure S1. Hygromycin resistance phenotypes of *Se* and *Ai* enable use of *hph* as a selectable
916 marker in *Se.* Growth of *Ai* (top) and *Se* (bottom) in co-cultures containing the indicated
917 concentrations of hygromycin.

918

919 Figure S2. Optimization of the *Se* transformation protocol and transformation of a second
920 Saccharibacteria species. (A, B) Quantification of *Se, Ai* and transformed *Se* populations over the
921 course of our transformation protocol (See Figure 2B) with varying lengths (A) or concentrations
922 (B) of transforming DNA. Transformations were performed in the presence (top panels) and
923 absence (bottom panels) of selection for transformants with hygromycin. (C) Schematic
924 depicting the intergenic neutral site (blue, NS1) targeted for insertion of a hygomycin resistance
925 cassette (yellow) in the *N. lyticus* ML1 genome and the linear DNA fragment employed in
926 transformation experiments with this species. Primer binding sites used for genotyping are
927 indicated (sites 1-3). (D) Genotyping of *N. lyticus–P. propionicum* co-cultures transformed with
928 the linear DNA fragment depicted in (C) (right) or parallel negative control co-cultures with no
929 DNA added (left) at the end of passage 4 (see Methods) with primers targeting NS1 (1,2) or *hph*
930 (3). Positive control *N. lyticus* primers target a genomic locus distant from NS1.

931

932 Figure S3. Population dynamics of *Se* and *Ai* during transposon mutagenesis. Transformation co-
933 culture was diluted and passaged into fresh media, with the addition of fresh *Ai,* after collection
934 of the T0, T1 and T2 samples (see Methods).

935

936 Figure S4. Distribution of 921 core protein families across CPR and non-CPR bacterial genomes,
937 including *Se* and *N. lyticus* ML1. Columns represent core families (derived from Meheust *et al.*[5])
938 and rows represent individual genomes from the indicated bacterial groups. CPR enriched
939 protein families indicated at top (blue), and dendrogram at left represents clustering of bacterial
940 strains based on protein family content.
941
942 Figure S5. The *Ai* genome and structural models for the *Se* proteome generated using
943 metagenomic sequence enriched MSAs for use in future studies of the *Se-Ai* system. (A)
944 Overview of the genome sequence of *A. israelii* F0345. (B-D) Example *Se* protein structure
945 models and associated predicted alignment matrices obtained using shallow (right) or
946 metagenomic sequence-improved (left) MSAs. *Se* proteins models (blue) are aligned to models
947 for top FS hits (light grey, AF database50 numbers A0A8B1YQG7 (B), A0A660LZS9 (C) and
948 A0A563CGZ5 (D)), when available. Some structures are trimmed to highlight the alignment.
949
950 Table S1. Tn-seq data and analysis. Shaded rows indicate genes found to be important for *Se*
951 fitness during in vitro co-culture with *Ai* by both TRANSIT and ALDEx2, at all time points
952 sampled. Blue text indicates genes within loci noted at the perimeter of Figure 4A.
953
954 Table S2. Sequence and structure-based annotation of the *Se* genome. Shaded rows indicate
955 genes encoding proteins for which the AlphaFold confidence score (pLDDT) was improved by
956 10 or more points upon inclusion of extensive metagenomic data in the MSA.
957
958 Table S3. Oligonucleotide primers employed in this study.

959

960 Table S4. Heterologous gene expression cassettes and linear fragments used in Saccharibacteria

961 transformations.

962 **References**

963 1. Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J.,
964 Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., et al. (2016). A new view of the tree of
965 life. Nat Microbiol *1*, 16048. 10.1038/nmicrobiol.2016.48.

966 2. Katz, M., Hover, B.M., and Brady, S.F. (2016). Culture-independent discovery of natural
967 products from soil metagenomes. J Ind Microbiol Biotechnol *43*, 129-141. 10.1007/s10295-015-
968 1706-6.

969 3. Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F.,
970 Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., et al. (2013). Insights into the phylogeny and
971 coding potential of microbial dark matter. Nature *499*, 431-437. 10.1038/nature12352.

972 4. Ji, Y., Zhang, P., Zhou, S., Gao, P., Wang, B., and Jiang, J. (2022). Widespread but
973 Poorly Understood Bacteria: Candidate Phyla Radiation. Microorganisms *10*.
974 10.3390/microorganisms10112232.

975 5. Meheust, R., Burstein, D., Castelle, C.J., and Banfield, J.F. (2019). The distinction of
976 CPR bacteria from other bacteria based on protein family content. Nat Commun *10*, 4173.
977 10.1038/s41467-019-12171-z.

978 6. Castelle, C.J., Brown, C.T., Anantharaman, K., Probst, A.J., Huang, R.H., and Banfield,
979 J.F. (2018). Biosynthetic capacity, metabolic variety and unusual biology in the CPR and
980 DPANN radiations. Nat Rev Microbiol *16*, 629-645. 10.1038/s41579-018-0076-2.

981 7. Batinovic, S., Rose, J.J.A., Ratcliffe, J., Seviour, R.J., and Petrovski, S. (2021).
982 Cocultivation of an ultrasmall environmental parasitic bacterium with lytic ability against
983 bacteria associated with wastewater foams. Nat Microbiol *6*, 703-711. 10.1038/s41564-021-
984 00892-1.

985 8. He, X., McLean, J.S., Edlund, A., Yooseph, S., Hall, A.P., Liu, S.Y., Dorrestein, P.C.,
986 Esquenazi, E., Hunter, R.C., Cheng, G., et al. (2015). Cultivation of a human-associated TM7
987 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. Proc Natl Acad Sci U S A
988 *112*, 244-249. 10.1073/pnas.1419038112.

989 9. Kuroda, K., Yamamoto, K., Nakai, R., Hirakata, Y., Kubota, K., Nobu, M.K., and
990 Narihiro, T. (2022). Symbiosis between Candidatus Patescibacteria and Archaea Discovered in
991 Wastewater-Treating Bioreactors. MBio *13*, e0171122. 10.1128/mbio.01711-22.

992 10. Yakimov, M.M., Merkel, A.Y., Gaisin, V.A., Pilhofer, M., Messina, E., Hallsworth, J.E.,
993 Klyukina, A.A., Tikhonova, E.N., and Gorlenko, V.M. (2022). Cultivation of a vampire:
994 'Candidatus Absconditicoccus praedator'. Environmental microbiology *24*, 30-49. 10.1111/1462-
995 2920.15823.

996 11. McLean, J.S., Bor, B., Kerns, K.A., Liu, Q., To, T.T., Solden, L., Hendrickson, E.L.,
997 Wrighton, K., Shi, W., and He, X. (2020). Acquisition and Adaptation of Ultra-small Parasitic

998 Reduced Genome Bacteria to Mammalian Hosts. Cell Rep *32*, 107939.
999 10.1016/j.celrep.2020.107939.

1000 12. Bor, B., Bedree, J.K., Shi, W., McLean, J.S., and He, X. (2019). Saccharibacteria (TM7)
1001 in the Human Oral Microbiome. J Dent Res *98*, 500-509. 10.1177/0022034519831671.

1002 13. Chipashvili, O., Utter, D.R., Bedree, J.K., Ma, Y., Schulte, F., Mascarin, G., Alayyoubi,
1003 Y., Chouhan, D., Hardt, M., Bidlack, F., et al. (2021). Episymbiotic Saccharibacteria suppresses
1004 gingival inflammation and bone loss in mice through host bacterial modulation. Cell Host
1005 Microbe *29*, 1649-1662 e1647. 10.1016/j.chom.2021.09.009.

1006 14. Adler, C.J., Dobney, K., Weyrich, L.S., Kaidonis, J., Walker, A.W., Haak, W., Bradshaw,
1007 C.J., Townsend, G., Soltysiak, A., Alt, K.W., et al. (2013). Sequencing ancient calcified dental
1008 plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial
1009 revolutions. Nature genetics *45*, 450-455, 455e451. 10.1038/ng.2536.

1010 15. Bor, B., McLean, J.S., Foster, K.R., Cen, L., To, T.T., Serrato-Guillen, A., Dewhirst,
1011 F.E., Shi, W., and He, X. (2018). Rapid evolution of decreased host susceptibility drives a stable
1012 relationship between ultrasmall parasite TM7x and its bacterial host. Proc Natl Acad Sci U S A
1013 *115*, 12277-12282. 10.1073/pnas.1810625115.

1014 16. Nie, J., Utter, D.R., Kerns, K.A., Lamont, E.I., Hendrickson, E.L., Liu, J., Wu, T., He, X.,
1015 McLean, J., and Bor, B. (2022). Strain-Level Variation and Diverse Host Bacterial Responses in
1016 Episymbiotic Saccharibacteria. mSystems *7*, e0148821. 10.1128/msystems.01488-21.

1017 17. Xie, B., Wang, J., Nie, Y., Tian, J., Wang, Z., Chen, D., Hu, B., Wu, X.L., and Du, W.
1018 (2022). Type IV pili trigger episymbiotic association of Saccharibacteria with its bacterial host.
1019 Proc Natl Acad Sci U S A *119*, e2215990119. 10.1073/pnas.2215990119.

1020 18. Bor, B., Collins, A.J., Murugkar, P.P., Balasubramanian, S., To, T.T., Hendrickson, E.L.,
1021 Bedree, J.K., Bidlack, F.B., Johnston, C.D., Shi, W., et al. (2020). Insights Obtained by
1022 Culturing Saccharibacteria With Their Bacterial Hosts. J Dent Res *99*, 685-694.
1023 10.1177/0022034520905792.

1024 19. Cross, K.L., Campbell, J.H., Balachandran, M., Campbell, A.G., Cooper, S.J., Griffen,
1025 A., Heaton, M., Joshi, S., Klingeman, D., Leys, E., et al. (2019). Targeted isolation and
1026 cultivation of uncultivated bacteria by reverse genomics. Nat Biotechnol *37*, 1314-1321.
1027 10.1038/s41587-019-0260-6.

1028 20. He, X. (2023). Culture-based approaches to studying "microbial dark matter". Proc Natl
1029 Acad Sci U S A *120*, e2219691120. 10.1073/pnas.2219691120.

1030 21. Stackenbrandt, E., and Goebel, B.M. (1994). Taxonomic Note: A Place for DNA-DNA
1031 Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in
1032 Bacteriology. Int. J. Syst. Evol. Microbiol. *44*, 846-849.

1033 22. Damke, P.P., Celma, L., Kondekar, S.M., Di Guilmi, A.M., Marsin, S., Depagne, J.,
1034 Veaute, X., Legrand, P., Walbott, H., Vercruyssen, J., et al. (2022). ComFC mediates transport

1035 and handling of single-stranded DNA during natural transformation. Nat Commun *13*, 1961.
1036 10.1038/s41467-022-29494-z.

1037 23. Dubnau, D., and Blokesch, M. (2019). Mechanisms of DNA Uptake by Naturally
1038 Competent Bacteria. Annu Rev Genet *53*, 217-237. 10.1146/annurev-genet-112618-043641.

1039 24. Sharma, D.K., Misra, H.S., Bihani, S.C., and Rajpurohit, Y.S. (2023). Biochemical
1040 Properties and Roles of DprA Protein in Bacterial Natural Transformation, Virulence, and Pilin
1041 Variation. J Bacteriol *205*, e0046522. 10.1128/jb.00465-22.

1042 25. Johnston, C., Martin, B., Fichant, G., Polard, P., and Claverys, J.P. (2014). Bacterial
1043 transformation: distribution, shared mechanisms and divergent control. Nat Rev Microbiol *12*,
1044 181-196. 10.1038/nrmicro3199.

1045 26. DeJesus, M.A., Ambadipudi, C., Baker, R., Sassetti, C., and Ioerger, T.R. (2015).
1046 TRANSIT--A Software Tool for Himar1 TnSeq Analysis. PLoS computational biology *11*,
1047 e1004401. 10.1371/journal.pcbi.1004401.

1048 27. Fernandes, A.D., Reid, J.N., Macklaim, J.M., McMurrough, T.A., Edgell, D.R., and
1049 Gloor, G.B. (2014). Unifying the analysis of high-throughput sequencing datasets: characterizing
1050 RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data
1051 analysis. Microbiome *2*, 15. 10.1186/2049-2618-2-15.

1052 28. Bedree, J.K., Bor, B., Cen, L., Edlund, A., Lux, R., McLean, J.S., Shi, W., and He, X.
1053 (2018). Quorum Sensing Modulates the Epibiotic-Parasitic Relationship Between Actinomyces
1054 odontolyticus and Its Saccharibacteria epibiont, a Nanosynbacter lyticus Strain, TM7x. Front
1055 Microbiol *9*, 2049. 10.3389/fmicb.2018.02049.

1056 29. Wu, C., and Ton-That, H. (2010). Allelic exchange in Actinomyces oris with mCherry
1057 fluorescence counterselection. Applied and environmental microbiology *76*, 5987-5989.
1058 10.1128/AEM.00811-10.

1059 30. Gane, A., Bileshi, M.L., Dohan, D., Speretta, E., Heliou, A., Meng-Papaxanthos, L.,
1060 Zellner, H., Brevdo, E., Parikh, A., Martin, M.J., et al. (2022). ProtNLM: Model-based Natural
1061 Language Protein
1062 Annotation. Preprint.

1063 31. Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS computational biology *7*,
1064 e1002195. 10.1371/journal.pcbi.1002195.

1065 32. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer,
1066 E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: The protein
1067 families database in 2021. Nucleic Acids Res *49*, D412-D419. 10.1093/nar/gkaa913.

1068 33. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang,
1069 J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures
1070 and interactions using a three-track neural network. Science *373*, 871-876.
1071 10.1126/science.abj8754.

1072 34. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O.,
1073 Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., et al. (2021). Highly accurate protein
1074 structure prediction with AlphaFold. Nature *596*, 583-589. 10.1038/s41586-021-03819-2.

1075 35. Remmert, M., Biegert, A., Hauser, A., and Soding, J. (2011). HHblits: lightning-fast
1076 iterative protein sequence searching by HMM-HMM alignment. Nature methods *9*, 173-175.
1077 10.1038/nmeth.1818.

1078 36. Steinegger, M., Mirdita, M., and Soding, J. (2019). Protein-level assembly increases
1079 protein sequence recovery from metagenomic samples manyfold. Nature methods *16*, 603-606.
1080 10.1038/s41592-019-0437-4.

1081 37. Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C.H. (2007). UniRef:
1082 comprehensive and non-redundant UniProt reference clusters. Bioinformatics *23*, 1282-1288.
1083 10.1093/bioinformatics/btm098.

1084 38. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Le, J., Gilchrist, C.L.M.,
1085 Soding, J., and Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek.
1086 bioRxiv.

1087 39. Horton, J.R., and Cheng, X. (2000). PvuII endonuclease contains two calcium ions in
1088 active sites. J Mol Biol *300*, 1049-1056. 10.1006/jmbi.2000.3938.

1089 40. Johnston, C.D., Cotton, S.L., Rittling, S.R., Starr, J.R., Borisy, G.G., Dewhirst, F.E., and
1090 Lemon, K.P. (2019). Systematic evasion of the restriction-modification barrier in bacteria. Proc
1091 Natl Acad Sci U S A *116*, 11454-11459. 10.1073/pnas.1820256116.

1092 41. Cheng, H., Schaeffer, R.D., Liao, Y., Kinch, L.N., Pei, J., Shi, S., Kim, B.H., and
1093 Grishin, N.V. (2014). ECOD: an evolutionary classification of protein domains. PLoS
1094 computational biology *10*, e1003926. 10.1371/journal.pcbi.1003926.

1095 42. Zhang, J., Schaeffer, R.D., Durham, J., Cong, Q., and Grishin, N.V. (2023). DPAM: A
1096 domain parser for AlphaFold models. Protein Sci *32*, e4548. 10.1002/pro.4548.

1097 43. Moreira, D., Zivanovic, Y., Lopez-Archilla, A.I., Iniesto, M., and Lopez-Garcia, P.
1098 (2021). Reductive evolution and unique predatory mode in the CPR bacterium Vampirococcus
1099 lugosii. Nat Commun *12*, 2454. 10.1038/s41467-021-22762-4.

1100 44. Murugkar, P.P., Collins, A.J., Chen, T., and Dewhirst, F.E. (2020). Isolation and
1101 cultivation of candidate phyla radiation Saccharibacteria (TM7) bacteria in coculture with
1102 bacterial hosts. J Oral Microbiol *12*, 1814666. 10.1080/20002297.2020.1814666.

1103 45. Cressler, C.E., Mc, L.D., Rozins, C., J, V.D.H., and Day, T. (2016). The adaptive
1104 evolution of virulence: a review of theoretical predictions and empirical tests. Parasitology *143*,
1105 915-930. 10.1017/S003118201500092X.

1106 46. Egido, J.E., Costa, A.R., Aparicio-Maldonado, C., Haas, P.J., and Brouns, S.J.J. (2022).
1107 Mechanisms and clinical importance of bacteriophage resistance. FEMS Microbiol Rev *46*.
1108 10.1093/femsre/fuab048.

1109 47. He, C., Keren, R., Whittaker, M.L., Farag, I.F., Doudna, J.A., Cate, J.H.D., and Banfield,
1110 J.F. (2021). Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR
1111 bacteria and DPANN archaea in groundwater ecosystems. Nat Microbiol *6*, 354-365.
1112 10.1038/s41564-020-00840-5.

1113 48. Ellison, C.K., Dalia, T.N., Vidal Ceballos, A., Wang, J.C., Biais, N., Brun, Y.V., and
1114 Dalia, A.B. (2018). Retraction of DNA-bound type IV competence pili initiates DNA uptake
1115 during natural transformation in Vibrio cholerae. Nat Microbiol *3*, 773-780. 10.1038/s41564-
1116 018-0174-y.

1117 49. Hepp, C., and Maier, B. (2016). Kinetics of DNA uptake during transformation provide
1118 evidence for a translocation ratchet mechanism. Proc Natl Acad Sci U S A *113*, 12467-12472.
1119 10.1073/pnas.1608110113.

1120 50. Tian, J., Utter, D.R., Cen, L., Dong, P.T., Shi, W., Bor, B., Qin, M., McLean, J.S., and
1121 He, X. (2022). Acquisition of the arginine deiminase system benefits epiparasitic
1122 Saccharibacteria and their host bacteria in a mammalian niche environment. Proc Natl Acad Sci
1123 U S A *119*. 10.1073/pnas.2114909119.

1124 51. Batty, I. (1958). Actinomyces odontolyticus, a new species of actinomycete regularly
1125 isolated from deep carious dentine. J Pathol Bacteriol *75*, 455-459. 10.1002/path.1700750225.

1126 52. Dewhirst, F.E., Chen, T., Izard, J., Paster, B.J., Tanner, A.C., Yu, W.H., Lakshmanan, A.,
1127 and Wade, W.G. (2010). The human oral microbiome. J Bacteriol *192*, 5002-5017.
1128 10.1128/JB.00542-10.

1129 53. Nikolaitchouk, N., Hoyles, L., Falsen, E., Grainger, J.M., and Collins, M.D. (2000).
1130 Characterization of Actinomyces isolates from samples from the human urogenital tract:
1131 description of Actinomyces urogenitalis sp. nov. Int J Syst Evol Microbiol *50 Pt 4*, 1649-1654.
1132 10.1099/00207713-50-4-1649.

1133 54. Collins, A.J., Murugkar, P.P., and Dewhirst, F.E. (2021). Establishing Stable Binary
1134 Cultures of Symbiotic Saccharibacteria from the Oral Cavity. J Vis Exp. 10.3791/62484.

1135 55. Wick, R.R., Judd, L.M., Cerdeira, L.T., Hawkey, J., Meric, G., Vezina, B., Wyres, K.L.,
1136 and Holt, K.E. (2021). Trycycler: consensus long-read assemblies for bacterial genomes.
1137 Genome biology *22*, 266. 10.1186/s13059-021-02483-z.

1138 56. Vaser, R., and Sikic, M. (2021). Time- and memory-efficient genome assembly with
1139 Raven. Nat Comput Sci *1*, 332-336.

1140 57. Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long
1141 sequences. Bioinformatics *32*, 2103-2110. 10.1093/bioinformatics/btw152.

1142 58. Zimin, A.V., Puiu, D., Luo, M.C., Zhu, T., Koren, S., Marcais, G., Yorke, J.A., Dvorak,
1143 J., and Salzberg, S.L. (2017). Hybrid assembly of the large and highly repetitive genome of
1144 Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm.
1145 Genome Res *27*, 787-792. 10.1101/gr.213405.116.

1146 59. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*,
1147 2068-2069. 10.1093/bioinformatics/btu153.

1148 60. Arkin, A.P., Cottingham, R.W., Henry, C.S., Harris, N.L., Stevens, R.L., Maslov, S.,
1149 Dehal, P., Ware, D., Perez, F., Canon, S., et al. (2018). KBase: The United States Department of
1150 Energy Systems Biology Knowledgebase. Nat Biotechnol *36*, 566-569. 10.1038/nbt.4163.

1151 61. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately maximum-
1152 likelihood trees for large alignments. PLoS One *5*, e9490. 10.1371/journal.pone.0009490.

1153 62. Dehlinger, B., Jurss, J., Lychuk, K., and Putonti, C. (2021). The Dynamic Codon Biaser:
1154 calculating prokaryotic codon usage biases. Microb Genom *7*. 10.1099/mgen.0.000663.

1155 63. Hall, M.P., Unch, J., Binkowski, B.F., Valley, M.P., Butler, B.L., Wood, M.G., Otto, P.,
1156 Zimmerman, K., Vidugiris, G., Machleidt, T., et al. (2012). Engineered luciferase reporter from a
1157 deep sea shrimp utilizing a novel imidazopyrazinone substrate. ACS Chem Biol *7*, 1848-1857.
1158 10.1021/cb3002478.

1159 64. Zalacain, M., Gonzalez, A., Guerrero, M.C., Mattaliano, R.J., Malpartida, F., and
1160 Jimenez, A. (1986). Nucleotide sequence of the hygromycin B phosphotransferase gene from
1161 Streptomyces hygroscopicus. Nucleic Acids Res *14*, 1565-1581. 10.1093/nar/14.4.1565.

1162 65. Cutler, K.J., Stringer, C., Lo, T.W., Rappez, L., Stroustrup, N., Brook Peterson, S.,
1163 Wiggins, P.A., and Mougous, J.D. (2022). Omnipose: a high-precision morphology-independent
1164 solution for bacterial cell segmentation. Nature methods *19*, 1438-1448. 10.1038/s41592-022-
1165 01639-4.

1166 66. Gallagher, L.A., Bailey, J., and Manoil, C. (2020). Ranking essential bacterial processes
1167 by speed of mutant death. Proc Natl Acad Sci U S A *117*, 18010-18017.
1168 10.1073/pnas.2001507117.

1169 67. Gallagher, L.A. (2019). Methods for Tn-Seq Analysis in Acinetobacter baumannii.
1170 Methods in molecular biology (Clifton, N.J *1946*, 115-134. 10.1007/978-1-4939-9118-1_12.

1171 68. Arrieta-Ortiz, M.L., Pan, M., Kaur, A., Pepper-Tunick, E., Srinivas, V., Dash, A.,
1172 Immanuel, S.R.C., Brooks, A.N., Shepherd, T.R., and Baliga, N.S. (2023). Disrupting the ArcA
1173 Regulatory Network Amplifies the Fitness Cost of Tetracycline Resistance in Escherichia coli.
1174 mSystems *8*, e0090422. 10.1128/msystems.00904-22.

1175 69. Anishchenko, I., Baek, M., Park, H., Hiranuma, N., Kim, D.E., Dauparas, J., Mansoor, S.,
1176 Humphreys, I.R., and Baker, D. (2021). Protein tertiary structure prediction and refinement using
1177 deep learning and Rosetta in CASP14. Proteins *89*, 1722-1733. 10.1002/prot.26194.

1178 70. Chen, I.A., Chu, K., Palaniappan, K., Ratner, A., Huang, J., Huntemann, M., Hajek, P.,
1179 Ritter, S., Varghese, N., Seshadri, R., et al. (2021). The IMG/M data management and analysis
1180 system v.6.0: new tools and advanced capabilities. Nucleic Acids Res *49*, D751-D763.
1181 10.1093/nar/gkaa939.

1182 71. Mitchell, A.L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G.,
1183 Crusoe, M.R., Kale, V., Potter, S.C., Richardson, L.J., et al. (2020). MGnify: the microbiome
1184 analysis resource in 2020. Nucleic Acids Res *48*, D570-D578. 10.1093/nar/gkz1035.
1185

Figure 1

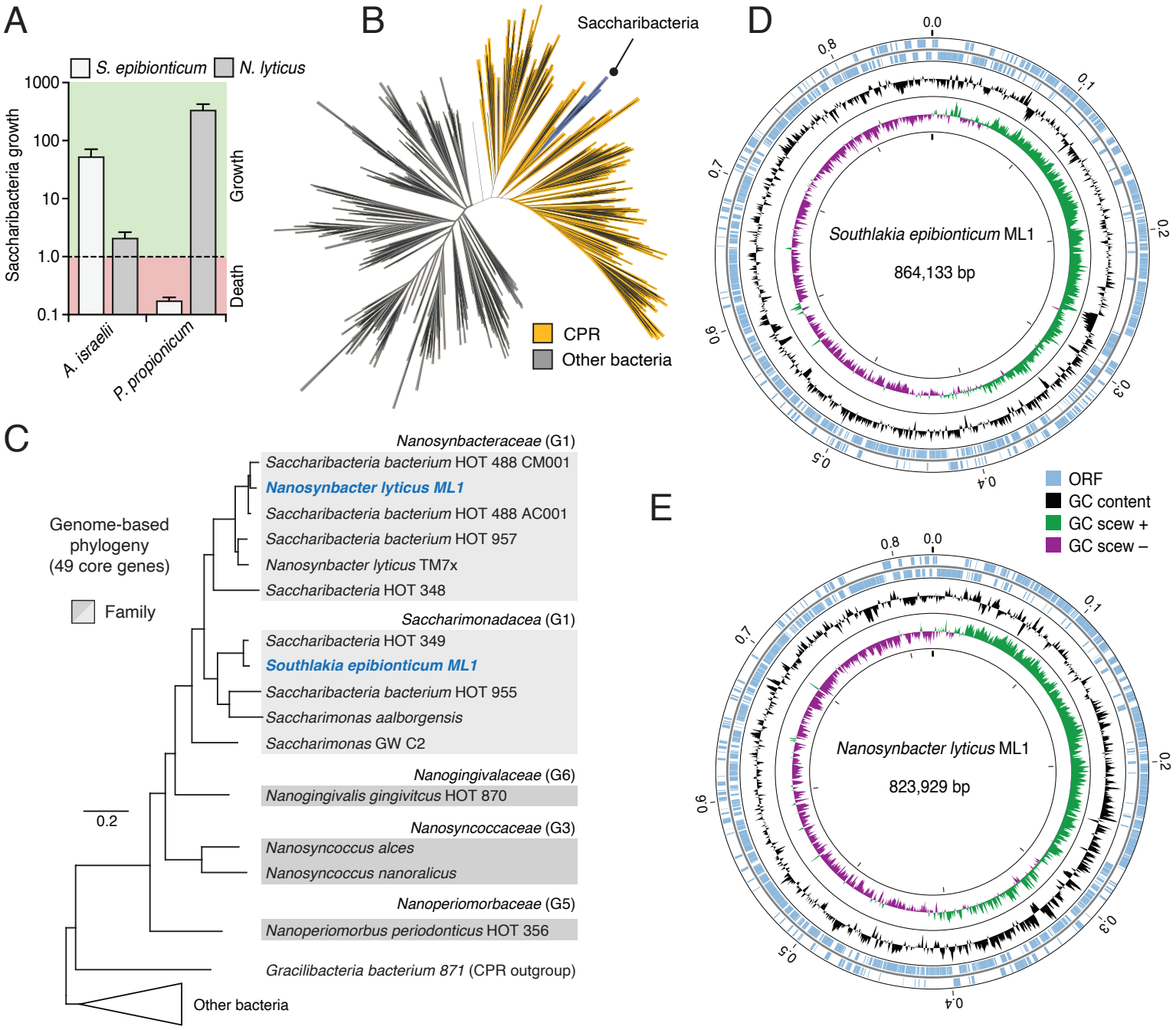

A
S. epibionticum
N. lyticus
Saccharibacteria growth
1000
100
10
1.0
0.1
Growth
Death
A. israelii
P. propionicum
B
Saccharibacteria
CPR
Other bacteria
C
Genome-based phylogeny (49 core genes)
Family
Nanosynbacteraceae (G1)
Saccharibacteria bacterium HOT 488 CM001
Nanosynbacter lyticus ML1
Saccharibacteria bacterium HOT 488 AC001
Saccharibacteria bacterium HOT 957
Nanosynbacter lyticus TM7x
Saccharibacteria HOT 348
Saccharimonadacea (G1)
Saccharibacteria HOT 349
Southlakia epibionticum ML1
Saccharibacteria bacterium HOT 955
Saccharimonas aalborgensis
Saccharimonas GW C2
Nanogingivalaceae (G6)
Nanogingivalis gingivitcus HOT 870
0.2
Nanosyncoccaceae (G3)
Nanosyncoccus alces
Nanosyncoccus nanoralicus
Nanoperiomorbaceae (G5)
Nanoperiomorbus periodonticus HOT 356
Gracilibacteria bacterium 871 (CPR outgroup)
Other bacteria
D
Southlakia epibionticum ML1
864,133 bp
ORF
GC content
GC scew +
GC scew –
E
Nanosynbacter lyticus ML1
823,929 bp

Figure 2

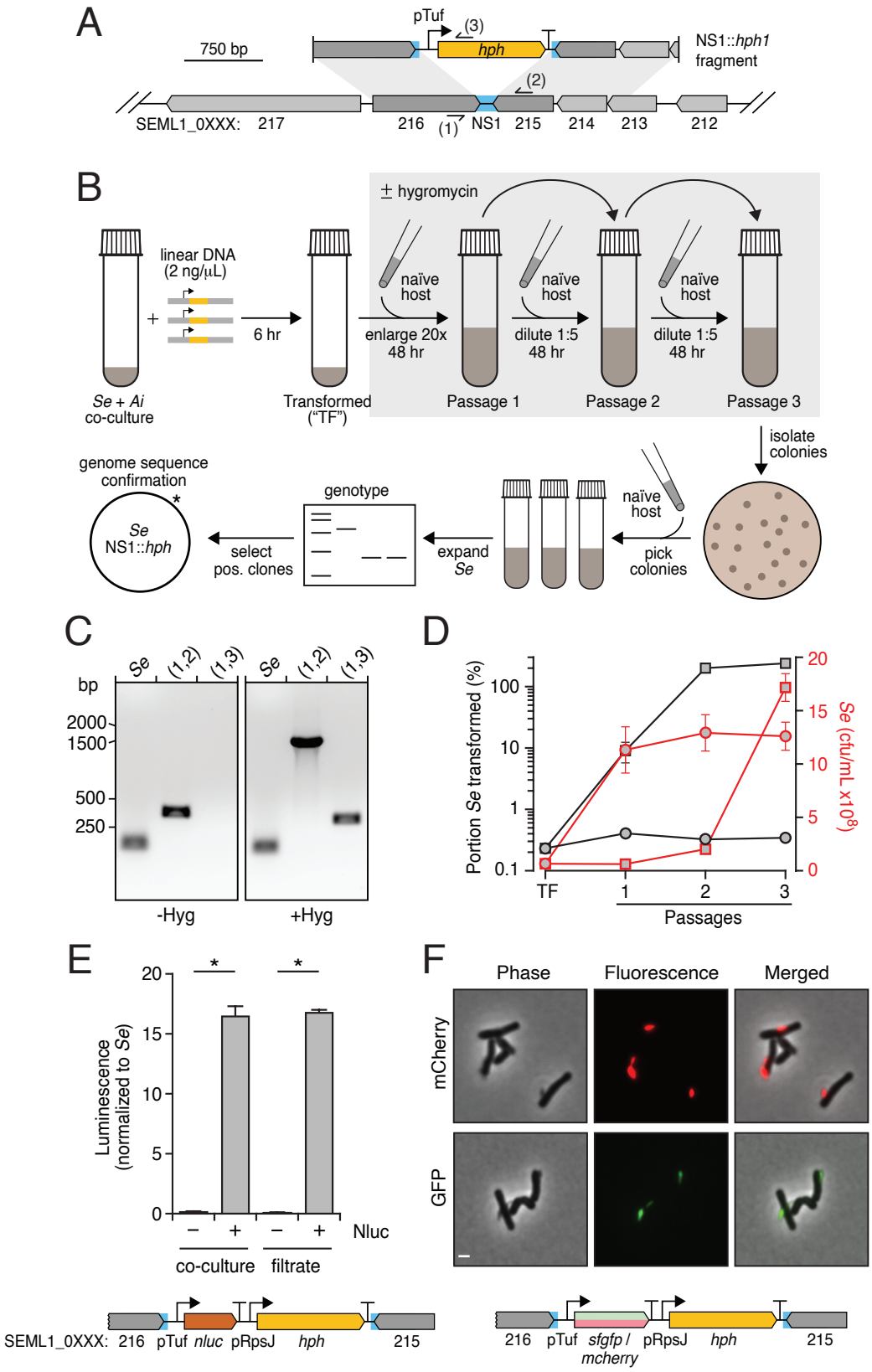

A
750 bp
pTuf
(3)
hph
NS1::hph1
fragment
(2)
SEML1_0XXX: 217
216
(1)
NS1
215
214
213
212
B
± hygromycin
linear DNA
(2 ng/µL)
6 hr
naïve
host
enlarge 20x
48 hr
dilute 1:5
48 hr
Se + Ai
co-culture
Transformed
("TF")
Passage 1
Passage 2
Passage 3
isolate
colonies
genome sequence
confirmation
Se
NS1::hph
select
pos. clones
genotype
expand
Se
pick
colonies
C
Se
(1,2)
(1,3)
bp
2000
1500
500
250
-Hyg
+Hyg
D
Portion Se transformed (%)
100
10
1
0.1
TF
1
2
3
Passages
Se (cfu/mL x10^8)
20
15
10
5
0
E
Luminescence
(normalized to Se)
20
15
10
5
0
*
*
–
+
–
+
Nluc
co-culture
filtrate
F
Phase
Fluorescence
Merged
mCherry
GFP
SEML1_0XXX: 216 pTuf nluc pRpsJ hph 215
216 pTuf sfgfp / mcherry pRpsJ hph 215

Figure 3

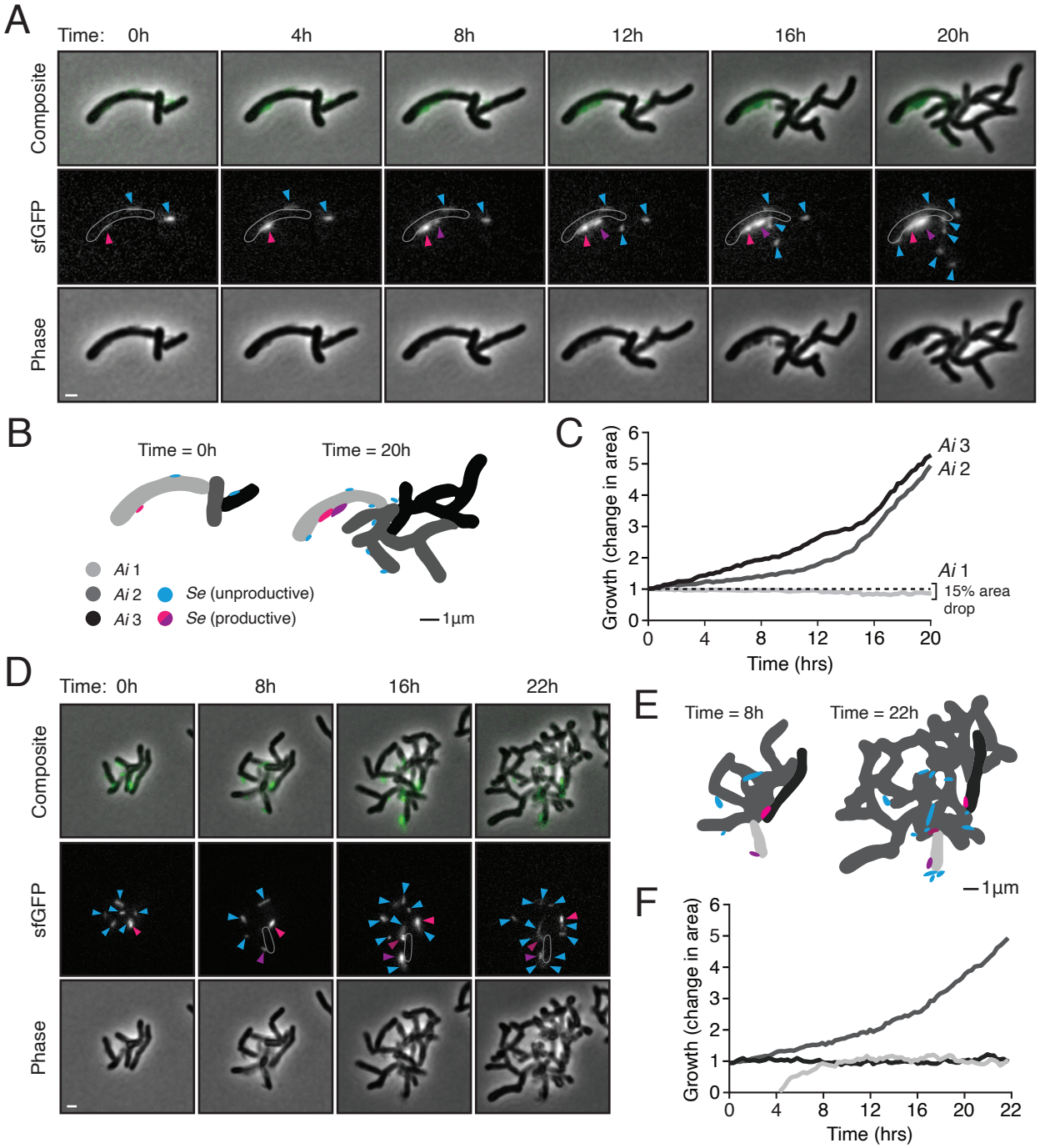
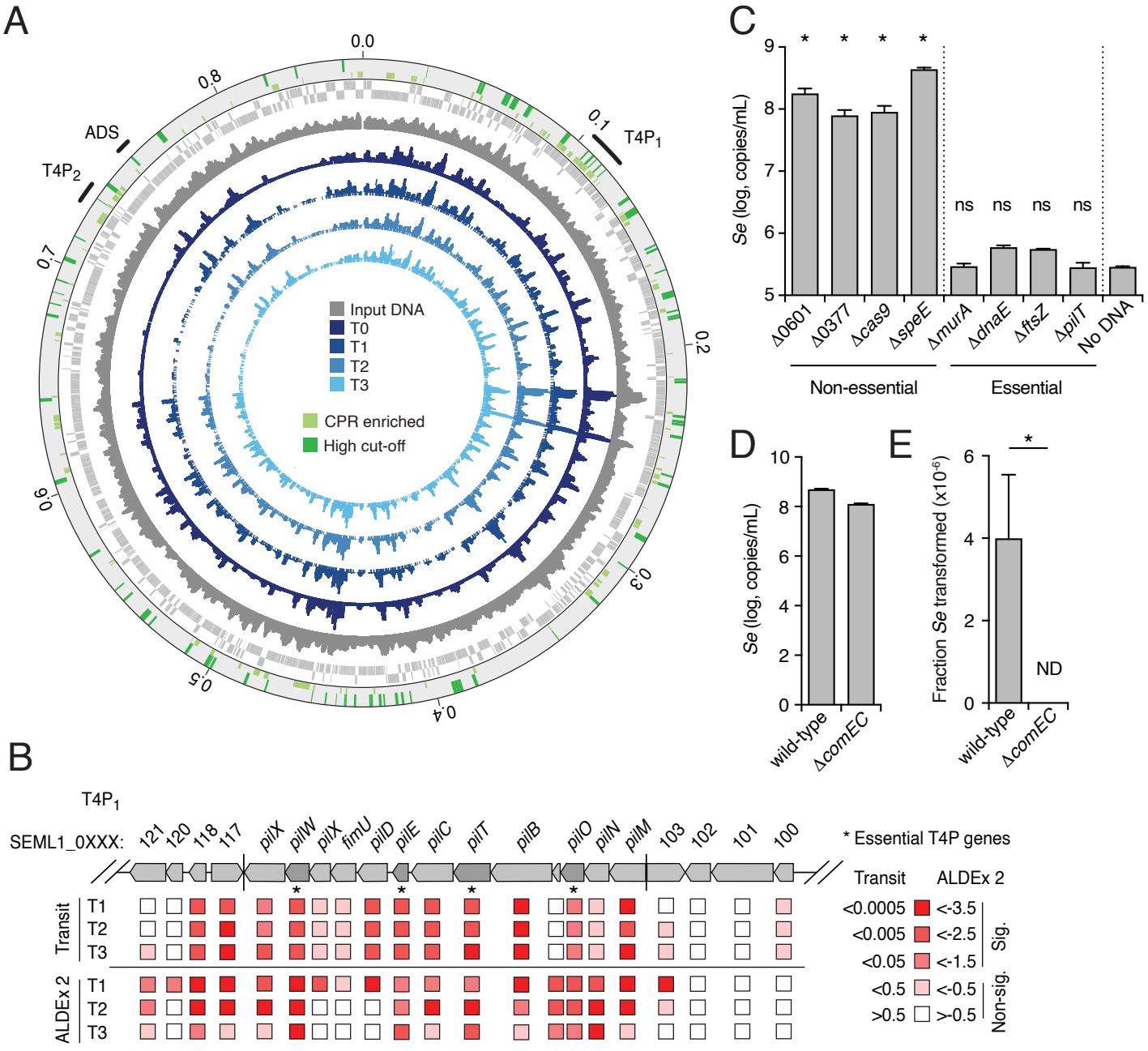

A
Time: 0h
4h
8h
12h
16h
20h
Composite
sfGFP
Phase
B
Time = 0h
Time = 20h
Ai 1
Ai 2
Ai 3
Se (unproductive)
Se (productive)
1µm
C
Growth (change in area)
Time (hrs)
Ai 3
Ai 2
Ai 1
15% area drop
D
Time: 0h
8h
16h
22h
Composite
sfGFP
Phase
E
Time = 8h
Time = 22h
1µm
F
Growth (change in area)
Time (hrs)

Figure 4


A
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
ADS
T4P2
T4P1
Input DNA
T0
T1
T2
T3
CPR enriched
High cut-off
C
Se (log, copies/mL)
Δ0601
Δ0377
Δcas9
ΔspeE
ΔmurA
ΔdnaE
ΔftsZ
ΔpilT
No DNA
ns
Non-essential
Essential
D
Se (log, copies/mL)
wild-type
ΔcomEC
E
Fraction Se transformed (x10-6)
ND
wild-type
ΔcomEC

B
T4P1
SEML1_0XXX:
121
120
118
117
pilX
pilW
pilX
fimU
pilD
pilE
pilC
pilT
pilB
pilO
pilN
pilM
103
102
101
100
* Essential T4P genes
Transit
ALDEx 2
T1
T2
T3
Transit
ALDEx 2
<0.0005
<0.005
<0.05
<0.5
>0.5
<-3.5
<-2.5
<-1.5
<-0.5
>-0.5
Sig.
Non-sig.

Figure 5

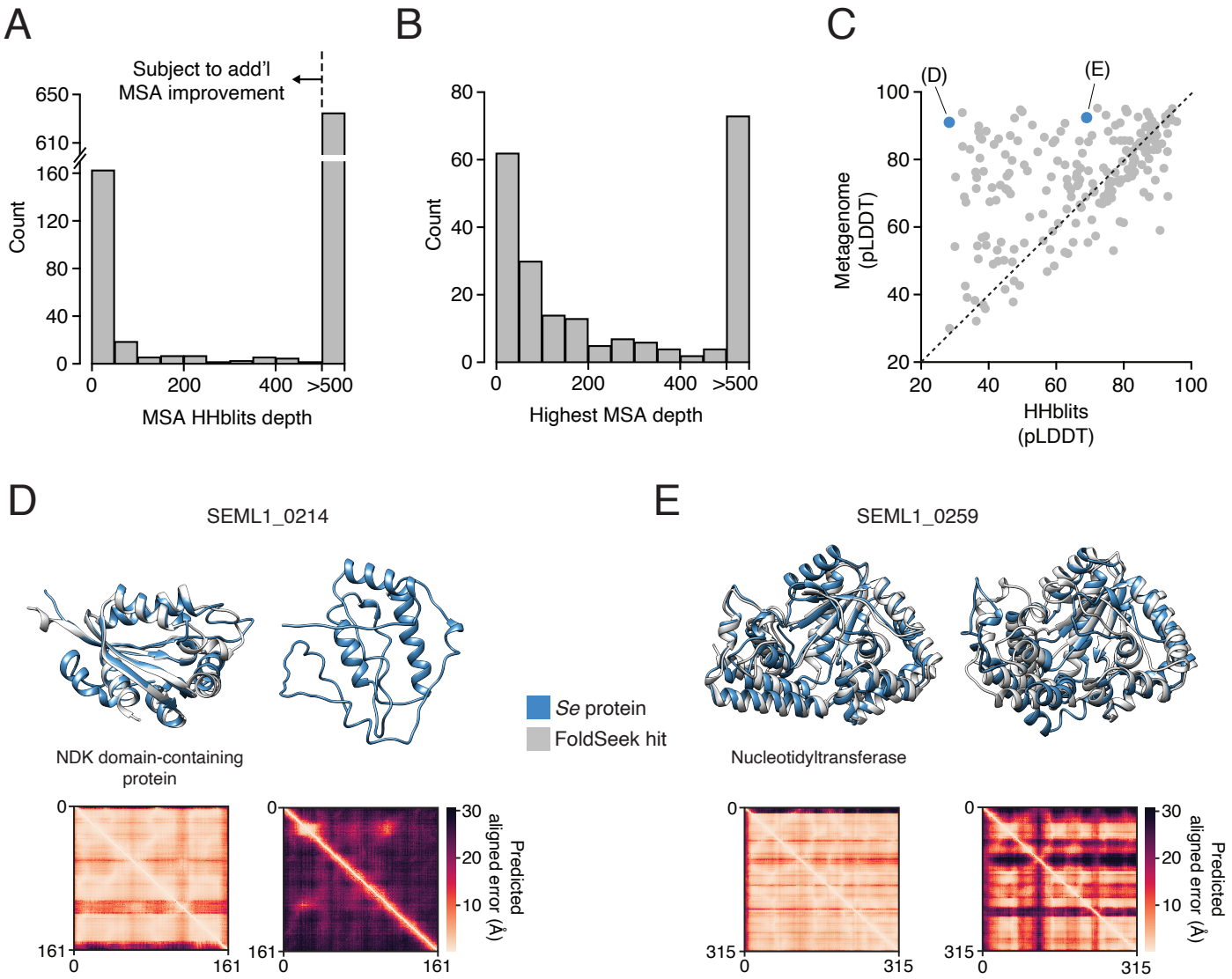

A
Subject to add'l
MSA improvement
650
610
160
120
80
40
0
Count
0
200
400
>500
MSA HHblits depth
B
80
60
40
20
0
Count
0
200
400
>500
Highest MSA depth
C
(D)
(E)
100
80
60
40
20
Metagenome
(pLDDT)
20
40
60
80
100
HHblits
(pLDDT)
D
SEML1_0214
NDK domain-containing
protein
0
161
0
161
0
161
0
161
30
20
10
Predicted
aligned error (Å)
Se protein
FoldSeek hit
E
SEML1_0259
Nucleotidyltransferase
0
315
0
315
0
315
0
315
30
20
10
Predicted
aligned error (Å)

Figure S1

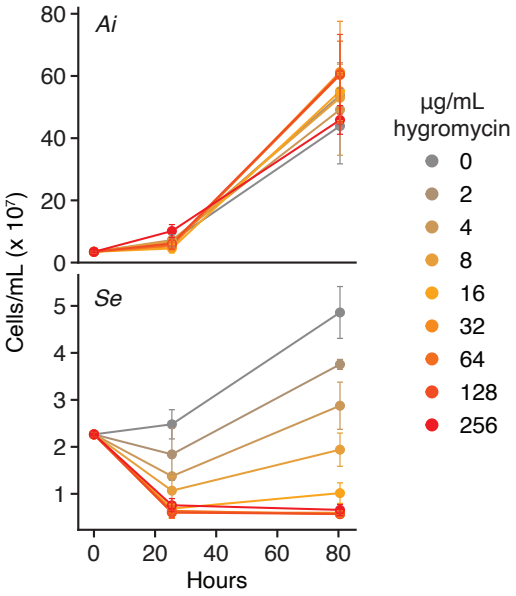

Ai
Se
Cells/mL (x 10^7)
Hours
0
20
40
60
80
1
2
3
4
5
µg/mL
hygromycin
0
2
4
8
16
32
64
128
256

Figure S2

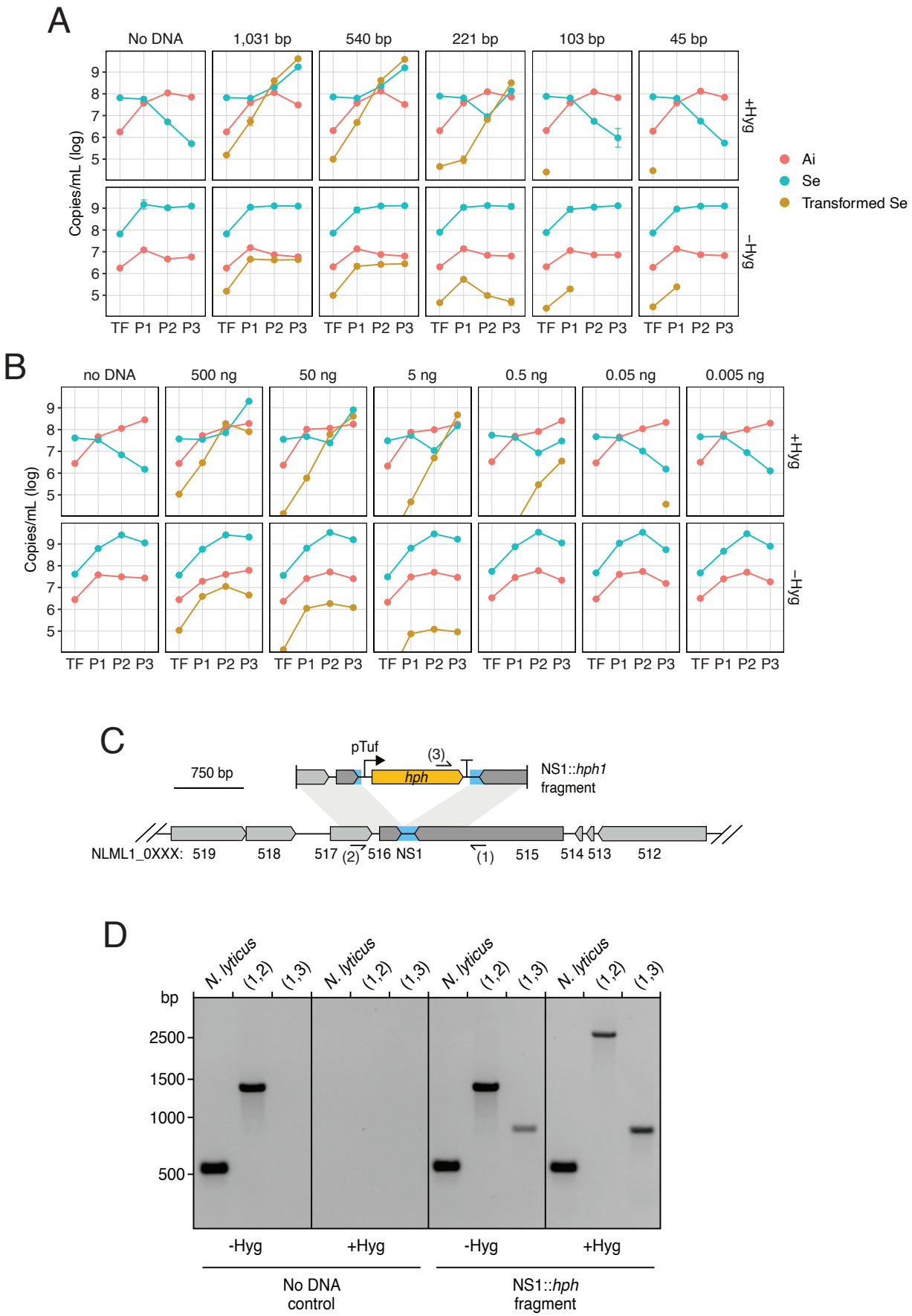

A
No DNA
1,031 bp
540 bp
221 bp
103 bp
45 bp
Copies/mL (log)
+Hyg
−Hyg
TF P1 P2 P3
Ai
Se
Transformed Se
B
no DNA
500 ng
50 ng
5 ng
0.5 ng
0.05 ng
0.005 ng
Copies/mL (log)
+Hyg
−Hyg
TF P1 P2 P3
C
750 bp
pTuf
(3)
hph
NS1::hph1
fragment
NLML1_0XXX: 519 518 517 (2) 516 NS1 (1) 515 514 513 512
D
N. lyticus
(1,2)
(1,3)
bp
2500
1500
1000
500
-Hyg
+Hyg
No DNA
control
NS1::hph
fragment

Figure S3

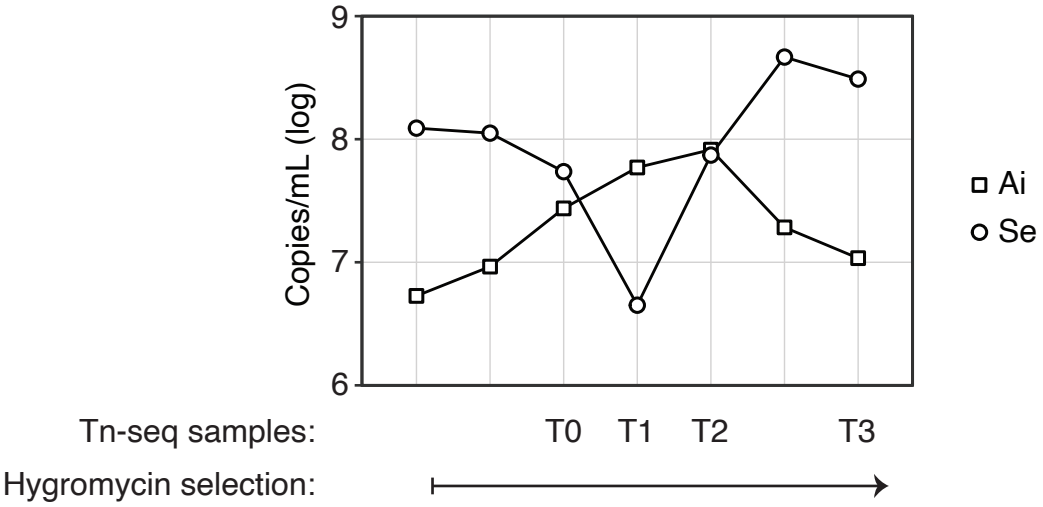

9
8
7
6
Copies/mL (log)
Ai
Se
Tn-seq samples:
T0
T1
T2
T3
Hygromycin selection:

Figure S4

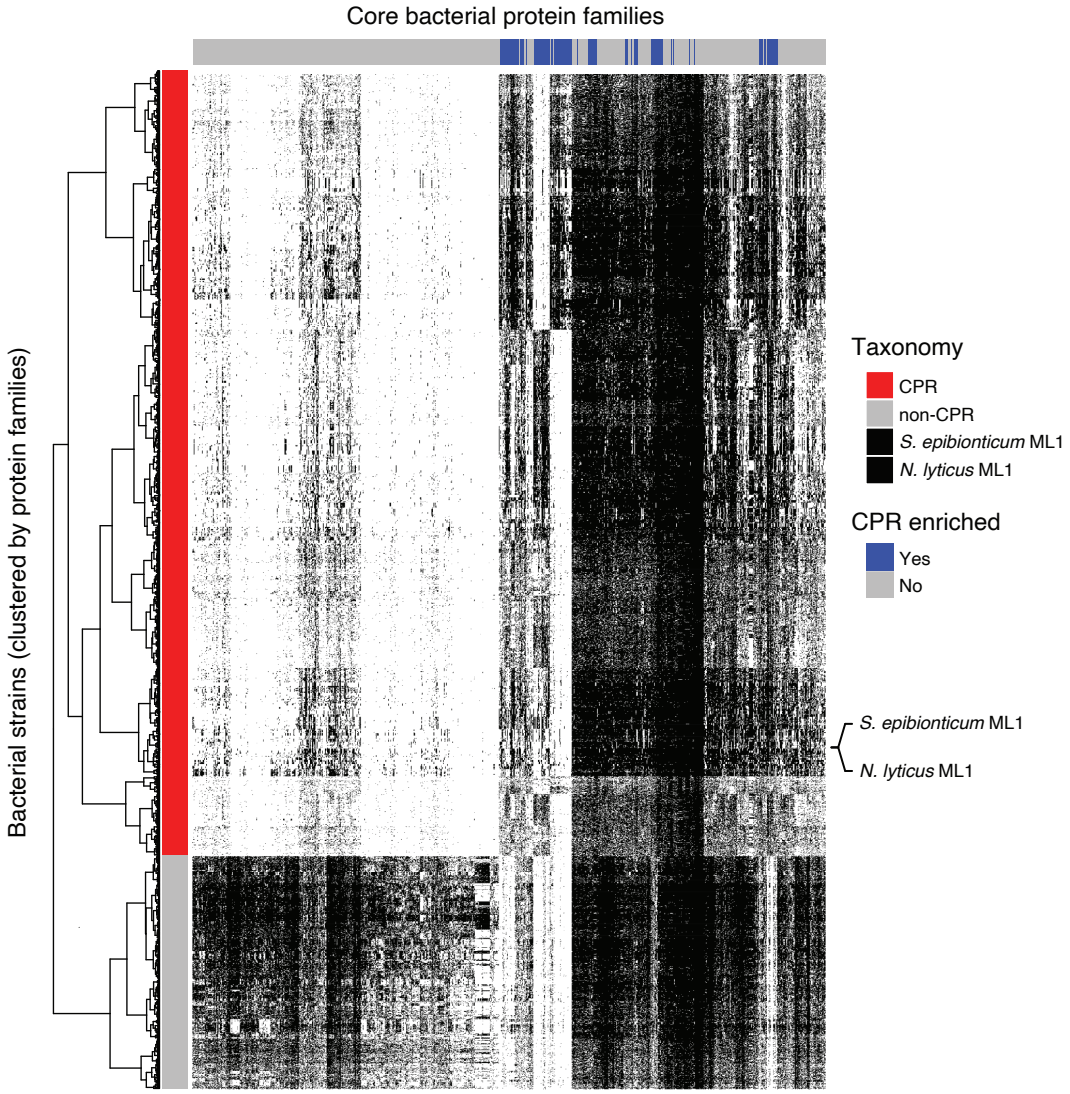
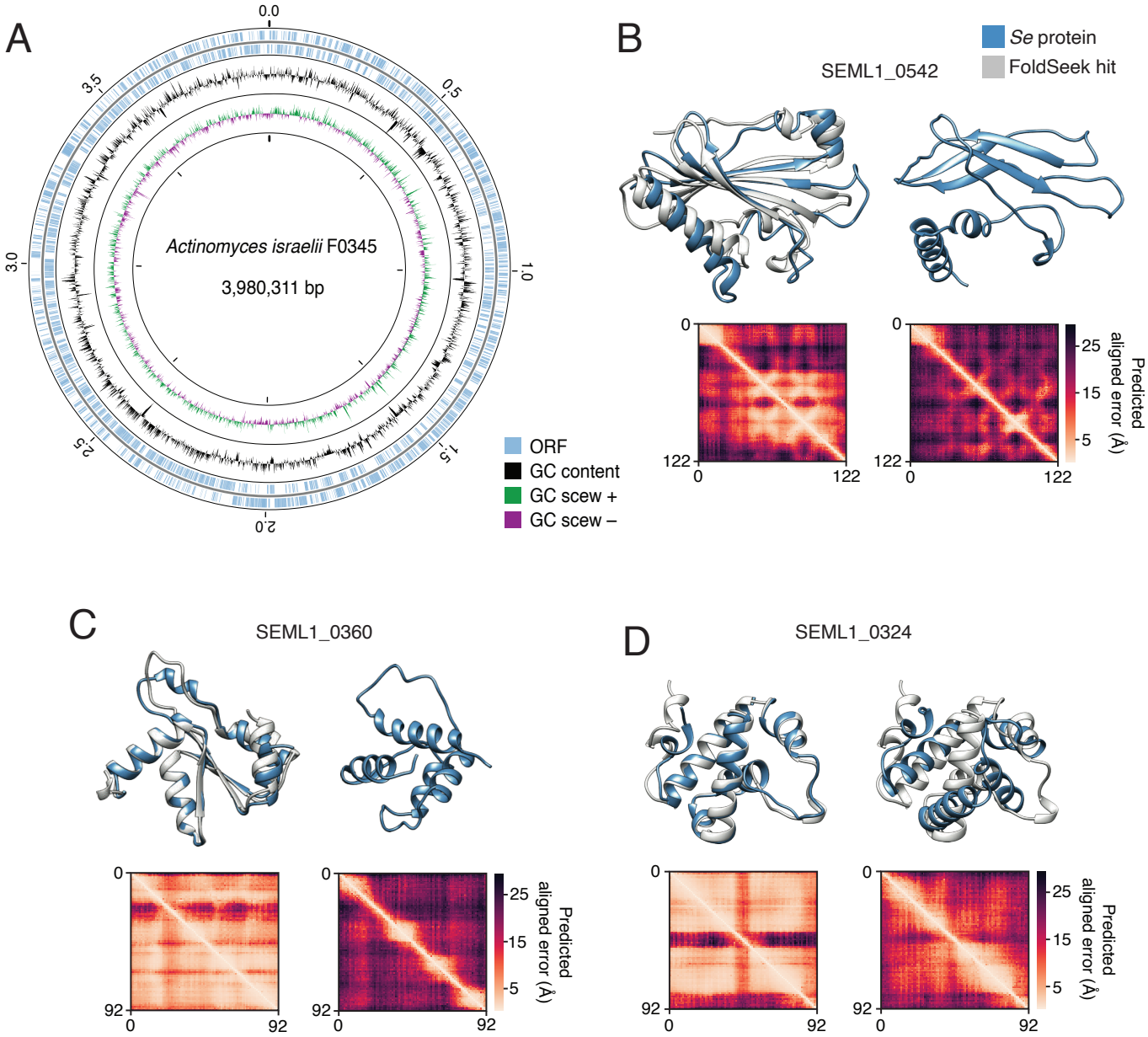

Core bacterial protein families
Bacterial strains (clustered by protein families)
Taxonomy
CPR
non-CPR
S. epibionticum ML1
N. lyticus ML1
CPR enriched
Yes
No
S. epibionticum ML1
N. lyticus ML1

Figure S5


A
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Actinomyces israelii F0345
3,980,311 bp
ORF
GC content
GC scew +
GC scew –
B
SEML1_0542
Se protein
FoldSeek hit
0
122
Predicted aligned error (Å)
25
15
5
C
SEML1_0360
92
D
SEML1_0324