Title

Statistical design of a synthetic microbiome that clears a multi-drug resistant gut pathogen

Authors

Rita A. Oliveira[1,7], Bipul Pandey[1,9], Kiseok Lee[2], Mahmoud Yousef[2], Robert Y. Chen[3], Conrad Triebold[4], Emma McSpadden[1], Fidel Haro[5], Valeryia Aksianiuk[6], Ramaswamy Ramanujam[1], Seppe Kuehn[2,8], Arjun S. Raman[1,8,9,*]

Affiliations

[1]Duchossois Family Institute, University of Chicago, Chicago, IL, 60637
[2]Department of Ecology and Evolution, University of Chicago, Chicago, IL, 60637
[3]Department of Psychiatry, University of Washington, Seattle, WA, 98195
[4]Section of Genetic Medicine, University of Chicago, Chicago, IL, 60637
[5]Pritzker School of Medicine, University of Chicago, Chicago, IL, 60637
[6]Institute for Cell Biology, University of Bern, Bern, Switzerland
[7]Department of Internal Medicine, University of Chicago, Chicago, IL, 60637
[8]Center for Physics of Evolving Systems, University of Chicago, Chicago, IL, 60637
[9]Department of Pathology, University of Chicago, Chicago, IL, 60637

*Correspondence to: araman@bsd.uchicago.edu

Abstract

Microbiomes perform critical functions across many environments on Earth[1–3]. However, elucidating principles of their design is immensely challenging[4–7]. Using a diverse bank of human gut commensal strains and clearance of multi-drug resistant *Klebsiella pneumoniae* as a target, we engineered a functional synthetic microbiome using a process that was agnostic to mechanism of action, bacterial interactions, or compositions of natural microbiomes. Our strategy was a modified 'Design-Build-Test-Learn' approach ('DBTL+') coupled with statistical inference that learned design principles by considering only the strain presence-absence of designed communities. In just a single round of DBTL+, we converged on a generative model of *K. pneumoniae* suppression. Statistical inference performed on our model identified 15 strains that were key for community function. Combining these strains into a community ('SynCom15') suppressed *K. pneumoniae* across unrelated *in vitro* environments and matched the clearance ability of a whole stool transplant in a pre-clinically relevant mouse model of infection. Considering metabolic profiles of communities instead of strain presence-absence yielded a poor generative model, demonstrating the advantage of using strain presence-absence for deriving principles of community design. Our work introduces the concept of 'statistical design' for engineering synthetic microbiomes, opening the possibility of synthetic ecology more broadly.

52 **Main**

53 Engineering communities of microbes for desired functions ('synthetic ecology') is of
54 fundamental importance and holds great practical promise for addressing many problems facing
55 humanity[5,8,9]. So called 'top-down' approaches—reducing an already functional, whole
56 microbiome to key microbes—and 'bottom-up' approaches—designing communities one
57 bacterium at a time—have found success in creating functional communities[10–13]. However, the
58 ability to create new communities that predictably execute a desired function according to
59 principles of design, i.e. deriving 'generative' models of microbiome engineering, remains
60 immensely challenging. In large part, this is due to the daunting complexity of ecosystems: they
61 are comprised of many parts that interact with each other and the environment in dynamic and
62 unintuitive manners to give rise to emergent, collective function[6,7,14–17]. Recognition of this
63 complexity has driven recent interest in using new statistical approaches such as statistical
64 learning, deep learning, and artificial intelligence for engineering synthetic microbiomes[18–23].

65 Using a diverse collection of human gut commensal strains, we sought to engineer a
66 bacterial microbiome that could clear multi-drug resistant (MDR) *K. pneumoniae*—a pathogen
67 classified in the 'Priority 1: Critical' category of antibiotic resistant organisms by the World
68 Health Organization[24]. Towards this goal, we implemented a 'Design-Build-Test-Learn' (DBTL)
69 approach that was different from a traditional DBTL framework in two ways. First, the initial
70 round of community design was subject to a constraint: maximizing genomic diversity of
71 constituent bacterial strains. Our rationale in implementing this constraint was to minimize
72 potential functional redundancy in constructed communities. Second, a model of community
73 function was statistically learned by considering only the pattern of strain presence-absence for
74 designed communities, thereby remaining agnostic to many parameters that could influence
75 community structure and function. We term our approach 'DBTL+'. Implementing just a single
76 round of DBTL+ wherein 96 'designed microbial communities' (DMCs) were built and tested
77 resulted in an accurate generative statistical model of community design for suppressing *K.*

78 *pneumoniae* in an *in vitro* setting. Statistical inference performed on our model identified a set of
79 15 key strains that when combined into a community ('SynCom15') (i) sustainably suppressed
80 *K. pneumoniae* across various diverse *in vitro* environments, (ii) matched the clearance ability of
81 a fecal microbial transplant (FMT) in a pre-clinically relevant mouse model of infection, (iii) was
82 a safe intervention *in vivo*, (iv) could not be obviously deconstructed into a functional subset of
83 strains, and (v) did not resemble the composition of natural human gut microbiotas. We found
84 that considering the metabolic capacity of DMCs including fatty acid and nutrient metabolism—
85 appreciated mechanisms of *K. pneumoniae* suppression—instead of strain presence-absence
86 resulted in a poor generative model, highlighting the advantage of describing DMCs by their
87 strain content for deriving generative models of community design[25–27]. Our work describes a
88 potentially therapeutic, sparse synthetic microbiome made of human gut commensal bacteria for
89 treatment of MDR *K. pneumoniae* infections and, more generally, introduces the concept of
90 'statistical design' for microbial ecosystems.
91
92 *A generative model of community design for suppressing K. pneumoniae*
93 To begin our DBTL+ approach, we first isolated and whole-genome sequenced 848 gut
94 commensal strains from fecal samples of healthy donors (**Fig. 1A**,**B**; **Supplementary Table 1**)
95 (Methods). Our strain bank was enriched for the phyla Bacteroidota, Bacillota, Actinomycetota,
96 Pseudomonadota, and Verrucomicrobia, and contained a richness of diversity at the genus and
97 species levels reflecting the diversity of donor microbiomes (**Fig. 1B**, **Extended Data Fig. 1** and
98 **2; Supplementary Information**) (Methods). The possible combinatorial space of DMCs we
99 could synthesize was $2^{848}/2$—an insurmountable number. As such, we reduced the size of the
100 strain bank while maintaining the genomic diversity of the resulting set (**Fig. 1C**;
101 **Supplementary Table 2A**). We chose 46 strains as the size of our reduced strain bank
102 because creating a nearly 50-member community was the practical limit we could achieve
103 without compromising the fitness of bacteria in culture. Despite substantially reducing the size of

104 the strain bank, the possible combinatorial space of communities was still ~$2^{46}/2$, or 35 trillion,
105 possibilities. We therefore implemented a constraint to design the first round of communities.
106 Our rationale was to create communities comprised of a diverse set of strains, rather than a set
107 of strains that were closely related to each other, to maximize the potential for functional
108 diversity of a given community. Therefore, we used the UMAP space of the 46 strains to design
109 diverse communities (**Fig. 2C**; **Supplementary Table 2B**).

110 To create a diverse community of size *N*, one option could be to choose the set of *N*
111 strains that maximize dispersion across the UMAP space. This problem has been encountered
112 in the field of facilities optimization and is known as 'the discrete p-dispersion problem'[28–30].
113 However, this problem is considered 'NP complete'—a class of problem in computer science
114 that is formally hard to solve and whose solutions can be verified only in non-polynomial time.
115 Therefore, we created an algorithm to generate diverse communities (Methods). First, for a
116 community consisting of *N* out of the 46 strains in our strain bank, 10,000 communities of size *N*
117 were randomly created. Second, for each of the communities, all pairwise distances (dispersal)
118 between constituent strains were computed based on their respective distances in the UMAP
119 space. Third, for each of the communities, the dispersal values between strains were ordered
120 from largest to smallest. Finally, the community with the maximum mean dispersal of the lowest
121 30% of all dispersal values between strains was chosen as a DMC to build and test. By
122 choosing the community with the maximal mean dispersal of the most closely related strains
123 (i.e. the lowest 30%), this algorithm enforces the constraint of diversity across the whole
124 community (**Extended Data Fig. 3A**). As an example, implementing this algorithm to engineer a
125 five-member DMC would result in a bacterial community spanning different regions of the UMAP
126 space (**Fig. 1D**).

127 We created 96 DMCs in total—92 diverse DMCs, 3 replicates, and one DMC with all 46
128 strains (**Fig. 1E**; **Supplementary Table 3A**). As we had no prior for constraining the size of the
129 DMCs, our rationale was to span a wide range of membership sizes. We designed the 96 DMCs

130 to span two to 46 strains with the average size being 15 bacterial strains with 5 strains as the
131 standard deviation. As the size of the DMC increased, the Shannon diversity increased as well,
132 illustrating that our strategy of design resulted in metagenomically diverse communities
133 (**Extended Data Fig. 3B**). All DMCs were tested for their ability to suppress *K. pneumoniae*
134 MH258—an MDR strain isolated from a patient sample obtained from Memorial Sloan Kettering
135 Hospital (MH) representative of the epidemic multilocus sequence type (ST) 258 clone
136 harboring the bla_{KPC}-encoded carbapenemase. We chose this strain to use as our target for
137 suppression because it was amongst the most multi-drug resistant strains that have been
138 previously characterized, exhibiting resistance against a diversity of antibiotics[31]. DMCs were
139 co-cultured with a GFP-tagged *K. pneumoniae* strain MH258 in Brain-Heart-Infused media with
140 cysteine (BHIS) for 120 hours in an anaerobic chamber (Methods). The abundance of *K.*
141 *pneumoniae* during co-culture with DMCs was quantified through time by plating (**Extended**
142 **Data Fig. 4A**) (Methods).

143 We found that across all DMCs, *K. pneumoniae* grew for the first 24 hours from an
144 abundance between 10^6 and 10^7 to an abundance of 10^8 on average and remained constant
145 through the next 24 hours (**Extended Data Fig. 4B**). After the first 48 hours of co-culture and up
146 to 120 hours, the 96 DMCs reproducibly exhibited a range of capacity to suppress *K.*
147 *pneumoniae* spanning no suppression to suppression greater than seven orders of magnitude
148 equivalent to clearing *K. pneumoniae* given the lower limit of detection for our assay (**Fig. 1E**,
149 **Extended Data Fig. 4B-E**; **Supplementary Table 3B**). The DMC containing all 46 strains
150 ('DMC46') suppressed *K. pneumoniae* the most, while *K. pneumoniae* alone maintained the
151 highest abundance. Moreover, we found that the suppressive capacity of DMCs was unrelated
152 to the size of community composition or the presence or absence of a stereotyped taxonomic
153 signature (**Fig. 1E**, **Extended Data Fig. 4F,G**). This result suggested it was likely not the
154 presence or absence of a single strain that mediated the suppression of *K. pneumoniae*, but
155 rather a complex set of microbial interactions.

156 We trained and validated a Random-Forests (RF) machine-learning algorithm to learn a
157 statistical relationship between DMC design—defined by only the designed pattern of strain
158 presence and absence of DMCs as represented by the matrix shown in **Fig. 1E**—and the
159 ending *K. pneumoniae* abundance after co-culture with the community[32,33]. Thus, no information
160 about which strains engrafted or survived in the culture, strain dynamics during the experiment,
161 ending configuration of the community, information regarding the nature of microbial
162 interactions, or information regarding mechanism of *K. pneumoniae* suppression was
163 considered when training or validating this model. The RF model was trained on 90% of the
164 data and validated on the remaining 10% 100 times for bootstrap support, resulting in an in-
165 sample validation r^2 value of 0.98 (**Supplementary Table 4A**) (Methods).
166 We then tested the predictive capacity of our RF model for newly constructed DMCs that
167 the model had never seen as a true 'out-of-sample' test. We created 60 new DMCs spanning
168 different membership sizes that were not a part of the initial 96 DMCs and were predicted by the
169 trained RF model to span a large dynamic range of *K. pneumoniae* clearance in our assay
170 (**Supplementary Table 5A**). Thus the 60 new DMCs defined a true 'out-of-sample' set
171 generated by our RF model. We compared the abundance of *K. pneumoniae* for the 60 new
172 DMCs as predicted by our RF model versus the *K. pneumoniae* abundance we experimentally
173 observed after co-culture of each of the 60 new DMCs with *K. pneumoniae* for 120 hours. We
174 found that our RF model was predictive of the resulting *K. pneumoniae* abundance to an r^2
175 value of 0.6 ($p < 10^{-3}$) (**Fig. 1F**, **Supplementary Table 5B**).
176 Collectively, our results showed that our RF model could accurately predict the capacity
177 of a complex microbial community defined by our 46 strains to suppress *K. pneumoniae*,
178 thereby enabling engineering of new communities with desired suppressive capacity. Thus, in a
179 single round of DBTL where the first round of design was constrained by genomic diversity of
180 strain combinations (DBTL+), we derived a generative model of community design for
181 suppressing *K. pneumoniae* in BHIS media.

182

183 *Defining and characterizing SynCom15*

184 We sought to define the critical strains responsible for clearing *K. pneumoniae*. Current
185 experimental and computational approaches used to define key sets of strains responsible for
186 community function are limited in their abilities to consider higher-order, emergent microbial
187 interactions. In addition, the distribution of feature importance scores generated from our
188 predictive RF model were continuous and therefore unable to delineate groups of important
189 strains (**Extended Data Fig. 5A**) (**Supplementary Table 4B**). Moreover, because RF models
190 are tree-based, they are designed to identify individual features important for prediction, not
191 groups of features. We therefore implemented a statistical-inference based strategy initially
192 developed in the field of quantitative finance and then applied to the study of protein evolution
193 as well as to longitudinal analysis of human microbiomes for identifying groups of collectively
194 interacting parts critical for defining system function. The underlying idea is to first use statistical
195 co-variation between component parts as a proxy for interactions, then to define groups of
196 components that robustly co-vary with each other amongst systems that survive a selective
197 process. Implementing this approach has successfully identified collectives across different
198 scales of complexity: groups of stocks defining economic 'sectors', groups of amino acids
199 defining functional units of proteins ('protein sectors'), and groups of microbes within
200 microbiomes defining covarying units of therapeutic importance ('ecogroups')[15,34–39]. We
201 adapted this approach to help identify a collective group of strains critical for suppressing *K.*
202 *pneumoniae*.

203 We scored 100,000 *in silico*-generated DMCs for their predicted capacity to suppress *K.*
204 *pneumoniae* after co-culture using our RF model. We then selected the set of DMCs predicted
205 to suppress *K. pneumoniae* at least five orders of magnitude (**Fig. 2A**). The number of DMCs in
206 the resulting set was 5,752. We created an alignment of these DMCs defined by their designed
207 strain presence-absence and labeled each DMC by its *K. pneumoniae* abundance predicted

208 from the RF model (**Fig. 2B**, **Supplementary Table 6A**). We next performed Principal
209 Components Analysis (PCA) on the alignment of communities, yielding 46 principal components
210 (PCs) of data-variance. We regressed the contribution of each of the 5,752 DMCs onto each PC
211 against the predicted *K. pneumoniae* abundance to identify PCs that most associate with *K.*
212 *pneumoniae* suppression in a data-driven, unbiased manner. We found that PC46, containing
213 <0.1% data-variance, was the most associated with *K. pneumoniae* abundance (**Fig. 2C**,
214 **Supplementary Table 6B-D**) (Methods).
215 Similar to the distribution of RF importance scores, the contribution of strains onto PC46
216 was continuous precluding the ability to define groups of strains to construct communities (**Fig.**
217 **2D**, left panel; **Supplementary Table 6E**). Interestingly, the contribution of strains onto PC46
218 did not resemble the distribution of feature importance scores from the RF model, suggesting
219 that PC46 contained information that was different from the RF model (**Extended Data Fig. 5B**).
220 To use the information in PC46 to define groups of statistically interacting strains, we computed
221 the statistical similarity between all pairs of strains on PC46 (Methods). The concept behind this
222 measure is that two strains that significantly contribute to PC46 and are close together along
223 PC46 are, on average, co-present in DMCs predicted to suppress *K. pneumoniae*. Hierarchical
224 clustering of the pairwise similarity between strains illustrated a distinct block structure amongst
225 five separate groups (**Fig. 2D**, right panel; **Supplementary Table 6F**). Five strains that
226 contributed the most to defining PC46—*Clostridium innocuum*, *Clostridium symbosium*,
227 *Colinsella aerofaciens*, *Escherichia coli*, and *Bacteroides xylanisolvens*—formed a group that
228 we term 'Block 1' (**Fig. 2D**, right panel, orange group). The following ten strains that contributed
229 to PC46—*Lacrimispora celerecrescens*, *Bacteroides caccae*, *Blautia faecis*, *Blautia obeum*,
230 *Clostridium scindens*, a *Bifidobacterium* species, *Megasphera massiliensis*, *Coprococcus*
231 *comes*, *Mitsuokella jalaludinii*, and *Blautia producta*—formed a group that we term 'Block 5'
232 (**Fig. 2D**, right panel, green group). Blocks 1 and 5 exhibited collective similarity amongst each
233 other; we term this group of strains 'SynCom15'—a 15-member group comprised of statistically

234 interacting strains that are co-present in communities predicted to clear *K. pneumoniae*. In
235 contrast to SynCom15, three other groups of strains were statistically inferred to be co-absent in
236 communities predicted to clear *K. pneumoniae*. These groups were comprised of 7, 10, and 14
237 strains; we term these groups 'Block 2', 'Block 3', and 'Block 4' respectively (**Fig. 2D,** right
238 panel, red group, brown group, and yellow group) (**Supplementary Table 6G**).
239 We hypothesized that SynCom15 would be efficacious at clearing *K. pneumoniae* across
240 different environments because it was predicted to contain the key, critical species for DMC
241 function. We built and tested SynCom15 as well as all other Blocks for their capacity to clear *K.*
242 *pneumoniae* across three unrelated media conditions: BHIS, media created from the cecal
243 extracts of germ-free (GF) mice, and media created from the cecal extracts of specific-
244 pathogen-free (SPF) mice treated with broad spectrum antibiotics (Ab-treated SPF) (**Fig. 2E**, left
245 panel) (Methods). As a comparator to SynCom15 and the other Blocks, we also tested
246 DMC46—the community that suppressed *K. pneumoniae* the most in BHIS media. Notably, our
247 results clearly illustrated the environmental dependence of community efficacy. Blocks 2 and 3
248 were consistently ineffective at suppressing *K. pneumoniae* across environments while Blocks
249 1, 4 and 5 were able to suppress *K. pneumoniae* depending on the environment in which they
250 were tested—Block 1 in BHIS and Blocks 4 and 5 in GF cecal extract. Thus, Blocks 1, 4 and 5
251 were conditionally effective. In contrast, we found that DMC46 and SynCom15 suppressed *K.*
252 *pneumoniae* across all three conditions and were therefore unconditionally effective. DMC46
253 cleared *K. pneumoniae* across all environments. SynCom15 suppressed *K. pneumoniae* five
254 orders of magnitude in BHIS, cleared *K. pneumoniae* in GF cecal extracts, and suppressed *K.*
255 *pneumoniae* greater than four orders of magnitude in Ab-SPF cecal extracts. (**Fig. 2E**, right
256 panels) (**Supplementary Table 7**).
257 Thus, our strategy of statistical inference performed on the RF model of community
258 design defined SynCom15—a phylogenetically diverse 15-member community—that

259 suppressed *K. pneumoniae* across diverse environmental contexts in a manner similar to
260 DMC46—the community containing all 46 strains.
261
262 *SynCom15 clears K. pneumoniae in a pre-clinically relevant mouse model of infection*
263 Because DMC46 and SynCom15 were unconditionally effective at clearing *K.*
264 *pneumoniae in vitro*, we sought to test the ability of both communities to clear *K. pneumoniae* in
265 a more complex, clinically relevant environment. We evaluated the efficacy of DMC46 and
266 SynCom15 in a mouse model of infection. To mimic a clinically relevant scenario, we did not
267 use germ-free mice (mice without a microbiome). Rather, we treated SPF mice with broad
268 spectrum antibiotics to deplete their gut microbiota then infected them with *K. pneumoniae*—a
269 sequence of events commonly encountered in patients who acquire MDR *K. pneumoniae*
270 infection. Additionally, we singly-housed mice to ensure that no sharing of microbes by
271 coprophagia amongst animals would affect microbiome composition during and post-antibiotic
272 treatment[40]. Singly-housed antibiotic-treated SPF mice infected with *K. pneumoniae* MH258
273 were given either (i) saline (PBS), (ii) a heterologous whole stool transplant derived from mice
274 ('Fecal Microbial Transplant', FMT), (iii) Block 1, (iv) Block 2, (v) DMC46, or (vi) SynCom15 as
275 interventions for three sequential days after infection (Methods). Blocks 1 and 2 were given as
276 bacterial communities that were either conditionally efficacious across *in vitro* conditions or
277 unable to clear *K. pneumoniae* across any *in vitro* condition respectively. Fecal samples were
278 collected and *K. pneumoniae* abundances were tracked through the course of the experiment
279 by plating (**Fig. 3A**).
280 We found that Block 1, and Block 2 did not suppress *K. pneumoniae* relative to saline.
281 The FMT suppressed *K. pneumoniae* three orders of magnitude one day after the last gavage
282 and up to six orders of magnitude from four days after the last gavage until the end of the
283 experiment. DMC46 suppressed *K. pneumoniae* three orders of magnitude one day after the
284 last gavage, four orders magnitude four days after the last gavage, and six orders of magnitude

285 nine days after the last gavage. Thus, DMC46 was able to suppress *K. pneumoniae* but
286 exhibited slow kinetics of response compared to the FMT. In contrast, SynCom15 rapidly
287 suppressed *K. pneumoniae*, resulting in a reduction of abundance by five orders of magnitude
288 one day after the last gavage. Additionally, SynCom15 cleared *K. pneumoniae* four days after
289 the last gavage and maintained clearance through nine days after the last gavage (**Fig. 3B**)
290 (**Supplementary Table 8**). These results highlighted the rapid and sustained efficacy of
291 SynCom15 in clearing *K. pneumoniae in vivo* as well as the utility of reducing the community
292 size from the 46 strains defining DMC46 to the inferred key 15 strains defining SynCom15.
293 Taxonomic profiling of fecal samples procured through the experiment revealed that 10
294 of the 15 strains in SynCom15 engrafted in at least one of the mice within the cohort (**Fig. 3C**)
295 (Methods). Dynamics of SynCom15 strains showed that 5 of the 10 strains were present at
296 detectable fractional abundances throughout the course of the experiment—*C. symbiosum*, *B.*
297 *xylanisolvens*, *C. innocuum*, *B. obeum*, and *B. caccae* (**Extended Data Fig. 6**). Together, these
298 results illustrated that the engraftment and strain dynamics of SynCom15 in mice did not follow
299 obvious phylogenetic trends.
300 Dynamics of microbiota diversity and structure within the infected mice treated with
301 SynCom15 mirrored that of the FMT and returned the state of the microbiota to that observed
302 prior to antibiotic treatment (**Fig. 3D,E**; **Supplementary Table 9A,B**) (Methods). At the
303 phylogenetic description of phylum, class or family, we observed that treatment via FMT and
304 SynCom15 resulted in similar ending configurations of the microbiota (**Fig. 3F**). However, at the
305 genus-level description, we observed differences between the ending microbiota configuration
306 of mice treated with FMT or SynCom15. Treatment with FMT resulted in the return of
307 *Duncaniella* and *Paramuribaculum* (genera belonging to the order Bacteroidales). Treatment
308 with SynCom15 resulted in detectable presence of the genera *Bacteroides*, derived from the *B.*
309 *xylanisolvens* strain in SynCom15, in addition to a bloom of *Bifidobacterium* (**Fig. 3G**, **Extended**
310 **Data Fig. 6**). These results illustrated that treatment with SynCom15 yields a return to a diverse

311 microbiota that resembles a more human-like signature despite being engrafted in mice.
312 Histology of the mouse colon showed that SynCom15 was well tolerated as an intervention
313 showing no evidence of inflammation or tissue insult (**Extended Data Fig. 7)**.
314 Collectively, our results demonstrated that SynCom15 successfully cleared *K.*
315 *pneumoniae* in a pre-clinical mouse model of infection—a result consistent with our findings
316 showing that SynCom15 is unconditionally effective across *in vitro* environments. Additionally,
317 we found that treatment with SynCom15 was safe from the standpoint of microbiota recovery
318 and tissue injury. Together, these results point towards the therapeutic potential of SynCom15
319 for clearing *K. pneumoniae* from the gut.
320
321 *Compositional characterization of SynCom15*
322 Given the safety and efficacy of SynCom15, we sought to further characterize its
323 compositional content. First, we tested each strain of SynCom15 individually for its ability to
324 suppress *K. pneumoniae* in BHIS. We found that no individual strain suppresses *K. pneumoniae*
325 greater than two orders of magnitude and eleven of the strains suppressed *K. pneumoniae* only
326 up to one order of magnitude (**Extended Data Fig. 8A, Supplementary Table 10A**). Moreover,
327 the four strains that suppressed *K. pneumoniae* two orders of magnitude were found in Block 1,
328 a Block that suppressed *K. pneumoniae* comparable to SynCom15 in BHIS but was less
329 efficacious by several orders of magnitude in other environments without the addition of the
330 other ten strains comprising SynCom15. Thus, including the eleven strains that have only a
331 modest individual effect on suppressing *K. pneumoniae* in BHIS media was important for
332 achieving the suppressive capacity of SynCom15 in other environments. These findings
333 highlight the complex nature of the ability of SynCom15 to suppress *K. pneumoniae* across
334 environments.
335 Next, we interrogated whether data from our mouse experiment could inform which
336 strains of SynCom15 are important for functionality. We built two communities—(i) a community

337 constituting strains that consistently engrafted the mice (10 species) and (ii) a community
338 constituting strains that were consistently detected in mice across all timepoints (5 species)
339 (**Extended Data Fig. 8B**). The first community suppressed *K. pneumoniae* two orders of
340 magnitude in BHIS and did not suppress *K. pneumoniae* in GF cecal extract media; the second
341 community suppressed *K. pneumoniae* one order of magnitude in BHIS and did not suppress *K.*
342 *pneumoniae* in GF cecal extract media (**Extended Data Fig. 8C,D**, **Supplementary Table**
343 **10B**). Thus, the inclusion of strains constituting SynCom15 that were not statistically detectable
344 in the mouse fecal pellets was important for achieving the clearance of *K. pneumoniae* we
345 observed across environments.

346 Recent results have claimed the critical importance of *E. coli* in clearing *K.*
347 *pneumoniae*[25]. This motivated us to test the importance of our *E. coli* strain for SynCom15. We
348 therefore built two more communities—SynCom15 without *E. coli* and the community comprised
349 of strains that engrafted the mouse without *E. coli* (**Extended Data Fig. 8B**). Removing *E. coli*
350 from either community resulted in a decrease in *K. pneumoniae* suppression by just half an
351 order of magnitude in BHIS and no difference in suppression in GF cecal extract media
352 (**Extended Data Fig. 8C,D**, **Supplementary Table 10B**). Additionally, we note that the Block 1
353 community—a five-member community containing *E. coli*—was unable to suppress *K.*
354 *pneumoniae* in mice more than a saline intervention at day 11 and day 16 post infection (**Fig.**
355 **3B**). Recent studies have also suggested augmenting *E. coli* with large, diverse communities to
356 clear *K. pneumoniae*[25]. Our data provide a contrasted result: DMC46, a diverse community
357 comprised of 46 strains, contained a strain of *E. coli* but was not as effective as SynCom15,
358 comprised of 15 strains including the same *E. coli* strain, at suppressing *K. pneumoniae* in mice
359 (**Fig. 3B**).

360 Collectively, these observations illustrated that the efficacy of SynCom15 as a
361 community that suppresses *K. pneumoniae* across different environments cannot be solely
362 ascribed to the presence of any single strain, including *E. coli*, or an obvious subset of strains

363 gleaned from analysis of our mouse experiments. Moreover, coarse community descriptions,
364 like community diversity for instance, do not provide an explanation for our results. In contrast,
365 our findings highlight the utility of evaluating community function through our statistical approach
366 that considers emergent, and potentially non-obvious properties of the structure-function
367 relationship for communities.
368
369 *Comparison of SynCom15 with composition of healthy human fecal microbiomes*
370 We next explored the extent to which SynCom15 was represented across healthy
371 humans who provided FMTs from which we created our strain bank. We first interrogated the
372 prevalence of the genera constituting SynCom15 in fecal samples from healthy donors. We
373 found that the genera represented in SynCom15 reflected a diverse minority of the totality of
374 genera observed across the set of healthy gut microbiomes (**Fig. 4A**). Next, we interrogated the
375 prevalence of the SynCom15 species across the fecal samples of the healthy donors (Methods).
376 We found that no healthy human microbiome contained more than eleven of the SynCom15
377 species above a fractional abundance of 0.1% (**Fig. 4B**; **Supplementary Table 11**). Moreover,
378 we found certain SynCom15 species to be remarkably sparse in their prevalence across donors.
379 *M. jalaludinii* was not detectable in any donor; *M. massiliensis* was detectable in two donors; *C.*
380 *symbiosum* in three donors; and *C. scindens* in four donors. Amongst strains that were most
381 prevalent, *B. obeum* and *B. faecis* were detectable in 20 donors; *L. celerecrescens* in 14
382 donors; *C. comes* in 13 donors; *B. caccae* in 12 donors. Finally, we interrogated the fractional
383 abundance of SynCom15 species across the fecal samples of the 22 healthy donors. We found
384 SynCom15 species were present at a relative abundance of less than 5% across all donors,
385 with a majority of species being found at a relative abundance of less than 0.5% (**Fig. 4C**).
386 Together, these results illustrated two conclusions. First, the composition of SynCom15
387 was distinct from that found across healthy human gut microbiotas. This is either because
388 SynCom15 does not exist in the healthy samples from our cohort or because several of

389 SynCom15 strains are undetectable by our sequencing methods due to their low abundance.
390 Second, the strains comprising SynCom15 were low prevalence and abundance amongst fecal
391 samples of healthy donors. This result highlights the power of generating and using broadly
392 diverse strain banks for engineering synthetic bacterial communities as compared to strain
393 banks reflecting the compositional abundance and prevalence distributions gleaned from
394 analysis of natural human microbiomes.
395
396 *Community metabolism poorly predicts K. pneumoniae suppression*
397 Engineering SynCom15 was based on statistical analysis of a model that described
398 DMCs by their pattern strain presence-absence and their capacity to clear *K. pneumoniae*.
399 Thus, the model was not constructed using any information about mechanism of action.
400 Previous results have suggested the importance of media acidification and nutrient competition
401 as mechanisms by which complex bacterial communities could suppress *K. pneumoniae*[25–27].
402 Therefore, we compared the metabolic profiles of the five DMCs that suppressed *K.*
403 *pneumoniae* the most against the five DMCs that suppressed *K. pneumoniae* the least amongst
404 the 96 DMCs we had previously tested in BHIS (**Fig. 5A**, left panel; **Extended Data Fig. 9**;
405 **Supplementary Table 12A**) (Methods). We analyzed the profile of 118 metabolites across the
406 most and least suppressive DMCs after being co-cultured with *K. pneumoniae* for 72, 96, and
407 120 hours.
408 The metabolite patterns that distinguished DMCs that suppressed *K. pneumoniae* from
409 those that did not centered around two metabolic axes: concentrations of fatty acids (FAs) with
410 an emphasis on short-chain fatty acids and amino acids (**Supplementary Table 12B**). With
411 respect to FAs, the most suppressive DMCs produced phenylacetic acid, valeric acid, hexanoic
412 acid, and 5-aminovaleric acid and consumed lactic acid as well as succinic acid. With respect to
413 amino acids, the most suppressive DMCs consumed either (i) amino acids with non-polar side
414 chains (phenylalanine, alanine, isoleucine, leucine, valine) or (ii) glutamic acid and its

415 associated derivative 5-oxoproline (**Fig. 5A**, right panel). Metabolic profiling of SynCom15 co-
416 cultured with *K. pneumoniae* in BHIS revealed a similar trend. SynCom15 produced the same
417 FAs as the most suppressive DMCs, but also produced lactic acid as opposed to consuming it.
418 SynCom15 also consumed all the amino acids that the most suppressive DMCs consumed (**Fig.**
419 **5A**, right panel; **Supplementary Table 12C**). We also performed metabolic profiling of fecal
420 pellets collected from mice treated with either SynCom15 or saline in the experiment described
421 in **Fig. 3A**. Consistent with our *in vitro* results, we found a statistically significant increase in FA
422 production on day 10 and amino acid depletion on day 12 in infected mice given SynCom15
423 (**Fig. 5B**, **Supplementary Table 12D**). Our *in vitro* and *in vivo* results were in accordance with
424 previously published studies demonstrating the importance of environmental acidification and
425 nutrient competition as mechanisms by which MDR *K. pneumoniae* could be suppressed.
426 Furthermore, these results point to metabolic axes that are shared between the function of
427 SynCom15 in *in vitro* and *in vivo* conditions, suggesting a way that translatability of suppressive
428 capacity across distinct environments could be manifest.
429 We reasoned that if the mechanism of suppression was exclusively related to FA
430 production and amino acid depletion, we could build a generative statistical model of community
431 design based on the metabolite profile of a large number of DMCs spanning a range of *K.*
432 *pneumoniae* suppression. This would represent a more thorough test of the sufficiency of FA
433 production and nutrient depletion to explain how DMCs clear *K. pneumoniae*. Thus, we
434 performed metabolic profiling of 81 DMCs that we had designed and tested in BHIS media for
435 their capacity to suppress *K. pneumoniae* (**Supplementary Table 13A**). We removed 15 DMCs
436 from our analysis because they were poorly profiled across metabolite features. Metabolite
437 profiles were measured at 72, 96, and 120 hours of co-culture with *K. pneumoniae*. We also
438 performed metabolic profiling of the 60 DMCs that previously served as the 'out-of-sample'
439 DMCs at 72, 96, and 120 hours of co-culture with *K. pneumoniae* (**Supplementary Table 13B**).
440 We trained and validated an RF model on the metabolic profiles of the 96 DMCs to predict *K.*

441 *pneumoniae* abundance after 120 hours of co-culture (Methods). We then evaluated the
442 capacity of our trained model to predict the *K. pneumoniae* abundance of the 60 'out-of-sample'
443 DMCs after 120 hours of co-culture using their metabolic profile. We found that the RF model
444 trained on metabolite profiles was a markedly poor predictor of the *K. pneumoniae* abundance
445 of the 60 out-of-sample DMCs, attaining no predictive power with an r^2 value of 0.0048 (**Fig. 5C**,
446 **Supplementary Table 13C**). Following this result, as expected the predictive capacity of the RF
447 model built on metabolite profiles shared no similarity in predictive capacity with the RF model
448 built on strain presence-absence of DMCs that was highly predictive of *K. pneumoniae*
449 abundance (**Fig. 5D**).

450 To understand why the metabolite profile of a community was a poor predictor of *K.*
451 *pneumoniae* abundance, we interrogated the structure of metabolite profiles across the DMCs
452 used to train the model. We found that the neighborhood of metabolite space where there were
453 DMCs that suppressed *K. pneumoniae* also contained poorly suppressive DMCs. That is, the
454 metabolic landscape of DMCs was 'rugged'—interspersed with peaks and valleys of
455 suppressive capacity—rather than smooth (**Fig. 5E**, left panel; **Supplementary Table 14A**).
456 This result demonstrated there was a degeneracy of different, unrelated metabolite profiles
457 associated with clearing *K. pneumoniae*, resulting in a predictive model that was overfit to the
458 training set and therefore unable to generate new functional communities (**Extended Data Fig.**
459 **10**, **Supplementary Table 14B**). Consistent with this result, we found DMCs that were highly
460 suppressive of *K. pneumoniae* shared similar metabolite profiles with DMCs that exhibited
461 intermediate to low suppression of *K. pneumoniae* (**Extended Data Fig. 11**, **Supplementary**
462 **Table 14C,D**). In contrast, the landscape of DMCs defined by strain presence-absence was
463 smooth, increasing in the capacity to suppress *K. pneumoniae* from negative to positive along
464 the first principal component (**Fig. 5E**, right panel; **Supplementary Table 14E**). Thus,
465 describing DMCs by their strain presence-absence defined a space that was co-linear with *K.*
466 *pneumoniae* suppression thereby enabling learning an accurate statistical model of design.

467 Collectively, these results show that design based on a metabolic profile comprising our
468 targeted panel of features (amino acids, aromatics, branch-chained fatty acids, indoles, phenolic
469 aromatics, and short-chained fatty acids) may not be a reliable strategy for engineering
470 communities that clear *K. pneumoniae* in a predictable manner. Our findings highlight the utility
471 of considering the more coarse-grained description of strain presence-absence in creating
472 generative models of community design.
473
474 *Discussion*
475 Using clearance of MDR *K. pneumoniae* as a target function, we engineered a defined,
476 sparse microbiome—SynCom15—that is complex, safe, efficacious, and distinct from natural
477 human gut microbiome compositions using a statistical approach for community design. Our
478 results shed light on several notable findings.
479 First, merely designing genetically diverse communities did not guarantee creating
480 functional communities. However, imposing the constraint of genetic diversity on the 'Design'
481 portion of DBTL was crucial for reducing the space of possible DMCs and was a particularly
482 informative space for learning a generative statistical model. Indeed, extremely limited sampling
483 (building and testing 96 out of the immense number of possible DMCs) was sufficient to
484 converge on an accurate model of design *in vitro*. These results suggest a deep connection
485 between the phylogenies of strains and the collective functions encoded by microbial
486 communities, opening the possibility of phylogenetic-based 'bottom-up' design. The
487 development of emerging methods for parametrizing functional differences amongst strain-level
488 variants through considering their evolutionary history across the bacterial tree-of-life will be
489 useful for testing this idea in the future[41].
490 Second, accurately translating microbiome function from specific *in vitro* settings to other
491 *in vitro* and *in vivo* environments has historically been a significant challenge. Our data showed
492 that the generative model resulting from DBTL+ was insufficient for translating community

493 function across different environments. However, the constraints of the model were sufficient for
494 engineering a microbiome—SynCom15—that successfully translated function across
495 environments. To understand why this may be, we draw a parallel to learning theory in
496 computer science. A well-known problem in building models is creating statistical
497 representations that are 'overfit' to training environments. Analogously, performing DBTL+ in a
498 single environment, like BHIS, resulted in a generative model that was 'overfit' to the
499 environment in which DMCs were tested. A key insight that results from our work is that learning
500 the constraints on the model in a single environment enabled generalization of function to new
501 environments (e.g. cecal extract medias and SPF-infected mice). This finding is consistent with
502 emerging evidence suggesting that a way that the evolutionary process can generate adaptable
503 systems is not selecting for individual systems that function per se, but by selecting for
504 underlying structural regularities amongst ensembles of systems that function[42]. Using structural
505 regularities across functional systems as a criteria for design may create new systems where
506 variance in a core function is far lower than the variance encountered across different
507 environments, thereby enabling translatability. By inferring conserved statistical patterns across
508 thousands of DMCs that were predicted to highly suppress *K. pneumoniae*, our approach of
509 statistical inference may be an analytical manifestation of this principle.

510 Third, our results demonstrate how using metabolite information spanning previously
511 appreciated mechanisms by which *K. pneumoniae* can be suppressed results in a poor
512 generative model of community design. These findings suggest that likely, there are a myriad of
513 mechanisms by which the clearance of *K. pneumoniae* can be realized. These mechanisms
514 may be included in metabolic panels encompassing a broader set of features than ours or
515 revealed by other '-omics'-based panels that are becoming more common in microbiome
516 studies such as proteomics or transcriptomics. While future efforts aimed at collecting such
517 large datasets may be warranted to further elucidate mechanisms of *K. pneumoniae*

518 suppression and clearance, our results demonstrate that such information is unnecessary for
519 creating generative models of community design.

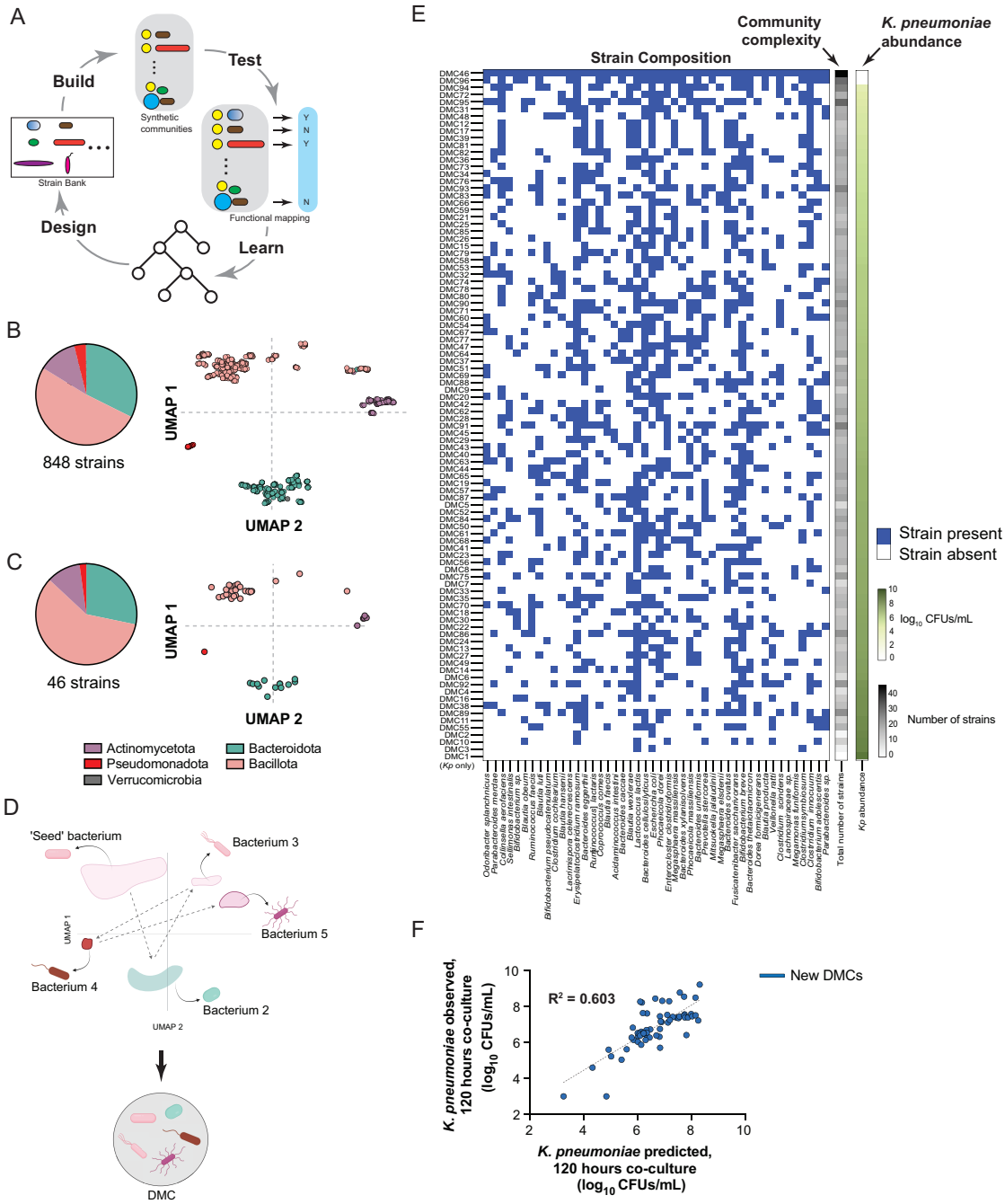
520 Fourth, SynCom15 was more efficacious at suppressing *K. pneumoniae* in mice
521 compared to DMC46—a 46-member community that contained the 15 strains defining
522 SynCom15. This result highlights the functional power of defined small bacterial communities in
523 contrast to recent studies advocating engineering large communities spanning 50 to greater
524 than 100 strains[10,25]. In addition to the gain in clearance capacity of *K. pneumoniae*, we stress
525 that the ability to engineer sparse, functional bacterial communities is a tremendous advantage
526 from a manufacturing and regulatory standpoint for creating therapeutic consortia for clinical
527 use[43]. Using DBTL+ coupled with statistical inference could be a procedure for achieving this
528 goal in an efficient manner.

529 Given previous studies highlighting the immense complexity between structure-function
530 relationships in microbial ecosystems, it may be expected that lots of high-content
531 measurements or complex computational models trained on many parameters are necessary
532 pre-requisites for deriving generative design principles of functional microbial
533 communities[10,14,16,44,45]. Consistent with this notion, existing efforts have utilized several different
534 avenues of knowledge to inform community design. These include (i) sophisticated modeling of
535 dynamical interactions between microbes and of the community as a whole, (ii) detailed
536 mechanistic knowledge of microbial interactions or mechanisms underlying a desired target
537 function, (iii) knowledge about the presence or absence of specific biological pathways encoded
538 within bacterial genomes comprising communities, (iv) knowledge about existing human
539 microbiome composition and structure, or (v) using the existence of natural communities with
540 desired functional traits (e.g. a fecal sample that resists colonization of gut pathogens) to reduce
541 community size by serial iterative rounds of screening[10–13,15,22,23,25,27,46–49]. Our results paint a
542 substantially different picture. We find that merely the pattern of strain presence-absence
543 coupled with the performance of a remarkably small number of designed diverse communities is

544 sufficient to (i) derive statistical generative models of community design *de novo* using relatively
545 simple learning algorithms (e.g. an RF machine-learning model) and (ii) engineer communities
546 whose functional capacity is translatable into new and markedly more complex environments. In
547 analogy to the evaluation of computational algorithms, our two-step approach—(i) using
548 proteome content to reduce our strain bank from 848 to 46 strains and (ii) implementing DBTL+
549 with statistical inference—is substantially compressive, able to navigate a remarkably high-
550 dimensional space to converge on SynCom15 with little information relative to the starting
551 combinatorial complexity (**Supplementary Discussion**, **Extended Data Fig. 12**). A likely
552 driving force behind our results for the target function of *K. pneumoniae* suppression is that in
553 contrast to the apparent complexity of microbial ecosystems, profoundly low-dimensional
554 representations of structure-function relationships exist and can be discovered in a facile
555 manner by placing statistical patterns of phenomenology before biological understanding—an
556 emerging viewpoint that has been the subject of some recent efforts in microbiome studies and
557 has rapidly found immense success in the form of deep-learning models at other scales of
558 biology, namely synthetic protein design[15,20,21,50–54]. Following this we note that our approach
559 does not consider mechanisms of action at any scale nor compositional information about
560 natural microbiomes and their associated functions. As the test ('T') module in our DBTL+
561 framework can be swapped out for theoretically any function with an assay, we pose that our
562 approach could, in principle, enable the statistical design of functional microbial communities
563 distinct from those found in nature and the pursuit of synthetic ecology more broadly.

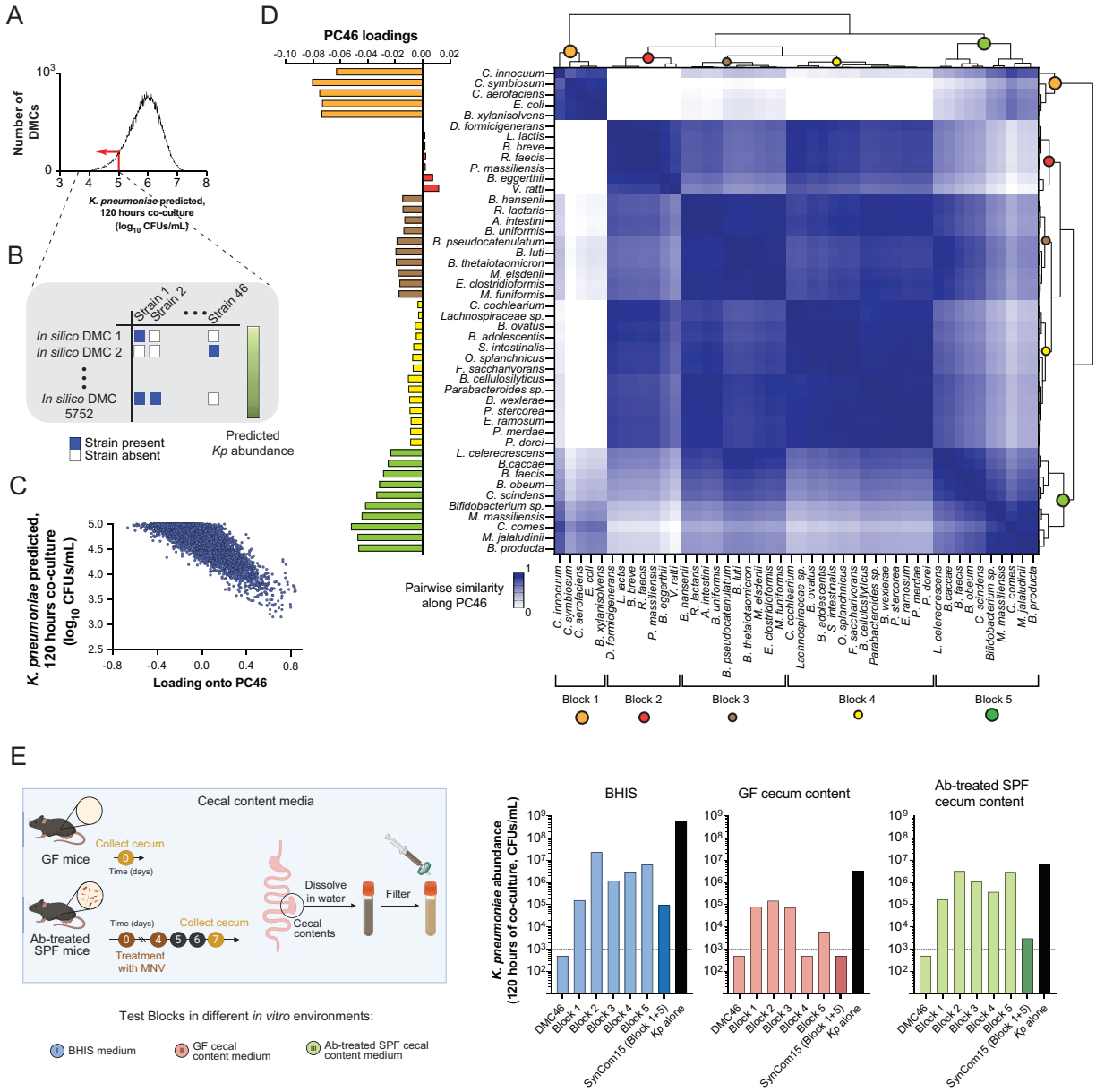
Figures and Extended Figures

Oliveira et al., Figure 1



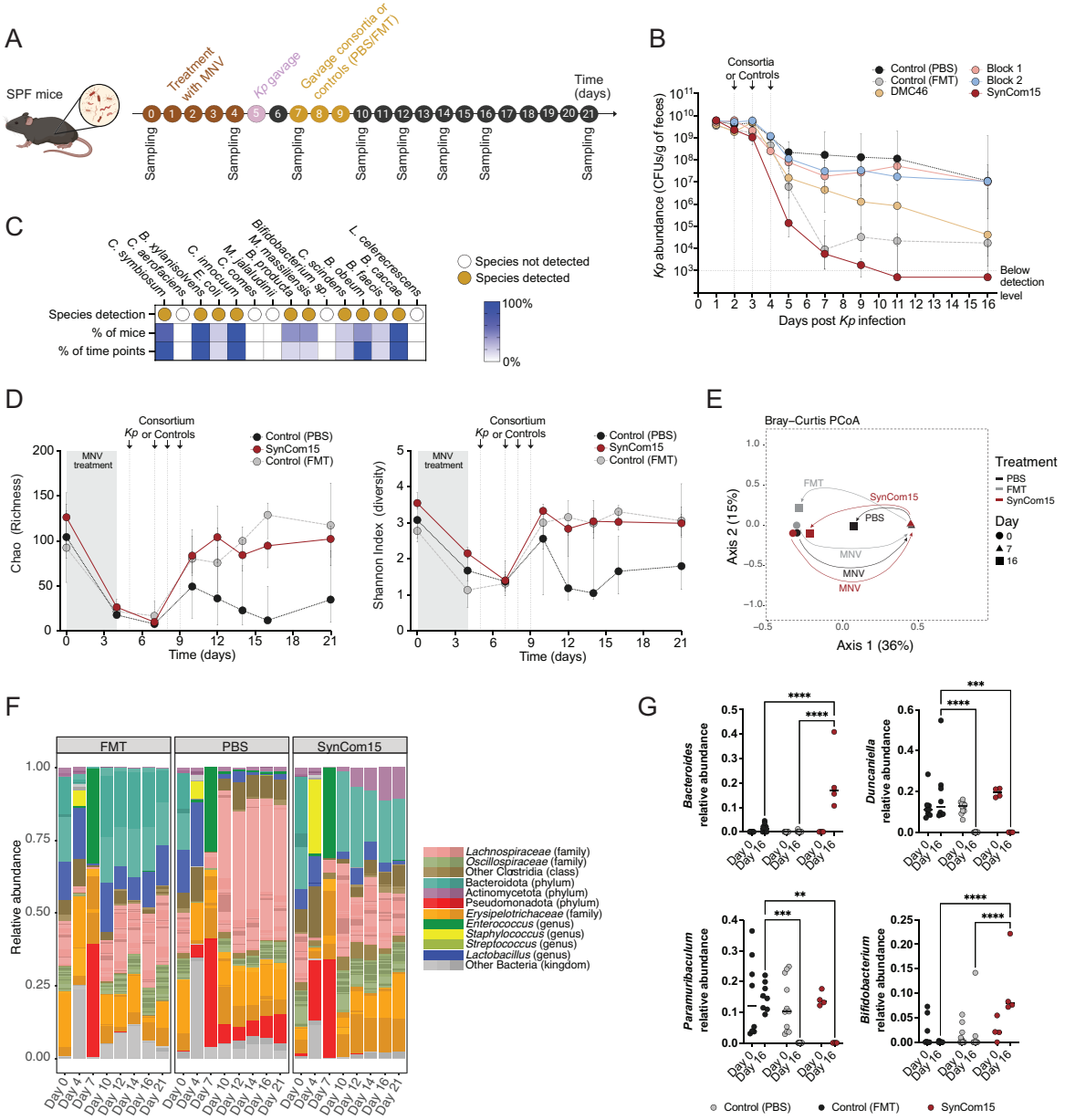
572 **Fig. 1. A generative model for engineering communities that suppress *K. pneumoniae*.**
573 **(A)** Workflow of a standard Design-Build-Test-Learn (DBTL) framework. Communities are
574 designed (D) and built (B) from a strain bank, tested (T) for desired function, and a statistical
575 model mapping community composition with function is learned (L). New communities are then
576 designed based on the learned model and the process is iterated. **(B,C)** Diversity of full strain
577 bank (panel B) and subset of strain bank used to make Designed Microbial Communities
578 (DMCs) (panel C) described at phylogenetic level of phylum. **(D)** Schematic for design of a five-
579 member DMC. 'Seed' bacterium is a randomly chosen member of our strain bank. **(E)**
580 Engineered DMCs (rows) described by strain composition (columns). Blue pixels mean that
581 strain is included in designed community; white pixels mean that strain is not included in the
582 designed community. Each row is labeled by the number of strains within the DMC ('Community
583 complexity') and the *K. pneumoniae* abundance after 120 hours of co-culture in BHIS media ('*K.*
584 *pneumoniae* abundance'). Rows are ordered by their ability to suppress *K. pneumoniae* after
585 120 hours of co-culture. 'DMC1', the last row, is *K. pneumoniae* in monoculture ('*Kp* only'). **(F)**
586 *K. pneumoniae* abundance predicted by RF model for 60 new DMCs not included in panel E (x-
587 axis) versus *K. pneumoniae* abundance observed after 120 hours of co-culture with the 60 new
588 DMCs (y-axis). RF model was trained and validated to predict *K. pneumoniae* abundance after
589 120 hours of co-culture using only the designed strain presence-absence matrix in panel E as
590 data.
591

Oliveira et al., Figure 2



604 **Fig. 2. Defining SynCom15 and evaluating its capacity to suppress *K. pneumoniae* across**
605 **different environments. (A,B)** Histogram of predicted *K. pneumoniae* abundance for 100,000
606 *in silico* generated DMCs. Red arrow is a threshold of predicted *K. pneumoniae* suppression;
607 DMCs to the left of the arrow were selected to create an alignment of 5,752 DMCs defined by
608 their pattern of strain presence-absence (panel A). Each *in silico* DMC is labeled by the
609 predicted *K. pneumoniae* abundance after 120 hours of co-culture (green bar) (panel B). **(C)**
610 Contribution of each of the 5,752 DMCs onto the 46th principal component (PC46) of the matrix
611 in panel B (x-axis) versus predicted *K. pneumoniae* abundance associated with each DMC (y-
612 axis). **(D)** Contribution of each strain onto PC46 (left panel). Right panel shows hierarchically
613 clustered strain-strain matrix where each entry is the similarity in contribution onto PC46
614 between two strains. Blocks 1 through 5 are defined according to the clustering pattern (colored
615 dots in dendrogram). Bars in left panel are colored according to which Block each strain
616 belongs. **(E)** Workflow for creating cecal extract media from germ-free ('GF') and antibiotic
617 treated specific pathogen free ('Ab-treated SPF) mice (left panel). *K. pneumoniae* abundance
618 (y-axis) for DMC46, all Blocks, and SynCom15 (darker shade) after 120 hours of co-culture in
619 BHIS (blue), GF cecal extract media (salmon), and Ab-treated SPF cecal extract media (green).
620 *K. pneumoniae* abundance after 120 hours of monoculture ('*Kp* alone') in each media is shown
621 in black. Dashed line is detection limit of assay.
622

Oliveira et al., Figure 3



628 **Fig. 3. SynCom15 sustainably clears *K. pneumoniae* in a pre-clinically relevant mouse**
629 **model of infection. (A)** Specific pathogen free (SPF) mice are treated with metronidazole,
630 neomycin, and vancomycin (MNV) (brown), then infected with *K. pneumoniae* ('*Kp* gavage',
631 pink), then given either a mouse fecal microbial transplant (FMT), saline (PBS), DMC46, Block
632 1, Block 2, or SynCom15 (beige). Fecal samples are collected at select days delineated in the
633 schematic as 'Sampling'; mice are sacrificed after day 21. **(B)** Median fecal abundance of *K.*
634 *pneumoniae* (y-axis) versus time (x-axis). Vertical dashed lines on days 2, 3, and 4 reflect
635 gavage of bacterial communities or controls ('Consortia or Controls'). Error bars indicate
636 interquartile range. **(C)** Engraftment statistics and relative presence of SynCom15 strains
637 through the experiment. **(D)** Median Chao and Shannon diversity indices (y-axes) versus time
638 (x-axes) for SPF mice treated with MNV, infected with *K. pneumoniae* ('*Kp*'), and given PBS,
639 FMT, or SynCom15. Error bars indicate interquartile range. **(E)** PCoA of fecal microbiota for
640 SPF mice on day 0, 7, and 16 of experiment; colored shape is centroid for indicated cohort. **(F)**
641 Distribution of average relative abundance for fecal microbiota through time (x-axis) for infected
642 mice treated with FMT (left panel), PBS (middle panel), or SynCom15 (right panel). Distributions
643 are defined spanning kingdom to genera-level descriptions. **(G)** Relative abundance of
644 *Bacteroides*, *Duncaniella, Paramuribaculum* and *Bifidobacterium* genera that are differentially
645 abundant amongst infected mice treated with FMT, saline, or SynCom15 prior to antibiotic
646 treatment (day 0) and at day 16 after treatment (equivalent to 11 days after infection with *K.*
647 *pneumoniae*). Statistical tests performed are two-way ANOVA; **$p < 0.01$; ***$p < 0.001$; ****$p <$
648 0.0001.
649

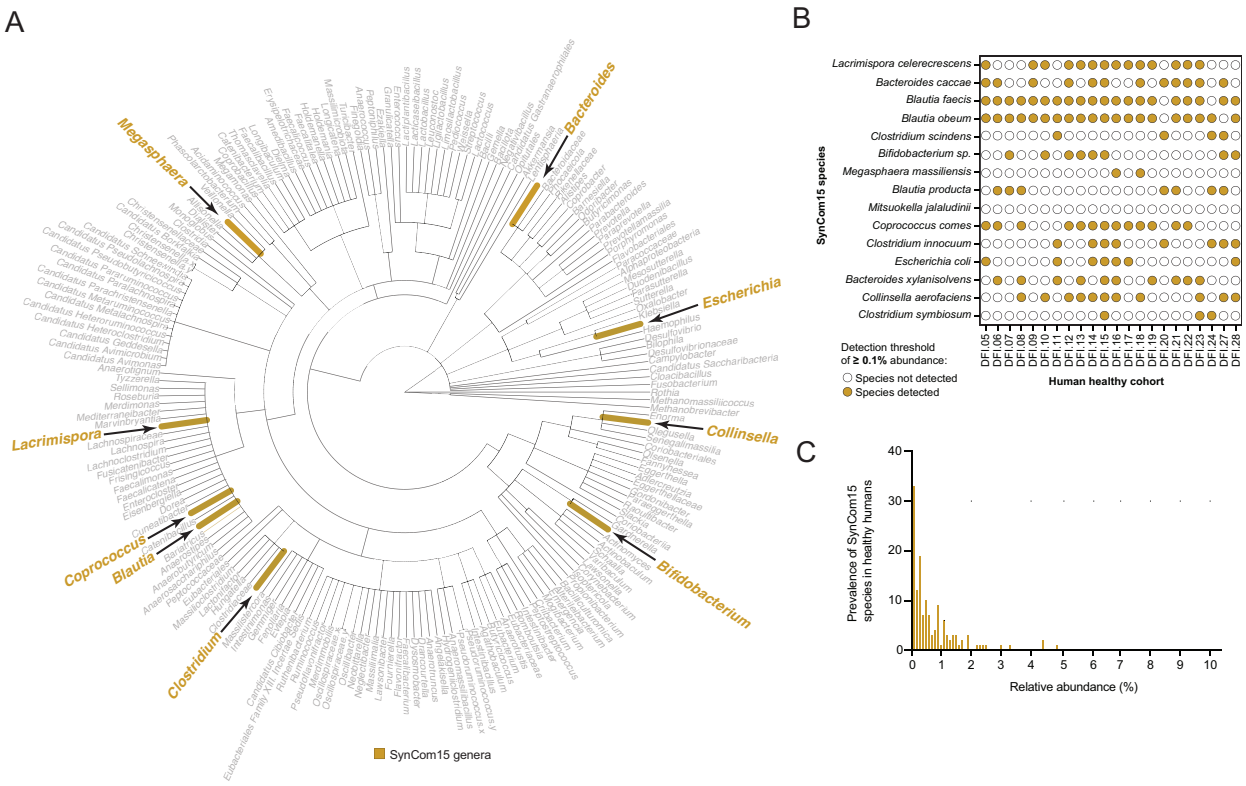
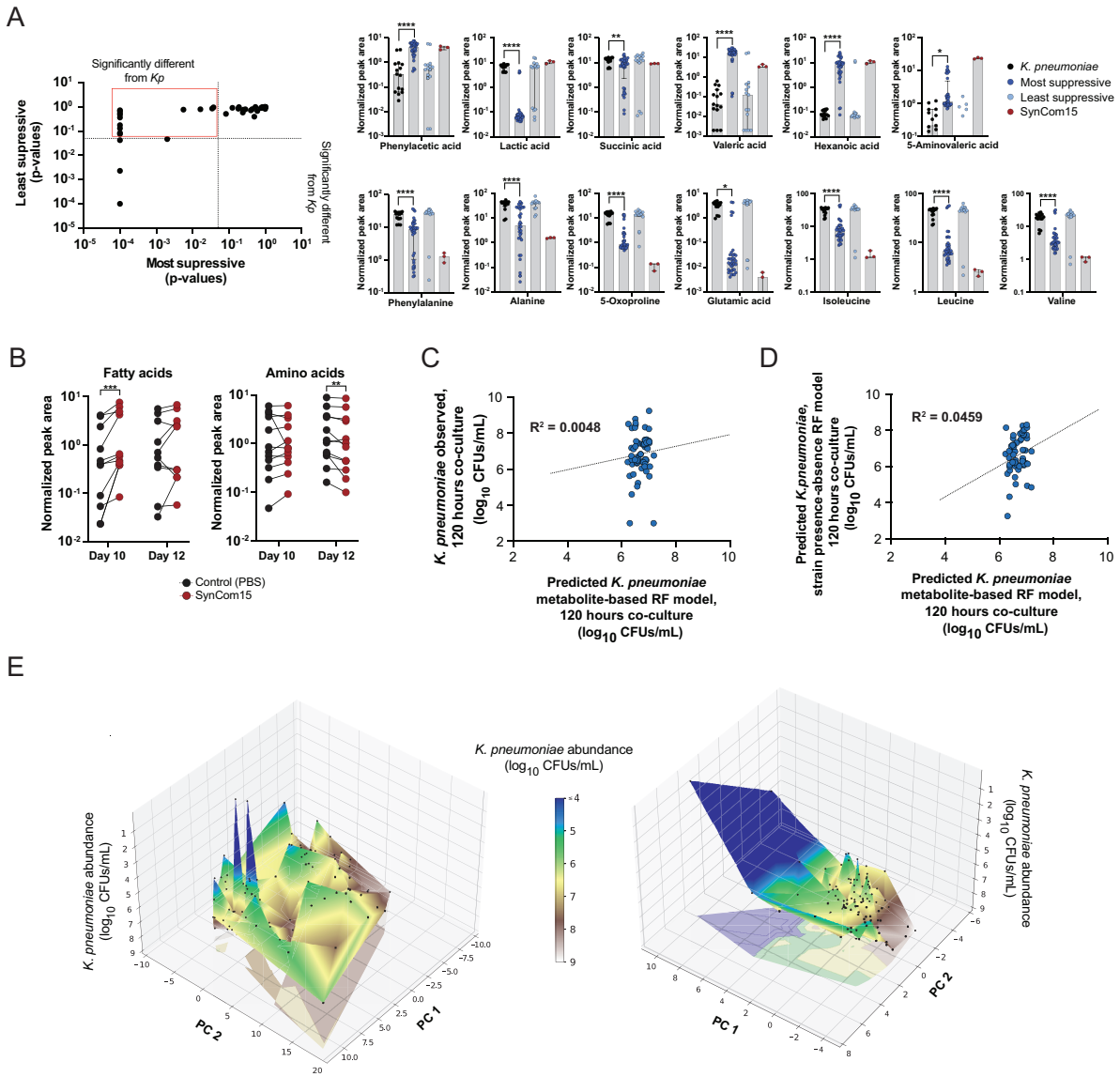
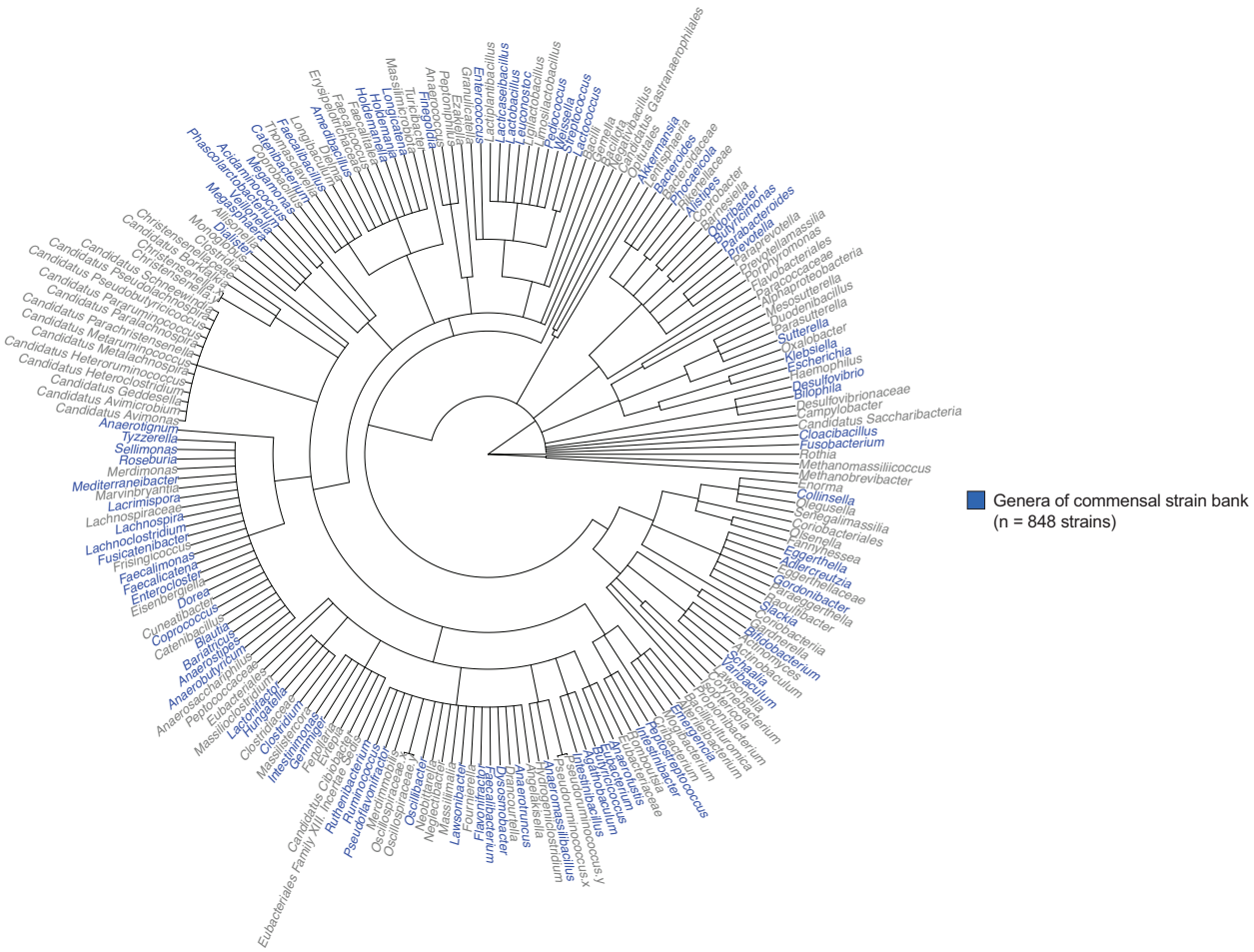


Fig. 4. Comparison of SynCom15 composition with composition of healthy human microbiomes. **(A)** Phylogenetic tree of genera present across fecal microbiomes of human donors. Brown genera are those found across SynCom15 strains (genera names are according to annotation by Metaphlan). **(B)** Prevalence pattern for species of SynCom15 (rows) across donor fecal microbiomes (DFI is Duchossois Family Institute; columns). **(C)** Histogram of relative abundance for SynCom15 species (x-axis) across all fecal samples from population of healthy human donors.

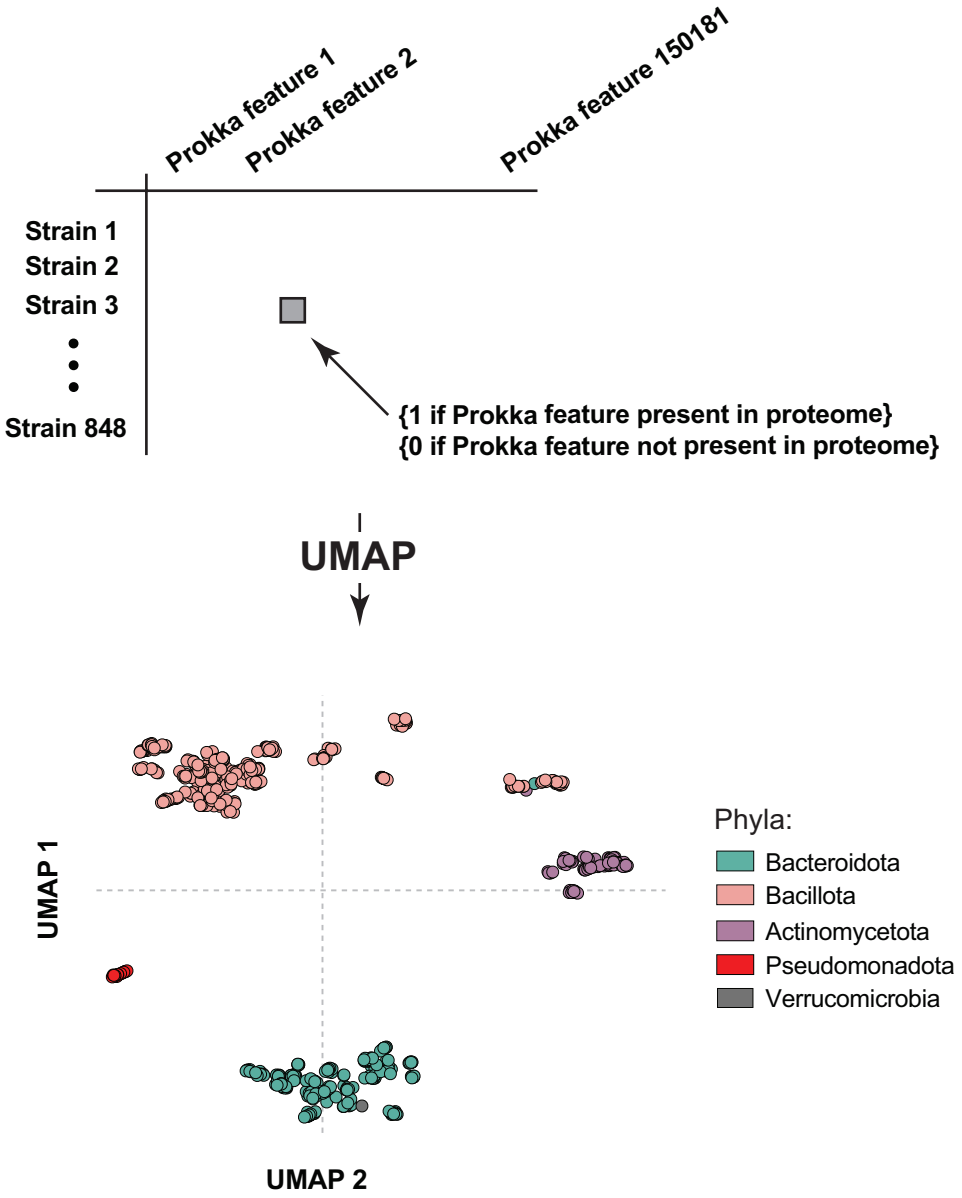
Oliveira et al., Figure 5



673 **Fig. 5. Comparing metabolite-based and strain-based models of community design. (A)** p-
674 values for differential enrichment of metabolites between the 5 most suppressive DMCs and *K.*
675 *pneumoniae* alone (x-axis); p-values for differential enrichment of metabolites between the 5
676 least suppressive DMCs and *K. pneumoniae* alone (y-axis) (p-values computed by two-way
677 ANOVA). Red box indicates features that are significantly differentially enriched in the most
678 suppressive DMCs but not in the least suppressive DMCs. Bar plots show distribution of
679 normalized peak areas (y-axis) for each metabolite feature in the red box in the left panel (x-
680 axes) for *K. pneumoniae* alone (black), the five most suppressive DMCs (blue), the five least
681 suppressive DMCs (light blue), and SynCom15 (maroon) at the 72, 96, and 120 hour culture
682 timepoint. *$p < 0.05$; **$p < 0.01$; ****$p < 0.0001$. **(B)** Distributions of normalized peak areas (y-
683 axes) of fatty acids and amino acids from fecal samples collected on Day 10 and Day 12 of
684 mouse experiment shown in **Fig. 3A** for mice gavaged with saline (PBS) or SynCom15
685 (maroon). ***$p<0.001$. **(C,D)** Correlation between predicted *K. pneumoniae* from RF model
686 trained on metabolite profile of DMCs (x-axis) and observed *K. pneumoniae* abundance after
687 120 hours of co-culture with DMCs (y-axis) (panel C). Correlation between predicted *K.*
688 *pneumoniae* abundance from RF model trained on metabolite profile of DMCs (x-axis) and RF
689 model trained on pattern of strain presence-absence in DMCs (y-axis) (panel D). Dots shown
690 are 60 'out of-sample' DMCs. **(E)** Structure of metabolite profiles for DMCs (PC1 vs. PC2)
691 versus *K. pneumoniae* abundance after 120 hours of co-culture (z-axis) (left panel). Structure of
692 strain presence-absence for DMCs (PC1 vs PC2) versus *K. pneumoniae* abundance after 120
693 hours of co-culture (z-axis) (right panel). Each dot on the surfaces is a DMC; surfaces are
694 interpolated.
695

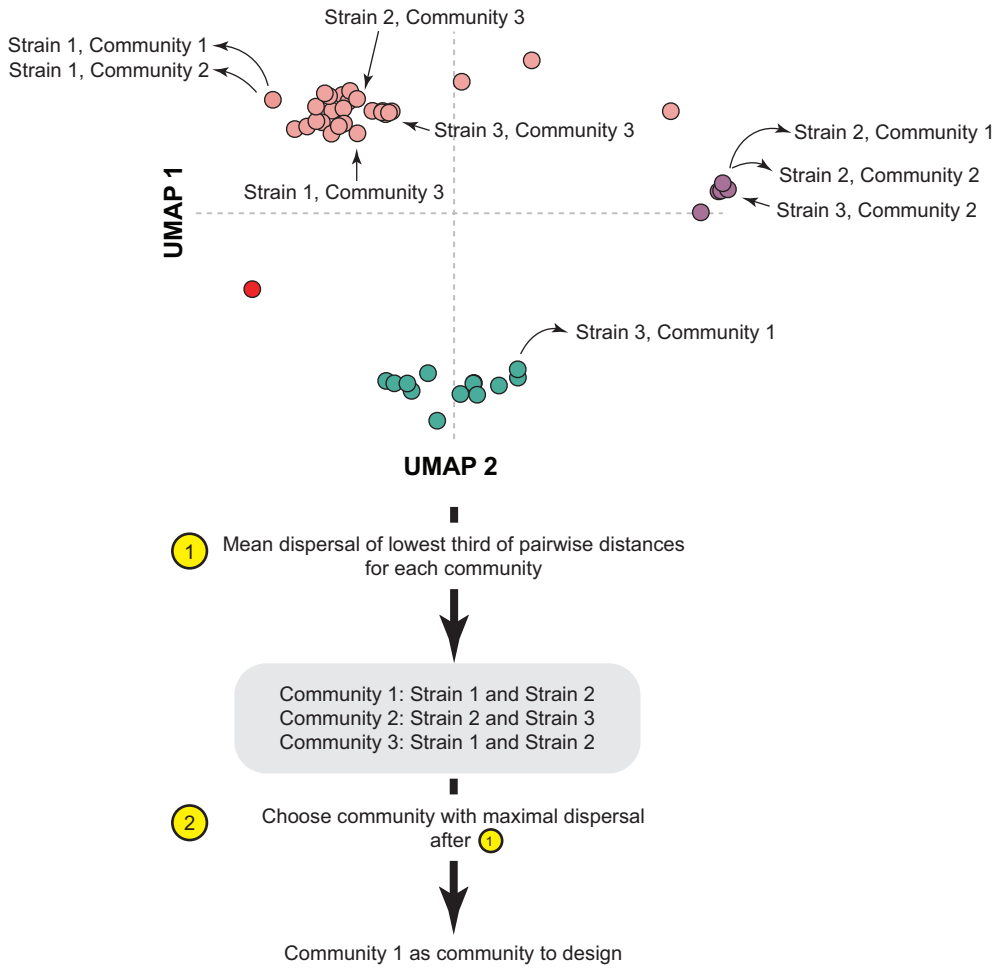


Extended Data Fig. 1. Fecal samples collected from 22 healthy human donors were subject to shotgun metagenomic sequencing (Methods). Tree of the genera comprising all fecal microbiomes (annotations per Metaphlan) is shown here (Methods). Colored in blue are the distribution of genera observed in our bank of 848 commensal strains.

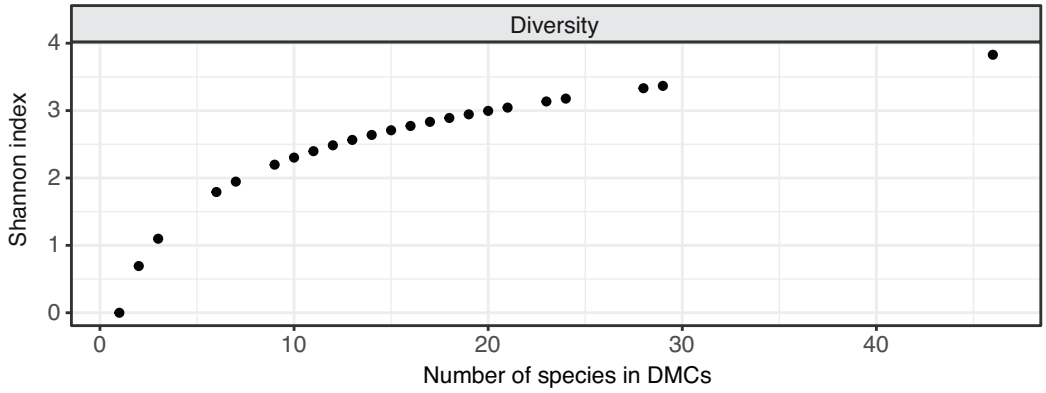


Extended Data Fig. 2. Matrix of strain by Prokka feature for all 848 gut commensals was created where entries are a '1' if the Prokka feature is present in the strain proteome, and '0' if Prokka feature is absent in the strain proteome. Matrix was subject to UMAP visualization; UMAP plot is shown in **Fig. 1B**, right panel.

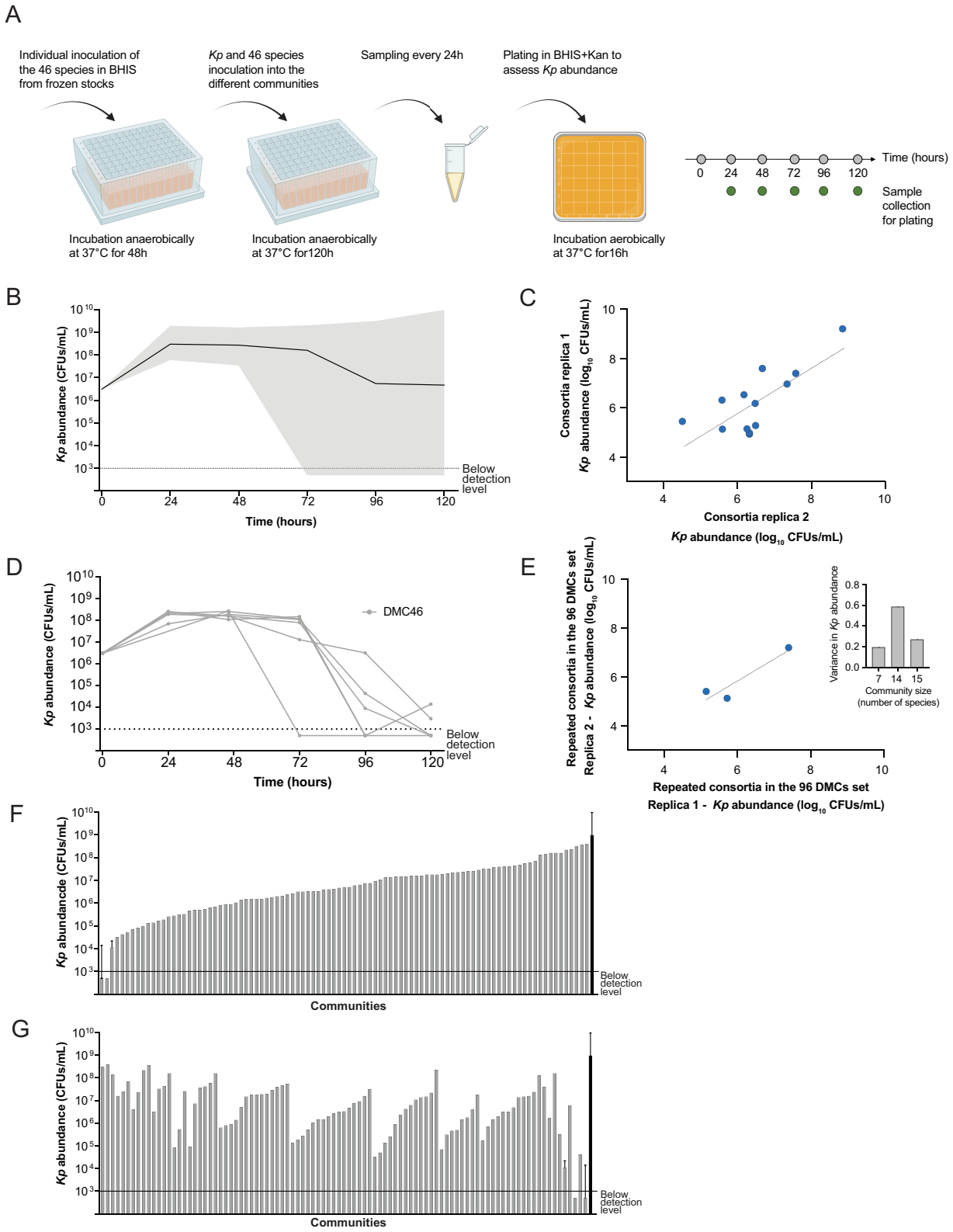
A


Strain 2, Community 3
Strain 1, Community 1
Strain 1, Community 2
Strain 3, Community 3
Strain 2, Community 1
Strain 2, Community 2
Strain 3, Community 2
Strain 1, Community 3
UMAP 1
Strain 3, Community 1
UMAP 2
1
Mean dispersal of lowest third of pairwise distances for each community
Community 1: Strain 1 and Strain 2
Community 2: Strain 2 and Strain 3
Community 3: Strain 1 and Strain 2
2
Choose community with maximal dispersal after 1
Community 1 as community to design

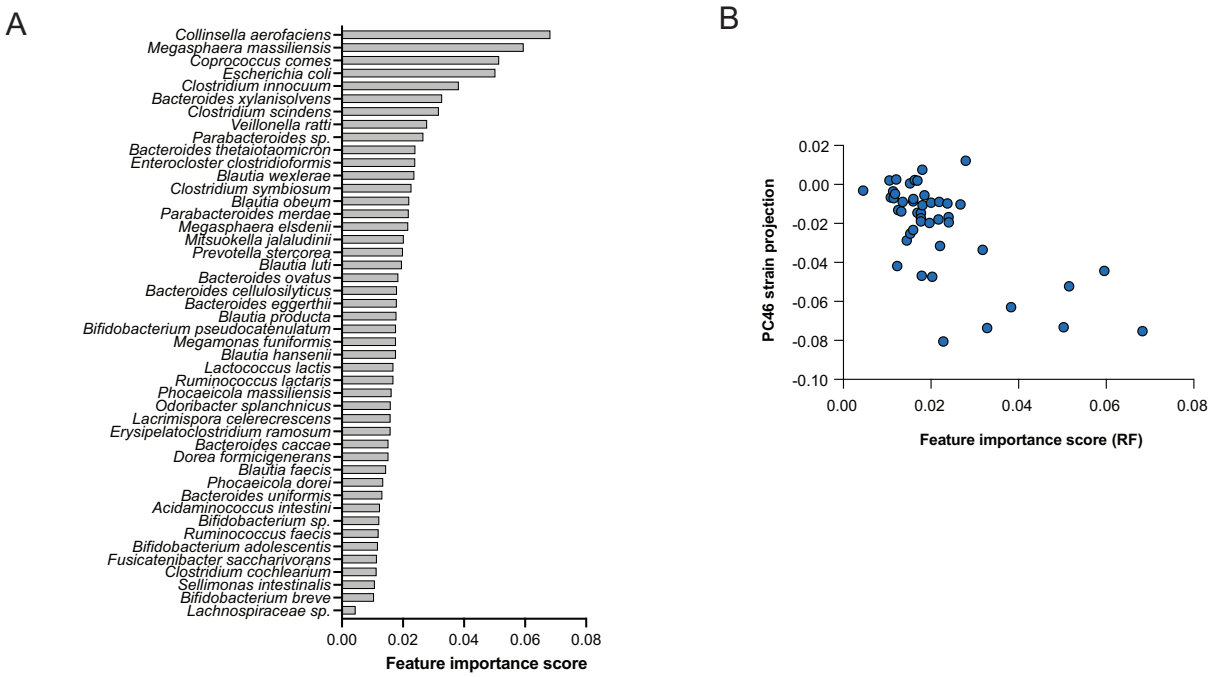
B


Diversity
Shannon index
0
1
2
3
4
0
10
20
30
40
Number of species in DMCs

Extended Data Fig. 3. (A) Workflow for algorithm used to design DMCs. Communities with three bacterial strains are shown as an example. Given three possible communities that could be created, the first step is to choose the mean dispersal of the lowest third of pairwise distances between strains for each community. In the example shown here, the lowest third is equivalent to the minimum pairwise distance for each community due to the communities being comprised of only three strains (gray box). The second step is to the choose the community with the maximal dispersion per Step 1. In the case shown here, 'Community 1' would be chosen as a DMC for incorporation into our DBTL framework. **(B)** Average Shannon diversity (y-axis) versus number of species in DMCs (x-axis). The maximum possible Shannon diversity is set by the DMC containing all 46 strains used to engineer DMCs.

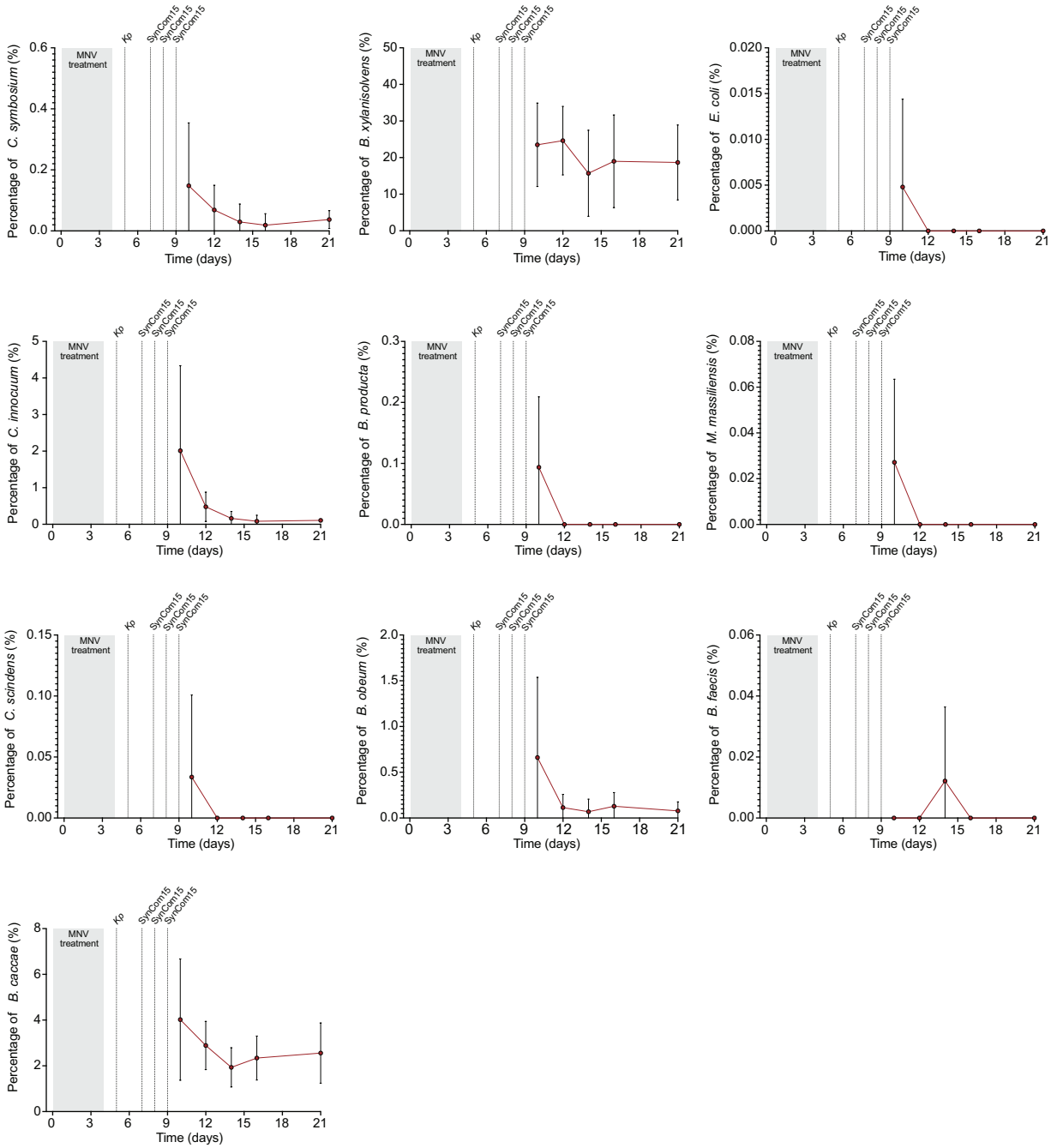

A
Individual inoculation of the 46 species in BHIS from frozen stocks
Kp and 46 species inoculation into the different communities
Sampling every 24h
Plating in BHIS+Kan to assess Kp abundance
Incubation anaerobically at 37°C for 48h
Incubation anaerobically at 37°C for120h
Incubation aerobically at 37°C for16h
Time (hours)
0 24 48 72 96 120
Sample collection for plating
B
Kp abundance (CFUs/mL)
Below detection level
Time (hours)
C
Consortia replica 1
Kp abundance (log10 CFUs/mL)
Consortia replica 2
Kp abundance (log10 CFUs/mL)
D
DMC46
Kp abundance (CFUs/mL)
Below detection level
Time (hours)
E
Repeated consortia in the 96 DMCs set
Replica 2 - Kp abundance (log10 CFUs/mL)
Repeated consortia in the 96 DMCs set
Replica 1 - Kp abundance (log10 CFUs/mL)
Variance in Kp abundance
Community size (number of species)
F
Kp abundancde (CFUs/mL)
Below detection level
Communities
G
Kp abundance (CFUs/mL)
Below detection level
Communities

741 **Extended Data Fig. 4. (A)** Workflow for evaluating clearance capacity of DMCs for *K.*
742 *pneumoniae* ('*Kp*') *in vitro* ('BHIS' is Brain Heart Infused media supplemented with cysteine,
743 'kan' is kanamycin). **(B)** Timecourse of *K. pneumoniae* abundance (y-axis) for all 96
744 communities shown in **Fig. 1E**. Solid line represents median, shade represents range. **(C-E)**
745 Reproducibility of assay. Panel C; correlation between the suppressive capacity of several
746 different DMCs across two experimental replicates. Panel D; Timecourse of DMC containing all
747 46 bacterial strains (DMC46) across five experimental replicates. Panel E; variation in three
748 DMCs that were replicated within the 96 DMCs shown in **Fig. 1E**; inset shows variance in *K.*
749 *pneumoniae* abundance for all three DMCs as a function of their respective community size.
750 **(F,G)** Panel F; *K. pneumoniae* abundance (y-axis) after co-culture with each DMC (x-axis) for
751 120 hours where the x-axis is ordered by most suppressive (left) to least suppressive (right)
752 DMCs. Panel G; *K. pneumoniae* abundance (y-axis) after co-culture with each DMC (x-axis) for
753 120 hours where the x-axis is ordered by least complex (left) to most complex (right) DMCs. For
754 both plots, the black bar is *K. pneumoniae* grown in monoculture in BHIS.
755

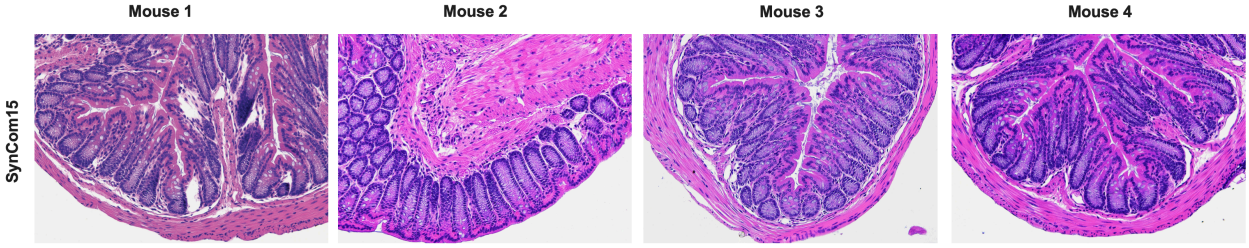


Extended Data Fig. 5. (A) Feature importance score for each strain of SynCom15 resulting from RF model built on strain presence-absence. **(B)** Feature importance scores (x-axis) for each strain (dots) versus the projection of each strain onto PC46 (y-axis) of matrix defined in **Fig. 2B**. Projection of strains onto PC46 are also shown in **Fig. 2D**, left panel.

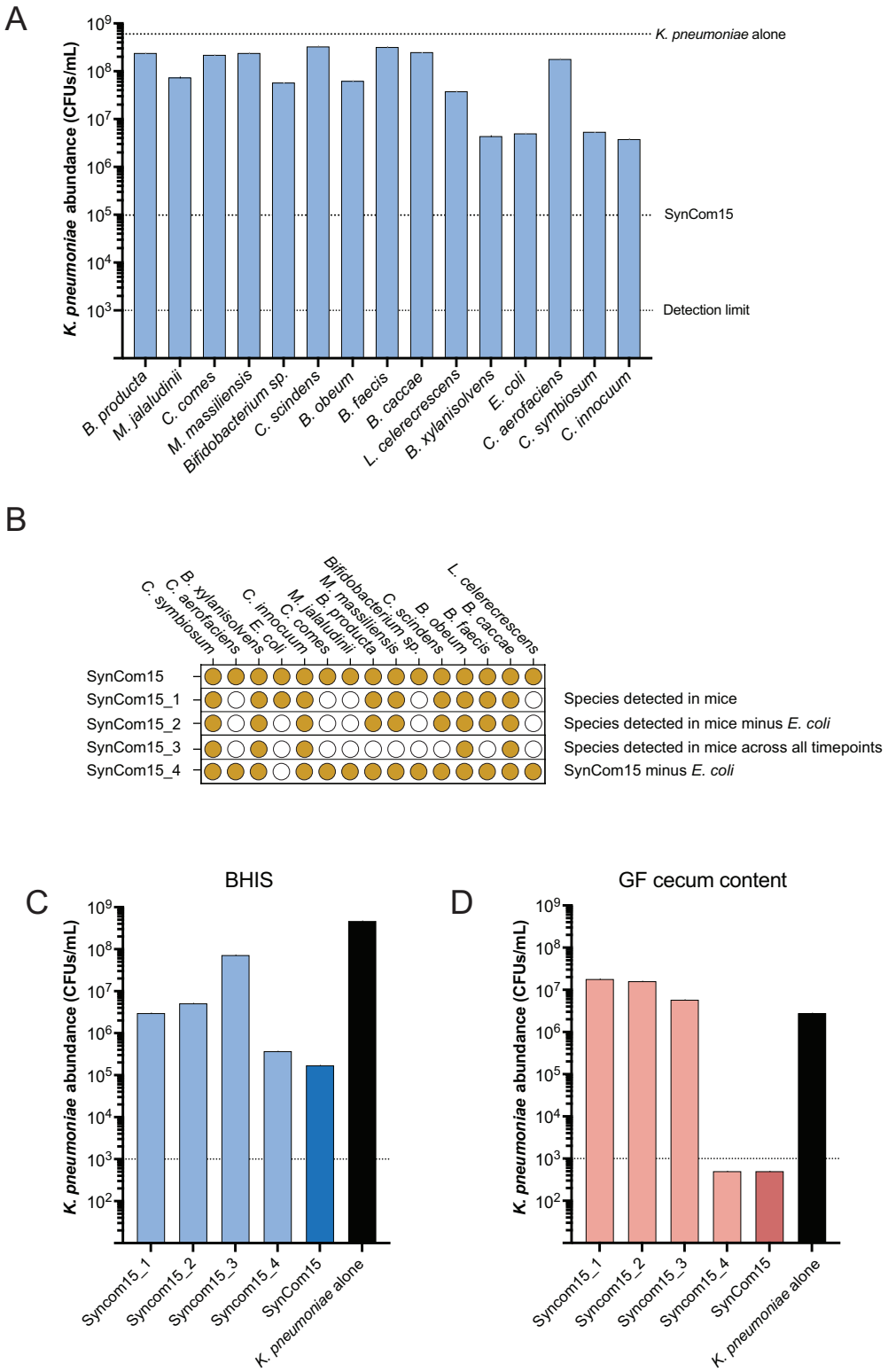
Oliveira et al., Extended Data Figure 6



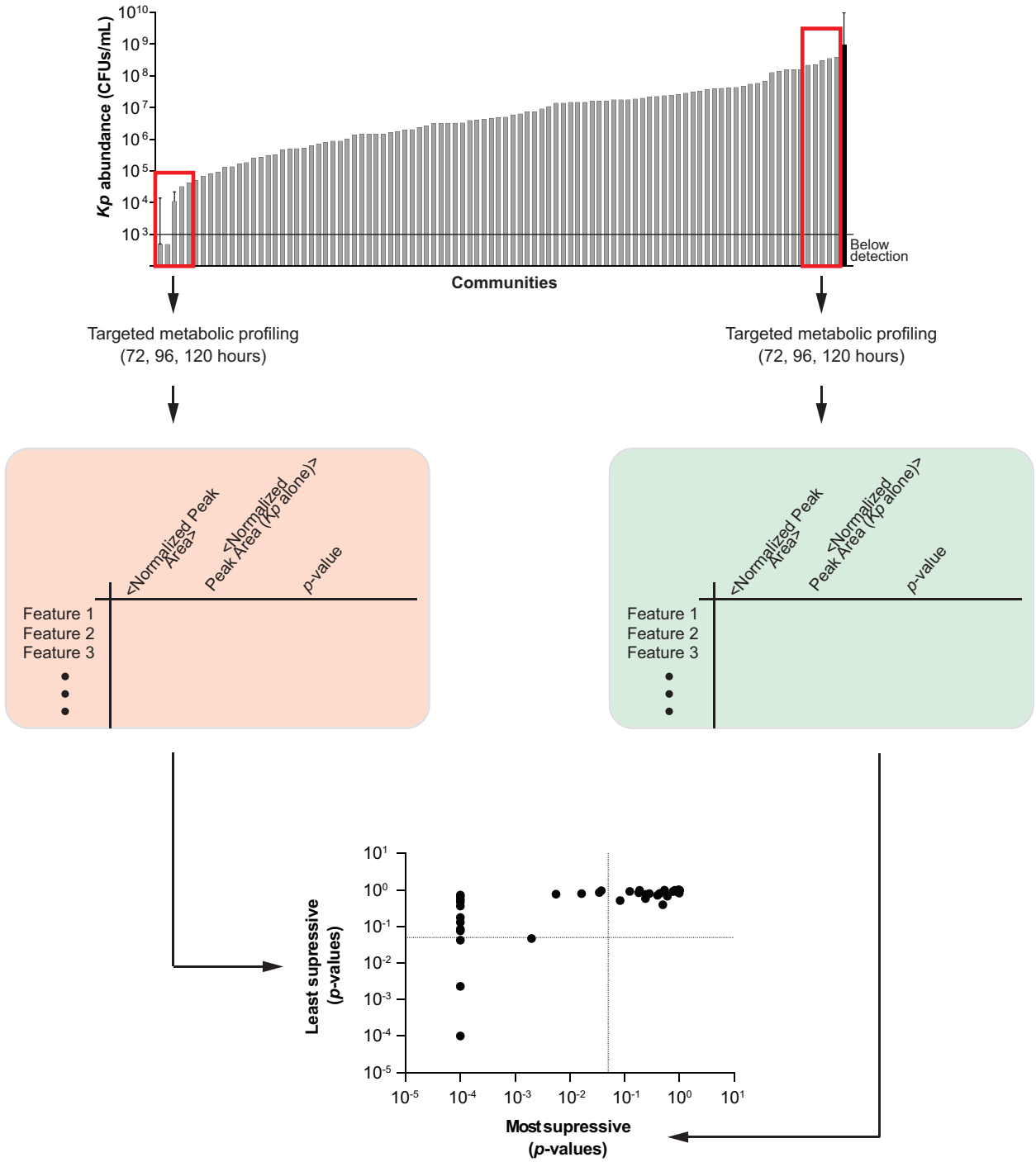
Extended Data Fig. 6. Dynamics of SynCom15 strains (y-axes) through time (x-axes) after serial triple gavage of SynCom15 in SPF mice pre-treated with broad spectrum antibiotics ('MNV treatment') and then infected with *K. pneumoniae* MH258 ('*Kp*'). Error bars represent +/- 1 standard deviation across cohort of mice.



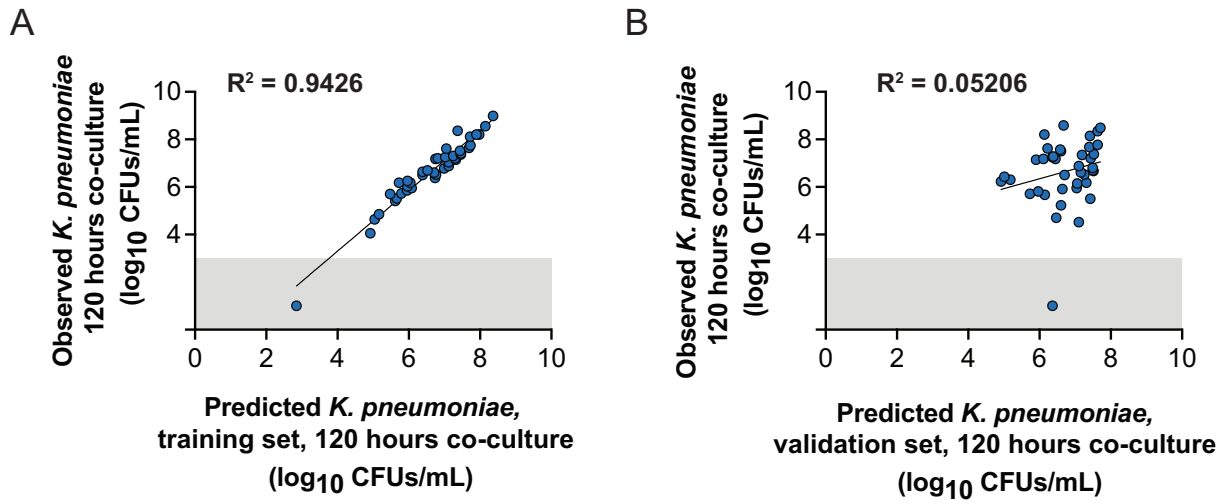
Extended Data Fig. 7. Hematoxylin and eosin stain of colon for infected mice given SynCom15.


A
K. pneumoniae abundance (CFUs/mL)
K. pneumoniae alone
SynCom15
Detection limit
B. producta
M. jalaludinii
C. comes
M. massiliensis
Bifidobacterium sp.
C. scindens
B. obeum
B. faecis
B. caccae
L. celerecrescens
B. xylanisolvens
E. coli
C. aerofaciens
C. symbiosum
C. innocuum
B
SynCom15
SynCom15_1
SynCom15_2
SynCom15_3
SynCom15_4
Species detected in mice
Species detected in mice minus E. coli
Species detected in mice across all timepoints
SynCom15 minus E. coli
C
BHIS
D
GF cecum content
Syncom15_1
Syncom15_2
Syncom15_3
Syncom15_4
SynCom15
K. pneumoniae alone

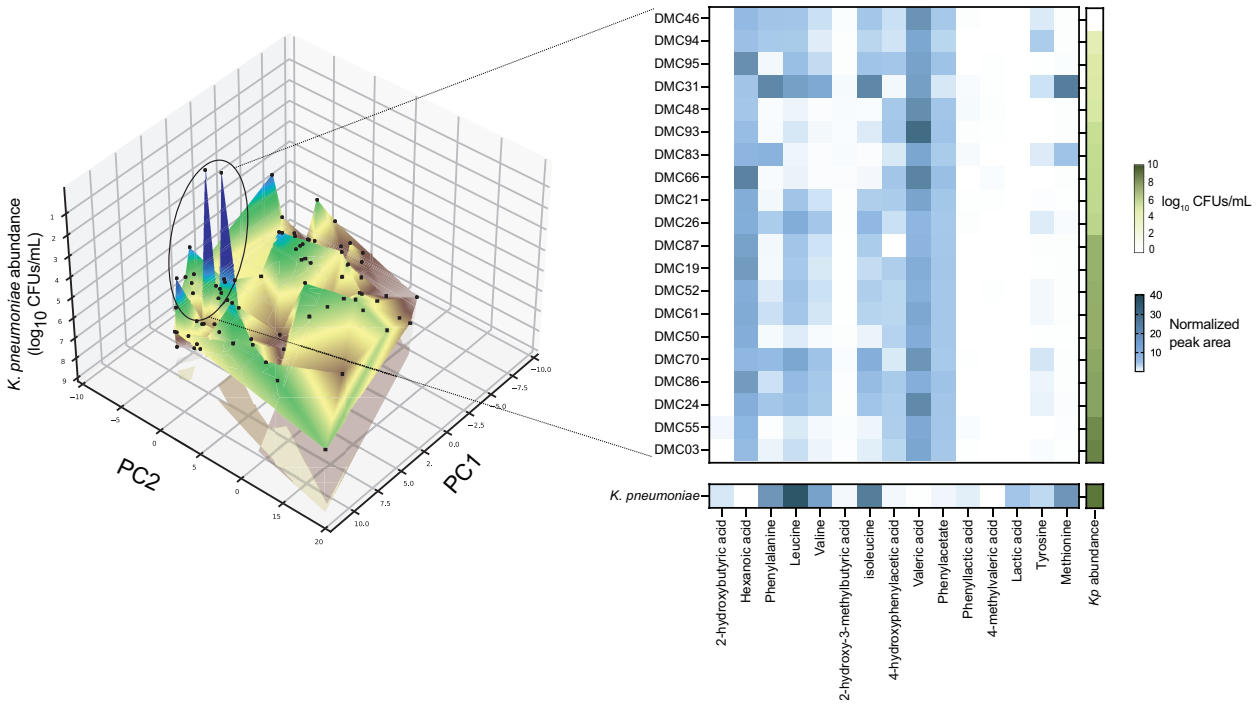
782 **Extended Data Fig. 8. (A)** *K. pneumoniae* abundance (y-axis) when co-cultured with each of
783 the SynCom15 strains (x-axis) for 120 hours in BHIS media. '*K. pneumoniae* alone' labels the *K.*
784 *pneumoniae* abundance in monoculture for 120 hours in BHIS shown in **Fig. 2E**. 'SynCom15'
785 labels the *K. pneumoniae* abundance of *K. pneumoniae* in co-culture with SynCom15 for 120
786 hours in BHIS shown in **Fig. 2E**. 'Detection limit' labels the lower limit of *K. pneumoniae*
787 abundance for our assay. **(B)** Composition of (i) SynCom15, (ii) strains of SynCom15 that
788 engrafted mice ('SynCom15_1'), (iii) strains of SynCom15 that engrafted mice without *E. coli*
789 ('SynCom15_2'), (iv) strains of SynCom15 found in mice across all timepoints of the experiment
790 shown in **Fig. 3** ('SynCom15_3'), (v) SynCom15 without *E. coli* ('SynCom15_4'). **(C,D)** All
791 communities defined in panel B were assayed for *K. pneumoniae* clearance in BHIS (panel C)
792 and GF cecum content (panel D) media. Black bars indicate *K. pneumoniae* abundance in BHIS
793 and GF cecum content when cultured alone.
794



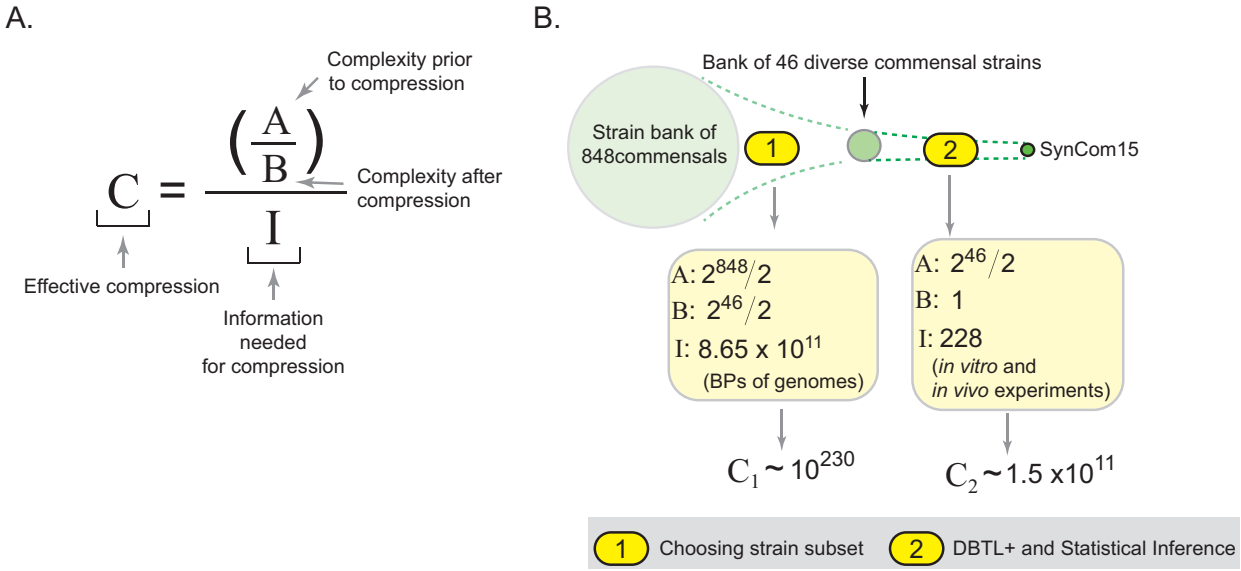
Extended Data Fig. 9. Workflow for detecting metabolite features distinguishing most and least suppressive DMCs when co-cultured with *K. pneumoniae* compared to *K. pneumoniae* in monoculture. We used the distribution of *K. pneumoniae* abundances after 120 hours of co-culture with DMCs to select the five 'most suppressive' and 'least suppressive' DMCs (red boxes in barplot) for further metabolic analysis.



Extended Data Fig. 10. Predictive capacity of RF model trained on metabolite profiles of DMCs for the training set of data (panel A) and the validation set of data (panel B).



Extended Data Fig. 11. Metabolite profiles after 120 hours of co-culture with *K. pneumoniae* and *K.* pneumoniae ('*Kp*') abundance for communities delineated in inset. Metabolite features are chosen as those that most contribute to variance along the main principal component (PC1) of metabolite variation across all 81 DMCs used to train the RF model. Surface shown here is the same surface as shown in **Fig. 5E**.



Extended Data Fig. 12. (A) Computing 'Effective compression'. **(B)** Compression from starting point—bank of 848 gut commensal strains—to SynCom15 across two steps performed serially. Step 1 was using the genomes of the strain bank to reduce complexity from 848 to 46 strains. Step 2 was performing DBTL+ and statistical inference to engineer SynCom15. For the 'I' value in Step 1, the total information used were the full genomes of all 848 strains. As a conservative measure of compressive power, we considered each base pair for all genomes as a unique piece of information. The total number of basepairs for our commensal strain bank was 8.65 x 10^{11}. For the 'I' value in Step 2, 96 DMCs were tested for their capacity to suppress *K. pneumoniae*; 12 DMCs were tested to validate the reproducibility of our assay for evaluating the suppressive capacity of a DMC; 60 DMCs were tested as 'out-of-sample' communities to test the predictive capacity of our RF model built on DMC presence-absence; 21 experiments were performed to evaluate the capacity of all Blocks, DMC46, and SynCom15 to suppress *K. pneumoniae* in BHIS, germ-free cecal extract, and antibiotic-treated specific-pathogen free (Ab-SPF) cecal extract media; 39 experiments were performed to evaluate the capacity of saline, an FMT, Block 1, Block 2, and SynCom15. In total, this yields 228 experiments performed to compress the space of $2^{46}/2$ to SynCom15.

835 **Methods**

836 *Creation and whole genome sequencing of strain bank*

837 Fecal samples were obtained from 28 human donors that fell within the age range of 18
838 to 63 with a median age of 35. Donors were selected as those with no antibiotic use in the past
839 year, no known history of diabetes, colitis, autoimmune disease, cancer, pneumonia, dysentery,
840 or cellulitis at time of consent. Institutions that approved protocols of fecal sample collection were
841 Memorial Sloan Kettering (MSK) and the University of Chicago. Fresh fecal samples were
842 immediately reduced in an anaerobic chamber upon collection and diluted and cultured on various
843 growth media. Agar media types vary, but include any of the following: Columbia Blood Agar,
844 Brain Heart Infusion + Yeast, Brain Heart Infusion + Mucin, Brain Heart Infusion + Yeast + Acetate
845 or N-acetylglucosamine, reinforced Clostridial Agar, Peptone Yeast Glucose, Yeast Casitone
846 Fatty Acids, Defined media M5. Colonies were selected and grown to be sufficiently turbid, 20%
847 glycerol/PBS stocks were created and stored in a -80°C freezer.

848 Colonies were selected for whole-genome based on pyro-sequencing of the 16S region
849 which provides a rough estimate of genus level designation. For each donor, only colonies that
850 had a sequence identity threshold of less than 99% from CD-Hit (v. 4.8.1) were selected for whole-
851 genome sequencing[55,56]. Bacterial genomic DNA was extracted using QIAamp DNA Mini Kit
852 (QIAGEN) according to manufacturer's manual. The purified DNA was quantified using a Qubit
853 2.0 fluorometer. 1000ng of each sample was prepared for sequencing using the QIAseq FX DNA
854 Library Kit (QIAGEN). The protocol was carried out for a targeted fragment size of 550bp.
855 Sequencing was performed on the MiSeq or NextSeq platform (Illumina) with a paired-end (PE)
856 kit in pools designed to provide 1-3 million PE reads per sample with read length of 250 or 150
857 bp.

858 Adapters were trimmed off with Trimmomatic (v0.39) with following parameters: the
859 leading and trailing 3 bp of the sequences were trimmed off, quality was controlled by a sliding
860 window of 4, with an average quality score of 15 (default parameters of Trimmomatic)[57]. Moreover,

861 any read that was less than 50 bp long after trimming and quality control were discarded. The
862 remaining high-quality reads were assembled into contigs using SPAdes (v3.15.4)[58].
863 Taxonomic classification of the assembled contigs was performed with the following
864 methods: (a) Kraken2 (v2.1.2; (b) full/partial length 16S rRNA gene from each isolated colony's
865 assembled contigs is extracted and input into BLASTn (v2.10.1+) to query against NCBI's RNA
866 RefSeq database[59–61]. Top five hits for each query are manually curated to determine an isolate's
867 identity, with identity and coverage cutoff both at 95%; (c) GTDB-Tk (v1.5.1)[62]. The final taxonomy
868 is determined by the consensus of the three methods. Any colony that did not match initial pyro-
869 sequencing taxonomy or lacked consensus was excluded from the commensal strain bank.
870
871 *Construction of tree of bacterial genera across fecal microbiomes of healthy donors*
872 From the metagenomic sequencing data of the fecal samples collected across healthy
873 donors, bacterial genera present were identified by Metaphlan4[63]. Names were then extracted
874 and cross-referenced with NCBI taxonomy using the taxize application in R[64]. The resulting tree
875 was constructed based on NCBI taxonomic classification.
876
877 *Construction of UMAP plot shown in **Fig. 1C**.*
878 All gut commensal strains were annotated by their Prokka annotations and an alignment
879 was created (848 rows comprising commensal strains, 150181 columns comprising Prokka
880 annotated features). Each entry in the alignment is a '1' or a '0' indicating the presence or absence
881 of a specific feature in a particular bacterial proteome.
882
883 *Shotgun metagenomics of fecal samples from healthy human donors*
884 Procedure for acquiring metagenomic data from fecal samples of healthy donors followed
885 the same protocol as that described by Odenwald *et al*[65].
886

887 *Design strategy for bacterial communities*

888 To design a bacterial comprised of *N* strains, we perform the following steps using the
889 UMAP plot based on bacterial genomes of 46 strains shown in **Fig. 1C** as the basis of our
890 approach.

891

892 *Step 1*: Create 10,000 communities randomly of size *N*. The ensemble of all 10,000
893 communities of size *N* is represented as

894 $$C_{size\ N} = \{c_1, \dots, c_{10,000}\} \qquad (1)$$

895 *Step 2*: Each community, c_i, is defined by a set of *N* bacterial strains:

896 $$c_i = \{s_1, \dots, s_N\} \qquad (2)$$

897 where s_j is strain *j* in c_i. Compute all pairwise distances in the UMAP space for all strains
898 in C_i. For instance, the pairwise distance between strain 1 and 2 is:

899 $$pd_{1,2} = dist(s_1, s_2) \qquad (3)$$

900 where 'dist' is the function that computes the distance between s_1 and s_2 in the UMAP
901 space. We define the distribution of all pairwise distances for c_i as

902 $$PD_i = \{pd_{1,2}, pd_{1,3}, \dots, pd_{N-1,N}\} \qquad (4)$$

903 *Step 3*: Order PD_i for a given c_i from largest to smallest values, then compute the mean
904 pairwise distance across the lower 30% of values comprising PD_i. We term this value the
905 'mean adjusted dispersal'.

906

907 *Step 4*: Compute the mean adjusted dispersal for all communities in $C_{size\ N}$.

908

909 *Step 5*: Identify the community within the 10,000 communities comprising $C_{size\ N}$ with the
910 maximum mean adjusted dispersal. This community is the designed community
911 comprising *N* strains.

912

913 This process is outlined for a community comprised of three strains in **Extended Data**

914 **Fig. 3A**.

915

916 *Creating the Klebsiella pneumoniae MH258 strain used in experiments*

917 The *K. pneumoniae*-MH258 isolate was previously described elsewhere[31]. For better in

918 vitro and in vivo selection of this strain, *K. pneumoniae* -MH258 was transformed by

919 electroporation with pmCherry-sfGFP (86441; addgene).

920

921 *Experimental workflow for Kp clearance assay*

922 The 46 bacterial strains described in **Supplementary Table 1B** were individually

923 inoculated from a frozen stock into 900μL of BHI supplemented with cysteine 0.1% (BHIS)

924 previously reduced. Strains were incubated at 37°C in static conditions for 48h in anaerobiosis to

925 ensure that the most fastidious strains reach stationary phase. *K. pneumoniae*-MH258 sGFP was

926 also inoculated in the same conditions, but only 24h after commensal isolates inoculation due to

927 the fast growth capacity of this species and was incubated for 24h. All strain densities were

928 assessed by taking 100 μL of each culture and measuring OD_{600} in a Biotek Cytation 5. To build

929 all DMCs, isolates were inoculated in 900 μL of BHIS previously reduced in different combinations

930 with an initial OD_{600} of 0.001, so that the densest community reaches a maximum total initial OD_{600}

931 of approximately 0.05. *K. pneumoniae* was added at the same initial OD_{600} of 0.001 to all DMCs.

932 Cultures were incubated at 37°C in static conditions and anaerobiosis for 5 days. To *assess K.*

933 pneumoniae abundance, 10 μL of each culture were collected daily and homogenized in 90 μL of

934 PBS and serially diluted. Diluted samples were plated in BHIS with kanamycin (50μg/mL). Plates

935 were incubated at 37°C overnight in aerobiosis. GFP expressing *K. pneumoniae-MH258* colony

936 forming units (CFUs) were enumerated. In parallel, 100uL of each culture was also collected to

937 recover the cell phase and the supernatants at 72h, 96h, and 120h. These samples were stored
938 at -80°C to be later processed for shotgun metagenomics and metabolomics.
939
940 *Training and validation of Random Forest (RF) model*
941 We used a RandomForestRegressor, available with scikit-learn python package[33]. Tree
942 Depth was set to 12 levels per tree, the number of trees was set to 100, and the maximum number
943 of features was set to "sqrt" (square-root of the number of strains total). Out-of-bag error was
944 measured by a combination of R^2 (where numbers less than 1 indicate more error) and Mean
945 Squared Error (where larger numbers indicate more error). To train and validate our model, we
946 randomly split our dataset into 90% training and 10% true-out-of-sample 100 times. The input
947 data was a vector of 46 1's and 0's as shown in the matrix displayed in **Fig. 1E** corresponding to
948 the pattern of presence-absence for each DMC. In each iteration, the RandomForestRegressor
949 was fit to the training set via 6-fold cross validation. Cross-validation accuracy was measured
950 through Pearson Correlation. The true out of sample set was then predicted, and prediction
951 accuracy was measured by computing Mean Squared Error and Pearson Correlation of the
952 predicted versus measured *K. pneumoniae* abundances after 120 hours of co-culture with the
953 DMC. Feature Importance Scores for all features were observed and stored. This process was
954 repeated 100 times, and prediction accuracies and feature significance scores were averaged.
955 An additional RandomForestRegressor model was then trained on the entirety of the dataset with
956 6-fold cross-validation. Cross-validation accuracy was measured by calculating Mean Squared
957 Error, Pearson Correlation, and R^2. Averaged prediction accuracies and feature significance
958 scores were used to estimate prediction error.
959
960 *Statistical analysis of matrix in **Fig. 2B***
961 The matrix in **Fig. 2B** was subject to PCA resulting in 46 principal components of data-
962 variance (eigenvectors). We found that the first principal component (PC1) was significantly

963 associated with community complexity (**Supplementary Table 6C**). To isolate the effect of *K.*
964 *pneumoniae* clearance from community size, we first performed a series of steps to 'regress out'
965 the effect of community size. First, let $\boldsymbol{x_i}$ be the community size of DMC i. Let $\boldsymbol{y_i}$ be the predicted
966 *K. pneumoniae* clearance from the RF model for DMC i. A linear model is then created regressing
967 community size against *K. pneumoniae* clearance taking the form:

968 $$\boldsymbol{y_i} = \boldsymbol{\beta_1 x_i} + \boldsymbol{\beta_0} + \varepsilon \quad (5)$$

969 $$\widehat{\boldsymbol{y_i}} = \boldsymbol{\beta_1 x_i} + \boldsymbol{\beta_0} \quad (6)$$

970 where $\widehat{\boldsymbol{y_i}}$ is the *K. pneumoniae* clearance of DMC i as a function of its size. The residuals of this
971 linear model are given by

972 $$\boldsymbol{\varepsilon_i} = \boldsymbol{y_i} - \widehat{\boldsymbol{y}}_i \quad (7)$$

973 where ε_i is the degree of *K. pneumoniae* clearance of DMC i after removing linearly modeled
974 information related to the size of the DMC. All principal components were regressed against $\boldsymbol{r}_i$
975 and principal component 46 (PC46) was found to be the most significantly associated with
976 predicted, residualized *K. pneumoniae* clearance (**Supplementary Table 6D**).
977
978 <u>*Defining the matrix in **Fig. 2D***</u>
979 Let $\boldsymbol{u} \in \mathbb{R}^{46}$ be the vector of column projections of each strain on PC46 of the matrix
980 defined in **Fig. 2B**.
981 Let s be the scalar value denoting the maximum value of $\boldsymbol{u}$

982 $$A^{46x46} = \left(a_{ij}\right) \quad (8)$$

983 $$a_{ij} = d_{ij}^2 = ||u_i - u_j||^{\,2} \quad (9)$$

984 Where $||\cdot||$ denotes the Euclidian norm on $\mathbb{R}^{46}$

985 $$A_{i,j} = \begin{bmatrix} 0 & \cdots & d_{46,1}^2 \\ \vdots & \ddots & \vdots \\ d_{1,46}^2 & \cdots & 0 \end{bmatrix} \quad (10)$$

986

987 $$S_{i,j} = s - A_{i,j} = \begin{bmatrix} s & \cdots & s - d_{46,1}^2 \\ \vdots & \ddots & \vdots \\ s - d_{1,46}^2 & \cdots & s \end{bmatrix} \quad (11)$$

988 The resulting symmetric similarity matrix, $S_{i,j}$, with rows and columns indicating each strain and
989 each element representing the similarity between strain i and strain j describes how strains are
990 related to one another based on their projections along PC46. Hierarchical clustering on the
991 resulting similarity matrix was then performed to identify groups of strains. Strains that are more
992 similar are often found in communities that suppress *K. pneumoniae* and those that are more
993 distant are rarely found in communities that suppress *K. pneumoniae*.
994
995 *Characterization of mice used for all experiments spanning **Fig. 2** and **Fig. 3**.*
996 All mouse experiments were performed in accordance with and approved by the
997 Institutional Animal Care and Use Committee of the University of Chicago under protocol 72599.
998 Male specific-pathogen-free C57BL/6J mice, aged 8 weeks to 10 weeks, from Jackson
999 Laboratories were used for all experiments. Mice were kept within a facility that maintained a
1000 12 hour light and 12 hour dark cycle and controlled humidity (30–70%) and temperature (68–
1001 79 °F). Mice were housed in sterile, autoclaved cages with irradiated feed (LabDiets 5K67) and
1002 acidified, autoclaved water upon arriving at the on-site mouse facility. Mouse handling and cage
1003 changes were performed by investigators wearing sterile gowns, masks and gloves in a sterile
1004 biosafety hood. Mice were cohoused with their original shipment group until starting the
1005 experiment.
1006 For germ-free (GF) studies, 8–10-week-old wild-type male C57BL/6J mice were used for
1007 all studies. Mice were initially obtained from The Jackson Laboratory and subsequently bred and
1008 raised in a GF isolator. After removal from the GF isolator, mice were handled in a sterile manner
1009 and individually housed in sealed negative pressure bio-containment unit isolators. Throughout
1010 breeding, mice were housed within the University of Chicago Gnotobiotic Research Animal

1011 Facility (GRAF) and maintained at a 12 hour light and 12 hour dark cycle and controlled humidity
1012 (30–70%) and temperature (68–79 °F). Gnotobiotic mice were fed an ad libitum diet of autoclaved
1013 Teklad Global 18% Protein Rodent Diet (Sterilizable) (2018S/2018SC).
1014
1015 *Creating GF and antibiotic (Ab)-SPF cecal extract media*
1016 To create GF cecal extract media, 8–10-week-old wild-type male C57BL/6J GF mice were
1017 euthanized and cecal contents were collected, weighted, and homogenized in 10mL of sterile
1018 distilled water on a of per gram of content. Cecal suspension was centrifuged, and supernatants
1019 were filtered through a 0.22 mm filter. GF cecal extract media was stored at -80°C.
1020 To create ab-SPF cecal extract media C57BL/6J SPF male mice at 8-10 weeks of age
1021 were singly housed and placed under an antibiotic regime (0.25g MNV – metronidazole,
1022 neomycin, vancomycin) in the drinking water (day 0). Four days later, antibiotic treatment was
1023 halted and mice were placed on normal acidified water (day 4). Cages and food were also
1024 changed. On day 7 were euthanized and cecal contents were collected, weighted, and
1025 homogenized in 10mL of sterile distilled water on a of per gram of content. Cecal suspension was
1026 centrifuged, and supernatants were filtered through a 0.22 mm filter. Ab-SPF cecal extract media
1027 was stored at -80°C.
1028
1029 *K. pneumoniae clearance in cecal extract media*
1030 DMCs capacity to inhibit *Kp* was tested by individually inoculated the 46 isolates from a
1031 frozen stock into 900μL of BHIS previously reduced. Strains were incubated at 37°C in static
1032 conditions for 48h in anaerobiosis. *Kp* was also inoculated in the same conditions, but only 24h
1033 after commensal isolates inoculation, and was incubated for 24h. All isolates density were
1034 assessed by taking 100 μL of each culture and measuring OD_{600} in Biotek Cytation 5. To build all
1035 defined bacterial consortia, isolates were inoculated in 900 μL of either GF or Ab-SPF cecal

1036 extract media previously reduced in different combinations with an initial OD_{600} of 0.001, so that
1037 the densest community reaches a maximum total initial OD_{600} of approximately 0.05. To all
1038 defined communities, *K. pneumoniae* was added at the same initial OD_{600} of 0.001. Cultures were
1039 incubated at 37°C in static conditions and anaerobiosis for 5 days. To assess for *K. pneumoniae*
1040 levels 10 μL of each culture were collected daily and homogenized in 90 μL of PBS and serially
1041 diluted. Diluted samples were plated in BHIS with kanamycin (50μg/mL). Plates were incubated
1042 at 37°C overnight in aerobiosis. GFP expressing *K. pneumoniae* CFUs were enumerated. In
1043 parallel, 100uL of each culture was also collected to recover the cell phase and the supernatants
1044 at 72h, 96h, and 120h. These samples were stored at -80°C to be later processed for shotgun
1045 metagenomics and metabolomics.
1046
1047 *Preparation of mice stool samples for fecal microbiota transplant (FMT)*
1048 Fecal samples from 15-20 mice SPF mice from different cages (to increase sample diversity)
1049 were collected to a 50 mL tube. Samples were transferred immediately to the anaerobic chamber
1050 (anaerobic exposure was kept under 30 min). Samples were dissolved in 1 mL of PBS 20%
1051 glycerol 0.1% cysteine (previously filtered and reduced) per fecal pellet (1mL per ~20 mg of fecal
1052 sample) using a mechanical pestle and vortexing. Samples were aliquoted in cryovials and stored
1053 -80°C until use.
1054
1055 *SPF mouse model of K. pneumoniae infection*
1056 C57BL/6J male at 8-10 weeks of age were singly housed and placed under an antibiotic
1057 regime (0.25g MNV – metronidazole, neomycin, vancomycin) in the drinking water (day 0). Four
1058 days later, antibiotic treatment was halted and mice were placed on normal acidified water (day
1059 4). Cages and food were also changed. On day 5 all mice were gavaged with 100μL of PBS
1060 containing 500 CFUs of *K. pneumoniae*, prepared as previously explained. On days 7, 8, and 9

1061 mice were gavaged with 100uL of either selected defined bacterial consortia, a fecal microbiota
1062 transplant from naïve healthy mice, or PBS. Fecal samples were collected on days 0, 4, 7, 10,
1063 12, 14, 16, and 21 (final day of the experiment) for 16s rRNA sequencing and on day 10 and 12
1064 for metabolomics. These were immediately place on dry ice after collection and later stored at -
1065 80°C. To assess for *K. pneumoniae* levels, fecal samples were collected on days 7, 10, 12, 14,
1066 16, and 21. Fecal samples were homogenized in 1mL of PBS and serially diluted. Undiluted and
1067 diluted samples were plated in BHIS and kanamycin (50μg/mL).
1068
1069 *Determining engraftment of SynCom15 strains in SPF mice*
1070 To determine SynCom15 strain engraftment, 16s rRNA sequences from all 15 strains were
1071 blasted against 16S rRNA sequences derived from fecal samples of antibiotic-treated SPF mice
1072 gavaged with SynCom15 consortium. Fecal-derived sequences were assigned to a SynCom15
1073 strain if their 16s rRNA percentage sequence identity was 100% with a minimum of a 95%
1074 coverage.
1075
1076 *Determining structure of microbiota in infected SPF mice given saline, FMT, or SynCom15*
1077 DNA was extracted using the QIAamp PowerFecal Pro DNA kit (Qiagen). Before
1078 extraction, samples were subjected to mechanical disruption using a bead beating method.
1079 Briefly, samples were suspended in a bead tube (Qiagen) along with lysis buffer and loaded on a
1080 bead mill homogenizer (Fisherbrand). Samples were then centrifuged, and supernatant was
1081 resuspended in a reagent that effectively removed inhibitors. DNA was then purified routinely
1082 using a spin column filter membrane and quantified using Qubit.
1083 16S sequencing was performed for murine studies, where V4–V5 region within 16S rRNA
1084 gene was amplified using universal bacterial primers—563F (5′-nnnnnnnn-NNNNNNNNNNNN-
1085 AYTGGGYDTAAA-GNG-3′) and 926R (5′-nnnnnnnn-NNNNNNNNNNNN-CCGTCAATTYHT-

1086 TTRAGT-3′), where ‘N’ represents the barcodes and ‘n’ are additional nucleotides added to offset
1087 primer sequencing. Approximately 412-bp region amplicons were then purified using a spin
1088 column-based method (Minelute, Qiagen), quantified and pooled at equimolar concentrations.
1089 Illumina sequencing-compatible Unique Dual Index adapters were ligated onto the pools using
1090 the QIAseq 1-step amplicon library kit (Qiagen). Library quality control was performed using Qubit
1091 and TapeStation and sequenced on Illumina MiSeq platform to generate 2 × 250 bp reads.
1092 Raw V4–V5 16S rRNA gene sequence data were demultiplexed and processed through
1093 the dada2 pipeline (v1.18.0) into amplicon sequence variants (ASVs) with minor modifications in
1094 R (v4.0.3)[66]. Specifically, reads were first trimmed at 190 bp for both forward and reverse reads
1095 to remove low-quality nucleotides. Chimeras were detected and removed using the default
1096 consensus method in the dada2 pipeline. Then, ASVs with length between 320 bp and 365 bp
1097 were kept and deemed as high-quality ASVs. Taxonomy of the resultant ASVs was assigned to
1098 the genus level using the RDP Classifier (v2.13) with a minimum bootstrap confidence score of
1099 80[67].
1100
1101 *Comparison of SynCom15 with microbiotas of healthy human donors*
1102 To investigate the presence of SynCom15 strains in samples from healthy human donors,
1103 SynCom15 strains taxonomic names were searched in the 22 fecal samples obtained from the
1104 DFI 22 human donors. For SynCom15 strain unclassified to species level *Bifidobacterium* sp., the
1105 most closely related species annotated by GTDB with an 98.21% ANI (*Bifidobacterium*
1106 *pseudocatenulatum*) was used[62,68].
1107
1108 *Metabolic profiling of designed communities*
1109 For metabolite extraction from liquid cultures, samples were incubated at −80 °C between
1110 1 h and 12 h. Four volumes of methanol spiked with internal standards were added to each culture
1111 supernatant. Samples were then centrifuged at −10 °C and 20,000 × g for 15 min followed by the

1112 transfer of 100 µL of supernatant to pre-labelled mass spectrometer autosampler vials (MicroLiter,
1113 09-1200).

1114 For metabolite extraction from fecal samples, extraction solvent (80% methanol spiked
1115 with internal standards and stored at -80 °C) was added at a ratio of 100 mg of material/mL of
1116 extraction solvent in beadruptor tubes (Fisherbrand; 15-340-154). Samples were homogenized at
1117 4 °C on a Bead Mill 24 Homogenizer (Fisher; 15-340-163), set at 1.6 m/s with 6 thirty-second
1118 cycles, 5 seconds off per cycle. Samples were then centrifuged at -10 °C, 20,000 x g for 15 min
1119 and the supernatant was used for subsequent metabolomic analysis.

1120 Short chain fatty acids were derivatized as described by Haak *et al*. with the following
1121 modifications[69]. The metabolite extract (100 µL) was added to 100 µL of 100 mM borate buffer
1122 (pH 10) (Thermo Fisher, 28341), 400 µL of 100 mM pentafluorobenzyl bromide (Millipore Sigma;
1123 90257) in Acetonitrile (Fisher;A955-4), and 400 µL of n-hexane (Acros Organics; 160780010) in
1124 a capped mass spec autosampler vial (Microliter; 09-1200). Samples were heated in a
1125 thermomixer C (Eppendorf) to 65 °C for 1 hour while shaking at 1300 rpm. After cooling to RT,
1126 samples were centrifuged at 4 °C, 2000 x g for 5 min, allowing phase separation. The hexanes
1127 phase (100 µL) (top layer) was transferred to an autosampler vial containing a glass insert and
1128 the vial was sealed. Another 100 µL of the hexanes phase was diluted with 900 µL of nhexane
1129 in an autosampler vial. Concentrated and dilute samples were analyzed using a GC-MS (Agilent
1130 7890A GC system, Agilent 5975C MS detector) operating in negative chemical ionization mode,
1131 using a HP-5MSUI column (30 m x 0.25 mm, 0.25 µm; Agilent Technologies 19091S-433UI),
1132 methane as the reagent gas (99.999% pure) and 1 µL split injection (1:10 split ratio). Oven ramp
1133 parameters: 1 min hold at 60 °C, 25 °C per min up to 300 °C with a 2.5 min hold at 300 °C. Inlet
1134 temperature was 280 °C and transfer line was 310 °C. A 10-point calibration curve was prepared
1135 with acetate (100 mM), propionate (25 mM), butyrate (12.5 mM), and succinate (50 mM), with 9
1136 subsequent 2x serial dilutions.

1137 Metabolites were also analyzed using GC-MS with electron impact ionization. The
1138 metabolite extract (100 µL) mass spec autosampler vials (Microliter; 09-1200) and dried down
1139 completely under nitrogen stream at 30 L/min (top) 1 L/min (bottom) at 30 °C (Biotage SPE Dry
1140 96 Dual; 3579M). To dried samples, 50 µL of freshly prepared 20 mg/mL methoxyamine (Sigma;
1141 226904) in pyridine (Sigma; 270970) was added and incubated in a thermomixer C (Eppendorf)
1142 for 90 min at 30 °C and 1400 rpm. After samples are cooled to room temperature, 80 µL of
1143 derivatizing reagent (BSTFA + 1% TMCS; Sigma; B-023) and 70 µL of ethyl acetate (Sigma;
1144 439169) were added and samples were incubated in a thermomixer at 70 °C for 1 hour and 1400
1145 rpm. Samples were cooled to RT and 400 µL of Ethyl Acetate was added to dilute samples. Turbid
1146 samples were transferred to microcentrifuge tubes and centrifuged at 4 °C, 20,000 x g for 15 min.
1147 Supernatants were then added to mass spec vials for GCMS analysis. Samples were analyzed
1148 using a GC-MS (Agilent 7890A GC system, Agilent 5975C MS detector) operating in electron
1149 impact ionization mode, using a HP-5MSUI column (30 m x 0.25 mm, 0.25 µm; Agilent
1150 Technologies 19091S- 433UI) and 1 µL injection. Oven ramp parameters: 1 min hold at 60 °C, 16
1151 °C per min up to 300 °C with a 7 min hold at 300 °C. Inlet temperature was 280 °C and transfer
1152 line was 300 °C.
1153 Data analysis was performed using MassHunter Quantitative Analysis software (version
1154 B.10, Agilent Technologies) and confirmed by comparison to authentic standards. Normalized
1155 peak areas were calculated by dividing raw peak areas of targeted analytes by averaged raw
1156 peak areas of internal standards.
1157
1158 *Training an RF model on metabolic content*
1159 First, Z-scores of all metabolites were centered and normalized. This was done by
1160 subtracting the mean Z score from the observed Z score and dividing it by the standard deviation
1161 of Z scores. This normalization ensured that for each metabolite, the distribution across all

1162 communities was zero and its standard deviation was one. With respect to output, a pseudocount
1163 of 10 was added to all *K. pneumoniae* values to enable prediction of the decadic logarithm (log_{10})
1164 of *K. pneumoniae* abundance.

1165 After standardization, 50% of the data was used for training and the remaining 50% for
1166 validation. A RF model was built with 10,000 trees with mean squared error minimization as the
1167 strategy for training. The number of features chosen by each tree was set to 10, based on the
1168 square root of the total number of metabolites available to profile. This feature selection was
1169 optimized by testing model performance with a feature range between 2 and 50. The model
1170 displayed stable performance when the number of features per tree was between 7 and 20. Below
1171 7, the model performance degraded due to insufficient information on relationships between
1172 metabolite features; above 20, the RF trees became too similar thereby impacting overall model
1173 effectiveness by skewing the final decision output by the model. Once trained, the RF model was
1174 tested on the training, test, and out-of-sample tests.
1175

Supplementary Information

Supplementary Data

The alignment of 848 gut commensal strains annotated by Prokka annotations can be found in dryad (link to repository to be determined pending review).

Supplementary Discussion

Assessing the compressive power of our approach

The process by which we converged on SynCom15 as a community that clears *K. pneumoniae* involved (i) reducing the complexity of the strain bank from 848 to 46 diverse strains and (ii) performing DBTL+ in BHIS and statistical inference with experimental validation *in vitro* and *in vivo*. Conceptualizing our two-step process as an algorithm, we sought to compute the equivalent of a 'compression' for converging on a single functional complex community from a bank of 848 strains. In evaluating computational algorithms, compression is a measure of data complexity prior to compression relative to after compression. As our process took into account biological information in the form of bacterial genome sequences and experiments, we normalized the compression ratio by the amount of information needed to perform the compression. We therefore defined an 'effective compression' as

$$C = \frac{\left(\frac{A}{B}\right)}{I} \qquad (1)$$

where C is the effective compression of a process, A is the complexity of data prior to compression, B is the complexity of data after compression, and I is the information needed for compression from A to B (**Extended Data Fig. 12A**).

For our first step, we reduced the strain bank from 848 strains to 46 strains representative of the full phylogenetic diversity by genome sequencing each of the 848 strains, annotating each genome by their gene content, and performing dimension-reduction via a UMAP analysis. Therefore, the total complexity prior to compression was $2^{848}/2$, the total

complexity after compression was $2^{46}/2$, and the information needed to be collected for compression were all base pairs of the 848 commensal strains (8.65 x 10^{11} basepairs). Considering these values, the effective compression of our first step was ~10^{230}—a substantial compression driven by the sizeable drop in complexity of the strain bank (**Extended Data Fig. 12B**). For our second step, we used the diversity of the 46 strains to create 96 DMCs, 60 'out-of-sample' DMCs, we learned an RF model and performed statistical inference to derive SynCom15; and we performed 72 more experiments to show that SynCom15 could generally clear *K. pneumoniae*. Therefore, the total complexity prior to compression was $2^{46}/2$, the total complexity after compression was 1 (SynCom15), and the information needed to be collected for compression was 228 total experiments. Considering these values, the effective compression for our second step was ~10^{11} (**Extended Data Fig. 12B**).

Collectively, this analysis showed that despite the apparently immense amount of data reflected in the whole genome sequences of 848 bacterial strains, this complexity is offset by many orders of magnitude through our approach of reducing combinatorial dimensionality by diversity-based design and DBTL+ with statistical inference. That is, the amount of compressive information held by the set of bacterial genomes is a markedly small fraction of the compressive information encoded by our two-step process. We comment on why our approach may be achieving a high compressive power in the Discussion.

1226 **References**

1227 1. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509
1228 (2021).
1229 2. Sunagawa, S. *et al.* Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev.*
1230 *Microbiol.* **18**, 428–445 (2020).
1231 3. Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome
1232 Project. *Nature* **569**, 641–648 (2019).
1233 4. Becks, L., Hilker, F. M., Malchow, H., Jürgens, K. & Arndt, H. Experimental demonstration of
1234 chaos in a microbial food web. *Nature* **435**, 1226–1229 (2005).
1235 5. Zomorrodi, A. R. & Segrè, D. Synthetic Ecology of Microbes: Mathematical Models and
1236 Applications. *J. Mol. Biol.* **428**, 837–861 (2016).
1237 6. Lawson, C. E. *et al.* Common principles and best practices for engineering microbiomes.
1238 *Nat. Rev. Microbiol.* **17**, 725–741 (2019).
1239 7. Vrancken, G., Gregory, A. C., Huys, G. R. B., Faust, K. & Raes, J. Synthetic ecology of the
1240 human gut microbiota. *Nat. Rev. Microbiol.* **17**, 754–763 (2019).
1241 8. Alivisatos, A. P. *et al.* MICROBIOME. A unified initiative to harness Earth's microbiomes.
1242 *Science* **350**, 507–508 (2015).
1243 9. Dubilier, N., McFall-Ngai, M. & Zhao, L. Microbiology: Create a global microbiome effort.
1244 *Nature Publishing Group UK* http://dx.doi.org/10.1038/526631a (2015)
1245 doi:10.1038/526631a.
1246 10. Cheng, A. G. *et al.* Design, construction, and in vivo augmentation of a complex gut
1247 microbiome. *Cell* **185**, 3617-3636.e19 (2022).
1248 11. Atarashi, K. *et al.* Treg induction by a rationally selected mixture of Clostridia strains from the
1249 human microbiota. *Nature* **500**, 232–236 (2013).
1250 12. Caballero, S. *et al.* Cooperating Commensals Restore Colonization Resistance to
1251 Vancomycin-Resistant Enterococcus faecium. *Cell Host Microbe* **21**, 592-602.e4 (2017).

1252 13. Clark, R. L. *et al.* Design of synthetic human gut microbiome assembly and butyrate
1253 production. *Nat. Commun.* **12**, 3254 (2021).
1254 14. Faith, J. J., Ahern, P. P., Ridaura, V. K., Cheng, J. & Gordon, J. I. Identifying gut microbe-
1255 host phenotype relationships using combinatorial communities in gnotobiotic mice. *Sci.*
1256 *Transl. Med.* **6**, 220ra11 (2014).
1257 15. Raman, A. S. *et al.* A sparse covarying unit that describes healthy and impaired human gut
1258 microbiota development. *Science* **365**, (2019).
1259 16. Chang, C.-Y., Bajić, D., Vila, J. C. C., Estrela, S. & Sanchez, A. Emergent coexistence in
1260 multispecies microbial communities. *Science* **381**, 343–348 (2023).
1261 17. Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* **10**,
1262 538–550 (2012).
1263 18. Hernández Medina, R. *et al.* Machine learning and deep learning applications in microbiome
1264 research. *ISME Commun* **2**, 98 (2022).
1265 19. Lawson, C. E. Retooling Microbiome Engineering for a Sustainable Future. *mSystems*
1266 e0092521 (2021).
1267 20. Skwara, A. *et al.* Statistically learning the functional landscape of microbial communities. *Nat*
1268 *Ecol Evol* **7**, 1823–1833 (2023).
1269 21. Gowda, K., Ping, D., Mani, M. & Kuehn, S. Genomic structure predicts metabolite dynamics
1270 in microbial communities. *Cell* **185**, 530-546.e25 (2022).
1271 22. Baranwal, M. *et al.* Recurrent neural networks enable design of multifunctional synthetic
1272 human gut microbiome dynamics. *Elife* **11**, (2022).
1273 23. Stein, R. R. *et al.* Computer-guided design of optimal microbial consortia for immune system
1274 modulation. *Elife* **7**, (2018).
1275 24. *Prioritization of pathogens to guide discovery, research and development of new antibiotics*
1276 *for drug-resistant bacterial infections, including tuberculosis*. (World Health Organization,
1277 2019).

1278 25. Spragge, F. *et al.* Microbiome diversity protects against pathogens by nutrient blocking.
1279 *Science* **382**, eadj3502 (2023).
1280 26. Sorbara, M. T. *et al.* Inhibiting antibiotic-resistant Enterobacteriaceae by microbiota-
1281 mediated intracellular acidification. *J. Exp. Med.* **216**, 84–98 (2019).
1282 27. Osbelt, L. *et al.* Klebsiella oxytoca causes colonization resistance against multidrug-resistant
1283 K. pneumoniae in the gut via cooperative carbohydrate competition. *Cell Host Microbe* **29**,
1284 1663-1679.e7 (2021).
1285 28. Kuby, M. J. Programming models for facility dispersion: The *p*-dispersion and maxisum
1286 dispersion problems. *Geogr. Anal.* **19**, 315–329 (1987).
1287 29. Erkut, E. The discrete p-dispersion problem. *Eur. J. Oper. Res.* **46**, 48–60 (1990).
1288 30. Drezner, Z. & Erkut, E. Solving the Continuous p-Dispersion Problem Using Non-Linear
1289 Programming. *J. Oper. Res. Soc.* **46**, 516–520 (1995).
1290 31. Xiong, H. *et al.* Distinct Contributions of Neutrophils and CCR2+ Monocytes to Pulmonary
1291 Clearance of Different Klebsiella pneumoniae Strains. *Infect. Immun.* **83**, 3418–3427 (2015).
1292 32. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
1293 33. Liaw, A. & Wiener, M. Classification and Regression by randomForest. (2007).
1294 34. Plerou, V. *et al.* Random matrix approach to cross correlations in financial data. *Phys. Rev.*
1295 *E Stat. Nonlin. Soft Matter Phys.* **65**, 066126 (2002).
1296 35. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein sectors: evolutionary units of
1297 three-dimensional structure. *Cell* **138**, 774–786 (2009).
1298 36. Rivoire, O., Reynolds, K. A. & Ranganathan, R. Evolution-Based Functional Decomposition
1299 of Proteins. *PLoS Comput. Biol.* **12**, e1004817 (2016).
1300 37. Lutsiv, T. *et al.* Compositional Changes of the High-Fat Diet-Induced Gut Microbiota upon
1301 Consumption of Common Pulses. *Nutrients* **13**, (2021).

1302 38. Jia, J. *et al.* Conserved Covarying Gut Microbial Network in Preterm Infants and Childhood
1303 Growth During the First 5 Years of Life: A Prospective Cohort Study. *Am. J. Clin. Nutr.* **118**,
1304 561–571 (2023).
1305 39. Dahirel, V. *et al.* Coordinate linkage of HIV evolution reveals regions of immunological
1306 vulnerability. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 11530–11535 (2011).
1307 40. Oliveira, R. A. *et al.* Klebsiella michiganensis transmission enhances resistance to
1308 Enterobacteriaceae gut invasion by nutrition competition. *Nat Microbiol* **5**, 630–641 (2020).
1309 41. Doran, B. A. *et al.* An evolution-based framework for describing human gut bacteria. *bioRxiv*
1310 (2023) doi:10.1101/2023.12.04.569969.
1311 42. Kouvaris, K., Clune, J., Kounios, L., Brede, M. & Watson, R. A. How evolution learns to
1312 generalise: Using the principles of learning theory to understand the evolution of
1313 developmental organisation. *PLoS Comput. Biol.* **13**, e1005358 (2017).
1314 43. Oliveira, R. A. & Pamer, E. G. Assembling symbiotic bacterial species into live therapeutic
1315 consortia that reconstitute microbiome functions. *Cell Host Microbe* **31**, 472–484 (2023).
1316 44. Wang, M. *et al.* Strain dropouts reveal interactions that govern the metabolic output of the
1317 gut microbiome. *Cell* **186**, 2839-2852.e21 (2023).
1318 45. Diener, C. & Gibbons, S. M. More is Different: Metabolic Modeling of Diverse Microbial
1319 Communities. *mSystems* **8**, e0127022 (2023).
1320 46. Winkler, E. S. *et al.* The Intestinal Microbiome Restricts Alphavirus Infection and
1321 Dissemination through a Bile Acid-Type I IFN Signaling Axis. *Cell* **182**, 901-918.e18 (2020).
1322 47. Hromada, S. *et al.* Negative interactions determine *Clostridioides difficile* growth in synthetic
1323 human gut communities. *Mol. Syst. Biol.* **17**, e10355 (2021).
1324 48. Venturelli, O. S. *et al.* Deciphering microbial interactions in synthetic human gut microbiome
1325 communities. *Mol. Syst. Biol.* **14**, e8157 (2018).
1326 49. Connors, B. M. *et al.* Control points for design of taxonomic composition in synthetic human
1327 gut communities. *Cell Systems* **14**, 1044-1058.e13 (2023).

1328 50. Russ, W. P. *et al.* An evolution-based model for designing chorismate mutase enzymes.
1329 *Science* **369**, 440–445 (2020).
1330 51. Yeh, A. H.-W. *et al.* De novo design of luciferases using deep learning. *Nature* **614**, 774–780
1331 (2023).
1332 52. Anishchenko, I. *et al.* De novo protein design by deep network hallucination. *Nature* **600**,
1333 547–552 (2021).
1334 53. Wang, J. *et al.* Scaffolding protein functional sites using deep learning. *Science* **377**, 387–
1335 394 (2022).
1336 54. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature*
1337 **620**, 1089–1100 (2023).
1338 55. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein
1339 or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
1340 56. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and
1341 comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).
1342 57. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence
1343 data. *Bioinformatics* **30**, 2114–2120 (2014).
1344 58. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De
1345 Novo Assembler. *Curr. Protoc. Bioinformatics* **70**, e102 (2020).
1346 59. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome*
1347 *Biol.* **20**, 257 (2019).
1348 60. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421
1349 (2009).
1350 61. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status,
1351 taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-45 (2016).
1352 62. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify
1353 genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).

1354 63. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with
1355 uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
1356 64. Chamberlain, S. A. & Szöcs, E. taxize: taxonomic search and retrieval in R. *F1000Res.* **2**,
1357 191 (2013).
1358 65. Odenwald, M. A. *et al.* Bifidobacteria metabolize lactulose to optimize gut metabolites and
1359 prevent systemic infection in patients with liver disease. *Nat Microbiol* **8**, 2033–2049 (2023).
1360 66. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data.
1361 *Nat. Methods* **13**, 581–583 (2016).
1362 67. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid
1363 assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*
1364 **73**, 5261–5267 (2007).
1365 68. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a
1366 phylogenetically consistent, rank normalized and complete genome-based taxonomy.
1367 *Nucleic Acids Res.* **50**, D785–D794 (2022).
1368 69. Haak, B. W. *et al.* Impact of gut colonization with butyrate-producing microbiota on
1369 respiratory viral infection following allo-HCT. *Blood* **131**, 2978–2986 (2018).
1370

Data Availability

The datasets generated in our study are available within Supplementary Tables. Metagenomic data generated from profiling of human fecal microbiomes used in this study are publicly available on NCBI under BioProject ID PRJNA838648. 16S data generated from mouse experiments used in this study will be publicly available on NCBI under BioProject ID PRJNA1074807. Raw data files associated with metabolomic data used in this study will be found on MassIVEW repository MSV000094183.

Code Availability

All code was written in either Python or R; code for all analysis will be found on Github (https://github.com/aramanlab/Oliveira_et_al_2024).

Figures

Figure panels associated with data were generated using either the Prism software (v10.2.0), various available packages in R, or Python. Figure schemes were generated using BioRender (BioRender.com) or Adobe Illustrator.

Acknowledgments

We thank members of the Pamer, Kuehn, and Raman laboratories for helpful discussion. We thank D. Pincus, M. Mani, A. Murugan, R. Ranganathan, and E. Pamer for helpful discussion. We thank members of the biobank, genomics, and metabolomics core services within the Duchossois Family Institute (DFI) and E. Pamer for their help in isolate collection, sequencing, and metabolomic profiling of all samples described in this manuscript. We thank the late E. Littman for his bioinformatic contribution for creating the UMAP plot of the commensal strain bank. This work is supported by the Duchossois Family Institute (DFI) at the University of Chicago and the Dr. Ralph and Marian Falk Medical Research Trust.

Author Information

B.P. performed all analysis associated with metabolomic profiling for the 96 DMCs and out-of-sample DMCs as well as building the RF model for metabolite-based community *K. pneumoniae* clearance prediction; K. L. wrote the code to implement the UMAP-based design strategy; M. Y. wrote the code to create the RF model based on strain presence-absence and scored 100,000 DMCs using the resulting RF model; R.Y.C. performed statistical analysis of the 100,000 DMCs to ultimately yield SynCom15; C.T. aided in the assay of all DMCs for evaluating their capacity to clear *K. pneumoniae* as well as the effect of select communities in an *in vivo* setting; E.M. aided in the assay of variants of SynCom15 for evaluating their capacity to clear *K. pneumoniae* across various *in vitro* conditions; F.H. and V.A. aided with establishing the plate-based assay to evaluate clearing *K. pneumoniae*; R.R. aided in analysis of taxonomic composition of fecal pellets procured from mice and analysis of fecal samples collected from healthy human donors; R.O. conducted all experiments involving DMCs across different conditions, all *in vivo* experiments, all experiments involving characterization of SynCom15, provided material for metabolomic and genomic analysis; S.K. and A.S.R. conceived of the statistical approach for community design; A.S.R. supervised all aspects of data collection and analysis; R.O. and A.S.R. conceived of the *in vitro* and *in vivo* experiments, the analysis of healthy human samples, the metabolic profiling of DMCs; and the evaluation of compression described in Supplementary Discussion; R.O. and A.S.R. wrote the manuscript.

Ethics Declarations

Patents (63/543,XXX & 63/543,XXX) related to this research have been filed by The University of Chicago with S.K., R.A.O, and A.S.R as inventors.

Materials and Correspondence

Author to whom correspondence and materials request should be addressed is A.S.R.