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.

Main

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

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

pneumoniae in an *in vitro* setting. Statistical inference performed on our model identified a set of 15 key strains that when combined into a community ('SynCom15') (i) sustainably suppressed *K. pneumoniae* across various diverse *in vitro* environments, (ii) matched the clearance ability of a fecal microbial transplant (FMT) in a pre-clinically relevant mouse model of infection, (iii) was a safe intervention *in vivo*, (iv) could not be obviously deconstructed into a functional subset of strains, and (v) did not resemble the composition of natural human gut microbiotas. We found that considering the metabolic capacity of DMCs including fatty acid and nutrient metabolism—appreciated mechanisms of *K. pneumoniae* suppression—instead of strain presence-absence resulted in a poor generative model, highlighting the advantage of describing DMCs by their strain content for deriving generative models of community design[25–27]. Our work describes a potentially therapeutic, sparse synthetic microbiome made of human gut commensal bacteria for treatment of MDR *K. pneumoniae* infections and, more generally, introduces the concept of 'statistical design' for microbial ecosystems.

A generative model of community design for suppressing K. pneumoniae

To begin our DBTL+ approach, we first isolated and whole-genome sequenced 848 gut commensal strains from fecal samples of healthy donors (**Fig. 1A**,**B**; **Supplementary Table 1**) (Methods). Our strain bank was enriched for the phyla Bacteroidota, Bacillota, Actinomycetota, Pseudomonadota, and Verrucomicrobia, and contained a richness of diversity at the genus and species levels reflecting the diversity of donor microbiomes (**Fig. 1B**, **Extended Data Fig. 1** and **2; Supplementary Information**) (Methods). The possible combinatorial space of DMCs we could synthesize was $2^{848}/2$—an insurmountable number. As such, we reduced the size of the strain bank while maintaining the genomic diversity of the resulting set (**Fig. 1C**; **Supplementary Table 2A**). We chose 46 strains as the size of our reduced strain bank because creating a nearly 50-member community was the practical limit we could achieve without compromising the fitness of bacteria in culture. Despite substantially reducing the size of

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

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

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

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

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

We trained and validated a Random-Forests (RF) machine-learning algorithm to learn a statistical relationship between DMC design—defined by only the designed pattern of strain presence and absence of DMCs as represented by the matrix shown in **Fig. 1E**—and the ending *K. pneumoniae* abundance after co-culture with the community[32,33]. Thus, no information about which strains engrafted or survived in the culture, strain dynamics during the experiment, ending configuration of the community, information regarding the nature of microbial interactions, or information regarding mechanism of *K. pneumoniae* suppression was considered when training or validating this model. The RF model was trained on 90% of the data and validated on the remaining 10% 100 times for bootstrap support, resulting in an in-sample validation r^2 value of 0.98 (**Supplementary Table 4A**) (Methods).

We then tested the predictive capacity of our RF model for newly constructed DMCs that the model had never seen as a true 'out-of-sample' test. We created 60 new DMCs spanning different membership sizes that were not a part of the initial 96 DMCs and were predicted by the trained RF model to span a large dynamic range of *K. pneumoniae* clearance in our assay (**Supplementary Table 5A**). Thus the 60 new DMCs defined a true 'out-of-sample' set generated by our RF model. We compared the abundance of *K. pneumoniae* for the 60 new DMCs as predicted by our RF model versus the *K. pneumoniae* abundance we experimentally observed after co-culture of each of the 60 new DMCs with *K. pneumoniae* for 120 hours. We found that our RF model was predictive of the resulting *K. pneumoniae* abundance to an r^2 value of 0.6 ($p < 10^{-3}$) (**Fig. 1F**, **Supplementary Table 5B**).

Collectively, our results showed that our RF model could accurately predict the capacity of a complex microbial community defined by our 46 strains to suppress *K. pneumoniae*, thereby enabling engineering of new communities with desired suppressive capacity. Thus, in a single round of DBTL where the first round of design was constrained by genomic diversity of strain combinations (DBTL+), we derived a generative model of community design for suppressing *K. pneumoniae* in BHIS media.

Defining and characterizing SynCom15

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

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

from the RF model (**Fig. 2B**, **Supplementary Table 6A**). We next performed Principal Components Analysis (PCA) on the alignment of communities, yielding 46 principal components (PCs) of data-variance. We regressed the contribution of each of the 5,752 DMCs onto each PC against the predicted *K. pneumoniae* abundance to identify PCs that most associate with *K. pneumoniae* suppression in a data-driven, unbiased manner. We found that PC46, containing <0.1% data-variance, was the most associated with *K. pneumoniae* abundance (**Fig. 2C**, **Supplementary Table 6B-D**) (Methods).

Similar to the distribution of RF importance scores, the contribution of strains onto PC46 was continuous precluding the ability to define groups of strains to construct communities (**Fig. 2D**, left panel; **Supplementary Table 6E**). Interestingly, the contribution of strains onto PC46 did not resemble the distribution of feature importance scores from the RF model, suggesting that PC46 contained information that was different from the RF model (**Extended Data Fig. 5B**). To use the information in PC46 to define groups of statistically interacting strains, we computed the statistical similarity between all pairs of strains on PC46 (Methods). The concept behind this measure is that two strains that significantly contribute to PC46 and are close together along PC46 are, on average, co-present in DMCs predicted to suppress *K. pneumoniae*. Hierarchical clustering of the pairwise similarity between strains illustrated a distinct block structure amongst five separate groups (**Fig. 2D**, right panel; **Supplementary Table 6F**). Five strains that contributed the most to defining PC46—*Clostridium innocuum*, *Clostridium symbosium*, *Colinsella aerofaciens*, *Escherichia coli*, and *Bacteroides xylanisolvens*—formed a group that we term ‘Block 1’ (**Fig. 2D**, right panel, orange group). The following ten strains that contributed to PC46—*Lacrimispora celerecrescens*, *Bacteroides caccae*, *Blautia faecis*, *Blautia obeum*, *Clostridium scindens*, a *Bifidobacterium* species, *Megasphera massiliensis*, *Coprococcus comes*, *Mitsuokella jalaludinii*, and *Blautia producta*—formed a group that we term ‘Block 5’ (**Fig. 2D**, right panel, green group). Blocks 1 and 5 exhibited collective similarity amongst each other; we term this group of strains ‘SynCom15’—a 15-member group comprised of statistically

interacting strains that are co-present in communities predicted to clear *K. pneumoniae*. In contrast to SynCom15, three other groups of strains were statistically inferred to be co-absent in communities predicted to clear *K. pneumoniae*. These groups were comprised of 7, 10, and 14 strains; we term these groups 'Block 2', 'Block 3', and 'Block 4' respectively (**Fig. 2D,** right panel, red group, brown group, and yellow group) (**Supplementary Table 6G**).

We hypothesized that SynCom15 would be efficacious at clearing *K. pneumoniae* across different environments because it was predicted to contain the key, critical species for DMC function. We built and tested SynCom15 as well as all other Blocks for their capacity to clear *K. pneumoniae* across three unrelated media conditions: BHIS, media created from the cecal extracts of germ-free (GF) mice, and media created from the cecal extracts of specific-pathogen-free (SPF) mice treated with broad spectrum antibiotics (Ab-treated SPF) (**Fig. 2E**, left panel) (Methods). As a comparator to SynCom15 and the other Blocks, we also tested DMC46—the community that suppressed *K. pneumoniae* the most in BHIS media. Notably, our results clearly illustrated the environmental dependence of community efficacy. Blocks 2 and 3 were consistently ineffective at suppressing *K. pneumoniae* across environments while Blocks 1, 4 and 5 were able to suppress *K. pneumoniae* depending on the environment in which they were tested—Block 1 in BHIS and Blocks 4 and 5 in GF cecal extract. Thus, Blocks 1, 4 and 5 were conditionally effective. In contrast, we found that DMC46 and SynCom15 suppressed *K. pneumoniae* across all three conditions and were therefore unconditionally effective. DMC46 cleared *K. pneumoniae* across all environments. SynCom15 suppressed *K. pneumoniae* five orders of magnitude in BHIS, cleared *K. pneumoniae* in GF cecal extracts, and suppressed *K. pneumoniae* greater than four orders of magnitude in Ab-SPF cecal extracts. (**Fig. 2E**, right panels) (**Supplementary Table 7**).

Thus, our strategy of statistical inference performed on the RF model of community design defined SynCom15—a phylogenetically diverse 15-member community—that

suppressed *K. pneumoniae* across diverse environmental contexts in a manner similar to DMC46—the community containing all 46 strains.

SynCom15 clears K. pneumoniae in a pre-clinically relevant mouse model of infection

Because DMC46 and SynCom15 were unconditionally effective at clearing *K. pneumoniae in vitro*, we sought to test the ability of both communities to clear *K. pneumoniae* in a more complex, clinically relevant environment. We evaluated the efficacy of DMC46 and SynCom15 in a mouse model of infection. To mimic a clinically relevant scenario, we did not use germ-free mice (mice without a microbiome). Rather, we treated SPF mice with broad spectrum antibiotics to deplete their gut microbiota then infected them with *K. pneumoniae*—a sequence of events commonly encountered in patients who acquire MDR *K. pneumoniae* infection. Additionally, we singly-housed mice to ensure that no sharing of microbes by coprophagia amongst animals would affect microbiome composition during and post-antibiotic treatment[40]. Singly-housed antibiotic-treated SPF mice infected with *K. pneumoniae* MH258 were given either (i) saline (PBS), (ii) a heterologous whole stool transplant derived from mice ('Fecal Microbial Transplant', FMT), (iii) Block 1, (iv) Block 2, (v) DMC46, or (vi) SynCom15 as interventions for three sequential days after infection (Methods). Blocks 1 and 2 were given as bacterial communities that were either conditionally efficacious across *in vitro* conditions or unable to clear *K. pneumoniae* across any *in vitro* condition respectively. Fecal samples were collected and *K. pneumoniae* abundances were tracked through the course of the experiment by plating (**Fig. 3A**).

We found that Block 1, and Block 2 did not suppress *K. pneumoniae* relative to saline. The FMT suppressed *K. pneumoniae* three orders of magnitude one day after the last gavage and up to six orders of magnitude from four days after the last gavage until the end of the experiment. DMC46 suppressed *K. pneumoniae* three orders of magnitude one day after the last gavage, four orders magnitude four days after the last gavage, and six orders of magnitude

nine days after the last gavage. Thus, DMC46 was able to suppress *K. pneumoniae* but exhibited slow kinetics of response compared to the FMT. In contrast, SynCom15 rapidly suppressed *K. pneumoniae*, resulting in a reduction of abundance by five orders of magnitude one day after the last gavage. Additionally, SynCom15 cleared *K. pneumoniae* four days after the last gavage and maintained clearance through nine days after the last gavage (**Fig. 3B**) (**Supplementary Table 8**). These results highlighted the rapid and sustained efficacy of SynCom15 in clearing *K. pneumoniae in vivo* as well as the utility of reducing the community size from the 46 strains defining DMC46 to the inferred key 15 strains defining SynCom15.

Taxonomic profiling of fecal samples procured through the experiment revealed that 10 of the 15 strains in SynCom15 engrafted in at least one of the mice within the cohort (**Fig. 3C**) (Methods). Dynamics of SynCom15 strains showed that 5 of the 10 strains were present at detectable fractional abundances throughout the course of the experiment—*C. symbiosum*, *B. xylanisolvens*, *C. innocuum*, *B. obeum*, and *B. caccae* (**Extended Data Fig. 6**). Together, these results illustrated that the engraftment and strain dynamics of SynCom15 in mice did not follow obvious phylogenetic trends.

Dynamics of microbiota diversity and structure within the infected mice treated with SynCom15 mirrored that of the FMT and returned the state of the microbiota to that observed prior to antibiotic treatment (**Fig. 3D,E**; **Supplementary Table 9A,B**) (Methods). At the phylogenetic description of phylum, class or family, we observed that treatment via FMT and SynCom15 resulted in similar ending configurations of the microbiota (**Fig. 3F**). However, at the genus-level description, we observed differences between the ending microbiota configuration of mice treated with FMT or SynCom15. Treatment with FMT resulted in the return of *Duncaniella* and *Paramuribaculum* (genera belonging to the order Bacteroidales). Treatment with SynCom15 resulted in detectable presence of the genera *Bacteroides*, derived from the *B. xylanisolvens* strain in SynCom15, in addition to a bloom of *Bifidobacterium* (**Fig. 3G**, **Extended 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

gleaned from analysis of our mouse experiments. Moreover, coarse community descriptions, like community diversity for instance, do not provide an explanation for our results. In contrast, our findings highlight the utility of evaluating community function through our statistical approach that considers emergent, and potentially non-obvious properties of the structure-function relationship for communities.

Comparison of SynCom15 with composition of healthy human fecal microbiomes

We next explored the extent to which SynCom15 was represented across healthy humans who provided FMTs from which we created our strain bank. We first interrogated the prevalence of the genera constituting SynCom15 in fecal samples from healthy donors. We found that the genera represented in SynCom15 reflected a diverse minority of the totality of genera observed across the set of healthy gut microbiomes (**Fig. 4A**). Next, we interrogated the prevalence of the SynCom15 species across the fecal samples of the healthy donors (Methods). We found that no healthy human microbiome contained more than eleven of the SynCom15 species above a fractional abundance of 0.1% (**Fig. 4B**; **Supplementary Table 11**). Moreover, we found certain SynCom15 species to be remarkably sparse in their prevalence across donors. *M. jalaludinii* was not detectable in any donor; *M. massiliensis* was detectable in two donors; *C. symbiosum* in three donors; and *C. scindens* in four donors. Amongst strains that were most prevalent, *B. obeum* and *B. faecis* were detectable in 20 donors; *L. celerecrescens* in 14 donors; *C. comes* in 13 donors; *B. caccae* in 12 donors. Finally, we interrogated the fractional abundance of SynCom15 species across the fecal samples of the 22 healthy donors. We found SynCom15 species were present at a relative abundance of less than 5% across all donors, with a majority of species being found at a relative abundance of less than 0.5% (**Fig. 4C**).

Together, these results illustrated two conclusions. First, the composition of SynCom15 was distinct from that found across healthy human gut microbiotas. This is either because SynCom15 does not exist in the healthy samples from our cohort or because several of

SynCom15 strains are undetectable by our sequencing methods due to their low abundance. Second, the strains comprising SynCom15 were low prevalence and abundance amongst fecal samples of healthy donors. This result highlights the power of generating and using broadly diverse strain banks for engineering synthetic bacterial communities as compared to strain banks reflecting the compositional abundance and prevalence distributions gleaned from analysis of natural human microbiomes.

Community metabolism poorly predicts K. pneumoniae suppression

Engineering SynCom15 was based on statistical analysis of a model that described DMCs by their pattern strain presence-absence and their capacity to clear *K. pneumoniae*. Thus, the model was not constructed using any information about mechanism of action. Previous results have suggested the importance of media acidification and nutrient competition as mechanisms by which complex bacterial communities could suppress *K. pneumoniae*[25–27]. Therefore, we compared the metabolic profiles of the five DMCs that suppressed *K. pneumoniae* the most against the five DMCs that suppressed *K. pneumoniae* the least amongst the 96 DMCs we had previously tested in BHIS (**Fig. 5A**, left panel; **Extended Data Fig. 9**; **Supplementary Table 12A**) (Methods). We analyzed the profile of 118 metabolites across the most and least suppressive DMCs after being co-cultured with *K. pneumoniae* for 72, 96, and 120 hours.

The metabolite patterns that distinguished DMCs that suppressed *K. pneumoniae* from those that did not centered around two metabolic axes: concentrations of fatty acids (FAs) with an emphasis on short-chain fatty acids and amino acids (**Supplementary Table 12B**). With respect to FAs, the most suppressive DMCs produced phenylacetic acid, valeric acid, hexanoic acid, and 5-aminovaleric acid and consumed lactic acid as well as succinic acid. With respect to amino acids, the most suppressive DMCs consumed either (i) amino acids with non-polar side chains (phenylalanine, alanine, isoleucine, leucine, valine) or (ii) glutamic acid and its

associated derivative 5-oxoproline (**Fig. 5A**, right panel). Metabolic profiling of SynCom15 co-cultured with *K. pneumoniae* in BHIS revealed a similar trend. SynCom15 produced the same FAs as the most suppressive DMCs, but also produced lactic acid as opposed to consuming it. SynCom15 also consumed all the amino acids that the most suppressive DMCs consumed (**Fig. 5A**, right panel; **Supplementary Table 12C**). We also performed metabolic profiling of fecal pellets collected from mice treated with either SynCom15 or saline in the experiment described in **Fig. 3A**. Consistent with our *in vitro* results, we found a statistically significant increase in FA production on day 10 and amino acid depletion on day 12 in infected mice given SynCom15 (**Fig. 5B**, **Supplementary Table 12D**). Our *in vitro* and *in vivo* results were in accordance with previously published studies demonstrating the importance of environmental acidification and nutrient competition as mechanisms by which MDR *K. pneumoniae* could be suppressed. Furthermore, these results point to metabolic axes that are shared between the function of SynCom15 in *in vitro* and *in vivo* conditions, suggesting a way that translatability of suppressive capacity across distinct environments could be manifest.

We reasoned that if the mechanism of suppression was exclusively related to FA production and amino acid depletion, we could build a generative statistical model of community design based on the metabolite profile of a large number of DMCs spanning a range of *K. pneumoniae* suppression. This would represent a more thorough test of the sufficiency of FA production and nutrient depletion to explain how DMCs clear *K. pneumoniae*. Thus, we performed metabolic profiling of 81 DMCs that we had designed and tested in BHIS media for their capacity to suppress *K. pneumoniae* (**Supplementary Table 13A**). We removed 15 DMCs from our analysis because they were poorly profiled across metabolite features. Metabolite profiles were measured at 72, 96, and 120 hours of co-culture with *K. pneumoniae*. We also performed metabolic profiling of the 60 DMCs that previously served as the 'out-of-sample' DMCs at 72, 96, and 120 hours of co-culture with *K. pneumoniae* (**Supplementary Table 13B**). We trained and validated an RF model on the metabolic profiles of the 96 DMCs to predict *K.*

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

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

Collectively, these results show that design based on a metabolic profile comprising our targeted panel of features (amino acids, aromatics, branch-chained fatty acids, indoles, phenolic aromatics, and short-chained fatty acids) may not be a reliable strategy for engineering communities that clear *K. pneumoniae* in a predictable manner. Our findings highlight the utility of considering the more coarse-grained description of strain presence-absence in creating generative models of community design.

Discussion

Using clearance of MDR *K. pneumoniae* as a target function, we engineered a defined, sparse microbiome—SynCom15—that is complex, safe, efficacious, and distinct from natural human gut microbiome compositions using a statistical approach for community design. Our results shed light on several notable findings.

First, merely designing genetically diverse communities did not guarantee creating functional communities. However, imposing the constraint of genetic diversity on the 'Design' portion of DBTL was crucial for reducing the space of possible DMCs and was a particularly informative space for learning a generative statistical model. Indeed, extremely limited sampling (building and testing 96 out of the immense number of possible DMCs) was sufficient to converge on an accurate model of design *in vitro*. These results suggest a deep connection between the phylogenies of strains and the collective functions encoded by microbial communities, opening the possibility of phylogenetic-based 'bottom-up' design. The development of emerging methods for parametrizing functional differences amongst strain-level variants through considering their evolutionary history across the bacterial tree-of-life will be useful for testing this idea in the future[41].

Second, accurately translating microbiome function from specific *in vitro* settings to other *in vitro* and *in vivo* environments has historically been a significant challenge. Our data showed 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.

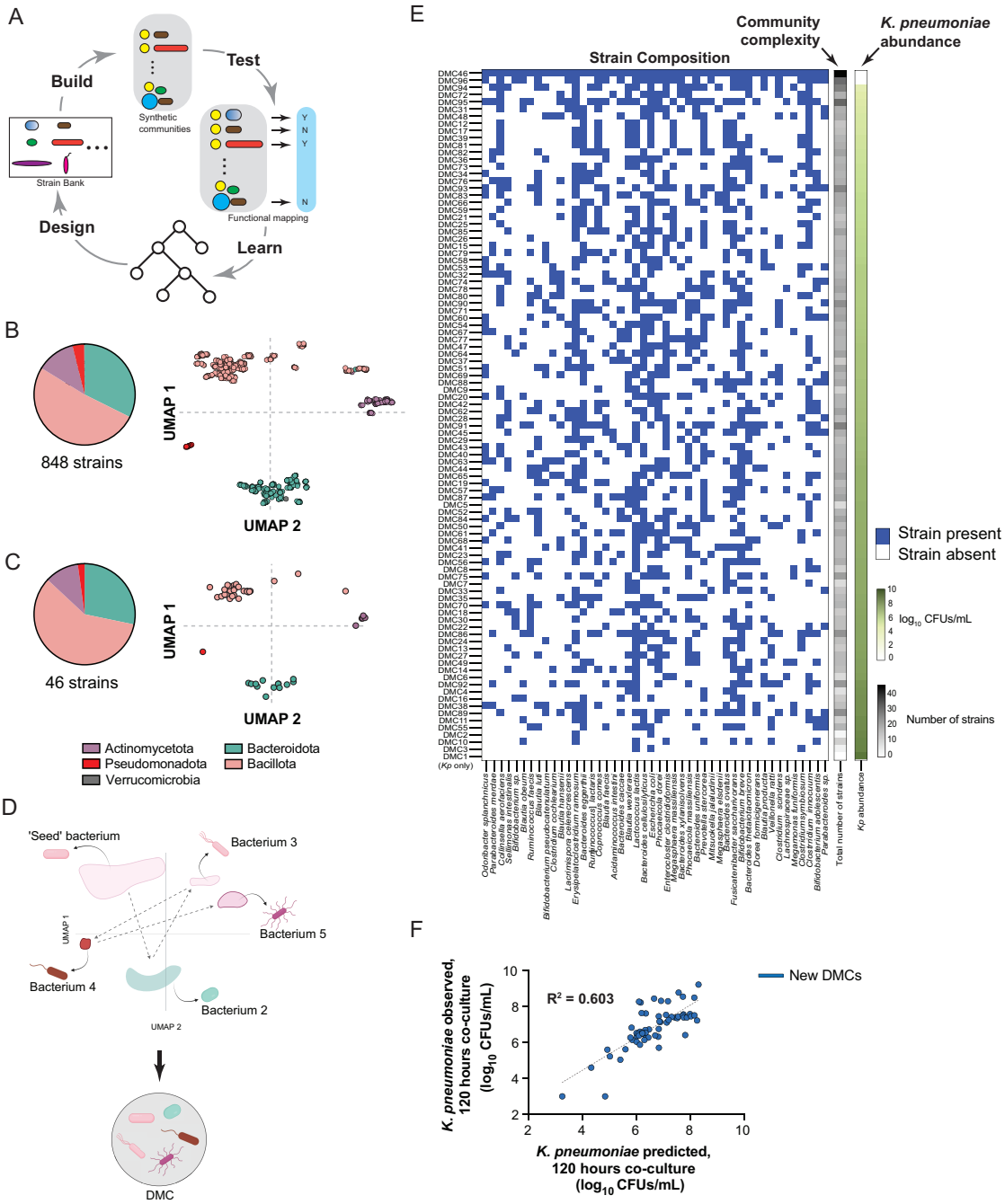
520 Fourth, SynCom15 was more efficacious at suppressing *K. pneumoniae* in mice
521 compared to DMC46—a 46-member community that contained the 15 strains defining
522 SynCom15. This result highlights the functional power of defined small bacterial communities in
523 contrast to recent studies advocating 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.

564 **Figures and Extended Figures**

Oliveira et al., Figure 1



565
566
567
568
569
570
571

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

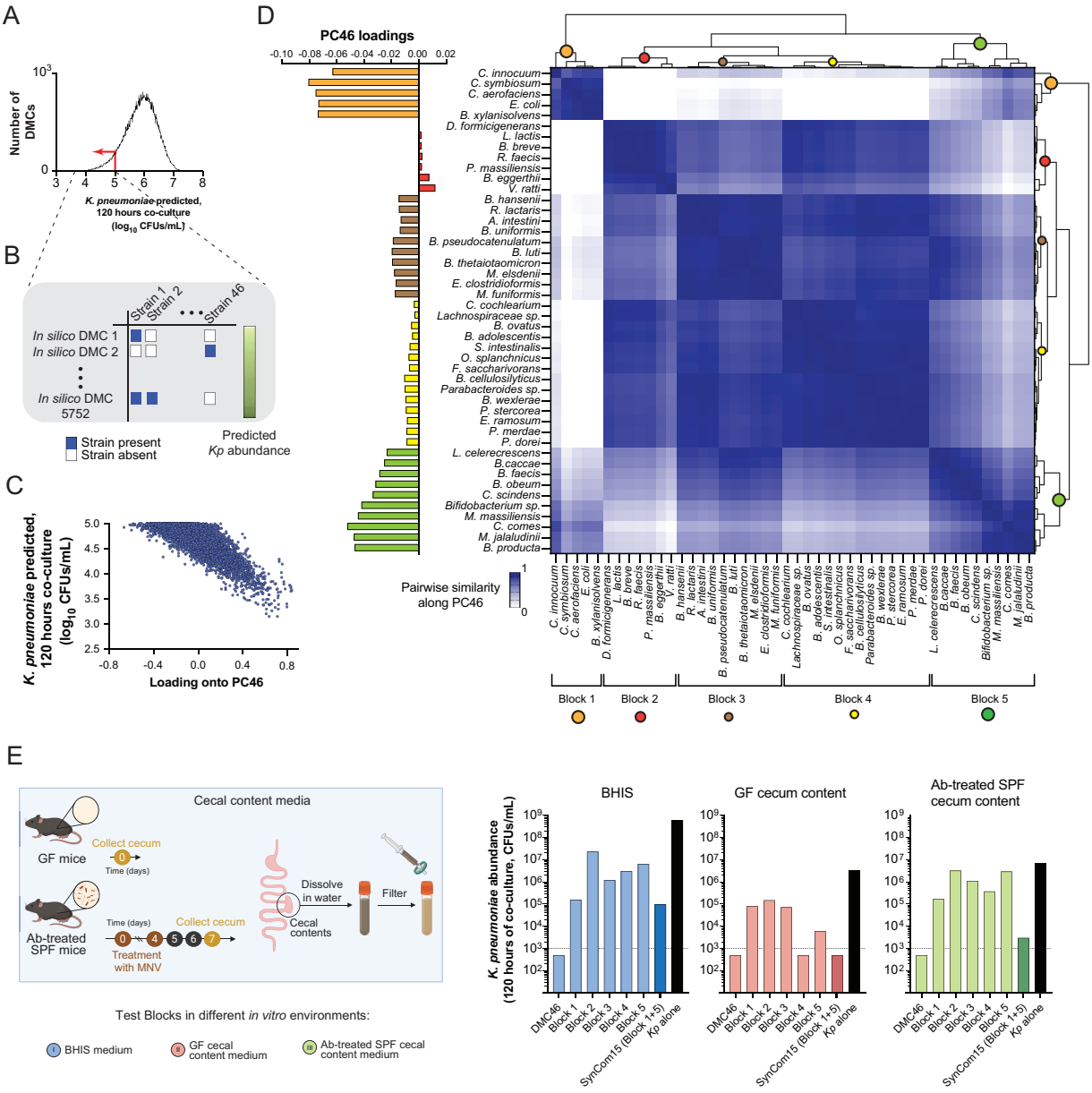
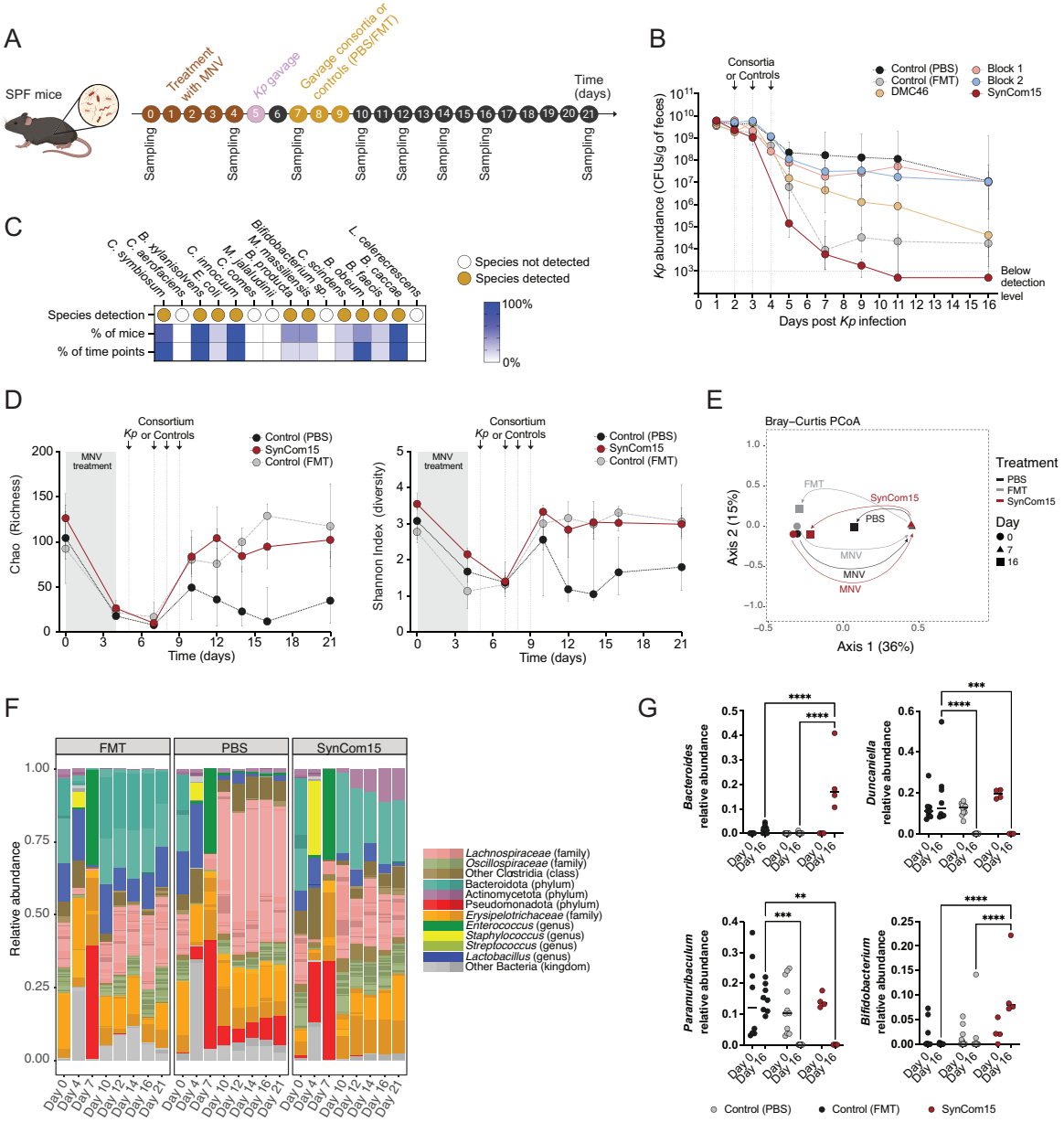


Fig. 2. Defining SynCom15 and evaluating its capacity to suppress *K. pneumoniae* across different environments. (A,B) Histogram of predicted *K. pneumoniae* abundance for 100,000 *in silico* generated DMCs. Red arrow is a threshold of predicted *K. pneumoniae* suppression; DMCs to the left of the arrow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

623

Oliveira et al., Figure 3



624
625
626
627

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

Oliveira et al., Figure 4

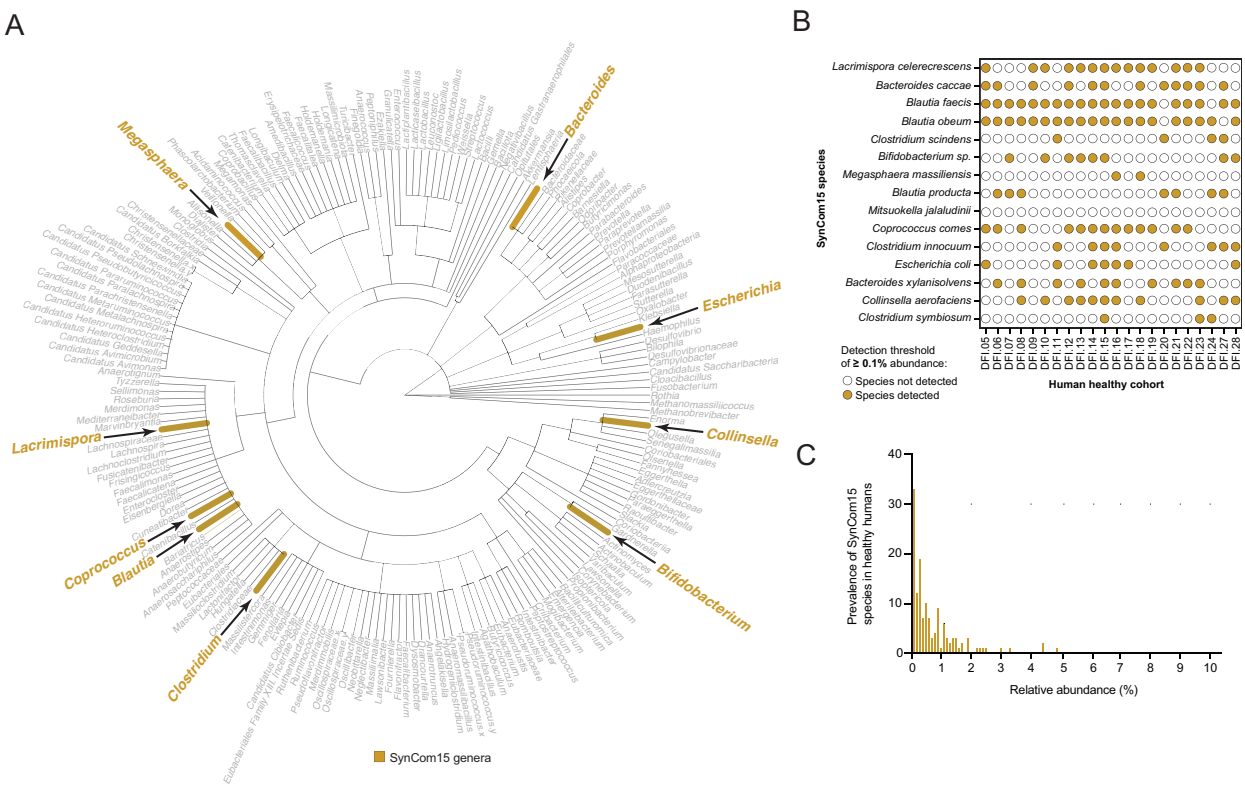
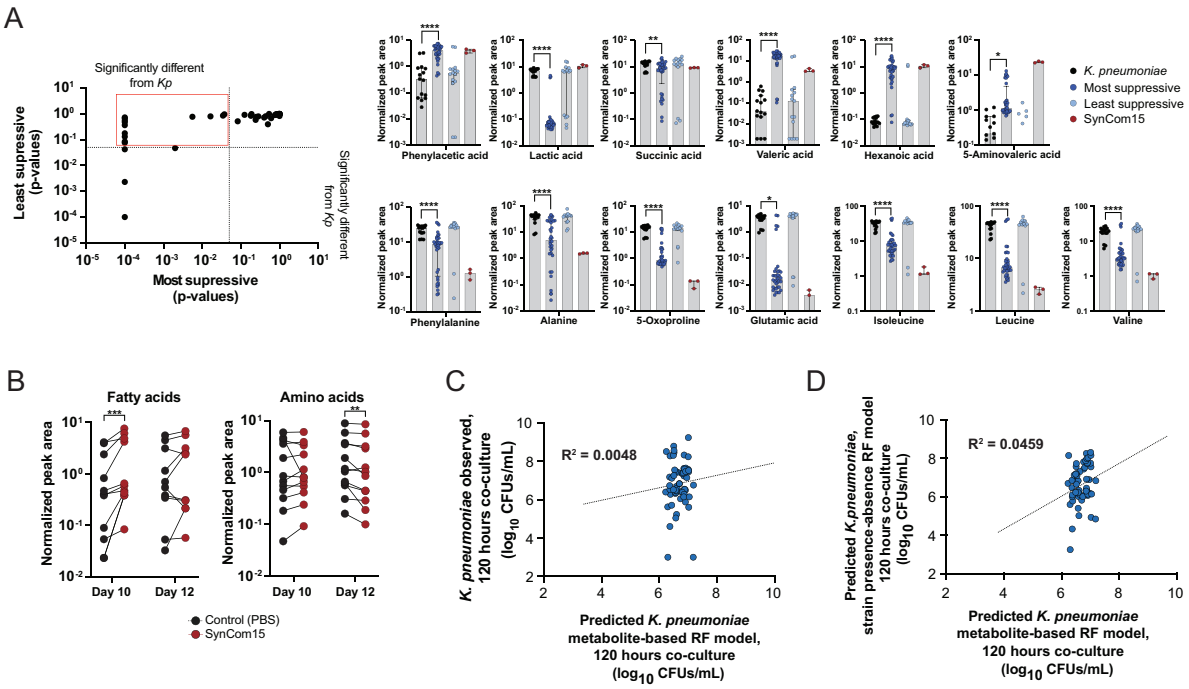
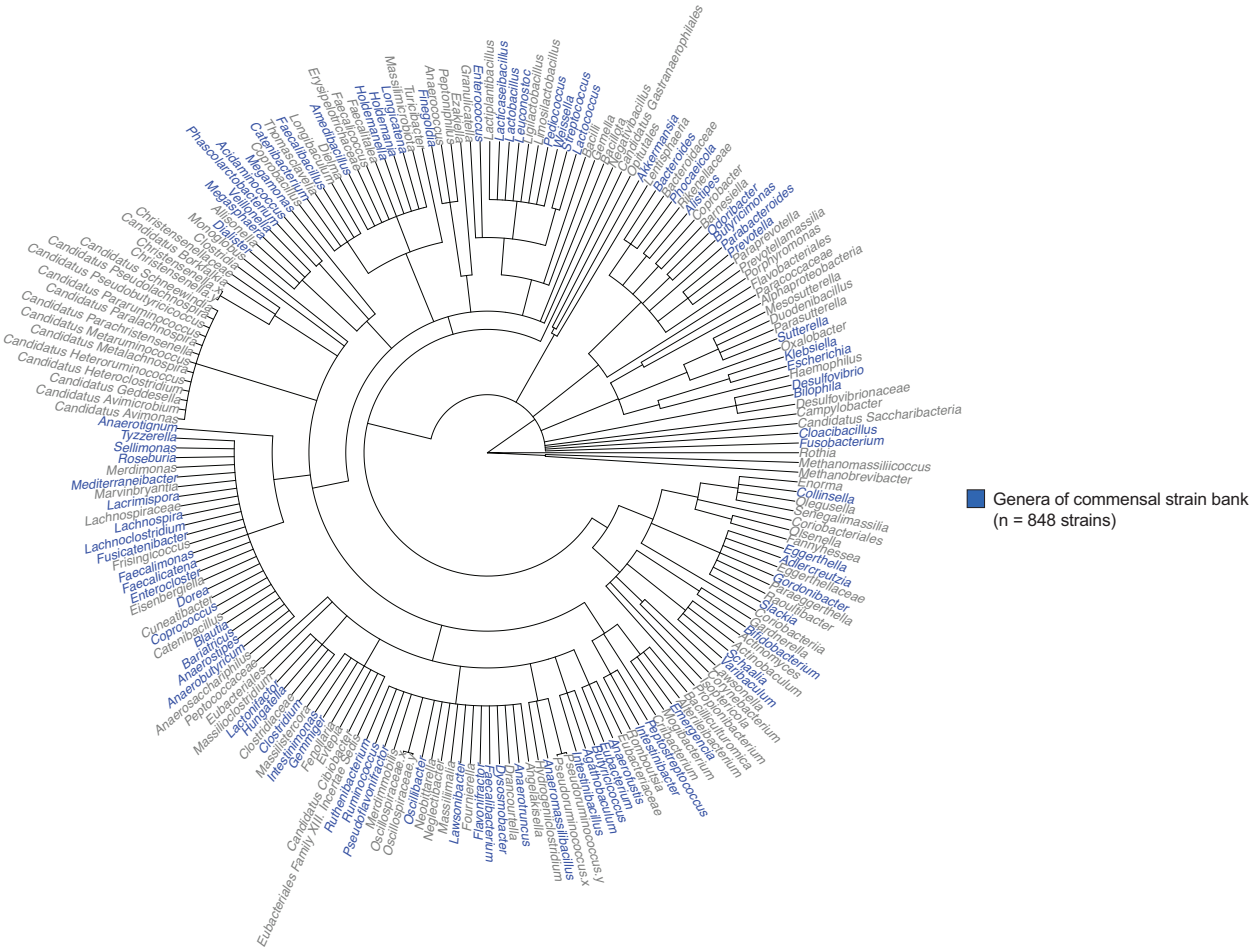


Fig. 4. Comparison of SynCom15 composition with composition of healthy human microbiomes. **(A)** Phylogenetic tree of genera present across fecal microbiomes of human donors. Brown genera are those found across SynCom15 strains (genera names are according to annotation by Metaphlan). **(B)** Prevalence 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.

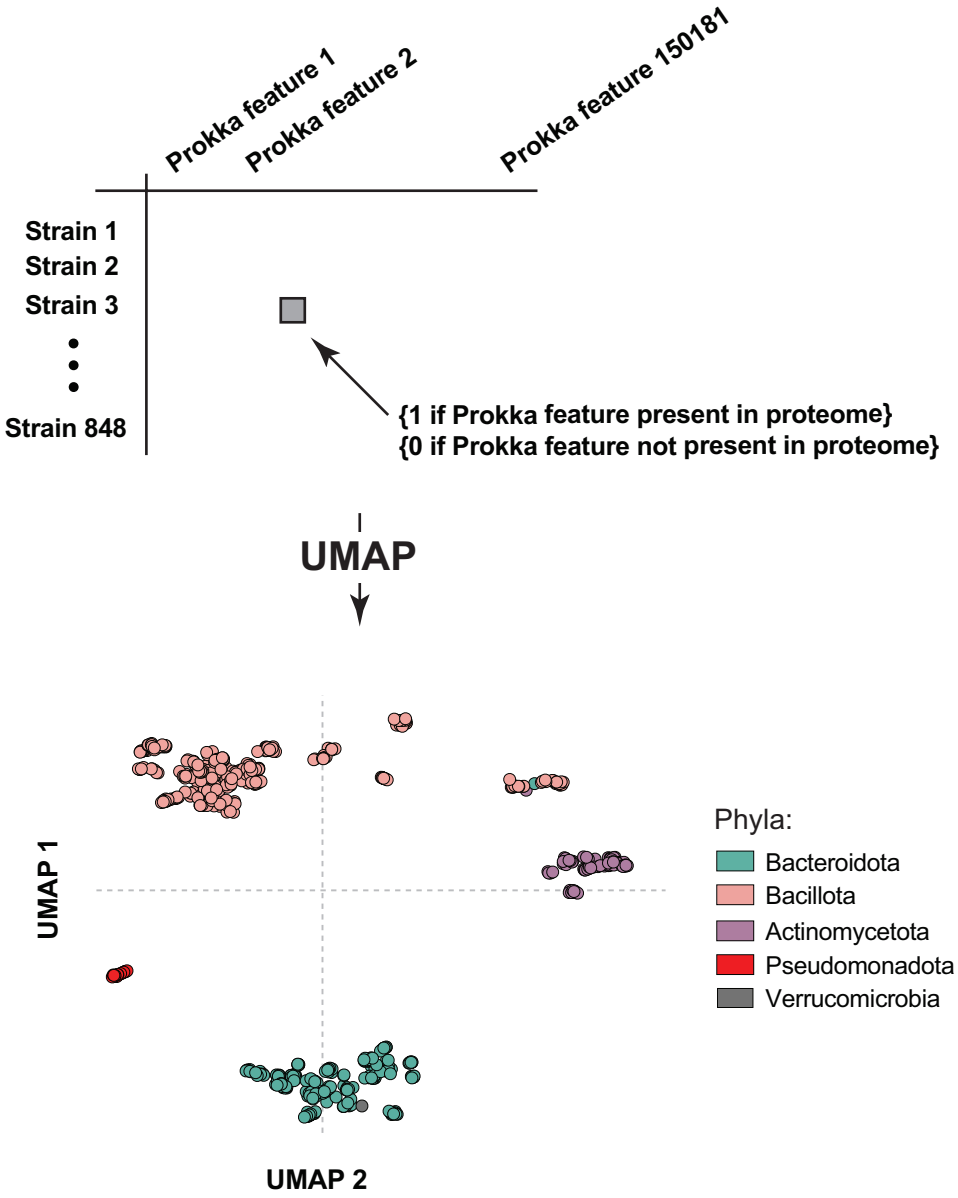
Oliveira et al., Figure 5



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

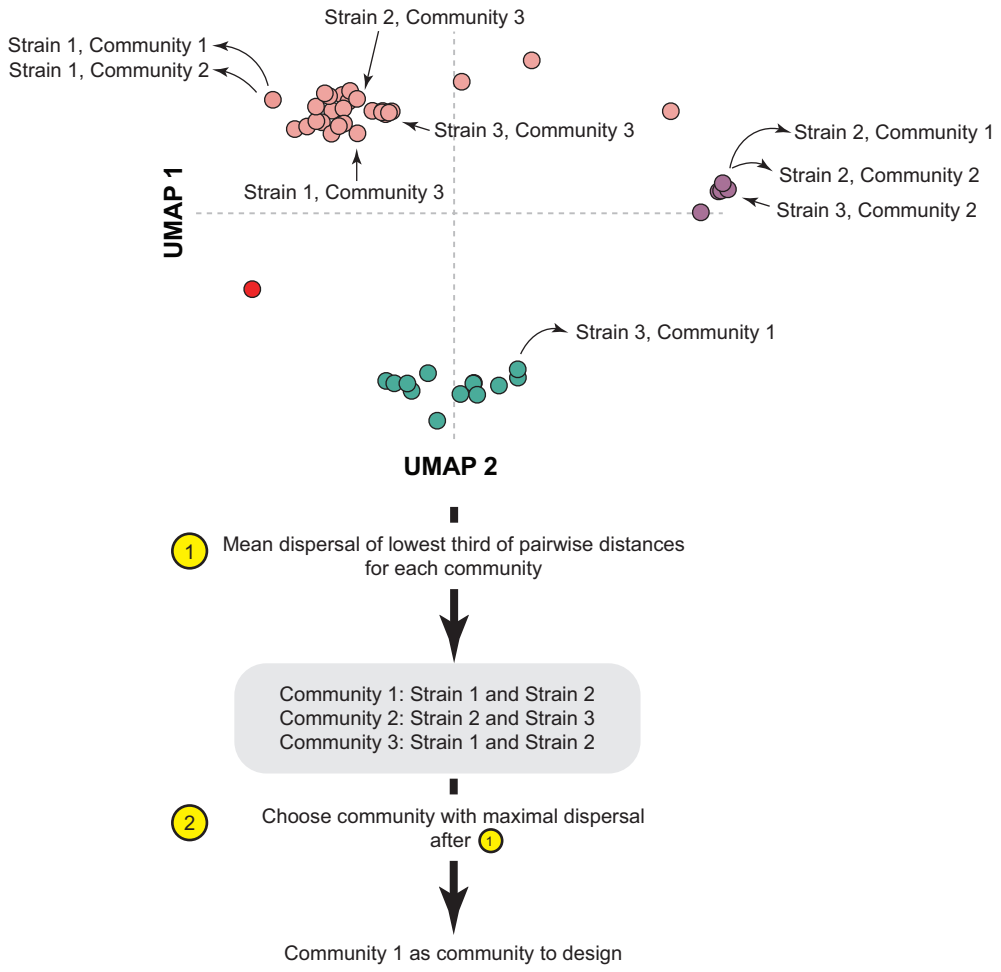


Extended Data Fig. 1. Fecal samples collected from 22 healthy human donors were subject to shotgun metagenomic sequencing (Methods). Tree of the genera comprising all fecal microbiomes (annotations per Metaphlan) is shown here (Methods). Colored in blue are the distribution of genera observed in our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

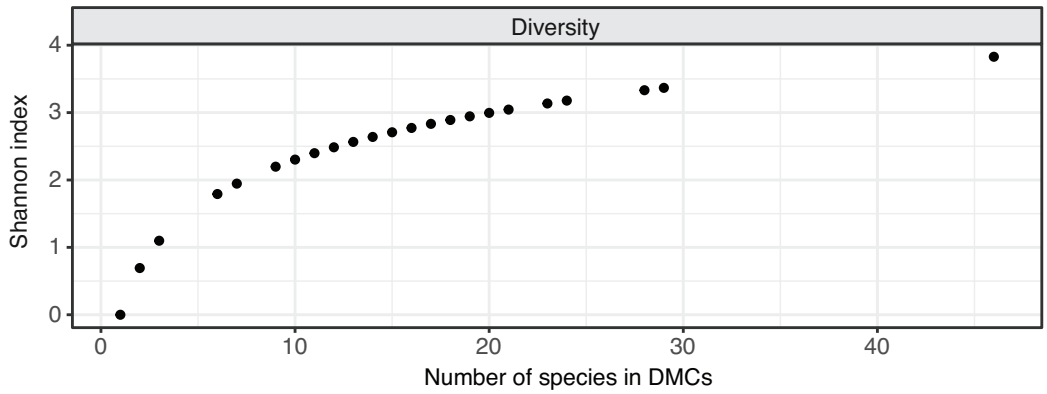


Extended Data Fig. 2. Matrix of strain by Prokka feature for all 848 gut commensals was created where entries are a '1' if the Prokka feature is present in the strain proteome, and '0' if Prokka feature is absent in the strain proteome. Matrix was subject to UMAP visualization; UMAP plot is shown in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

A

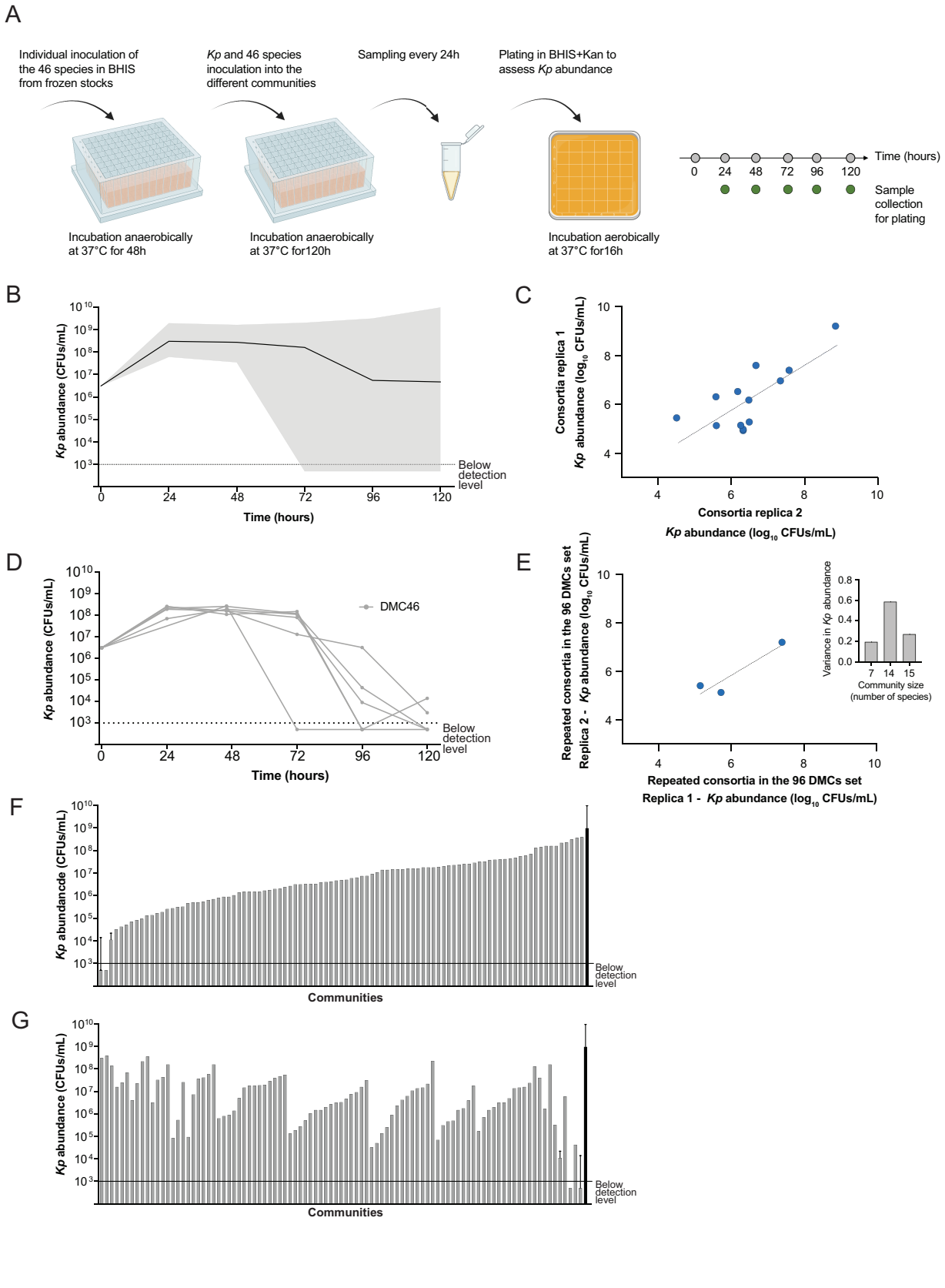


B

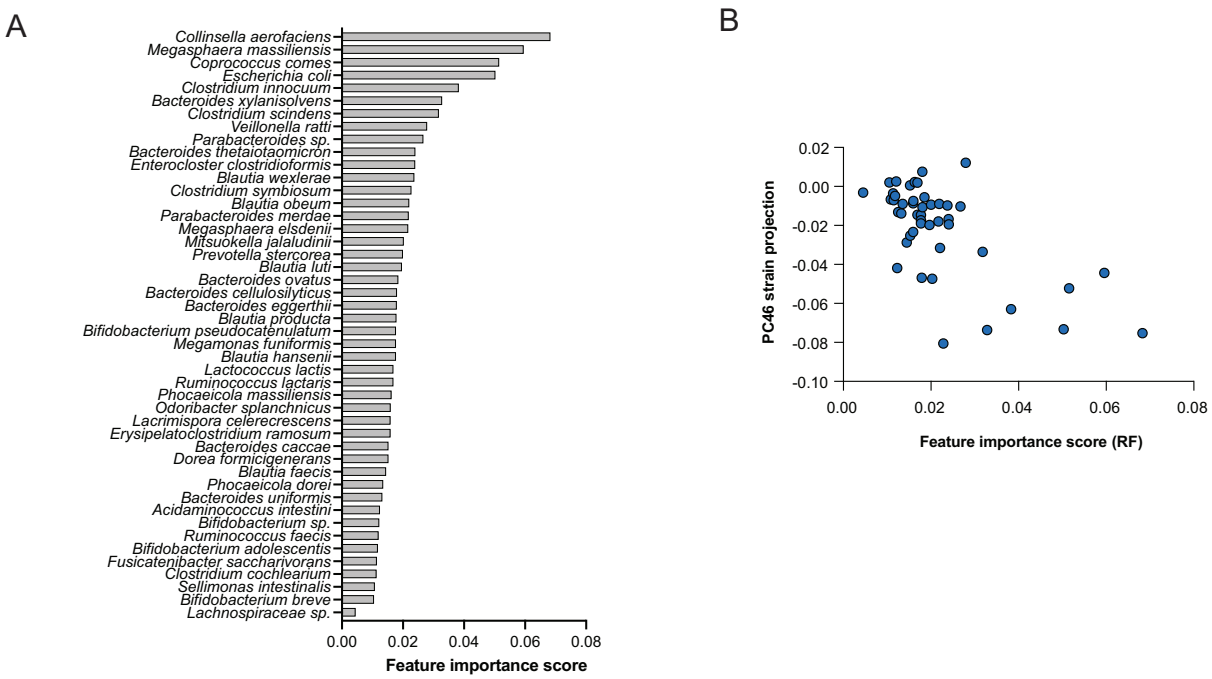


723
724
725

726 **Extended Data Fig. 3. (A)** Workflow for algorithm used to design DMCs. Communities with
727 three bacterial strains are shown as an example. Given three possible communities that could
728 be created, the first step is to choose the mean dispersal of the lowest third of pairwise
729 distances between strains for each community. In the example shown here, the lowest third is
730 equivalent to the minimum pairwise distance for each community due to the communities being
731 comprised of only three strains (gray box). The second step is to the choose the community with
732 the maximal dispersion per Step 1. In the case shown here, 'Community 1' would be chosen as
733 a DMC for incorporation into our DBTL framework. **(B)** Average Shannon diversity (y-axis)
734 versus number of species in DMCs (x-axis). The maximum possible Shannon diversity is set by
735 the DMC containing all 46 strains used to engineer DMCs.
736

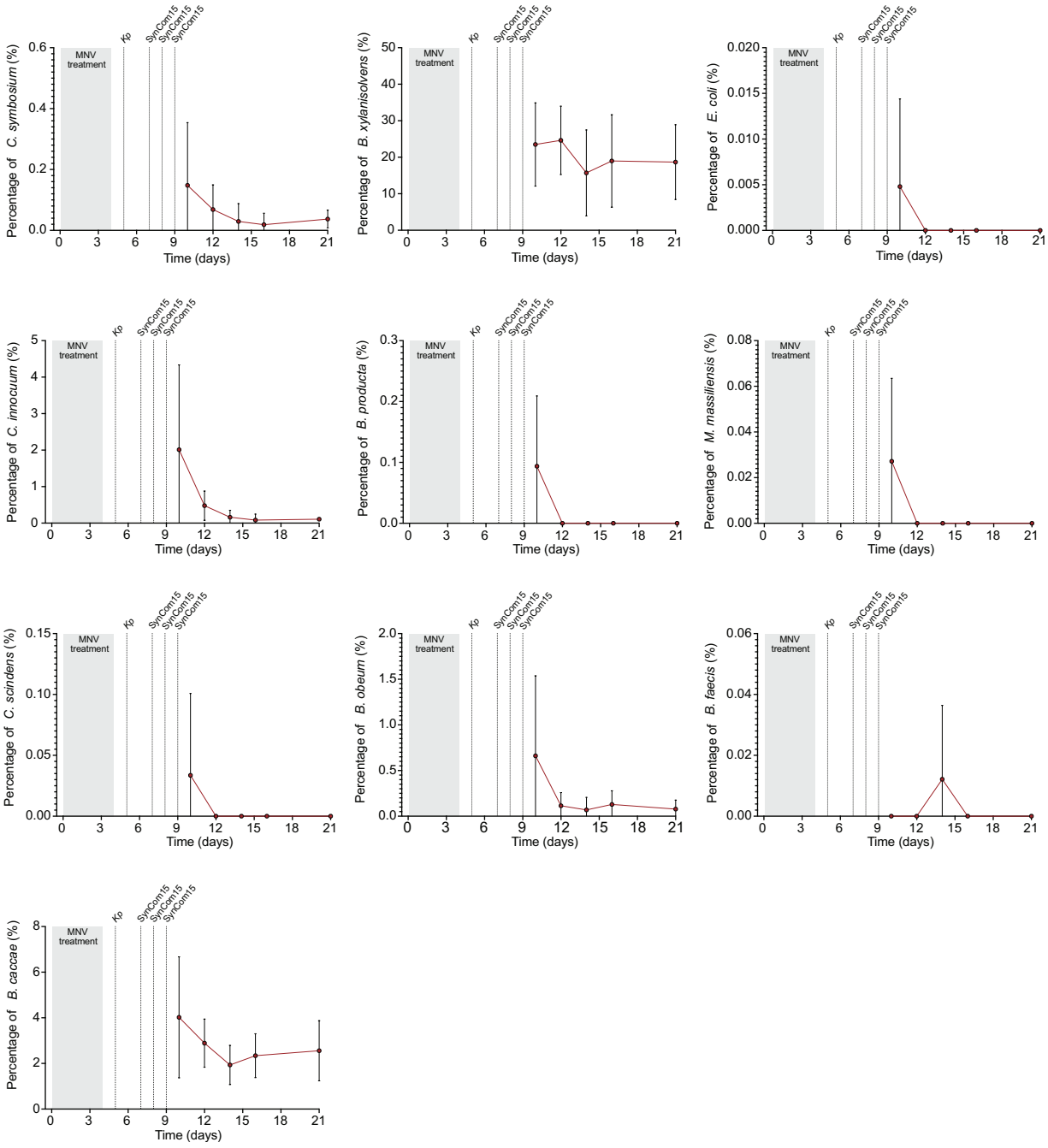


Extended Data Fig. 4. (A) Workflow for evaluating clearance capacity of DMCs for *K. pneumoniae* ('*Kp*') *in vitro* ('BHIS' is Brain Heart Infused media supplemented with cysteine, 'kan' is kanamycin). **(B)** Timecourse of *K. pneumoniae* abundance (y-axis) for all 96 communities shown in **Fig. 1E**.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

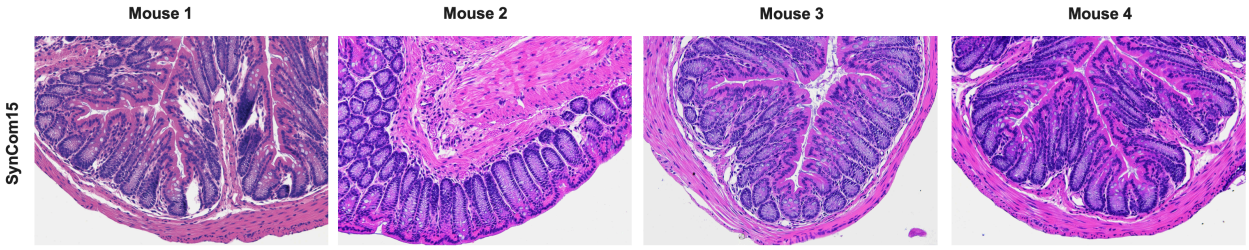


Extended Data Fig. 5. (A) Feature importance score for each strain of SynCom15 resulting from RF model built on strain presence-absence. **(B)** Feature importance scores (x-axis) for each strain (dots) versus the projection of each strain onto PC46 (y-axis) of matrix defined in **Fig. 2B**. Projection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

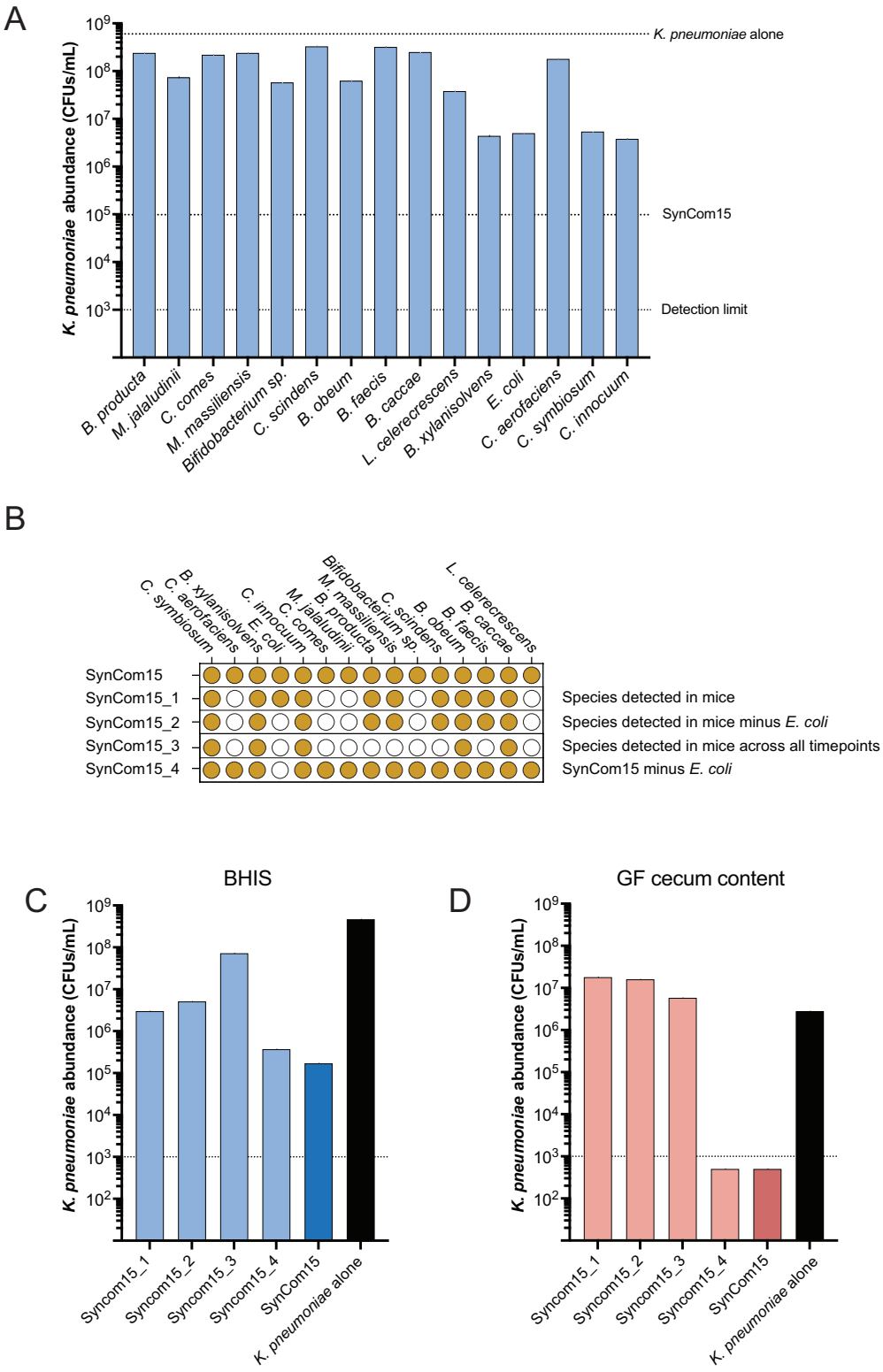
Oliveira et al., Extended Data Figure 6



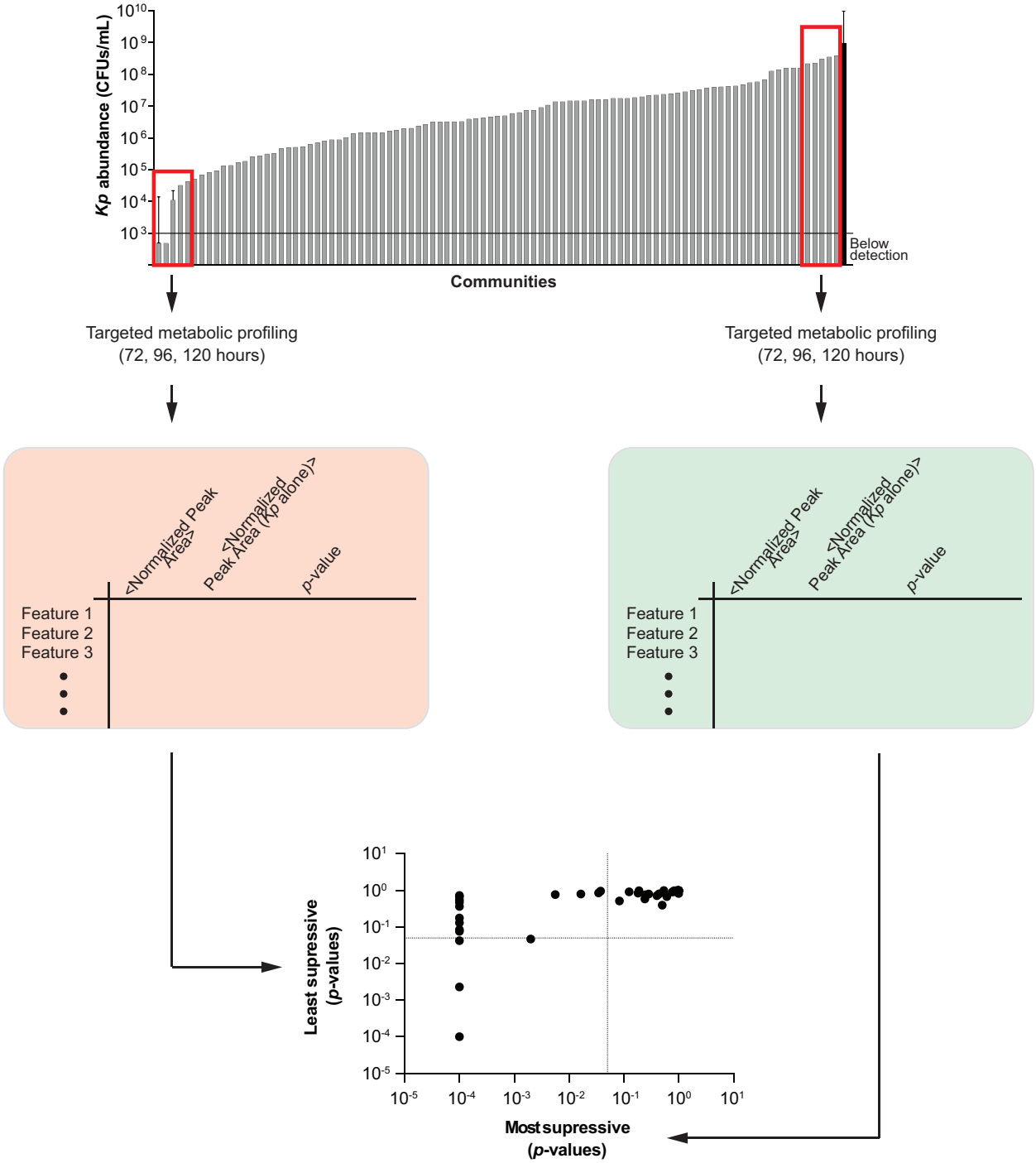
Extended Data Fig. 6. Dynamics of SynCom15 strains (y-axes) through time (x-axes) after serial triple gavage of SynCom15 in SPF mice pre-treated with broad spectrum antibiotics ('MNV treatment') and then infected with *K. pneumoniae* MH258 ('*Kp*'). Error bars represent +/- 1 standard deviation across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



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

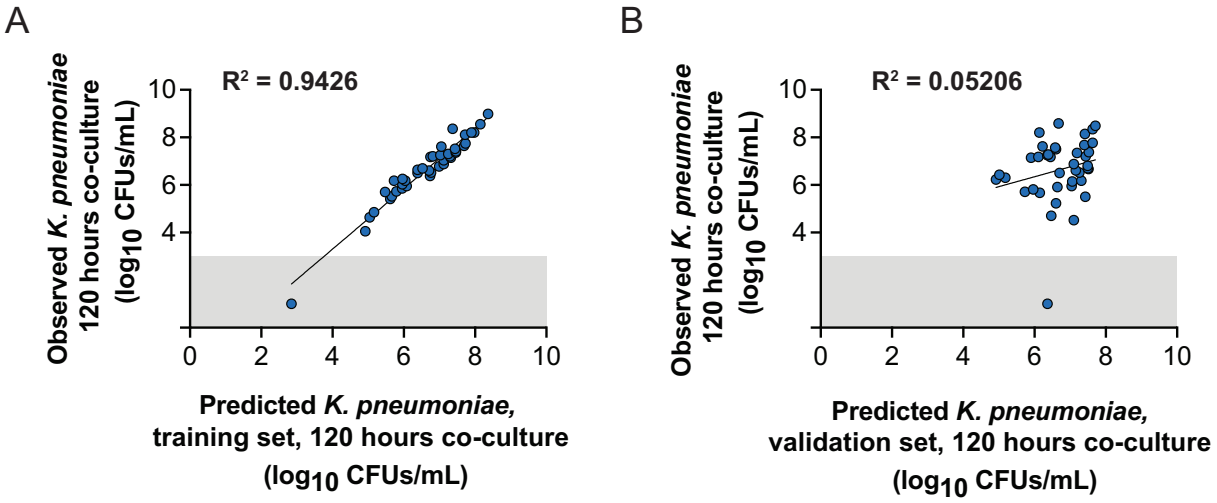


Extended Data Fig. 8. (A) *K. pneumoniae* abundance (y-axis) when co-cultured with each of the SynCom15 strains (x-axis) for 120 hours in BHIS media. '*K. pneumoniae* alone' labels the *K. pneumoniae* abundance in monoculture for 120 hours in BHIS shown in **Fig. 2E**. 'SynCom15' labels the *K. pneumoniae*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

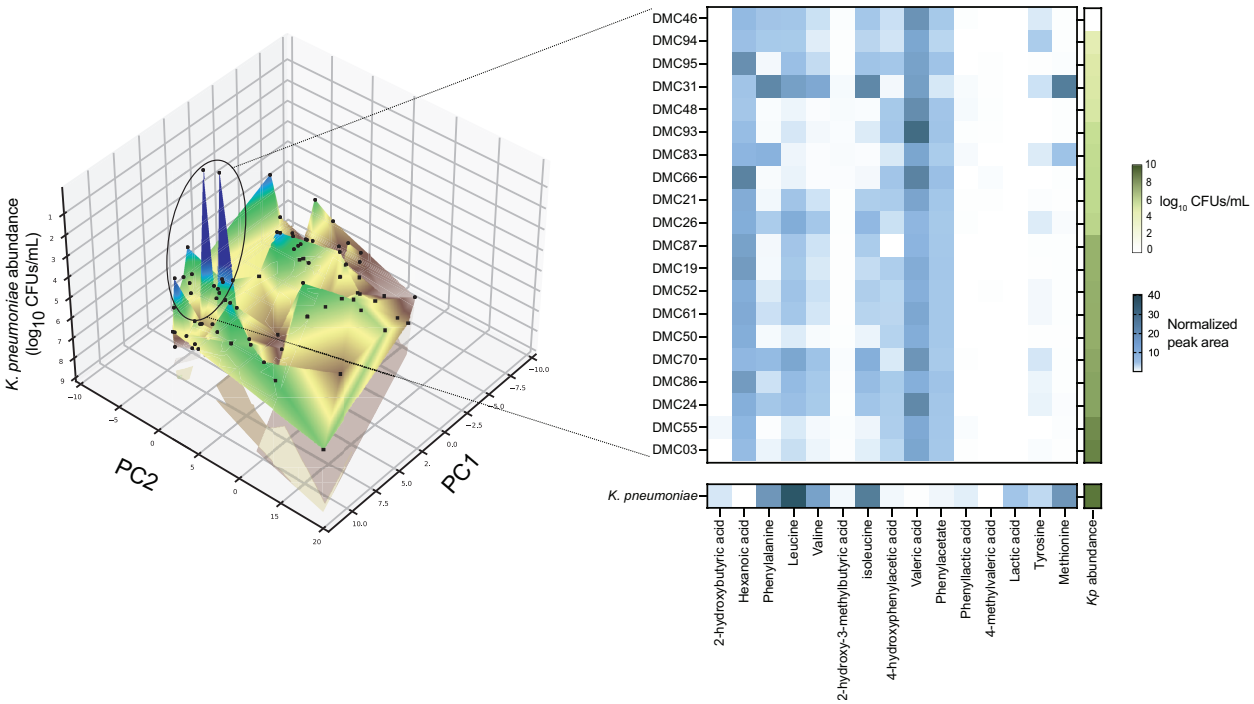


Extended Data Fig. 9. Workflow for detecting metabolite features distinguishing most and least suppressive DMCs when co-cultured with *K. pneumoniae* compared to *K. pneumoniae* in monoculture. We used the distribution of *K. pneumoniae* abundances after 120 hours of co-culture with DMCs to select the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

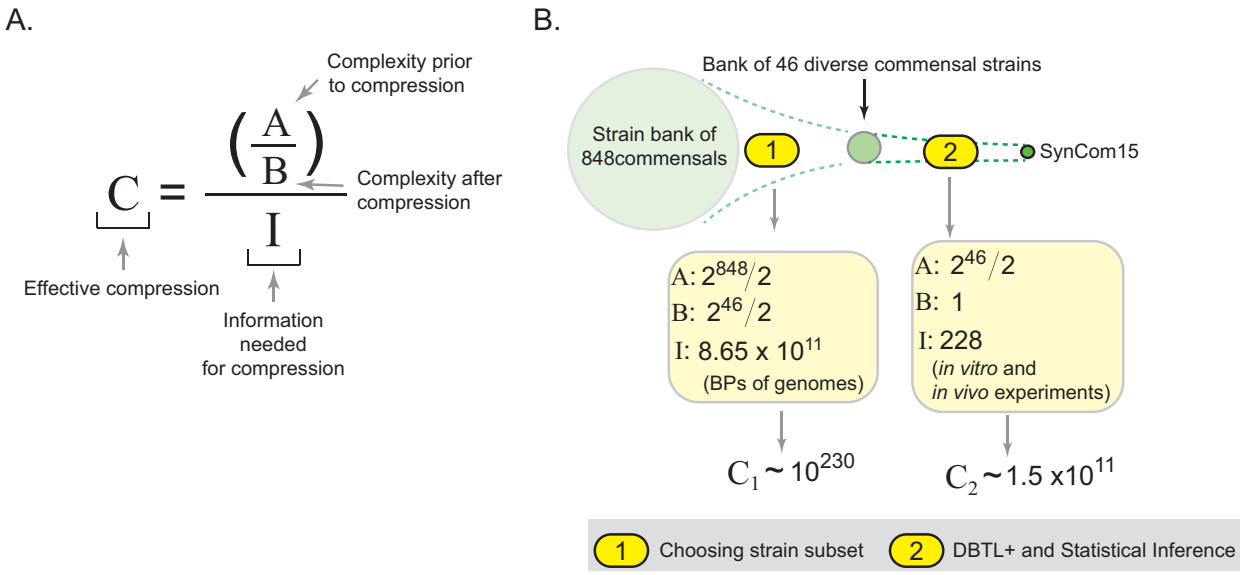
803



804
805
806 **Extended Data Fig. 10.** Predictive capacity of RF model trained on metabolite profiles of DMCs
807 for the training set of data (panel A) and the validation set of data (panel B).
808



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.

Methods

Creation and whole genome sequencing of strain bank

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

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

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

any read that was less than 50 bp long after trimming and quality control were discarded. The remaining high-quality reads were assembled into contigs using SPAdes (v3.15.4)[58].

Taxonomic classification of the assembled contigs was performed with the following methods: (a) Kraken2 (v2.1.2; (b) full/partial length 16S rRNA gene from each isolated colony's assembled contigs is extracted and input into BLASTn (v2.10.1+) to query against NCBI's RNA RefSeq database[59–61]. Top five hits for each query are manually curated to determine an isolate's identity, with identity and coverage cutoff both at 95%; (c) GTDB-Tk (v1.5.1)[62]. The final taxonomy is determined by the consensus of the three methods. Any colony that did not match initial pyro-sequencing taxonomy or lacked consensus was excluded from the commensal strain bank.

Construction of tree of bacterial genera across fecal microbiomes of healthy donors

From the metagenomic sequencing data of the fecal samples collected across healthy donors, bacterial genera present were identified by Metaphlan4[63]. Names were then extracted and cross-referenced with NCBI taxonomy using the taxize application in R[64]. The resulting tree was constructed based on NCBI taxonomic classification.

*Construction of UMAP plot shown in **Fig. 1C**.*

All gut commensal strains were annotated by their Prokka annotations and an alignment was created (848 rows comprising commensal strains, 150181 columns comprising Prokka annotated features). Each entry in the alignment is a '1' or a '0' indicating the presence or absence of a specific feature in a particular bacterial proteome.

Shotgun metagenomics of fecal samples from healthy human donors

Procedure for acquiring metagenomic data from fecal samples of healthy donors followed the same protocol as that described by Odenwald *et al*[65].

Design strategy for bacterial communities

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

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

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

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

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

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

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

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

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

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

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

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

This process is outlined for a community comprised of three strains in **Extended Data Fig. 3A**.

Creating the Klebsiella pneumoniae MH258 strain used in experiments

The *K. pneumoniae*-MH258 isolate was previously described elsewhere[31]. For better in vitro and in vivo selection of this strain, *K. pneumoniae* -MH258 was transformed by electroporation with pmCherry-sfGFP (86441; addgene).

Experimental workflow for Kp clearance assay

The 46 bacterial strains described in **Supplementary Table 1B** were individually inoculated from a frozen stock into 900μL of BHI supplemented with cysteine 0.1% (BHIS) previously reduced. Strains were incubated at 37°C in static conditions for 48h in anaerobiosis to ensure that the most fastidious strains reach stationary phase. *K. pneumoniae*-MH258 sGFP was also inoculated in the same conditions, but only 24h after commensal isolates inoculation due to the fast growth capacity of this species and was incubated for 24h. All strain densities were assessed by taking 100 μL of each culture and measuring OD_{600} in a Biotek Cytation 5. To build all DMCs, isolates were inoculated in 900 μL of BHIS previously reduced in different combinations with an initial OD_{600} of 0.001, so that the densest community reaches a maximum total initial OD_{600} of approximately 0.05. *K. pneumoniae* was added at the same initial OD_{600} of 0.001 to all DMCs. Cultures were incubated at 37°C in static conditions and anaerobiosis for 5 days. To *assess K.* pneumoniae abundance, 10 μL of each culture were collected daily and homogenized in 90 μL of PBS and serially diluted. Diluted samples were plated in BHIS with kanamycin (50μg/mL). Plates were incubated at 37°C overnight in aerobiosis. GFP expressing *K. pneumoniae-MH258* colony forming units (CFUs) were enumerated. In parallel, 100uL of each culture was also collected to

recover the cell phase and the supernatants at 72h, 96h, and 120h. These samples were stored at -80°C to be later processed for shotgun metagenomics and metabolomics.

Training and validation of Random Forest (RF) model

We used a RandomForestRegressor, available with scikit-learn python package[33]. Tree Depth was set to 12 levels per tree, the number of trees was set to 100, and the maximum number of features was set to "sqrt" (square-root of the number of strains total). Out-of-bag error was measured by a combination of R^2 (where numbers less than 1 indicate more error) and Mean Squared Error (where larger numbers indicate more error). To train and validate our model, we randomly split our dataset into 90% training and 10% true-out-of-sample 100 times. The input data was a vector of 46 1's and 0's as shown in the matrix displayed in **Fig. 1E** corresponding to the pattern of presence-absence for each DMC. In each iteration, the RandomForestRegressor was fit to the training set via 6-fold cross validation. Cross-validation accuracy was measured through Pearson Correlation. The true out of sample set was then predicted, and prediction accuracy was measured by computing Mean Squared Error and Pearson Correlation of the predicted versus measured *K. pneumoniae* abundances after 120 hours of co-culture with the DMC. Feature Importance Scores for all features were observed and stored. This process was repeated 100 times, and prediction accuracies and feature significance scores were averaged. An additional RandomForestRegressor model was then trained on the entirety of the dataset with 6-fold cross-validation. Cross-validation accuracy was measured by calculating Mean Squared Error, Pearson Correlation, and R^2. Averaged prediction accuracies and feature significance scores were used to estimate prediction error.

*Statistical analysis of matrix in **Fig. 2B***

The matrix in **Fig. 2B** was subject to PCA resulting in 46 principal components of data-variance (eigenvectors). We found that the first principal component (PC1) was significantly

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

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

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

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

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

where ε_i is the degree of *K. pneumoniae* clearance of DMC i after removing linearly modeled information related to the size of the DMC. All principal components were regressed against $\boldsymbol{r}_i$ and principal component 46 (PC46) was found to be the most significantly associated with predicted, residualized *K. pneumoniae* clearance (**Supplementary Table 6D**).

*Defining the matrix in **Fig. 2D***

Let $\boldsymbol{u} \in \mathbb{R}^{46}$ be the vector of column projections of each strain on PC46 of the matrix defined in **Fig. 2B**.

Let s be the scalar value denoting the maximum value of $\boldsymbol{u}$

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

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

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

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

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

The resulting symmetric similarity matrix, $S_{i,j}$, with rows and columns indicating each strain and each element representing the similarity between strain i and strain j describes how strains are related to one another based on their projections along PC46. Hierarchical clustering on the resulting similarity matrix was then performed to identify groups of strains. Strains that are more similar are often found in communities that suppress *K. pneumoniae* and those that are more distant are rarely found in communities that suppress *K. pneumoniae*.

*Characterization of mice used for all experiments spanning **Fig. 2** and **Fig. 3**.*

All mouse experiments were performed in accordance with and approved by the Institutional Animal Care and Use Committee of the University of Chicago under protocol 72599. Male specific-pathogen-free C57BL/6J mice, aged 8 weeks to 10 weeks, from Jackson Laboratories were used for all experiments. Mice were kept within a facility that maintained a 12 hour light and 12 hour dark cycle and controlled humidity (30–70%) and temperature (68–79 °F). Mice were housed in sterile, autoclaved cages with irradiated feed (LabDiets 5K67) and acidified, autoclaved water upon arriving at the on-site mouse facility. Mouse handling and cage changes were performed by investigators wearing sterile gowns, masks and gloves in a sterile biosafety hood. Mice were cohoused with their original shipment group until starting the experiment.

For germ-free (GF) studies, 8–10-week-old wild-type male C57BL/6J mice were used for all studies. Mice were initially obtained from The Jackson Laboratory and subsequently bred and raised in a GF isolator. After removal from the GF isolator, mice were handled in a sterile manner and individually housed in sealed negative pressure bio-containment unit isolators. Throughout breeding, mice were housed within the University of Chicago Gnotobiotic Research Animal

Facility (GRAF) and maintained at a 12 hour light and 12 hour dark cycle and controlled humidity (30–70%) and temperature (68–79 °F). Gnotobiotic mice were fed an ad libitum diet of autoclaved Teklad Global 18% Protein Rodent Diet (Sterilizable) (2018S/2018SC).

Creating GF and antibiotic (Ab)-SPF cecal extract media

To create GF cecal extract media, 8–10-week-old wild-type male C57BL/6J GF mice were euthanized and cecal contents were collected, weighted, and homogenized in 10mL of sterile distilled water on a of per gram of content. Cecal suspension was centrifuged, and supernatants were filtered through a 0.22 mm filter. GF cecal extract media was stored at -80°C.

To create ab-SPF cecal extract media C57BL/6J SPF male mice at 8-10 weeks of age were singly housed and placed under an antibiotic regime (0.25g MNV – metronidazole, neomycin, vancomycin) in the drinking water (day 0). Four days later, antibiotic treatment was halted and mice were placed on normal acidified water (day 4). Cages and food were also changed. On day 7 were euthanized and cecal contents were collected, weighted, and homogenized in 10mL of sterile distilled water on a of per gram of content. Cecal suspension was centrifuged, and supernatants were filtered through a 0.22 mm filter. Ab-SPF cecal extract media was stored at -80°C.

K. pneumoniae clearance in cecal extract media

DMCs capacity to inhibit *Kp* was tested by individually inoculated the 46 isolates from a frozen stock into 900μL of BHIS previously reduced. Strains were incubated at 37°C in static conditions for 48h in anaerobiosis. *Kp* was also inoculated in the same conditions, but only 24h after commensal isolates inoculation, and was incubated for 24h. All isolates density were assessed by taking 100 μL of each culture and measuring OD_{600} in Biotek Cytation 5. To build all defined bacterial consortia, isolates were inoculated in 900 μL of either GF or Ab-SPF cecal

extract media previously reduced in different combinations with an initial OD_{600} of 0.001, so that the densest community reaches a maximum total initial OD_{600} of approximately 0.05. To all defined communities, *K. pneumoniae* was added at the same initial OD_{600} of 0.001. Cultures were incubated at 37°C in static conditions and anaerobiosis for 5 days. To assess for *K. pneumoniae* levels 10 μL of each culture were collected daily and homogenized in 90 μL of PBS and serially diluted. Diluted samples were plated in BHIS with kanamycin (50μg/mL). Plates were incubated at 37°C overnight in aerobiosis. GFP expressing *K. pneumoniae* CFUs were enumerated. In parallel, 100uL of each culture was also collected to recover the cell phase and the supernatants at 72h, 96h, and 120h. These samples were stored at -80°C to be later processed for shotgun metagenomics and metabolomics.

Preparation of mice stool samples for fecal microbiota transplant (FMT)

Fecal samples from 15-20 mice SPF mice from different cages (to increase sample diversity) were collected to a 50 mL tube. Samples were transferred immediately to the anaerobic chamber (anaerobic exposure was kept under 30 min). Samples were dissolved in 1 mL of PBS 20% glycerol 0.1% cysteine (previously filtered and reduced) per fecal pellet (1mL per ~20 mg of fecal sample) using a mechanical pestle and vortexing. Samples were aliquoted in cryovials and stored -80°C until use.

SPF mouse model of K. pneumoniae infection

C57BL/6J male at 8-10 weeks of age were singly housed and placed under an antibiotic regime (0.25g MNV – metronidazole, neomycin, vancomycin) in the drinking water (day 0). Four days later, antibiotic treatment was halted and mice were placed on normal acidified water (day 4). Cages and food were also changed. On day 5 all mice were gavaged with 100μL of PBS containing 500 CFUs of *K. pneumoniae*, prepared as previously explained. On days 7, 8, and 9

mice were gavaged with 100uL of either selected defined bacterial consortia, a fecal microbiota transplant from naïve healthy mice, or PBS. Fecal samples were collected on days 0, 4, 7, 10, 12, 14, 16, and 21 (final day of the experiment) for 16s rRNA sequencing and on day 10 and 12 for metabolomics. These were immediately place on dry ice after collection and later stored at -80°C. To assess for *K. pneumoniae* levels, fecal samples were collected on days 7, 10, 12, 14, 16, and 21. Fecal samples were homogenized in 1mL of PBS and serially diluted. Undiluted and diluted samples were plated in BHIS and kanamycin (50μg/mL).

Determining engraftment of SynCom15 strains in SPF mice

To determine SynCom15 strain engraftment, 16s rRNA sequences from all 15 strains were blasted against 16S rRNA sequences derived from fecal samples of antibiotic-treated SPF mice gavaged with SynCom15 consortium. Fecal-derived sequences were assigned to a SynCom15 strain if their 16s rRNA percentage sequence identity was 100% with a minimum of a 95% coverage.

Determining structure of microbiota in infected SPF mice given saline, FMT, or SynCom15

DNA was extracted using the QIAamp PowerFecal Pro DNA kit (Qiagen). Before extraction, samples were subjected to mechanical disruption using a bead beating method. Briefly, samples were suspended in a bead tube (Qiagen) along with lysis buffer and loaded on a bead mill homogenizer (Fisherbrand). Samples were then centrifuged, and supernatant was resuspended in a reagent that effectively removed inhibitors. DNA was then purified routinely using a spin column filter membrane and quantified using Qubit.

16S sequencing was performed for murine studies, where V4–V5 region within 16S rRNA gene was amplified using universal bacterial primers—563F (5′-nnnnnnnn-NNNNNNNNNNNN-AYTGGGYDTAAA-GNG-3′) and 926R (5′-nnnnnnnn-NNNNNNNNNNNN-CCGTCAATTYHT-

TTRAGT-3′), where 'N' represents the barcodes and 'n' are additional nucleotides added to offset primer sequencing. Approximately 412-bp region amplicons were then purified using a spin column-based method (Minelute, Qiagen), quantified and pooled at equimolar concentrations. Illumina sequencing-compatible Unique Dual Index adapters were ligated onto the pools using the QIAseq 1-step amplicon library kit (Qiagen). Library quality control was performed using Qubit and TapeStation and sequenced on Illumina MiSeq platform to generate 2 × 250 bp reads.

Raw V4–V5 16S rRNA gene sequence data were demultiplexed and processed through the dada2 pipeline (v1.18.0) into amplicon sequence variants (ASVs) with minor modifications in R (v4.0.3)[66]. Specifically, reads were first trimmed at 190 bp for both forward and reverse reads to remove low-quality nucleotides. Chimeras were detected and removed using the default consensus method in the dada2 pipeline. Then, ASVs with length between 320 bp and 365 bp were kept and deemed as high-quality ASVs. Taxonomy of the resultant ASVs was assigned to the genus level using the RDP Classifier (v2.13) with a minimum bootstrap confidence score of 80[67].

Comparison of SynCom15 with microbiotas of healthy human donors

To investigate the presence of SynCom15 strains in samples from healthy human donors, SynCom15 strains taxonomic names were searched in the 22 fecal samples obtained from the DFI 22 human donors. For SynCom15 strain unclassified to species level *Bifidobacterium* sp., the most closely related species annotated by GTDB with an 98.21% ANI (*Bifidobacterium pseudocatenulatum*) was used[62,68].

Metabolic profiling of designed communities

For metabolite extraction from liquid cultures, samples were incubated at −80 °C between 1 h and 12 h. Four volumes of methanol spiked with internal standards were added to each culture supernatant. Samples were then centrifuged at −10 °C and 20,000 × g for 15 min followed by the

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

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

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

Metabolites were also analyzed using GC-MS with electron impact ionization. The metabolite extract (100 µL) mass spec autosampler vials (Microliter; 09-1200) and dried down completely under nitrogen stream at 30 L/min (top) 1 L/min (bottom) at 30 °C (Biotage SPE Dry 96 Dual; 3579M). To dried samples, 50 µL of freshly prepared 20 mg/mL methoxyamine (Sigma; 226904) in pyridine (Sigma; 270970) was added and incubated in a thermomixer C (Eppendorf) for 90 min at 30 °C and 1400 rpm. After samples are cooled to room temperature, 80 µL of derivatizing reagent (BSTFA + 1% TMCS; Sigma; B-023) and 70 µL of ethyl acetate (Sigma; 439169) were added and samples were incubated in a thermomixer at 70 °C for 1 hour and 1400 rpm. Samples were cooled to RT and 400 µL of Ethyl Acetate was added to dilute samples. Turbid samples were transferred to microcentrifuge tubes and centrifuged at 4 °C, 20,000 x g for 15 min. Supernatants were then added to mass spec vials for GCMS analysis. Samples were analyzed using a GC-MS (Agilent 7890A GC system, Agilent 5975C MS detector) operating in electron impact ionization mode, using a HP-5MSUI column (30 m x 0.25 mm, 0.25 µm; Agilent Technologies 19091S- 433UI) and 1 µL injection. Oven ramp parameters: 1 min hold at 60 °C, 16 °C per min up to 300 °C with a 7 min hold at 300 °C. Inlet temperature was 280 °C and transfer line was 300 °C.

Data analysis was performed using MassHunter Quantitative Analysis software (version B.10, Agilent Technologies) and confirmed by comparison to authentic standards. Normalized peak areas were calculated by dividing raw peak areas of targeted analytes by averaged raw peak areas of internal standards.

Training an RF model on metabolic content

First, Z-scores of all metabolites were centered and normalized. This was done by subtracting the mean Z score from the observed Z score and dividing it by the standard deviation of Z scores. This normalization ensured that for each metabolite, the distribution across all

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

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

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 (8.65×10^{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.

References

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

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

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

38. Jia, J. *et al.* Conserved Covarying Gut Microbial Network in Preterm Infants and Childhood Growth During the First 5 Years of Life: A Prospective Cohort Study. *Am. J. Clin. Nutr.* **118**, 561–571 (2023).

39. Dahirel, V. *et al.* Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 11530–11535 (2011).

40. Oliveira, R. A. *et al.* Klebsiella michiganensis transmission enhances resistance to Enterobacteriaceae gut invasion by nutrition competition. *Nat Microbiol* **5**, 630–641 (2020).

41. Doran, B. A. *et al.* An evolution-based framework for describing human gut bacteria. *bioRxiv* (2023) doi:10.1101/2023.12.04.569969.

42. Kouvaris, K., Clune, J., Kounios, L., Brede, M. & Watson, R. A. How evolution learns to generalise: Using the principles of learning theory to understand the evolution of developmental organisation. *PLoS Comput. Biol.* **13**, e1005358 (2017).

43. Oliveira, R. A. & Pamer, E. G. Assembling symbiotic bacterial species into live therapeutic consortia that reconstitute microbiome functions. *Cell Host Microbe* **31**, 472–484 (2023).

44. Wang, M. *et al.* Strain dropouts reveal interactions that govern the metabolic output of the gut microbiome. *Cell* **186**, 2839-2852.e21 (2023).

45. Diener, C. & Gibbons, S. M. More is Different: Metabolic Modeling of Diverse Microbial Communities. *mSystems* **8**, e0127022 (2023).

46. Winkler, E. S. *et al.* The Intestinal Microbiome Restricts Alphavirus Infection and Dissemination through a Bile Acid-Type I IFN Signaling Axis. *Cell* **182**, 901-918.e18 (2020).

47. Hromada, S. *et al.* Negative interactions determine *Clostridioides difficile* growth in synthetic human gut communities. *Mol. Syst. Biol.* **17**, e10355 (2021).

48. Venturelli, O. S. *et al.* Deciphering microbial interactions in synthetic human gut microbiome communities. *Mol. Syst. Biol.* **14**, e8157 (2018).

49. Connors, B. M. *et al.* Control points for design of taxonomic composition in synthetic human gut communities. *Cell Systems* **14**, 1044-1058.e13 (2023).

50. Russ, W. P. *et al.* An evolution-based model for designing chorismate mutase enzymes. *Science* **369**, 440–445 (2020).

51. Yeh, A. H.-W. *et al.* De novo design of luciferases using deep learning. *Nature* **614**, 774–780 (2023).

52. Anishchenko, I. *et al.* De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).

53. Wang, J. *et al.* Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).

54. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

55. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

56. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).

57. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

58. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. *Curr. Protoc. Bioinformatics* **70**, e102 (2020).

59. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

60. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

61. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-45 (2016).

62. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).

63. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).

64. Chamberlain, S. A. & Szöcs, E. taxize: taxonomic search and retrieval in R. *F1000Res.* **2**, 191 (2013).

65. Odenwald, M. A. *et al.* Bifidobacteria metabolize lactulose to optimize gut metabolites and prevent systemic infection in patients with liver disease. *Nat Microbiol* **8**, 2033–2049 (2023).

66. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).

67. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).

68. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).

69. Haak, B. W. *et al.* Impact of gut colonization with butyrate-producing microbiota on respiratory viral infection following allo-HCT. *Blood* **131**, 2978–2986 (2018).

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

1427 Author to whom correspondence and materials request should be addressed is A.S.R.