Novel genetic engineering tools for functional alteration of mammalian gut microbiomes

Sway P. Chen

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

ABSTRACT

Novel genetic engineering tools for functional alteration of mammalian gut microbiomes

Sway P. Chen

The gut microbiome is an integral component of the human body that plays a role in many physiological processes. Dysbiosis, an imbalance of the microbiome, has been associated with disease states including inflammatory bowel disease, type II diabetes, and obesity, and moreover, contributes to the pathogenesis of these states. Understanding the functional mechanisms governing microbial ecology and microbe-host interactions is essential to understanding the microbiome's role in health and disease. However, at present, functional genetic studies of diverse natural mammalian gut microbiomes remain challenging, due to a lack of genetic tools for bacteria outside of a handful of well-studied model organisms. Altering the metagenome of a complex microbial community requires novel platform technologies for genetic engineering which can operate in a generalized fashion across many different host organisms. In this thesis, I present two novel genetic tools designed for genetic modification of bacterial communities.

The first, the Cas-Transposon platform, is a host-independent targeted genome editing tool that utilizes programmable, targeted transposases to mediate site-specific gene insertions into user-defined loci. The Himar1 transposase naturally inserts transposases into random TA dinucleotides in a genome, but when fused to the dCas9 RNA-guided, DNA-binding protein, the fusion protein Himar1-dCas9 targets transposon insertions to a single TA site. The activity of Himar1-dCas9 was characterized using *in vitro* experiments, demonstrating that site-specific transposition is dependent on guide RNA (gRNA) orientation relative to the target site and the sequence surrounding the target site, but robust to variations in DNA and protein concentration, presence of background DNA, and temperature. We additionally showed that the Cas-Transposon platform is capable of performing

site-specific transposition into a plasmid *in vivo* in *E. coli*, although further optimization of the system may be necessary to effect site-specific transposition into a genomic locus. The Himar1-dCas9 protein is the first example of a transposase that inserts transposons into locations programmable by an RNA, making it a novel tool for gene insertion and knockout in potentially any organism, without relying on DNA repair by a host cell.

Metagenomic Alteration of Gut microbiome by In situ Conjugation (MAGIC) is an approach to directly modify gut bacteria in their native habitat by harnessing naturally occurring horizontal gene transfer activity to deliver engineered DNA. Because many gut bacteria are difficult to cultivate and thus difficult to genetically manipulate in the laboratory, MAGIC uses donor bacteria, delivered directly into the gut environment, to conjugate mobile vectors bearing engineered genetic payloads. Using payloads with selectable markers, we identified organisms across 4 major phyla of gut bacteria that were amenable to genetic modification with libraries of conjugative vectors we created. Using a lab-adapted *E. coli* strain as a donor, we achieved transient expression of the engineered payload in the microbiome. We also demonstrated that engineered native gut bacteria containing conjugative vectors could be deployed back into the gut to stably recolonize and mediate secondary transfer of the payload into other microbes, potentially enabling long-term infiltration of the payload into the metagenome. The results from this study suggest that both short-term and long-term genetic alteration of the metagenome are possible by choosing different donors, and that the MAGIC platform could enable development of more diverse microbial chasses for synthetic biology applications. MAGIC could also be used to create personalized engineered probiotics for diagnostic or therapeutic applications. In Chapter 4 of this thesis, we explored the targeted use of MAGIC to genetically modify Segmented Filamentous Bacteria, a gut commensal that is important for immune regulation but recalcitrant to *in vitro* cultivation.

The Cas-Transposon and MAGIC technologies expand our capabilities in the areas of targeted genome editing and gene delivery into bacteria, respectively. Together, they form a suite of

complementary approaches to genetically engineer undomesticated gut commensal bacteria and probe the functional genetic networks in the gut microbiome, which will enhance our understanding of microbiome ecology and host-microbiome interactions. In addition, the expanded range of genetic manipulations made possible by these tools may enable production of more diverse, perhaps personalized, probiotics containing engineered functions, such as sensing disease markers or drug delivery.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ACKNOWLEDGMENTS

I have many people to thank for helping me reach this point in my career and without whose support this dissertation would not have been possible. Thank you very much to my advisor, Harris Wang, for giving me the opportunity to work on several challenging but very rewarding projects and for mentoring me to become a better, more rigorous scientist. I greatly appreciate the time you took to teach me the basics of molecular biology and bacteriology when I first joined the lab, and over the last few years I have continued to learn so much from our discussions. Thank you for encouraging my creativity in designing experiments and projects, for pushing me to always think more critically, and for helping me stay motivated through difficult times when few experiments seemed to be working. I am confident that the critical reasoning, writing, presentation, and experimental skills I have learned during my time here will serve me well for the rest of my career.

Thank you to the members of the Wang lab, for being an amazing source of scientific knowledge and ideas and for your camaraderie. I learned a lot from you, and you made the lab a fun, exciting place to work. I'll miss the inside jokes, long discussions, and happy hours. I would especially like to thank Carlotta Ronda and Vitor Cabral, who worked closely with me on the MAGIC project. I am in awe of their tenacity and dedication to science, especially when it comes to literally running 24+ hour experiments between the mouse house, flow core, and lab.

I would also like to thank the members of my thesis committee, Saeed Tavazoie, Eric Greene, Ivo Ivanov, and Sam Sternberg, who provided helpful suggestions and guidance on my research. In particular, Ivo set up the collaboration with the Honda lab on genetically engineering SFB and provided a great deal of guidance on MAGIC animal experiments, Eric played a pivotal role in the Cas-Transposon project by repeatedly encouraging me to consider in vitro assays, and Saeed helped me get started with transposon sequencing and helped me strategize how to finish the PhD in a reasonable amount of time.

I thank Ron Liem and Zaia Sivo of the Integrated Program in CMBS and Steve Reiner, Patrice Spitalnik, and Jeffrey Brandt of the MD-PhD program for supporting my education and progress through the MD and PhD programs. My research was funded by grants from NIDDK, DARPA, and the Office of Naval Research.

A big thank you goes out to my friends, old and new, near and far, who have been a great source of support, humor, and shenanigans over the past several years. I'm glad to have had you in my life, and you've made it much easier to get through tough times during the PhD.

Finally, I would like to thank my family – my parents, Ruijue Peng and Xing Chen, and my grandparents, Wangchong Peng and Zhixiang Yuan – for raising me with a love for learning and problem-solving, encouraging me to dream big, and instilling in me the value of hard work and perseverance. I would not be in this position without your love and support, and I cannot thank you enough for the sacrifices you made to give me a life full of amazing opportunities.

Dedicated to my family

Chapter 1: Introduction

Functional studies of host-microbiome interactions require novel genetic engineering tools for microbial communities

Over the past two decades, human-associated microbiomes have increasingly been appreciated as essential components of the human body that play important roles in maintaining homeostasis. As a result of foundational advances in -omics technologies, it has become easier to catalog the diverse microbes that make up our microbiomes and to explore their genomes, transcriptomes, and proteomes[1-4]. Dysbiosis, a microbiome imbalance or impairment, has been associated with a wide range of disease states, including *Clostridium difficile* infection, inflammatory bowel disease, colorectal cancer, and metabolic syndrome[5, 6]. While it has become relatively straightforward to characterize microbiomes by sequencing approaches, the more challenging problem that remains is understanding the mechanisms by which specific genes and proteins function to shape microbiome ecology, influence host physiology, and cause disease. To gain a mechanistic understanding of microbial ecology and host-microbiome interactions and to engineer native commensal microbiota in useful ways, novel genetic tools for manipulating microbial communities are needed.

In humans, the gut microbiome forms the bulk of commensal microbes and modulates host physiology in a variety of ways. The gut microbiome contains up to 100 trillion cells comprising about 1000 species of bacteria, which collectively contain millions of genes, two orders of magnitude more than the genes encoded by the human genome[7, 8]. The vast majority (90-99%) of genes in the metagenome (the total genomic content of the microbiome) are bacterial, with the remainder being of archaeal, eukaryotic, or viral origin[8, 9]. Four bacterial phyla dominate the microbiome: Firmicutes, Bacteroidetes, Actinobacteria, and Proteobacteria[10]. The composition of the gut microbiome varies greatly between individuals and is influenced by age, diet, lifestyle,

environment, and disease, but the microbiomes of individual adults remain relatively stable over time[9, 11, 12]. Despite this divergence in microbiome compositions, studies have shown that broad functional categories of genes are shared between the metagenomes of individuals[4, 11, 13]. To give one example, a study of 18 subjects found no species-level phylotypes shared between all subjects, but 93% of functional gene categories were shared among all subjects[13]. This conservation of functional gene categories suggests that the gut microbiome performs similar functions across individuals with divergent microbiome compositions, although differences between microbiomes at finer scales (i.e., specific genes) contribute to functional differences between individuals[11].

The diverse functions encoded in the genomes of gut microbes enable microbes to perform essential roles in host physiology, including metabolism, immune system regulation, pathogen defense, and neural signaling. Gut microbes break down fibers in food, producing short chain fatty acids that nourish colonocytes and activate G-protein-coupled receptors to modulate hormonal signaling[14]. Gut microbes are also necessary for hydrolyzing and biotransforming bile salts in the intestine[15]. The gut microbiome alone is sufficient to drive major changes in host metabolism; for example, gut microbiota transplants from obese human subjects or patients with kwashiorkor (a form of protein malnutrition) into germ-free mice recapitulated these phenotypes in the mice[16, 17]. Gut microbes can affect drug metabolism, an important consideration in medication management. *Eggerthella lenta* is known to inactivate digoxin[18], while the chemotherapeutic drug CPT-11 is activated by gut bacteria, producing severe diarrhea as a side effect[19]. Commensal gut bacteria modulate the development of the mammalian immune system and provide resistance to pathogen invasion[20, 21]. The gut microbiome also affects neural signaling within the enteric and central nervous systems, working through the gut-brain axis to modulate depression-, anxiety-, and stress-related behavior[22, 23].

The insight that the gut microbiome is involved in host physiology raises the questions of which microbes are responsible for particular host-microbiome interactions, and by what functional mechanisms these microbes affect the host and/or each other. Metagenomic or metatranscriptomic sequencing studies of the microbiome during a given physiological state reveal associations of that state with specific microbes that are increased in abundance or microbial genes that are upregulated, but associations do not imply causation. Changes in the microbiome could be an effect of the host physiological state, rather than the cause. To definitively identify a particular microbe or gene as being causal for a host physiological state, it necessary to alter the abundance of that microbe or gene product and show that the associated host state occurs as a consequence of the change in the microbiome.

At present, several methods are used alter gut microbiomes to interrogate their function, but all are subject to limitations of specificity, magnitude of alteration, or biological relevance. Chemicals, such as prebiotics or xenobiotics, can be introduced into the gut environment to modify the abundance of microbes or microbial products. Prebiotics are chemical substrates that selectively induce microbial growth or activity. In the context of the gut microbiome, prebiotics are usually oligosaccharides or polysaccharides that promote growth of Bifidobacteria and Lactobacilli[24]. Xenobiotics are compounds foreign to the native environment that are intended to modulate microbial function or growth. Xenobiotics include antibiotics, which are commonly used to deplete the microbiome, as well as chemicals with more targeted functions. For example, β-glucuronidase inhibitors have been used to reduce toxicity of the drug CPT-11 in mice by inactivating the bacterial β-glucuronidase enzymes that reactivate the drug[19]. A structural analog of choline inhibits trimethylamine (TMA) production by gut microbes, which prevented development of atherosclerosis in mice caused by a downstream metabolite of TMA[25]. While chemical modifiers of the microbiome are useful for manipulating microbial function, they may act

nonspecifically, affecting wide swathes of bacteria, which makes their effect on the microbiome difficult to predict.

Another approach to modifying the gut microbiome is introduction of novel bacteria into the ecosystem. Probiotics are live bacteria which confer a benefit to the host when administered in adequate amounts[26]. Well-studied probiotics include *Lactobacillus rhamnosus*, which decreases stress-induced anxiety- and depression-related behaviors in mice by modulating signals through the vagus nerve[27], *Bifidobacterium* species, which break down plant carbohydrates to produce short-chain fatty acids[28], and *E. coli* Nissle 1917, which defends against colonization by enteric pathogens[29-31]. Engineered probiotics have also been developed to deliver synthetic gene products *in vivo*; in particular, *Lactococcus lactis* has been used as a probiotic chassis for expressing anti-tumor necrosis factor (TNF)-α antibodies, IL-10, and TGF-β1 in the gut[32-34]. However, probiotics are not native to natural gut microbiomes in animals, and they face resistance to long-term colonization of the gut by the indigenous microbiota[35-37]. Any observed effects on host physiology occur transiently for the duration of probiotic administration[35, 37] and do not necessarily reflect the function of native microbes within the gut without the perturbation of probiotic administration.

Studies to interrogate host-microbiome interactions have also utilized the engraftment of commensal gut bacteria into new hosts, in order to observe the effect of the bacteria. Fecal matter transplants (FMT), which contain the entire fecal microbiome, from human donors into germ-free were used to demonstrate the microbiome's role in causing obesity[16] and kwashiorkor[17]. FMT has also been successfully used in humans and mice to treat *Clostridium difficile* infection (CDI)[38, 39]. Engraftments of specific strains of bacteria have also been used to study the function of those strains in greater detail. For example, *Clostridium scindens*, a secondary bile acid-synthesizing bacterium, is associated with resistance to CDI in both mice and humans. When mice that were normally susceptible to CDI were colonized with *C. scindens*, the mice had a reduced rate of CDI,

likely due to production of secondary bile acids by *C. scindens* that were toxic to *C. difficile*[40]. Genetically engineered mutants of the gut symbiont *Clostridium sporogenes* were engrafted into mice to demonstrate that a metabolic pathway in this bacterium produces aromatic amino acid metabolites, which accumulate in the host serum and alter host immune activation and intestinal permeability[41]. Like probiotic studies, however, commensal bacteria engraftment studies are also potentially limited by colonization resistance[42]. Engraftments are typically performed on germ-free animals, on animals colonized with a few defined bacterial strains, or with co-administration of antibiotics to suppress the native microbiota, which facilitates the colonization of an exogenous strain but limits the biological applicability of these studies to animals with robust natural microbiomes.

Bacteriophages represent another route for modifying microbiomes. Phages can be used to infect and lyse their specific host cells to selectively deplete those hosts from the population; oral administration of phages has been used to selectively eliminate enteric pathogens such as *Shigella* and *Listeria* from mice without significantly affecting the rest of the microbiome[43, 44]. Engineered phages can also be used to genetically modify their hosts through lysogeny, delivering new genetic functions into the hosts[45]. However, because phages only infect specific hosts, targeting any given bacterial strain in the gut microbiome requires identification and cultivation of associated phages. At the present time, using phages to infect any given strain in the microbiome remains challenging, though the gut phage-ome is becoming better characterized and may yield useful phages for manipulating a wide variety of bacteria in the future[46].

Resistance to colonization makes engraftment of foreign bacteria into a diverse microbiome challenging, while approaches to increase or deplete bacteria or gene products using chemicals are non-specific. Direct genetic manipulation of gut microbes in their native environment would enable better targeted functional studies of gut microbiome ecology and host-microbiome interactions. Engrafting a novel genetic function or knocking out a genetic function

from the metagenome, without relying on introducing exogenous bacteria into the microbiome, would more directly show the effect of those functions. At present, however, genetic engineering of natural microbial communities remains an open challenge. Gut bacteria are obligate or facultative anaerobes, and many are difficult to cultivate *in vitro*, limiting their genetic tractability. Most genetic tools and protocols for manipulating bacteria are highly optimized for a small handful of well-studied model organisms, and are not easily generalizable to the diverse array of mammalian gut commensals.

This thesis aims to address the challenge of engineering gut microbes within their native context by introducing two novel platform technologies for genetically engineering bacteria. The first is the Cas-Transposon system for host-independent, programmable, targeted genomic integrations, described in Chapter 2. The second is a platform technology for genetically engineering gut microbes in their native environment, Metagenomic Alteration of Gut microbiome by *In situ* Conjugation (MAGIC), described in Chapter 3. In Chapter 4, I present an application of the MAGIC technology to genetically modify Segmented Filamentous Bacteria *in situ*, as SFB is a gut commensal that remains very difficult to cultivate *in vitro*. These technologies provide solutions for the challenges of targeted genome editing in prokaryotes and delivering engineered synthetic DNA into a diverse community of microbes. By bypassing the reliance on host DNA repair machinery and *in vitro* cultivation of microbes, which limit the efficacy of many currently available genetic engineering tools, Cas-Transposons and MAGIC are complementary tools that can be broadly applied to prokaryotes. These tools have the potential to greatly expand the range of possible genetic manipulations in the gut microbiome and the set of genetically tractable microbial chasses for synthetic biology applications.

Current approaches to targeted genome editing in bacteria

Experimentally probing the functions of gene networks and rationally engineering useful new functions in a microbiome requires platform technologies for genome editing that can be applied to many diverse species. Over the years, a number of methods have been developed for targeted modification of bacterial genomes[47]. Today, commonly used methods to edit bacterial genomes take one of two approaches: using host DNA replication and/or repair machinery to incorporate exogenous dsDNA or ssDNA, or relying on one of several host-independent enzymes to modify the genome.

Host-dependent methods for targeted genome editing include nuclease-guided genome editing, recombineering, and Multiplex Automated Genome Engineering (MAGE). Targeted nucleases are used in conjunction with host DNA repair systems to incorporate specific genomic edits. The nuclease creates a double-stranded break at the target site, which is repaired by homologous recombination (HR) to incorporate a synthetic template DNA containing the desired edit. Targeted nucleases include zinc-finger nucleases (ZFNs)[48] and transcription activator-like effector nucleases (TALENs)[49], which consist of a non-specific FokI nuclease domain fused to a DNA-binding domain that recognizes a specific motif. A more versatile and programmable class of nucleases are *Cas* nucleases associated with Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) systems, which originally evolved as an immune system for bacteria and archaea[50-52]. *Cas* nucleases cleave invading foreign DNA with the help of complementary RNAs encoded by CRISPR arrays[53]. A modified CRISPR system from *Streptococcus pyogenes* is commonly used for gene editing, in which a synthetic guide RNA (gRNA) targets Cas9 nuclease to an NGG protospacer adjacent motif (PAM) sequence[54]. Since the specificity of DNA cleavage is easily programmed by small RNA guide sequences, the CRISPR-Cas system has been used for genome editing in many eukaryotic systems including yeast[55], fruit flies[56], zebrafish[57], mice[58], and humans[59, 60], where the DSB can be repaired by HR or by non-homologous end joining (NHEJ) to create indels. In bacteria, which generally do not

have NHEJ systems[61], CRISPR-Cas genome editing utilizes HR[62]. CRISPR-Cas can also be used as antimicrobial system by cutting specific DNA sequences[63, 64]. While the nuclease/HR approach to genome editing has become ubiquitous, efficiency of HR remains a limiting factor in genetically modifying bacteria.

Recombineering uses host-specific phage proteins to recombine DNA with homologous flanking ends into the genome, increasing the efficiency of HR over native bacterial DNA repair. In *E. coli*, lambda red recombineering uses lambda phage proteins Exo, Beta, and Gam to enhance recombination efficiency, but the efficiency remains relatively low, particularly with larger DNA insert sizes of several kb[65]. Bacteria are able to incorporate short, synthetic ssDNA oligonucleotides during DNA replication with higher efficiency. MAGE utilizes this insight, delivering short (30 to 100 bp) ssDNAs that containing desired modifications into the cell during cell division[66]. The oligonucleotides anneal to the lagging strand of DNA replication forks and function as synthetic Okazaki fragments that get incorporated into daughter cells. Multiple genome edits can be incorporated simultaneously. Similarly to recombineering, the efficiency of oligo incorporation is enhanced by the lambda-red proteins. While recombineering and MAGE have been adapted to function in organisms other than *E. coli*, this process requires searching for phage proteins specific to each organism with the same functions as the lambda phage proteins[67] and laborious optimization for each organism, limiting the utility of these methods for genome editing across diverse species.

Host-independent approaches to genome editing may be expected to function more robustly across different host backgrounds and include recombinases, group II introns, and base editors. Recombinases such as Flp[68] and Cre[69] mediate recombination at defined recognition sequences to efficiently integrate large pieces of heterologous DNA, but have limited programmability since their DNA-recognition domains are challenging to retarget to other DNA motifs[70, 71]. Group II introns are retrotransposons that use an RNA intermediate to complement an

insertion site; the RNA is reverse-transcribed by the intron-encoded protein for incorporation into the DNA[72]. Group II introns function host-independently and have been used to make genomic insertions in a variety of Gram-negative and Gram-positive organisms[73, 74]. However, not all genes may contain suitable target sites for insertion, and insertions are limited in size to below 2kb[47, 74]. Recently, base editors, proteins consisting of a nucleotide deaminase fused to an RNA-guided, catalytically dead dCas9 protein, have been developed to make single-base substitutions at gRNA-targeted loci[75, 76]. This approach may be powerful going forward as a host-independent method for making controlled SNPs in bacteria and eukaryotes.

Presently, there is a lack of programmable host-independent methods for making targeted large insertions into bacterial genomes. To address this problem, we developed the Cas-Transposon platform, which uses dCas9/gRNA-guided Himar1 transposases to catalyze targeted transposition into the genome. Himar1 is a member of the mariner family of transposases originating from the horn fly *Haematobia irritans*[77] and naturally inserts multi-kb transposons randomly into TA dinucleotides, making it a useful tool for bacterial mutagenesis[78]. By fusing Himar1 to an RNA-guided DNA-binding protein, the transposase activity is limited to the locus at which it is tethered. In Chapter 2, we demonstrate that the Himar1-dCas9 fusion protein mediates site-specific transposon insertions robustly *in vitro* and show preliminary evidence of *in vivo* function in an *E. coli* plasmid assay. While further optimization is necessary to improve the Cas-Transposon system's function *in vivo*, this work represents the first demonstration of a novel mode of targeted genome editing applicable to any organism.

Gene delivery into diverse natural microbiomes

The first step in genetically engineering a microbe is introducing DNA into the cell. Traditional techniques for bacterial transformation include electroporation, chemical

transformation, and natural competence. While over 80 species of bacteria have been documented to be genetically engineerable by these *in vitro* methods, the necessary conditions for transformation in these microbes vary widely and are laborious to optimize for each individual species[79]. These transformation techniques also require microbes that are cultivable *in vitro*, which excludes a significant fraction of fastidious gut bacteria that remain uncultivable at present[80, 81].

In nature, microbes frequently acquire and propagate new DNA via horizontal gene transfer, the movement of DNA between different organisms. Horizontal gene transfer (HGT) occurs by one of three modalities: transformation, or uptake of DNA from the environment; conjugation, where DNA is transferred between cells in direct contact by type IV secretion systems[82]; and transduction, in which foreign DNA is introduced to a cell by an infecting virus. Past horizontal gene transfer events into an organism can be inferred by analyzing its genome; segments of DNA with high sequence similarity to DNA from otherwise distantly related organisms are more likely to have originated from HGT rather than vertical evolution[83]. Metagenomic studies have shown that the human gut microbiome has very high rates of HGT compared with other naturally occurring microbiomes, such as soil or marine microbiomes[84]. HGT allows microbes to rapidly acquire novel genes that enhance their fitness in the highly competitive gut environment. These horizontally transferred genes encode functions such as carbohydrate metabolism[85, 86], formation of cell surface structures that enhance colonization and cell-cell signaling[85], antibiotic resistance[87], and virulence factors[88]. The pool of mobile genes can confer very different phenotypes to different strains of the same species[89], and thus engineering the mobilome is potentially a powerful strategy for functionally altering a microbiome.

Conjugation can result in gene transfer events across very broad ranges of species. The vectors of conjugative gene transfer include plasmids, closed circles of DNA that self-replicate in a subset of bacteria, and integrative conjugative elements (ICEs), which propagate by stably

integrating into the recipient organism's genome. Conjugative plasmids have an essential backbone consisting of genes involved in self-replication and an origin of replication (oriR), as well as an origin of transfer (oriT) that binds to conjugative transfer proteins, which nick the plasmid and transfer it through a pilus as a single stranded DNA[90]. Plasmids may be self-transmissible, if they contain all necessary genes for conjugative transfer, or mobilizable, if they contain an oriT that can be mobilized by another conjugative element. ICEs include elements that are mobilized by integrases, which mediate recombination into a specific recognition sequence, and transposons, which integrate into random loci. Successful transfer of mobile elements between species or strains requires expression of replication and maintenance systems in the recipient, and avoidance of bacterial defense mechanisms against invading foreign DNA, such as CRISPR systems[91] and restriction/modification systems[92]. Experimental studies of the host ranges of these genetic elements, particularly plasmids, have historically focused on analyses of conjugation between two strains or species[93, 94]. Recently, advances in next-generation sequencing and high-throughput screening have enabled experimental tracking of plasmid transfer into diverse natural microbiomes, such as the soil microbiome[95]. The high frequency of conjugative HGT between distantly related bacteria observed makes conjugation a promising method of gene delivery into diverse microbiomes.

MAGIC harnesses HGT in the gut to deliver engineered gene circuits into native gut bacteria by *in situ* conjugation. We designed a set of broad host-range conjugative plasmid vectors containing a genetic payload consisting of an antibiotic resistance gene and a GFP gene and disseminated them into the gut microbiota of conventionally raised mice, using an engineered *E. coli* strain as a plasmid donor. From fecal samples, we were able to isolate and identify i*n situ* transconjugants from across 4 major phyla of gut bacteria using *in vitro* antibiotic selection and FACS enrichment of transconjugants. Though these transconjugants were only transiently present in feces, genetically modified native gut microbiota containing MAGIC vectors were

capable of long-term colonization of the gut when orally gavaged to mice and also mediated secondary HGT of the vectors into new recipients. By changing the donor strain (laboratory adapted *E. coli* vs. an individual host-adapted commensal strain) and vector parameters (origins of replication, regulatory elements), the MAGIC platform can be adjusted to actuate different scales of metagenome modification, in terms of time duration and range of transconjugants. This work is described in Chapter 3.

MAGIC potentially enables genetic engineering in organisms that we are unable to cultivate *in vitro*. Segmented Filamentous Bacteria (SFB) is difficult to cultivate, but is a key player in regulating intestinal mucosal immunity[21, 96] and has not yet been genetically manipulated. In Chapter 4, we explore the possibility of applying MAGIC to genetically modify SFB without *in vitro* cultivation. Here, we explain the design principles involved in designing genetic vectors for an organism with no prior genetic tools and show data from preliminary experiments.

Impact of novel genetic engineering tools

The overarching motivation for development of the Cas-Transposon and MAGIC technologies is to enable a wider range of genetic manipulations in more diverse bacteria, in order to genetically manipulate naturally occurring microbiomes. The ability to perform make specific modifications to the metagenome would be powerful for functional genetic screens of the microbiome as a unit. Additionally, the tools described here would facilitate engineering of native gut microbes as probiotics that perform useful functions, such as sensing inflammation[97, 98] or small molecules[99] and delivering therapeutic molecules[100] into the gut. To date, only a small handful of bacterial strains (*E. coli*, *Salmonella typhimurium*, several *Lactobacillus* strains, and several *Bacteroides* strains) have been genetically engineered as chasses for these purposes, but MAGIC could enable delivery of synthetic gene circuits into a wider range of microbes,

possibly personalized for individuals in order to actuate a longer-term effect. These improvements would advance our functional understanding of the gut microbiome as it relates to health and physiology and provide novel solutions for diagnosing and treating diseases of the microbiome. The technologies described here would also be broadly applicable to studying and manipulating other naturally occurring microbiomes as well.

Chapter 2: An engineered Cas-Transposon system enables programmable, precise insertions of large DNA constructs into user-defined genetic loci

Sway P. Chen[1,2], Harris H. Wang[1,4,#]

1. Department of Systems Biology, Columbia University Medical Center, New York, USA
2. Integrated Program in Cellular, Molecular and Biomedical Studies, Columbia University Medical Center, New York, USA
3. Department of Pathology and Cell Biology, Columbia University Medical Center, New York, USA

[#] Correspondences should be addressed to H.H.W. (hw2429@columbia.edu).

2.1 ABSTRACT

Efficient targeted insertion of DNA into a genome is a major challenge in genome editing. While CRISPR-Cas9 genome editing has become widely used in diverse organisms, reliance on the host organism's DNA repair machinery to incorporate DNA limits the efficiency of insertion in a size-dependent manner. Recombinases, which are commonly used to insert operon-sized constructs into genomes, are restricted to existing recognition sites in the genome and are difficult to target to other loci. Here, we describe the Cas-Transposon platform for genome editing, in which hyperactive Himar transposases fused to RNA-guided dCas9 proteins mediate programmable, site-specific transposon insertions. Himar transposases naturally cut and paste multi-kb DNA transposons into new, random TA dinucleotides; when the Himar transposase is fused to dCas9, which binds DNA complementary to a guide RNA (gRNA), the transposase becomes tethered to a target site specified by the gRNA and increases the rate of transposition into a nearby TA insertion site. Using *in vitro* assays, we demonstrated that Himar-dCas9 fusion proteins enriched transposon insertions at a single targeted TA site over 300-fold compared to a random transposase, and that site-specific transposition is dependent on target choice but robust to log-fold variations in protein, donor DNA, target DNA, and background DNA concentrations. We validated that Himar-dCas9 was also capable of mediating site-specific transposition into a target plasmid in *E. coli*, but did not find evidence of site-specific transposition into a targeted genomic locus in mammalian cells. While further optimization of the Cas-Transposon system is necessary to improve performance for genome editing, we have engineered the first synthetic transposase capable of host-independent, programmable targeted insertions, which represents a novel mechanism for DNA insertion into a targeted location.

2.2 INTRODUCTION

Genome engineering relies on molecular tools that enable targeted and specific modification of a genome to introduce insertions, deletions, and substitutions. While numerous advances have emerged over the last decade to produce programmable editing and deletion of bacterial and eukaryotic genomes[47], targeted genomic insertion has been a long outstanding challenge. Integration of desired heterologous DNA into the genome needs to be precise, programmable, and efficient; three key parameters of any genome integration methodology. Currently available genome integration tools are limited by one or more of these factors. Recombinases such as Flp[68] and Cre[69] that mediate recombination at defined recognition sequences to integrate heterologous DNA have limited programmability, since their DNA-recognition domains are challenging to retarget to other DNA motifs[70, 71]. Site-specific nucleases, such as Cas9[53, 59], zinc finger nucleases (ZFNs)[48], and transcription activator-like effector nucleases (TALENs)[49], can be programmed to generate double-strand DNA breaks that are then repaired to incorporate a template DNA, but this process relies on host homology-directed repair machinery, which are variable and often inefficient, especially as the size of the DNA insertion increases.

Transposable elements are widespread natural selfish genetic systems capable of integrating large pieces of DNA into both prokaryotic and eukaryotic genomes. Amongst various transposable elements described[101, 102], the Himar1 transposon from the horn fly *Haematobia irritans*[77] has been coopted as a popular tool for insertional mutagenesis. The Himar1 transposon is mobilized by the Himar1 transposase, which like other Tc1/mariner-family transposases, functions as a homodimer to bind the transposon DNA at the flanking inverted repeat sequences, excise the transposon, and paste it into a random TA dinucleotide on a target DNA[77, 103-105]. Himar1 transposition requires no host factors for transposition and functions *in vitro*[77], in bacteria[78], and in mammalian cells[106], and is capable of inserting transposons over 7kb in size[107]. A hyperactive mutant of the transposase, Himar1C9, which contains two amino acid substitutions and increases

transposition efficiency by 50-fold[108], has enabled the generation of genome-wide transposon insertion mutant libraries for genetic screens in diverse microbes[109-111]. However, since Himar1 transposons are inserted randomly into TA sites, their utility in targeted genome insertion applications has thus far been limited.

One approach to potentially increase the specificity of otherwise random transposon insertions is to increase affinity of the transposase to specific DNA motifs. Indeed previous studies have described fusing transposases to a DNA-binding protein (DBP) domain to increase targeting of transposon insertions to specific genetic loci. Fusing the Gal4 DNA-binding protein to Mos1 (a Tc1/mariner family member) and piggyBac transposases altered the distribution of integration sites[112] to near Gal4 recognition sites in plasmid-based assays in mosquito embryos. Fusion of DNA-binding zinc-finger or transcription activator-like effector proteins to piggyBac enabled integration into specified endogenous genomic loci in human cells[113-115]. ISY100 transposase (also a Tc1/mariner family member) has also been fused to a Zif268 Zinc-finger domain to increase specificity of transposon insertions to DNA adjacent to Zif268 binding sites[116]. While these studies demonstrated the feasibility of the transposase-DBP fusion approach for targeted genome insertion, the DBP used thus far are still challenging to reprogram and to limit off-targeting effects.

More recently, protein fusions using a catalytically inactive Cas9 nuclease (dCas9) as an RNA-guided, DNA-binding protein have enabled manipulation of genomic DNA and gene expression near user-defined loci. dCas9 targeted to a particular gene by a synthetic guide RNA (gRNA) blocks transcription by steric hindrance of RNA polymerase[117], while dCas9-transcription activator fusions targeted to a promoter enhance gene expression[118]. The FokI-dCas9 fusion nuclease is a directed nuclease that functions as a dimer, and thus provides increased site-specificity over the Cas9 nuclease[119, 120]. Base editors, fusion proteins consisting of a deaminase and dCas9, enable single-base substitutions at gRNA-targeted loci[75, 76]. dCas9 has also been tethered to the Gin serine recombinase to facilitate increased recombination specificity at a

targeted recombinase recognition site in the human genome[121]. At this time however, no dCas9-transposase fusions have been demonstrated for dCas9-targeted transposition into a genomic locus; a recent study by Luo et al. showed that a dCas9-piggyBac transposase protein did not have targeted activity[115].

In this study, we developed a novel platform technology, Cas-transposons, that unites the DNA integration capability of the Himar1 transposase and the programmable genome targeting capability of dCas9 to enable site-specific transposon insertions at user-defined loci in an easily generalizable fashion. The system uses Himar1-dCas9 fusion proteins, tethered by a gRNA to a specific locus, to integrate transposons carrying synthetic genetic payloads into that locus (**Figure 2.1a**). We demonstrated targeted transposon integration by Himar1-dCas9 proteins in both cell-free *in vitro* reactions and in *E. coli*. The Cas-transposon technology can potentially be optimized to function in any prokaryotic or eukaryotic cells, because the Himar1-dCas9 protein requires no host factors to function. Cas-transposons may be a useful tool for various applications such as metabolic pathway engineering[122] (e.g., simultaneous integration of operon-size pieces of DNA and knockout of an endogenous gene) or emergent gene drive technologies[123] (e.g., targeted propagation of engineered traits across populations).

2.3 RESULTS

Design of a targeted transposase protein

We designed proteins to mediate programmable, site-specific transposition using key insights from previous studies on Himar1 transposases and dCas9 fusion proteins. The dCas9 protein from *Streptococcus pyogenes*, which is the Cas9 nuclease with 2 amino acid substitutions in cleavage catalysis sites, has been well-characterized and used as the RNA-guided, DNA-binding domain in a number of fusion proteins, including DNA base editors and transcriptional activators. The Himar1C9 hyperactive transposase efficiently catalyzes transposition in diverse species and *in vitro*, indicating that it acts host-independently and robustly in a variety of

environments and may be amenable to protein modifications. For the purposes of exploring transposon insertion sites by next-generation sequencing, the Himar1 transposon features the advantage of a MmeI restriction site in its inverted repeat sequences, which enables isolation of transposon-insertion site junctions by restriction digest. Because the Himar1 transposase only inserts transposons into TA sites, a targeted Himar1-dCas9 fusion protein could potentially enable single-nucleotide precision in insertions, whereas transposases with less well-defined insertion sites (such as Tn5[124]) may not achieve that level of precision targeting.

Because protein fusions to the N-terminus of dCas9 and the C-terminus of mariner-family transposases have been previously described, we fused the C-terminus of Himar1C9 to the N-terminus of dCas9 using flexible protein linkers XTEN (N-SGSETPGTSESATPES-C) and L2 (N-GHGTGSTGSGSS-C), which were the two most optimal linkers used to construct FokI-dCas9 fusion proteins[119]. We performed preliminary tests to verify that both domains of the Himar1C9-dCas9 fusions were functional. To check that dCas9 was capable of binding a DNA target specified by a gRNA, we expressed the Himar1C9-L2-dCas9 fusion in an *E. coli* strain with a genomically integrated mCherry gene, along with 2 gRNAs targeting mCherry (gRNAs 5 and 16 in **Table 2.1**). If the Himar1C9-dCas9 fusion bound the mCherry genomic target, then the protein would sterically hinder RNA polymerase and decrease transcription of the gene, thereby decreasing mCherry fluorescence. Indeed, we observed mCherry knockdown by Himar1C9-dCas9, indicating that the DNA binding functionality was intact in the fusion protein (**Suppl. Figure S2.1a**). We verified that the transposase activity was also intact in Himar1C9-L2-dCas9 using an *E. coli* conjugation assay (**Suppl. Figure S2.1b**).

Design of an *in vitro* reporter system for targeted transposition events by Himar1C9-dCas9

We first explored the targeted transposition activity of a purified Himar1C9-dCas9 fusion protein within *in vitro* reactions. The fusion protein Himar1-XTEN-dCas9 (HdCas9) was cloned into a C-terminal 6xHis-tagged expression vector and purified from *E. coli* by nickel affinity

chromatography. Purified HdCas9 was mixed in a reaction buffer with a transposon donor plasmid pHimar6, which contained a Himar1 transposon bearing a chloramphenicol resistance marker; a transposon target plasmid pGT-B1, which contained the *GFP* gene; and one or more gRNAs targeted to various loci along *GFP* (**Figure 2.1b, Tables 2.1-2.2**).

We analyzed transposon insertions into the target plasmid by several assays. qPCR of target plasmid/transposon junctions, using one primer complementing the transposon DNA and one primer complementing the target plasmid, enabled quantification of transposons that were inserted into the target plasmid, as well as qualitative assessment of transposition specificity based on the enrichment of PCR products of the expected size (**Figure 2.1b-c**). For every transposon/target junction PCR, we also performed a control PCR amplifying the target plasmid's backbone to normalize for DNA input between samples. Relative Cq measurements shown in this study are the differences between these two Cq values. PCR primers used in this study are listed in **Table 2.3**. Next-generation transposon sequencing enabled quantitative measurement of the distribution of inserted transposons within the target plasmid (**Suppl. Figure S2.2**). Finally, transposition reaction products could be transformed into competent *E. coli*. Because the donor plasmid's R6K origin of replication inhibits replication in *E. coli* without the *pir* replication gene, we could select for transformants of the target plasmid and of target plasmids with an integrated transposon. Transposition efficiency was calculated by dividing the number of chloramphenicol-resistant transformants (recipients of target plasmid with a transposon) by the number of carbenicillin-resistant transformants (recipients of target plasmid). Transposition specificity among transformants was determined by Sanger sequencing of plasmids from chloramphenicol-resistant transformants.

Site-specificity of Himar1 transposon insertions is dependent on gRNA orientation.

Within the *in vitro* system, we assessed several gRNAs to determine how the spatial orientation of the gRNA relative to the target TA site affects site-specificity of transposition. We

designed gRNAs with spacings between 5 and 18 bp from a TA site, targeting either the template strand or non-template strand of the GFP gene (**Figure 2.2a, Table 2.1**). We tested each individual gRNA in triplicate reactions using 30 nM purified HdCas9 protein, 30 nM gRNA, 2.27 nM pHimar6, and 2.27 nM pGT-B1.

Based on the PCR assay, a single gRNA is sufficient to mediate site-specific transposition, but site-specific transposition is dependent on gRNA spacing to the TA target site (**Figure 2.2b**). All gRNA-targeted insertion events occurred at the nearest TA dinucleotide distal to the 5' end of the gRNA, independent of the targeted DNA strand (**Figure 2.2a**).We observed strong site-specificity for gRNAs with 7-9 bp spacings and with 16-18 bp spacings, based on the robust PCR band for transposon junctions of the expected size (red arrows). At very short spacings (5-6 bp), there was no PCR band at the expected junction size, suggesting that the HdCas9 protein sterically hinders transposition at locations less than 7bp from the TA target site. At spacings between 11-15 bp, there is a faint PCR band at the expected junction size, indicating that transposition at those sites may occur, but the site-specific transposition is not optimal. These findings are consistent with the previously observed spacing dependence for FokI-dCas9 proteins built using the same XTEN peptide linker[119]. The bimodal distribution of robustly targeting gRNA spacings may be due to the DNA double helix providing steric hindrance at intermediate spacings, since the spacing peaks are approximate 1 helix turn (~10 bp) apart. Comparison of relative Cq values between gRNA-targeted transposases and HdCas9 transposase in the absence of a gRNA suggests that transposition may occur at a higher rate in the presence of a targeting gRNA (**Figure 2.2b**). This may be due to circumvention of the need for a transposase dimer loaded with a transposon to search through space for a TA insertion site, as the gRNA-targeted transposases were already bound near an insertion site. We also tested a purified Tn5-dCas9 transposase with hyperactive E54K and L372P mutations in *in vitro* reactions using the same set of gRNAs, but did not find evidence of site-specific transposition in any of the reactions based on the PCR assay (**Suppl. Figure S2.3**).

To infer the complete distribution of transposon insertions around the target plasmid, we performed transposon sequencing on transposition products resulting from 3 GFP-targeted gRNAs (gRNAs 4, 8 and 12), a non-targeting gRNA (gRNA 5), and no gRNA (**Figure 2.2c**). The baseline distribution of random transposon insertion events was generated from quadruplicate *in vitro* reactions with no gRNA. In this distribution, transposon insertions were present throughout the 6.2 kb pGT-B1 plasmid, with a spike in abundance at position 5999, which is a TA site in the middle of a 12 bp stretch of T and A nucleotides. This observation is consistent with the observation that Himar1 transposase preferentially inserts transposons into flexible, T/A-rich DNA[125]. gRNA-targeted insertions, on the other hand, were less likely to be inserted into position 5999 and were enriched at their respective gRNA-adjacent TA site, compared with baseline, to varying degrees (**Figure 2.2b-c**). In particular, gRNA 4, which had an optimal spacing of 8 bp from the target TA site, had 42% of all transposon insertions being exactly at the target site, a 342-fold enrichment over baseline. Comparison of fold-enrichment across different gRNAs suggests that the specific target site and flanking DNA also plays a role in the specificity of transposon insertions. gRNA 12 had a higher fold-enrichment of insertions at its target site than gRNA 8, but a lower fraction of insertions, suggesting that the lack of specificity of gRNA12 may be attributable to its target site being disfavored for transposition.

Given that mariner transposases dimerize in solution in the absence of DNA[126], we hypothesized that HdCas9 dimerizes spontaneously, and then the active Himar dimer is guided to a gRNA-specific target locus by one of the dCas9 domains in the HdCas9 dimer (**Figure 2.1a**). This mechanism of activity is consistent with the observation that a single gRNA is sufficient to direct targeted transposition. Further support for this hypothesis comes from *in vitro* reactions containing pairs of gRNAs targeting the same TA site, but complementing opposite strands (**Suppl. Figure S2.4**). If Himar subunits did not spontaneously dimerize, then dimerization of HdCas9 would be enhanced by loading two HdCas9 monomers onto the same target plasmid, so that pairs of Himar monomers would be tethered in close proximity. We set up reactions in which

target DNA was preloaded with paired gRNA/HdCas9 complexes, vs. preloaded with single gRNA/HdCas9 complexes, and then mixed with transposon donor DNA (**Suppl. Figure S2.4a**). In all reactions, the final reaction contained 5 nM HdCas9 protein, 5 nM donor DNA, 5 nM target DNA, and 2.5 nM of each of two gRNAs. Under these conditions, there was no difference in transposition rate or specificity between the gRNA/HdCas9 complexes preloaded as pairs or as singletons (**Suppl. Figure S2.4b, c**). The observation that preloading of pairs of HdCas9 complexes does not improve transposition is consistent with the hypothesis that transposase dimers formed before one of the gRNA/dCas9 domains targeted the dimer to its final location.

Site-specific transposition activity of HdCas9 is robust to variations in protein and DNA concentrations.

To assess robustness of HdCas9 to various experimental conditions and determine the optimal parameters for site-specific transposition, we varied the concentrations of (1) protein-gRNA complexes, (2) transposon donor plasmid DNA, (3) target plasmid DNA, and (4) background off-target DNA within *in vitro* transposition reactions containing a single gRNA (gRNA 4). We also incubated the reactions at different temperatures and extended the reaction time from 3 hours up to 72 hours.

gRNA-guided HdCas9 mediated site-specific transposition events at protein concentrations between 3 and 100 nM in reactions with 5 nM donor and 5 nM target plasmids (**Figure 2.3a**). These reactions were also purified and transformed into electrocompetent *E. coli* to assess the efficiency and specificity of transposition (**Figure 2.3b**); the trend of higher transposition at higher transposase concentration is consistent with qPCR measurements. At a concentration of 30 nM of the gRNA-HdCas9 complex, the specificity of transposon insertion into the targeted TA site was 44% (11/25 colonies); at 100 nM of the gRNA-HdCas9 complex, the specificity was 47.5% (19/40 colonies).

Site-specific transposition was also robust to variations in DNA concentration. We observed site-specific transposition activity in HdCas9 reactions with 5 nM target plasmid DNA and between 0.05 to 5 nM donor plasmid DNA, which greater rates of transposition occurring at higher DNA concentrations (**Figure 2.3c**). Similarly, site-specific transposition occurred across a wide range of variation in target plasmid DNA (0.25 nM to 10 nM) while donor plasmid DNA was kept constant at 0.5 nM (**Figure 2.3d**). While the absolute rate of transposition (measured by Cq of the transposon-target junction qPCR) was higher at higher target DNA concentrations, the relative Cq held constant across target DNA concentrations, indicating that a similar proportion of target plasmids received a transposon in each reaction.

gRNA-guided HdCas9 was also able to transpose into a targeted site effectively in the presence of background DNA. We added up to 10 times more background *E. coli* genomic DNA than target plasmid DNA to a set of reactions containing 10 nM gRNA-HdCas9, 1 nM donor DNA, and 1 nM target DNA. Across all background DNA concentrations tested, HdCas9 was able to locate the gRNA-targeted site and insert transposons site-specifically within 3 hours at 30C, with no loss of specificity or efficiency (**Suppl. Figure S2.5a**). When this transposition reaction with background DNA was performed at 37C and over longer time courses, to mimic conditions in living cells growing at 37C, we observed similar results (**Suppl. Figure S2.5b-c, Figure 2.3e-f**). In the presence of 10 times more background DNA than target plasmid DNA, gRNA-HdCas9 at concentrations of 10 nM and 100 nM was capable of mediating site-specific transposition. The relative Cq and PCR band intensity of transposon-target junctions increased slightly between 3 and 16 hours, but PCR bands qualitatively remained at the same level of specificity, suggesting that gRNA-guide transposases are faster at locating the target site than catalyzing transposition and that the increase in site-specific transposon insertions over time is performed by gRNA-dCas9 bound transposases. After 16 hours, site-specific transposition events appeared to reach a plateau; the loss of specific transposon-target junctions at 72 hours on the PCR gels was most

likely due to degradation of the *in vitro* reaction components over time (**Suppl. Figure S2.5b, Figure 2.3e**).

HdCas9 mediates site-specific transposon insertions into plasmids *in vivo*

Given that HdCas9 robustly produces site-specific transposon insertions *in vitro*, we tested the ability of HdCas9 to mediate site-specific transposition in two *in vivo* systems in *E. coli* and in mammalian cells. In the first system, we transformed a set of 3 plasmids into S17 *E. coli*: pTarget, which contains a GFP target gene; pHimar6, the transposon donor plasmid; and a tet-inducible expression vector for HdCas9 that also contained 0 or 1 gRNA targeting GFP (**Figure 2.4a**). We grew the cells containing these 3 plasmids for 16 hours on selective agar plates with $MgCl_2$ and anhydrotetracycline (ATc) to allow for transposition to occur, and then extracted all plasmids from these cells. We measured transposition specificity in these plasmid pools by two methods: PCR of transposon-target plasmid junctions, and transformation of the pooled plasmids into competent cells and analysis of transposon insertions in individual transformants.

We first performed a series of validations that the HdCas9 system components functioned *in vivo*. The S17 *E. coli* strain was transformed with pTarget and one of several HdCas9/gRNA expression vectors. By measuring HdCas9-mediated knockdown of GFP in these strains, we confirmed that gRNAs (gRNA 1, gRNA 4, and non-targeting gRNA 5) targeted HdCas9 to the pTarget plasmid as expected and determined the optimal concentration of ATc for inducing HdCas9 expression (**Figure 2.4b**). Consistent with previously reported results[117], gRNA 1, which targets the non-template strand of the GFP gene, caused knockdown of GFP expression, but gRNA 4, which targets the template strand and does not sterically hinder RNA polymerase, did not cause GFP knockdown. HdCas9 concentrations reached saturation at ATc induction levels of 2 ng/mL, as further increasing the concentration of ATc did not result in further knockdown of GFP by gRNA 1. At an ATc level of 1 ng/mL, HdCas9 levels were high enough to bind to some, but not all, GFP target sites, based on the partial knockdown of GFP by gRNA 1. We also validated that

purified HdCas9 protein with gRNA 1 or gRNA 4 mediated targeted transposition into the GFP gene of pTarget *in vitro* (**Figure 2.4c**).

In the *in vivo* transposition assay, S17 *E. coli* containing pTarget and an HdCas9/gRNA expression vector were transformed with transposon donor plasmid pHimar6. After this transformation, cells were plated on selective agar plates containing a saturating concentration of $MgCl_2$ to enable transposase activity and 1 ng/mL ATC to induce non-saturating expression of HdCas9, in order to avoid overproduction inhibition of the Himar1C9 transposase[127]. After 16 hours of growth at 37C, we extracted plasmids from the pooled colonies and analyzed the plasmid pools for site-specific transposon insertion events. PCR for transposon-target plasmid junctions showed that gRNA 1 produced detectable site-specific transposon insertions into pTarget in 3 out of 5 independent replicates (**Figure 2.4d**). gRNA 4, however, did not produce an enrichment of PCR products corresponding to its targeted insertion site.

We also evaluated site-specificity of transposition by transforming the plasmid pools into electrocompetent MegaX *E. coli* and analyzing individual transformants by colony PCR and Sanger sequencing, to confirm that HdCas9 with gRNA 1 mediated precisely targeted transposon insertions into pTarget (**Figure 2.4a**). Analyzing 4 independent plasmid pools from gRNA 1-expressing cells (**Figure 2.4d**), we found that 3 out of 4 transformations produced colonies with mostly or all site-specific transposition products (**Figure 2.4e**). In transformations of 4 plasmid pools from cells without a gRNA, we did not obtain any transformants containing pTarget plasmids with an integrated transposon. This result suggests that *in vivo* transposition rates into the target plasmid were low, but precisely targeted by a gRNA complementary to the non-template strand of the GFP gene. Further analysis of transposition products from cells containing gRNA 4 or paired gRNAs targeting opposite strands at the same locus is required to determine whether transposition is dependent on gRNA strand complementarity and whether paired gRNAs function synergistically *in vivo*.

In a second *in vivo* test system, we tested the ability of gRNA-targeted HdCas9 to mediate site-specific transposition into a genomic locus in Chinese hamster ovary (CHO) cells. We transfected a line of CHO cells containing a single-copy, constitutively expressed eGFP gene in the genome with two plasmids, one containing a Himar transposon and 0, 1, or 2 gRNA expression operons, and the other being an expression vector for HdCas9 (**Suppl. Figure S2.6a**). In the mammalian expression vector, HdCas9 was fused to an N-terminal 3x-FLAG tag and SV40 nuclear localization signal (NLS) and a C-terminal 6x-His tag. Two gRNAs were designed to target the eGFP gene at the same TA insertion site, complementing opposite strands; we tested these gRNAs individually and as a pair *in vivo*, along with a non-targeting gRNA and no gRNA. Preliminary *in vitro* experiments demonstrated that the two gRNAs individually mediated site-specific transposition by the purified 3x-FLAG-NLS-HdCas9-6x-His protein (**Suppl. Figure S2.6b**).

The Himar transposon contained a promoterless puromycin resistance gene and mCherry gene as selectable markers, both of which would be inserted in-frame into the eGFP locus and expressed if correctly targeted by HdCas9 (**Suppl. Figure S2.6a**). Because the transposon genes would only be expressed if the transposon were integrated downstream of an endogenous genomic promoter, puromycin selection for transposon mutants was stringent against false positive clones resulting from random plasmid integration into the genome. We verified that transposon insertions into the target locus resulted in successful expression of puromycin resistance and mCherry by constructing a positive control cell line with the transposon cloned into that locus (**Suppl. Figure S2.6c**).

Following transfection, we performed antibiotic selection with puromycin to select for cells with an integrated transposon. From each transfection of approximately 10^6 cells, we obtained about 20 colonies representing independent transposition events. Negative controls for transposition, which were transfected with only the transposon donor plasmid, did not produce viable cells, indicating clean selection against background random plasmid integration events. All

colonies from each transfection were pooled for analysis by flow cytometry and PCR for transposon-target junctions. Flow cytometry showed that transfections with no gRNA resulted in few eGFP-negative cells, while some transfections with at least 1 gRNA (including the non-targeting gRNA) produced sizable proportions of eGFP-negative cells (**Suppl. Figure S2.6c-d**). However, PCR for the expected eGFP-transposon junction showed no evidence of targeted transposition in any of the transfections, suggesting that the eGFP-negative cells had lost expression of eGFP by some other mechanism (**Suppl. Figure S2.6e**). Although we did not observe targeted transposition by HdCas9 into a genomic locus in CHO cells, an optimized mammalian testbed for HdCas9 may enable screening for site-specific transposition events among larger samples of transposon insertions and shed light on the determinants of site-specific transposition in mammalian cells.

2.4 DISCUSSION

In this study, we demonstrated for the first time that a random transposase can be engineered to insert transposons site-specifically when targeted to a user-defined genetic locus by dCas9 and a gRNA. The HdCas9 fusion protein, consisting of a hyperactive Himar1 transposase and the *S. pyogenes* dCas9, functions both *in vitro* and *in vivo* in *E. coli* to mediate site-specific transposition into a targeted locus on a plasmid. We characterized the activity of HdCas9 *in vitro* across variations in protein and DNA concentration, as well as variations in reaction conditions such as temperature and duration, demonstrating that the site-specific transposition activity is robust to these parameters, but is dependent on the gRNA orientation relative to the target TA insertion site and the surrounding target DNA sequence.

While we demonstrated targeted insertion into a medium copy-number plasmid in *E. coli,* additional experiments are necessary to determine whether HdCas9 is capable of targeting transposon insertions into a genomic locus. A bacterial assay for targeted transposition into the genome could be performed by targeting the HdCas9 to a counterselectable genomic marker,

such as *galK*, *tetA*, or *sacB*, and selecting for the gene knockout phenotype[128, 129]. In such a scheme, differentiation between knockout and knockdown (via HdCas9 transcription blockade) phenotypes is essential, and could be facilitated by only transiently expressing the HdCas9/gRNA complex in cells to minimize the knockdown phenotype before initiating counterselection. Given the inherent leakiness of inducible expression systems, transient expression of the HdCas9/gRNA complex may be best accomplished by conjugation or transformation of a non-replicative vector (e.g., an R6K origin of replication plasmid) into the bacteria, so that transformants will lose the genes encoding the Cas-Transposon components within a defined period of time. Transient presence of the transposon donor vector would also be a useful feature in a genomic targeting assay, so that the pool of genomically integrated transposons (as opposed to transposons that remain intact in the donor vector) could be easily isolated and analyzed by Tn-seq/INSeq.

The Cas-Transposon system requires further optimization to enable targeted genomic transposon insertions in mammalian cells. We introduced the components of the Cas-Transposon system, with the HdCas9 protein and Himar transposon adapted for mammal cells, into CHO cells by transfection of 2 expression vectors. We obtained newly antibiotic-resistant clones of cells from transfections involving both the HdCas9 expression vector and the transposon donor vector, but not control transfections of only the transposon donor vector, indicating that the resistant cells had undergone transposition events that inserted the transposon into the genome. However, it remains unclear how well the HdCas9 protein and gRNAs were expressed, and how well the transposase, gRNAs, and transposon donor vectors were localized in the nucleus to enable genomic transposition. Antibody staining for the HdCas9 protein and fluorescence *in situ* hybridization (FISH) using probes for gRNAs and transposons would provide information on the localization and relative concentration of each Cas-Transposon system component. Modification of regulatory elements on the expression vectors and transfections of different quantities of each vector would allow for adjustment of component concentrations. Alternative methods of delivering these components into cells, such as electroporation of nucleoprotein complexes, may be

required to achieve optimal levels inside the nucleus. More sensitive screens for transposon insertions, as well as screening a larger pool of transposon insertions, may also provide greater insight into the mechanisms of HdCas9 targeting and transposition catalysis in mammalian cells.

The performance of the Cas-Transposon system in both bacterial and mammalian cells may be improved by engineering the HdCas9 protein for improved specificity and efficiency. DNA base editors consisting of a deaminase fused to a dCas9 protein have improved efficiency when two monomers of the deaminase, which operates as a homodimer, are fused to dCas9[75]. Because Himar1 likewise functions as a homodimer, the dimer-dCas9 fusion approach may be useful for targeting active Himar1 dimers to the desired insertion site. Modification of the Himar1 protein by directed evolution or by rational mutagenesis may also improve function. Previous studies have shown that Himar1, like other mariner transposases, is rate-limited by the synapsis of transposon ends to the protein dimer [127, 130]. Himar1 mutants with single amino-acid substitutions in the conserved WVPHEL motif have higher rates of transposition than the Himar1C9 mutant, because the allosteric inhibition of transposon synapsis is disrupted[127]. Thus, we speculate that a hyperactive WVPHEL mutation, which removes the natural rate-limiting step of transposon synapsis, combined with a mutation in the dimerization interface to slow down dimerization of Himar1 (normally a fast step), may result in a Himar1 protein that dimerizes more specifically at dCas9-targeted locations and then efficiently catalyzes transposition at those sites. In discussing protein engineering to optimize the Cas-Transposon system, we should also consider alternative transposases, alternative DNA-binding domains, such as dCas9 homologs from other species with smaller sizes and/or alternate sequence target sites, and alternative peptide linkers connecting the two domains. By using transposases with different insertion sites and DNA-binding domains with alternate PAM specificities and adjusting the linker length, the Cas-Transposon protein may be tailored for optimal function in different organisms with different GC contents and distributions of PAMs and insertion sites. Although the Tn5-dCas9 transposase described here did not produce site-specific transposon insertions, the gRNAs we tested were not optimized for

Tn5 insertion sites, and the protein itself was missing one canonical hyperactive mutation (M56A)[131], so there may be room for improvement of the Tn5-dCas9 transposase along with mariner-dCas9 fusions. If optimization of the Cas-Transposon technology is successful, the platform would enable a novel modality of site-specific DNA insertion for targeted genome editing.

2.5 MATERIALS AND METHODS

Measurement of fluorescence knockdown by Himar1C9-dCas9 fusion proteins in *E. coli*: We measured knockdown of mCherry in MG1655 galK::mCherry-specR *E. coli* and knockdown of GFP in s17 *E. coli* containing the pTarget plasmid. Each strain had a tet-inducible expression vector for Himar1C9-dCas9 or dCas9, which also contained up to constitutively expressed 2 gRNAs. Overnight cultures were diluted 1:40 into media containing various concentrations of anhydrotetracycline (ATc) to induce HdCas9 expression. 200 uL of each induced culture was grown in 96-well plates at 37C with shaking on a BioTek plate reader. Measurements of OD600 and fluorescence were taken every 5-10 minutes for 12 hours. Fluorescence measurements were reported as F/OD600.

Measurement of transposase activity in *E. coli*: Himar1C9-L2-dCas9 and Himar1C9 were expressed in MG1655 *E. coli* from a tet-inducible expression vector. These MG1655 strains were conjugated with diaminopimelic acid-auxotrophic donor strain EcGT2 (s17 asd::mCherry-specR), which contained transposon donor plasmid pHimar2. Donor and recipient cultures were grown overnight at 37C; donors were grown in LB media with DAP (50 uM) and kanamycin (50 ug/mL), and recipients were grown in LB with chloramphenicol (20 ug/mL). 100 uL of donor and recipient were diluted into 3 mL fresh media +/- 2 uM ATc and grown for 6 hours at 37C. 1 mL each of donor and recipient cultures were centrifuged and resuspended twice in PBS to wash the cells

and then centrifuged to form a pellet. Donor and recipient pellets were combined in 50 uL PBS and dropped onto LB agar with 50 uM DAP +/- 2 uM ATc. The cell droplets were dried at room temperature, and the conjugations were incubated for 10 hours at 30C. The resulting cells were scraped off, homogenized by pipetting, and plated on nonselective LB agar plates and LB + carbenicillin (50 ug/mL) plates to select for MG1655 cells with a transposon. Transposition rates were measured as the ratio of carbenicillin-resistant CFUs to total CFUs.

Purification of Himar-dCas9 protein for *in vitro* transposition assays: His-tagged Himar-dCas9 (HdCas9) was purified by nickel affinity chromatography. The Himar transposase gene was obtained from plasmid pSAM-BT, and the dCas9 gene sequence was obtained from pdCas9-bacteria, a gift from Stanley Qi (Addgene plasmid #44249). Linker sequences were obtained from previously described FokI-dCas9 fusions[119] and synthesized as gBlocks (Integrated DNA Technologies). Himar-dCas9 fusion constructs were cloned into a C-terminal 6xHis-tagged pET expression vector using NEBuilder® HiFi DNA Assembly Master Mix (New England Biolabs). pET-Himar-dCas9 vectors were electroporated into competent Rosetta2 cells (Novagen); transformants were selected on chloramphenicol 34 ug/uL and carbenicillin 50 ug/uL.

Rosetta2 protein expression strains were inoculated from a single colony into 4 mL of LB with chloramphenicol 34 ug/uL and carbenicillin 50 ug/uL and grown to saturation overnight at 37C with shaking. 1 mL of saturated culture was diluted 1:100 into 100 mL of fresh media and grown to OD 0.6-0.8 at 37C with shaking. 0.2 mM of IPTG was added to the culture, and the flask was moved to 18C and incubated overnight (16-18 hours) with shaking. The induced overnight culture was chilled on ice and centrifuged at 4C, 7197xg for 5 minutes to pellet the cells. The cell pellet was resuspended in 5 mL of Protein Resuspension Buffer (PRB) inside a 50 mL conical tube, which was kept chilled in an ice water bath. The cells were lysed by sonication using a Qsonica sonicator at 40% power for a total of 120 seconds in 20 second/20 second on/off intervals. The cell suspension was mixed by pipetting, and the sonication cycle was repeated to

ensure complete lysis. The cell lysate was centrifuged at 7197 x g for 10 minutes at 4C to pellet cell debris. The cleared cell lysate was collected.

To equilibrate the nickel affinity resin, 1 mL of Ni-NTA Agarose (Qiagen) was added to a 15 mL polypropylene gravity flow column (Qiagen), storage buffer was allowed to flow through by gravity, and the resin was washed in 5 mL of PRB and drained. The cleared cell lysate was added to the column and incubated at 4C on a rotating platform for 30 minutes. The cell lysate was flowed through, and the nickel resin was washed with 25 mL PWB in 5 mL increments. The protein was then eluted with PEB in 5 fractions of 0.5 mL each. Most protein eluted in fractions 2-4. Each elution fraction was analyzed by running an SDS-PAGE gel to check for the presence of the protein.

Elution fractions 2-4 were combined and dialyzed overnight in 500 mL of DB1 at 4C using 10K MWCO Slide-A-Lyzer™ Dialysis Cassettes (Thermo Fisher). The protein was dialyzed again in 500 mL DB2 for 6 hours. The dialyzed protein was quantified with the Qubit Protein Assay Kit (Thermo Fisher) and divided into single-use aliquots. Protein aliquots were snap frozen in dry ice and ethanol and stored at -80C. Tn5-dCas9 was purified in the same manner.

In vitro transposition reaction setup: Each *in vitro* HdCas9 transposition reaction was performed in a buffer consisting of 10% glycerol, 2 mM dithiothreitol, 250 ug/mL of bovine serum albumin, 25 mM HEPES (pH 7.9), 100 mM NaCl, and 10 mM MgCl2. Reactions contained 0-5 nM of transposon donor and target plasmids, 0-100 nM of transposase/gRNA complexes, and 0-800 ng of background *E. coli* genomic DNA in a final volume of 20 uL.

Plasmid DNA was purified using the ZymoPureII midiprep kit (Zymo Research) and purified again using the Zymo Clean and Concentrator-25 (Zymo Research) kit to remove all traces of RNAse. Genomic DNA was purified from *E. coli* using the MasterPure Gram Positive DNA Purification Kit (Epicentre) and purified again using the Zymo Clean and Concentrator-25

kit. All DNAs were eluted into nuclease-free water. DNA concentrations were quantified using the Qubit dsDNA BR Assay kit (Invitrogen).

gRNAs for *in vitro* experiments were synthesized and purified using the GeneArt™ Precision gRNA Synthesis Kit (Invitrogen) according to manufacturer's instructions. gRNA concentrations were measured using the Qubit RNA HS Assay Kit. gRNAs were aliquoted into single-use tubes, flash frozen in dry ice and ethanol, and stored at -80C until usage.

Frozen aliquots of transposase protein and gRNAs were thawed on ice. The protein was diluted to 20x final concentration in DB2 buffer, and gRNAs were diluted to the same molarity in nuclease-free water. The diluted protein and gRNA were mixed in equal volumes and incubated at room temperature for 15 minutes. Transposon donor DNA, target plasmid DNA, and background DNA (if applicable) were mixed on ice with 10 uL of a 2x buffer master mix and water to reach a volume of 18 uL; 2 uL of the protein/gRNA mixture was added last to the reaction. In reactions where the transposase/gRNA complex was preloaded onto the target plasmid, the target plasmid was mixed with protein and gRNA and incubated at 30C for 10 minutes, and the donor DNA was added last. Transposition reactions were incubated for 3-72 hours at 30-37C and then heat-inactivated at 75C for 20 minutes. Transposition products were purified using magnetic beads, eluted in 45 uL of nuclease-free water, and stored at -20C.

In vitro Tn5-dCas9 transposition reactions were set up similarly, except the reaction buffer contained 50 mM Tris-acetate (pH 7.5), 150 mM potassium acetate, 10 mM magnesium acetate, and 4 mM spermidine. Purified Tn5-dCas9 protein was diluted in 50% glycerol, 50 mM Tris-HCl (pH 7.6), 100 mM KCl, and 1 mM DTT.

qPCR assay for site-specific insertions: For each reaction, 2 qPCRs were performed to obtain the measure of the relative Cq: one PCR amplifying transposon-target plasmid junctions, and another PCR amplifying the target plasmid backbone to normalize for template DNA input across samples. Relative Cq measurements shown in this study are the differences between these two Cq values.

For in *vitro* Himar transposition into pGT-B1, primers p433 and p415 were used for junction PCRs, and primers p828 and p829 were used for control PCRs. For Tn5 *in vitro* transposition into pGT-B1, primers p858 and p859 were used for junction PCRs. For *in vitro* Himar transposition into pTarget or pZE41-eGFP, primers p898 and p415 were used for junction PCRs and primers p899 and p900 were used for control PCRs. All qPCR primers used in this study are listed in Table 3.

PCR reactions contained 1 uL each of 10 uM forward and reverse primers, 1 uL of purified transposition products as template DNA, 7 uL water, and 10 uL of Q5 2X Master Mix (NEB) + SYBR Green. Reactions were thermocycled using a Bio-Rad C1000 touch qPCR machine for 1 minute at 98C, followed by 35 cycles of 98C denaturation for 10 seconds, 67-68C annealing for 15 seconds, and 72C extension for 2 minutes. PCR products were quantified by SYBR Green fluorescence and qualitatively analyzed for specificity on agarose DNA gels.

Transformation assay for *in vitro* transposition reaction products: 5uL purified DNA from an *in vitro* transposition reaction was mixed with 45 uL distilled water and chilled on ice. 10 uL of thawed MegaX electrocompetent *E. coli* (Invitrogen) was added and pipetted up and down gently until just mixed. The mixture was transferred to a 0.1 cm gap electroporation cuvette (BioRad) and electroporated at a voltage of 1.8 kV. The cells were immediately recovered in 1 mL of SOC and incubated with shaking at 37C for 90 minutes. The cells were plated on antibiotic selection plates containing chloramphenicol 34 ug/mL to select for target plasmids containing transposons, and on plates with carbenicillin 50 ug/mL to measure the electroporation efficiency of target plasmids. The efficiency of transposition was measured as the ratio of chlor-resistant transformants to carb-resistant transformants. To assess specificity of inserted transposons, we performed colony PCR on transformants using the primer set p433/p415 with KAPA2G Robust HotStart ReadyMix (Kapa Biosystems) to amplify Himar-target plasmid junctions, which were analyzed by Sanger sequencing.

Transposon sequencing library preparation: Transposon junctions were PCR amplified from *in vitro* transposition reactions using primer sets p923/p433 and p923/p922, using Q5 HiFi 2x Master Mix (NEB) + SYBR Green. Reactions were thermocycled using a Bio-Rad C1000 touch qPCR machine for 1 minute at 98C, followed by cycles of 98C denaturation for 10 seconds, 68C annealing for 15 seconds, and 72C extension for 2 minutes. PCR reactions were stopped in late exponential phase in order to avoid oversaturation of PCR products. PCR products were purified using magnetic beads, and 100-200 ng of DNA per sample was digested with MmeI (NEB) for 1 hour in a reaction volume of 40 uL. The MmeI digestion products were purified using Dynabeads M-270 Streptavidin beads (ThermoFisher) according to manufacturer instructions. The digested transposon ends, bound to magnetic Dynabeads, were mixed with 1 ug sequencing adapter DNA (see next section), 1 uL T4 DNA ligase, and T4 DNA ligase buffer in a total reaction volume of 50 uL. The ligation reactions were incubated at room temperature (23C) for 1 hour; the ligase enzyme and excess sequencing adapter were removed by washing the beads according to Dynabead manufacturer instructions. Dynabeads were resuspended in 40 uL of water.

2 uL of the Dynabeads were used as template for the final PCR using barcoded P5 and P7 primers and Q5 HiFi 2x Master Mix (NEB) + SYBR Green. Reactions were thermocycled using a Bio-Rad C1000 touch qPCR machine for 1 minute at 98C, followed by cycles of 98C denaturation for 10 seconds, 67C annealing for 15 seconds, and 72C extension for 20 seconds. PCR reactions were stopped in late exponential phase in order to avoid oversaturation of PCR products. Equal amounts of DNA from all PCR reactions were combined into one sequencing library, which was purified and size-selected for 145 bp products using the Select-a-Size Clean and Concentrator kit (Zymo). The library was quantified with the Qubit dsDNA HS Assay Kit (Invitrogen) and combined in a 7:3 ratio with PhiX sequencing control DNA. The library was sequenced using a MiSeq V2 50 Cycle kit (Illumina) with custom read 1 and index 1 primers spiked into the standard read 1 and index 1 wells. Reads were mapped to the pGT-B1 plasmid using Bowtie 2[132].

Construction of sequencing adapter by annealing: Oligonucleotides Adapter_T and Adapter_B were diluted to 100 uM in nuclease-free water. 10 uL of each oligo was mixed with 2.5 uL water and 2.5 uL of 10x Annealing Buffer. Using a thermocycler, the mixture was heated to 95C and slowly cooled at 0.1C/second to 4C to yield 25 uL of 40 uM sequencing adapter. The adapter was stored at -20C until ready for use.

In vivo assays for transposition into a target plasmid: s17 *E. coli* were transformed with pTarget-GFP and pHdCas9-gRNA plasmids by electroporation; transformants were selected on LB with spectinomycin (240 ug/mL) and carbenicillin (50 ug/mL). *E. coli* containing these two plasmids were grown from a single colony to mid-log phase in liquid selective media, electroporated with 130 ng pHimar6 transposon donor plasmid DNA, and recovered in 1 mL LB for 1 hour at 37C with shaking post-electroporation. 100 uL of a 10^-3 dilution of the transformation was plated on LB agar plates with spectinomycin (240 ug/mL), carbenicillin (50 ug/mL), chloramphenicol (20 ug/mL), MgCl2 (20 mM), and ATC (0-2 ng/mL). Plates were grown at 37C for 16 hours. Between 10^3 and 10^4 colonies were scraped off each plate into 2 mL PBS and homogenized by pipetting. 500 uL of the cells were miniprepped using the QIAprep kit (Qiagen).

Minipreps from each transformation were evaluated by qPCR for transposon junctions and by a transformation assay. qPCR assays for transposon-target plasmid junctions were performed as described above, using primers p898 and p415 and 10 ng of miniprep DNA as PCR template. The control PCR to normalize for DNA input was performed with primers p899 and p900. In transformations, 150 ng of plasmid DNA was electroporated into 10 uL of MegaX electrocompetent cells diluted in 50 uL of ice-cold distilled water. Cells were immediately recovered in 1 mL of LB and incubated with shaking at 37C for 90 minutes. The cells were plated on LB agar with chloramphenicol (20 ug/mL) and spectinomycin (60 ug/mL) to select for target plasmids containing a Himar transposon. We performed colony PCR using the primer set

p898/p415 with KAPA2G Robust HotStart ReadyMix (Kapa Biosystems) to amplify Himar-target plasmid junctions, which were analyzed by Sanger sequencing.

Generation of CHO cell lines for transposition assays: CHO cells were cultured in Ham's F-12K (Kaighn's) Medium (ThermoFisher) with 10% fetal bovine serum and 1% penicillin-streptomycin. The eGFP+ CHO cell line was generated by transfection of plasmids pcDNA5/FRT/Hyg-eGFP and pOG44 into the Flp-In™-CHO cell line (ThermoFisher) followed by selection in media with hygromycin (500 ug/mL). An eGFP-, mCherry+, puromycin-resistant site-specific transposition positive control cell line was generated by transfection of plasmids pcDNA5/FRT/Hyg-Himar and pOG44 into the Flp-In™-CHO cell line followed by selection in media with puromycin (10 ug/mL). Transfections were performed on cells at 70% confluence in 6-well plates using 12 uL of Lipofectamine 2000 and 1000 ng of each plasmid. Antibiotic selection was initiated 48 hours after transfection. Polyclonal transfected cells were trypsinized and passaged for use in subsequent experiments.

In vivo transposition assays in mammalian cells: The eGFP+ CHO cell line was transfected with a pHP transposon donor plasmid and the pHdCas9-mammalian expression plasmid. Transfections were performed on cells at 70% confluence in 6-well plates using 12 uL of Lipofectamine 2000 and 1250 ng of each plasmid. In the transposition negative control, the pHP-M1-M2 plasmid was transfected without the pHdCas9-mammalian plasmid. Transfection efficiencies were 40-70% based on flow cytometry measurements of mCherry expression in cells 24 hours post-transfection of control pHP-on plasmids. Antibiotic selection with puromycin (10 ug/mL) was initiated 48 hours after transfection. Cells resulting from each transfection were trypsinized after 9 days of antibiotic selection, and the whole volume was transferred into a single well on a 12-well plate and grown for 4 more days in puromycin media. During the 13 days of

antibiotic selection, media was changed every 24 hours. Post-selection cells were trypsinized and diluted 1:5 in fresh media and analyzed for GFP and mCherry fluorescence on a Guava easyCyte flow cytometer (Millipore). Gates for mCherry and GFP fluorescence were set using mCherry-/eGFP- FLP-In CHO cells, mCherry-/eGFP+ CHO cells, and mCherry+/eGFP- transposition positive control CHO cells.

Genomic DNA from trypsinized cells was extracted using the Wizard Genomic DNA Purification Kit (Promega) for PCR analysis. qPCR for transposon-gDNA junctions was performed as described above, using primers p933 and p946. The control PCR to normalize for DNA input was performed using primers p931 and p932. 10 ng of purified gDNA per sample was used as PCR template.

Buffer compositions

Protein Resuspension Buffer (PRB): 20 mM Tris-HCl pH 8.0, 10 mM imidazole, 300 mM NaCl, 10% v/v glycerol. 1 tablet of cOmplete™, Mini, EDTA-free Protease Inhibitor Cocktail (Roche) was dissolved in 10 mL of buffer immediately before use.

Protein Wash Buffer (PWB): 20 mM Tris-HCl pH 8.0, 30 mM imidazole, 500 mM NaCl, 10% v/v glycerol

Protein Elution Buffer (PEB): 20 mM Tris-HCl pH 8.0, 500 mM imidazole, 500 mM NaCl, 10% v/v glycerol

Dialysis Buffer 1 (DB1): 25 mM Tris-HCl pH 7.6, 200 mM KCl, 10 mM $MgCl_2$, 2 mM DTT, 10% v/v glycerol

Dialysis Buffer 2 (DB2): 25 mM Tris-HCl pH 7.6, 200 mM KCl, 10 mM $MgCl_2$, 0.5 mM DTT, 10% v/v glycerol

10x Annealing Buffer: 100 mM Tris-HCl pH 8.0, 1M NaCl, 10 mM EDTA (pH 8.1)

ACKNOWLEDGEMENTS

We thank Saeed Tavazoie, Alexandra Ketcham, and Anupama Khare for guidance on transposon sequencing. Eric Greene, Justin Steinfeld, and Chu Jian Ma provided materials and guidance on *in vitro* protein expression and purification. We thank members of the Wang lab for helpful scientific discussions and feedback. Ross McBee, Andrew Kaufman, and Liyuan Liu assisted with mammalian cell transfections. H.H.W. acknowledges funding from ONR (N00014-15-1-2704), DARPA (W911NF-15-2-0065), NIH (1DP5OD009172), and Burroughs Wellcome PATH (1016691). S.P.C. is supported by a NIDDK F30 fellowship (F30 DK111145-01A1) and a NIH MSTP training grant (NIH T32GM007367).

AUTHOR CONTRIBUTIONS

S.P.C. and H.H.W. designed the study. S.P.C. performed the experiments. S.P.C. and H.H.W. analyzed the data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

A provisional patent application has been filed by The Trustees of Columbia University in the City of New York based on this work. The authors declare no additional competing financial interests.

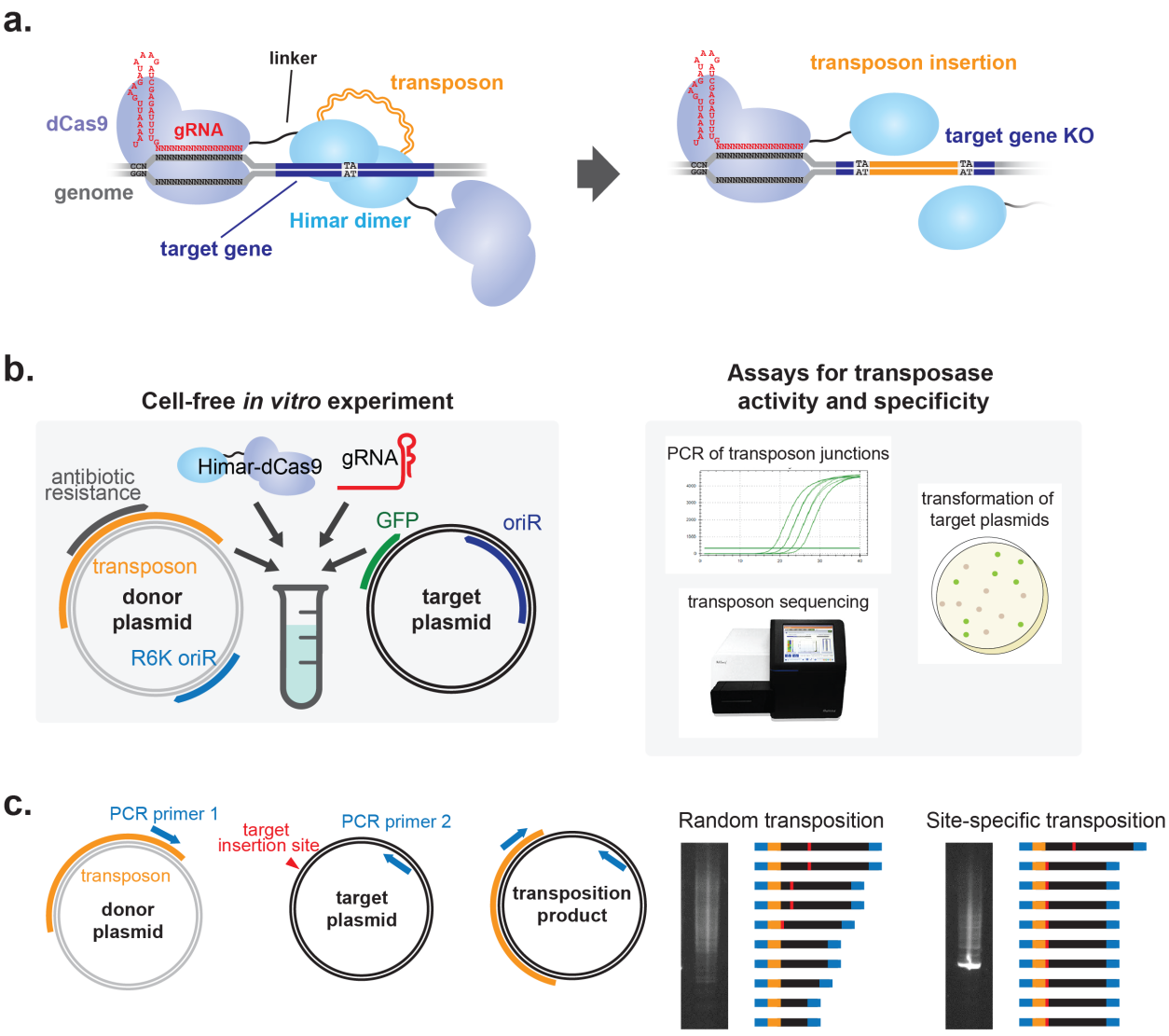


Figure 2.1. Schematics of *in vitro* Cas-Transposon test system. (a) Overview of Himar1-dCas9 protein function. The Himar1-dCas9 fusion protein is guided to the target insertion site by a gRNA, where it is tethered by the dCas9 domain. The Himar1 domain dimerizes with that of another fusion protein to cut-and-paste a Himar transposon into the target gene, which is knocked out in the same step. **(b)** Implementation of Cas-Transposon system *in vitro*. Transposon donor and target plasmids are mixed with purified protein and gRNA. Post-transposition target plasmids are analyzed by PCR for plasmid/transposon junctions, transformation and colony analysis, and transposon sequencing. **(c)** Schematic of target plasmid/transposon junction PCR. The PCR is performed using primer 1, which binds the transposon, and primer 2, which binds the target plasmid. Site-specific transposition should result in an enrichment for a PCR product corresponding with the expected transposition product.

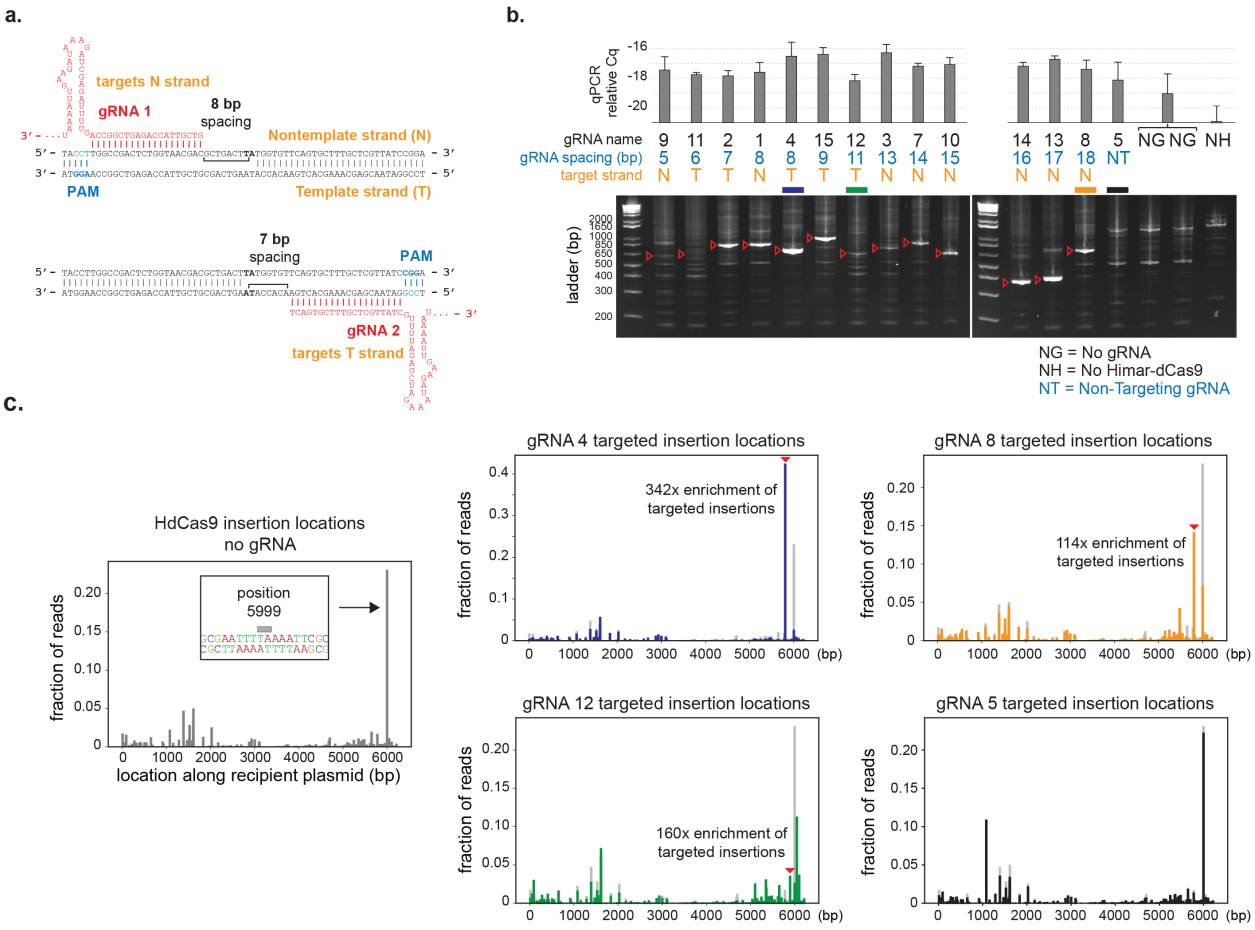


Figure 2.2. HdCas9 specificity is dependent on gRNA spacing and target site. (a) Illustration of gRNA strand orientation and spacings to TA insertion site. **(b)** PCR analysis of transposon/target junctions from *in vitro* reactions containing 30 nM HdCas9/gRNA complex, 2.27 nM transposon donor DNA, and 2.27 nM target DNA. Reactions (n=3) were run using gRNAs with spacings between 5 and 18 bp from the TA insertion site. Non-targeting gRNA (gRNA 5), no gRNA, and no transposase controls were also performed. Red arrows indicate expected site-specific PCR products for each gRNA. Error bars indicate standard deviation. **(c)** Transposon sequencing results for reactions with no gRNAs (left, gray, n=4) or with gRNAs 4 (blue, n=3), 8 (orange, n=3), 12 (green, n=3), or 5 (black, n=3). The baseline random distribution of transposons along the recipient plasmid in each panel with a gRNA is shown in light gray.

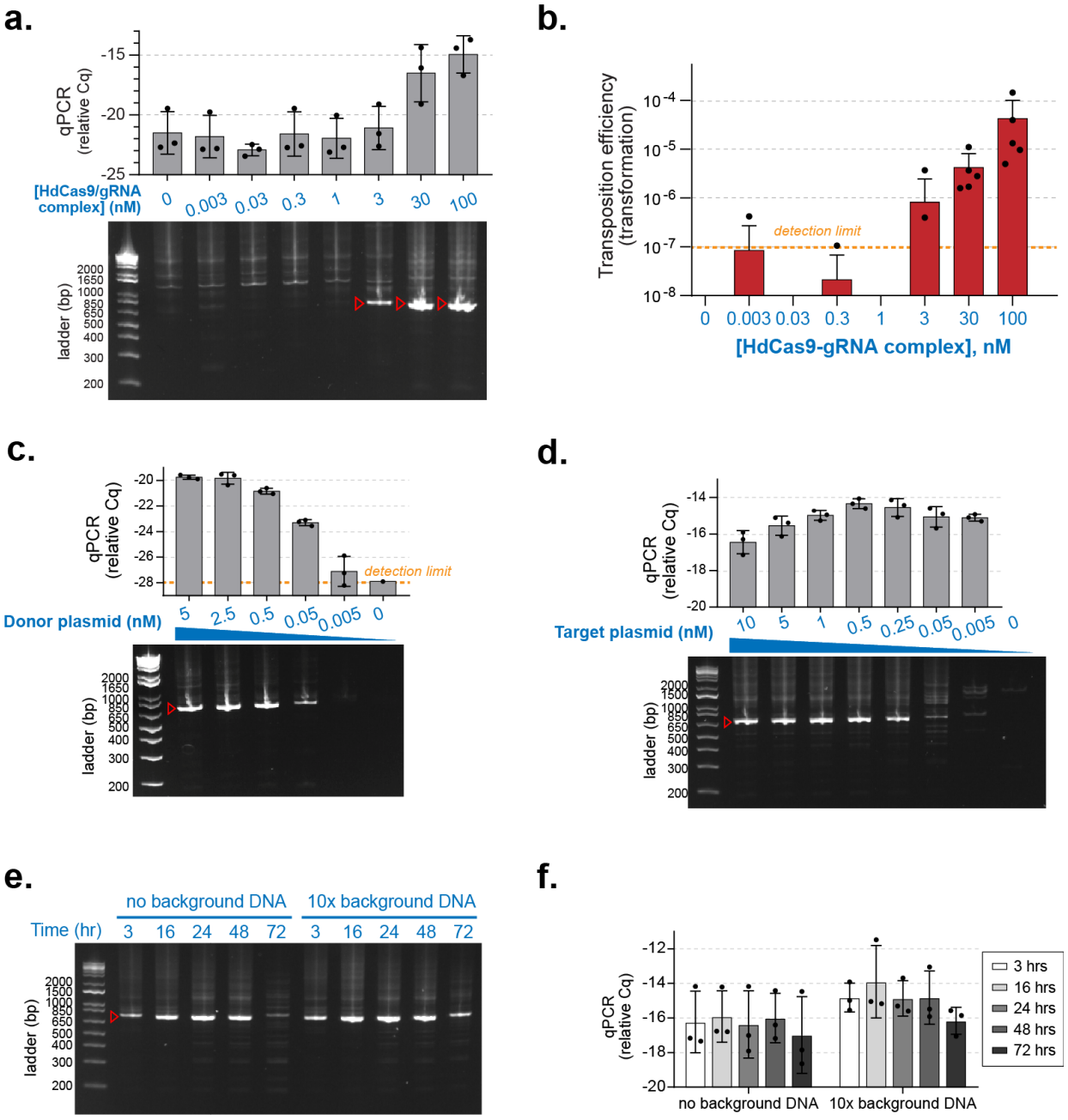


Figure 2.3. HdCas9-mediated site-specific transposition is robust to changes in ribonucleoprotein complex and DNA concentration. (a) PCR analysis of transposition reactions (n=3) using varying levels of HdCas-gRNA 4 complexes. Reactions were performed for 3 hours at 30C with 5 nM of donor and recipient plasmid DNA. **(b)** Transformation assay to measure transposition rates in reactions using varying levels of HdCas-gRNA 4 complexes. Reactions were performed for 3 hours at 30C with 5 nM of donor and recipient plasmid DNA. 3 independent reactions were performed per concentration, with 1-2 transformation replicates per reaction. **(c)** PCR analysis of transposition reactions (n=3) using varying levels of donor plasmid DNA. Reactions were performed for 3 hours at 30C with 5 nM of recipient plasmid DNA and 30

nM HdCas-gRNA 4 complex. **(d)** PCR analysis of transposition reactions (n=3) using varying levels of recipient plasmid DNA. Reactions were performed for 3 hours at 30C with 0.5 nM of donor plasmid DNA and 30 nM HdCas-gRNA 4 complex. **(e)** PCR analysis of transposition reactions (n=3) performed for different lengths of time, in the presence or absence of background non-specific DNA. Reactions were performed at 37C with 1 nM of recipient plasmid DNA, 1 nM of donor plasmid DNA, and 100 nM HdCas-gRNA 4 complex. Background *E. coli* genomic DNA was present at 10x the mass of recipient plasmid DNA. **(f)** qPCR measurement of transposition efficiency in reactions shown in panel (e). n=3 for each reaction condition.

In all panels, red arrows indicate the expected PCR product for gRNA 4, and error bars indicate standard deviation. Recipient plasmids were pGT-B1 and donor plasmids were pHimar6.

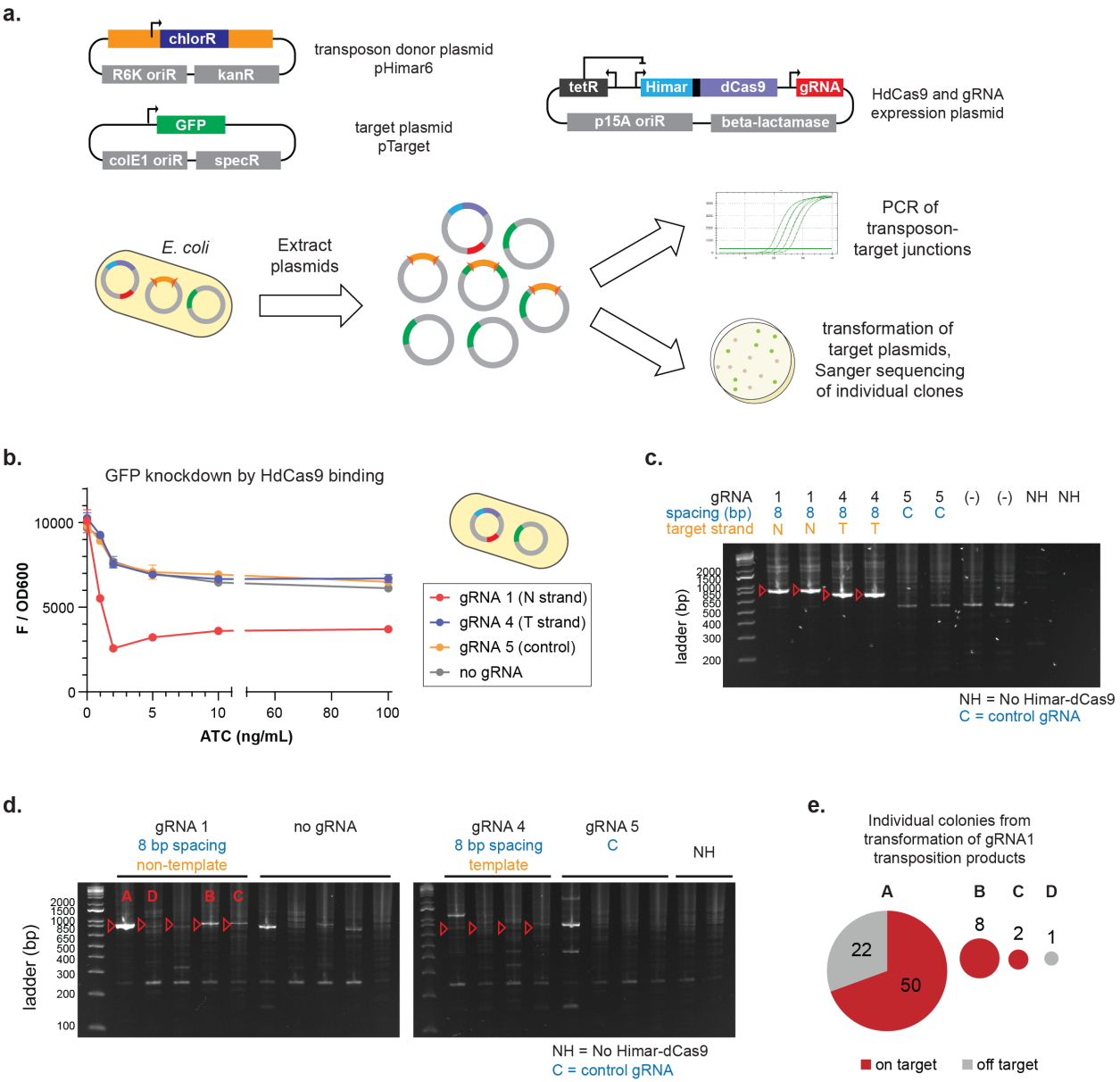
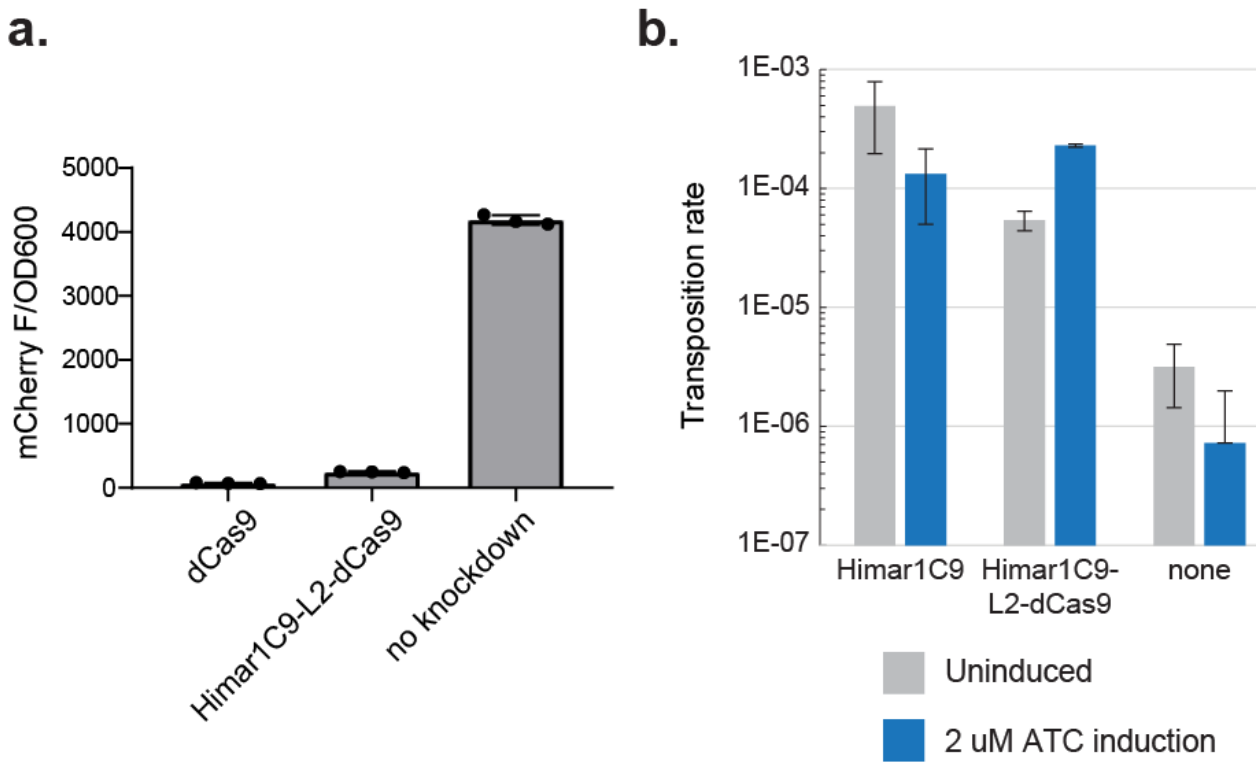
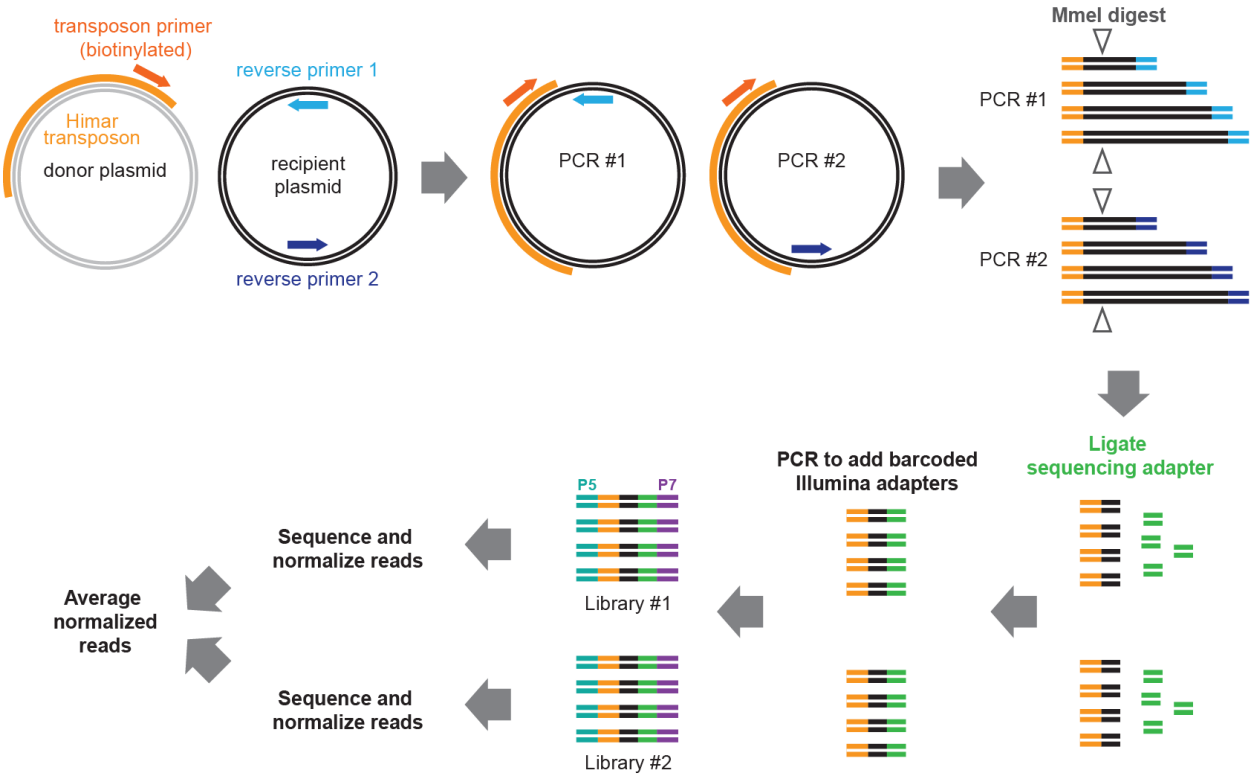


Figure 2.4. HdCas9 performs site-specific transpositions into plasmids in *E. coli*. (a) Three plasmids were transformed into S17 *E. coli* to create a testbed for HdCas9 transposition specificity *in vivo*. Post-transposition plasmids were extracted from the bacteria and analyzed by PCR and by transformation into competent *E. coli* with Sanger sequencing of plasmids from individual colonies. **(b)** HdCas9 knocks down GFP expression from the pTarget plasmid *in vivo* in *E. coli* with gRNA 1, which targets the non-template strand (N) of the GFP gene. HdCas9 does not knock down GFP fluorescence when expressed with a gRNA complementing the template strand (T) or with a non-targeting gRNA (NT) or no gRNA. n=2 per gRNA and ATC concentration. **(c)** PCR assay of *in vitro* transposition reactions using donor plasmid pHimar6 and recipient plasmid

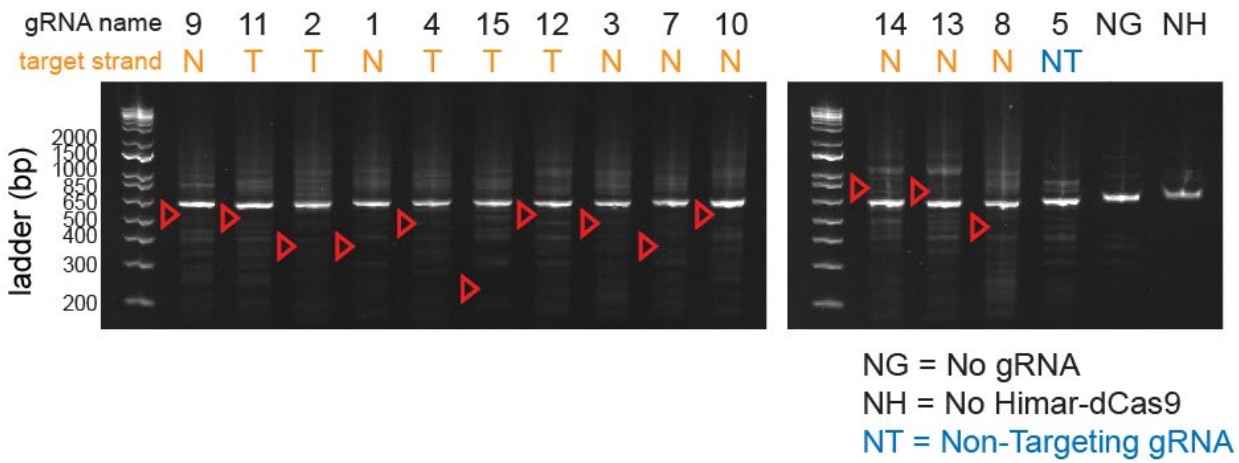
pTarget. 2.27 nM of donor and recipient plasmids along with 30 nM HdCas9-gRNA complex were incubated for 3 hours at 30C. Expected PCR products of targeted insertions are shown with red arrows. **(d)** PCR analysis of pTarget-transposon junctions resulting from *in vivo* transposition in bacteria. 3/5 of gRNA 1 PCR products show enrichment for the targeted insertion product. Transpositions A, B, C, and D with gRNA 1 were also analyzed by transformation and colony analysis. **(e)** Plasmid pools from 4 independent *in vivo* transposition experiments using gRNA 1 were transformed into *E. coli*, and the resultant colonies were analyzed by PCR and Sanger sequencing. The pie charts show the number of colonies containing on-target and off-target transposition products from each plasmid pool, with chart area proportional to the total number of colonies.



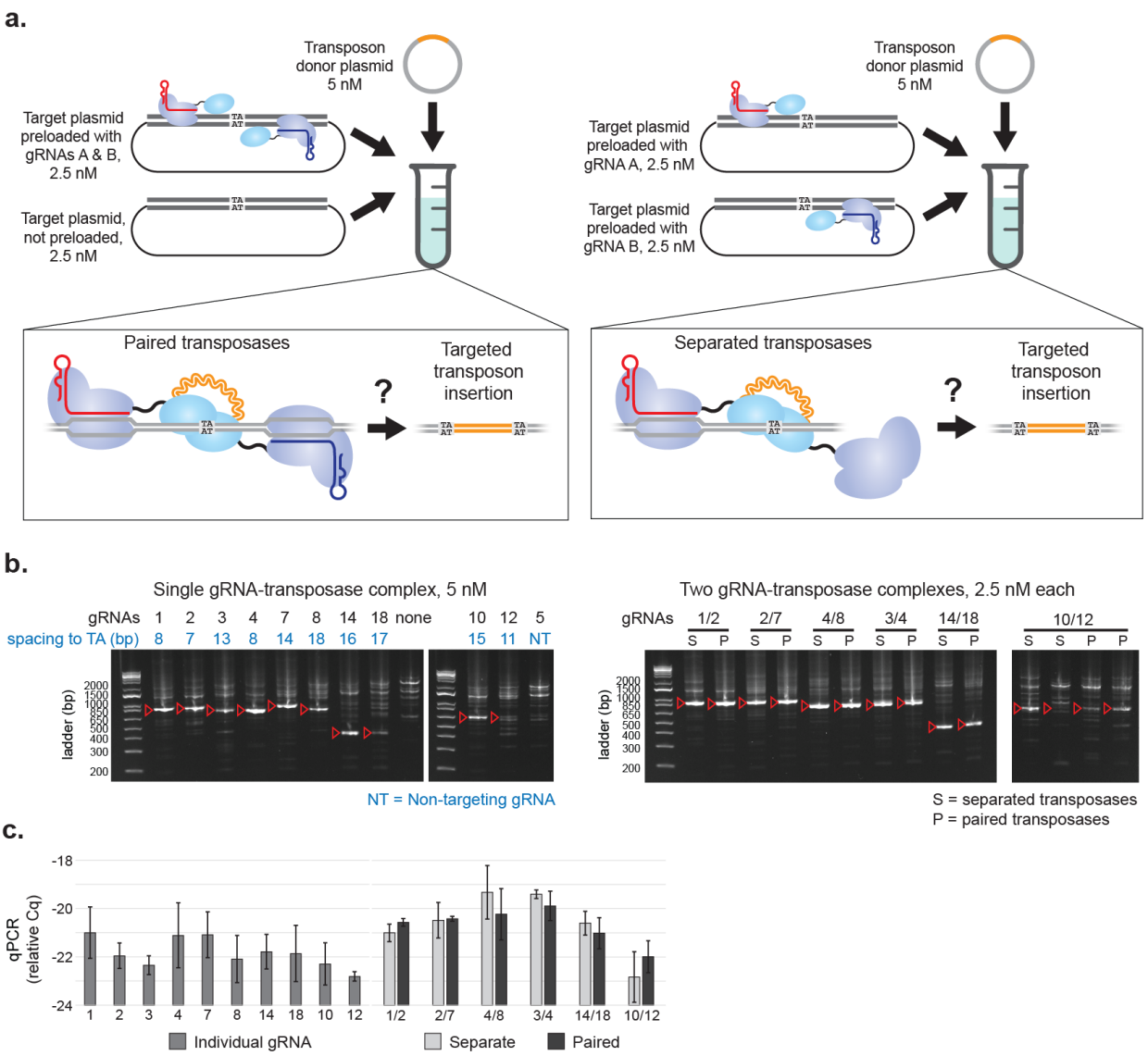
Supplementary Figure S2.1. Himar1-dCas9 fusions retain DNA binding and transposition functionalities. (a) dCas9 and Himar1C9-L2-dCas9 were expressed in MG1655 *galK::mCherry-specR E. coli* with gRNAs 5 and 16. Protein expression was induced with 2 uM ATC. Both proteins decreased mCherry expression, indicating that the Himar1C9-L2-dCas9 protein bound the mCherry gene specified by the gRNAs and blocked transcription. **(b)** Himar1C9 and Himar1C9-L2-dCas9 were measured for transposition activity in an *E. coli* conjugation assay. Both Himar1C9 and Himar1C9-L2-dCas9 mediated transposition at higher rates than the no transposase control.



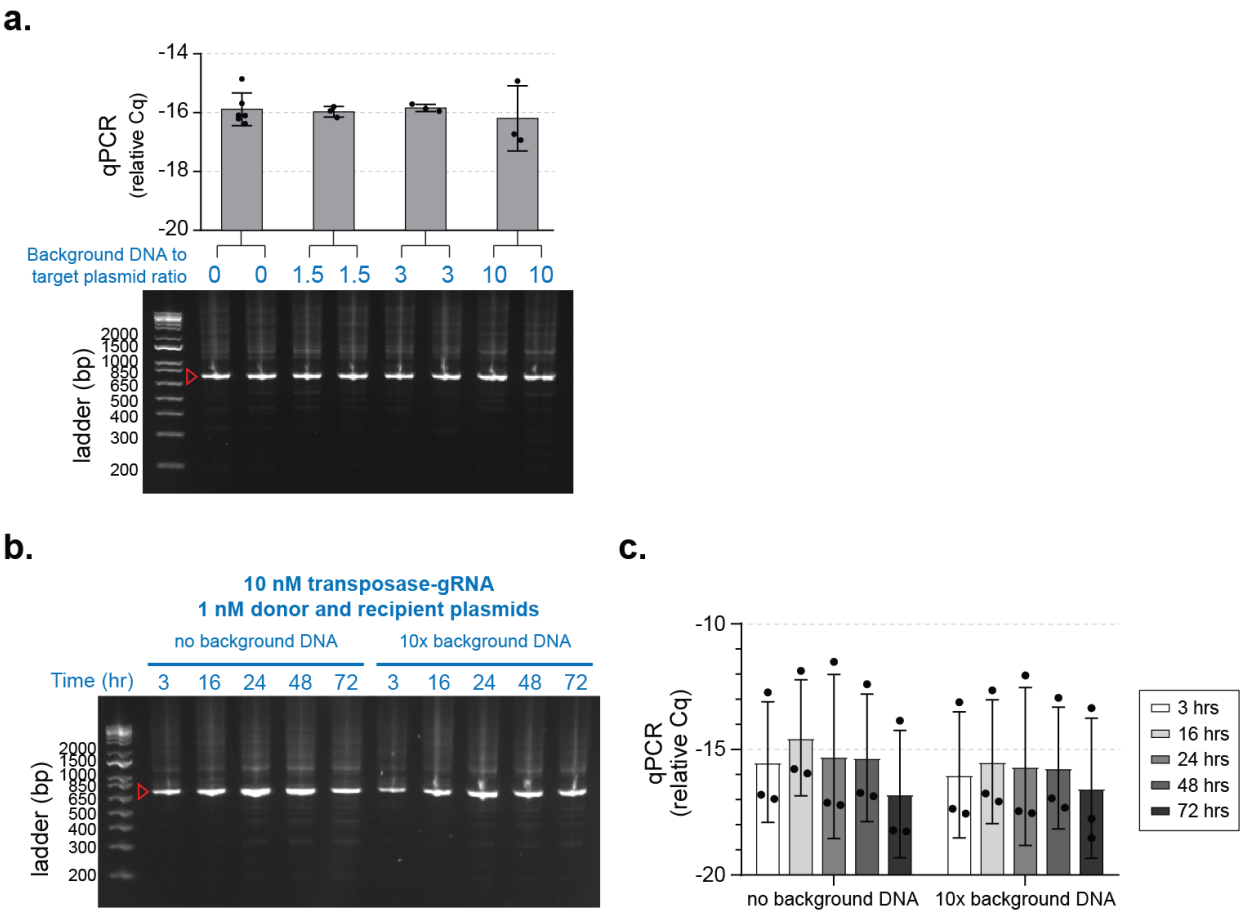
Supplementary Figure S2.2. Workflow for transposon sequencing library preparation from *in vitro* transposition reactions. To selectively isolate transposons that had become integrated into recipient plasmid for sequencing, we performed PCRs using a biotinylated primer complementing the transposon end and reverse primers complementing the recipient plasmid. Two PCRs using reverse primers on opposite sides of the recipient plasmid were performed to account for PCR size bias during amplification of transposon junction products. PCR products were isolated using streptavidin beads and digested with MmeI to isolate transposon ends with a ~17bp overhang. A sequencing adapter was ligated, and the DNA was PCR-amplified to add barcoded Illumina adapters. The resulting libraries from each PCR were sequenced independently and normalized for total reads, and the normalized libraries were averaged to obtain transposon insertion frequencies into each locus on the plasmid.



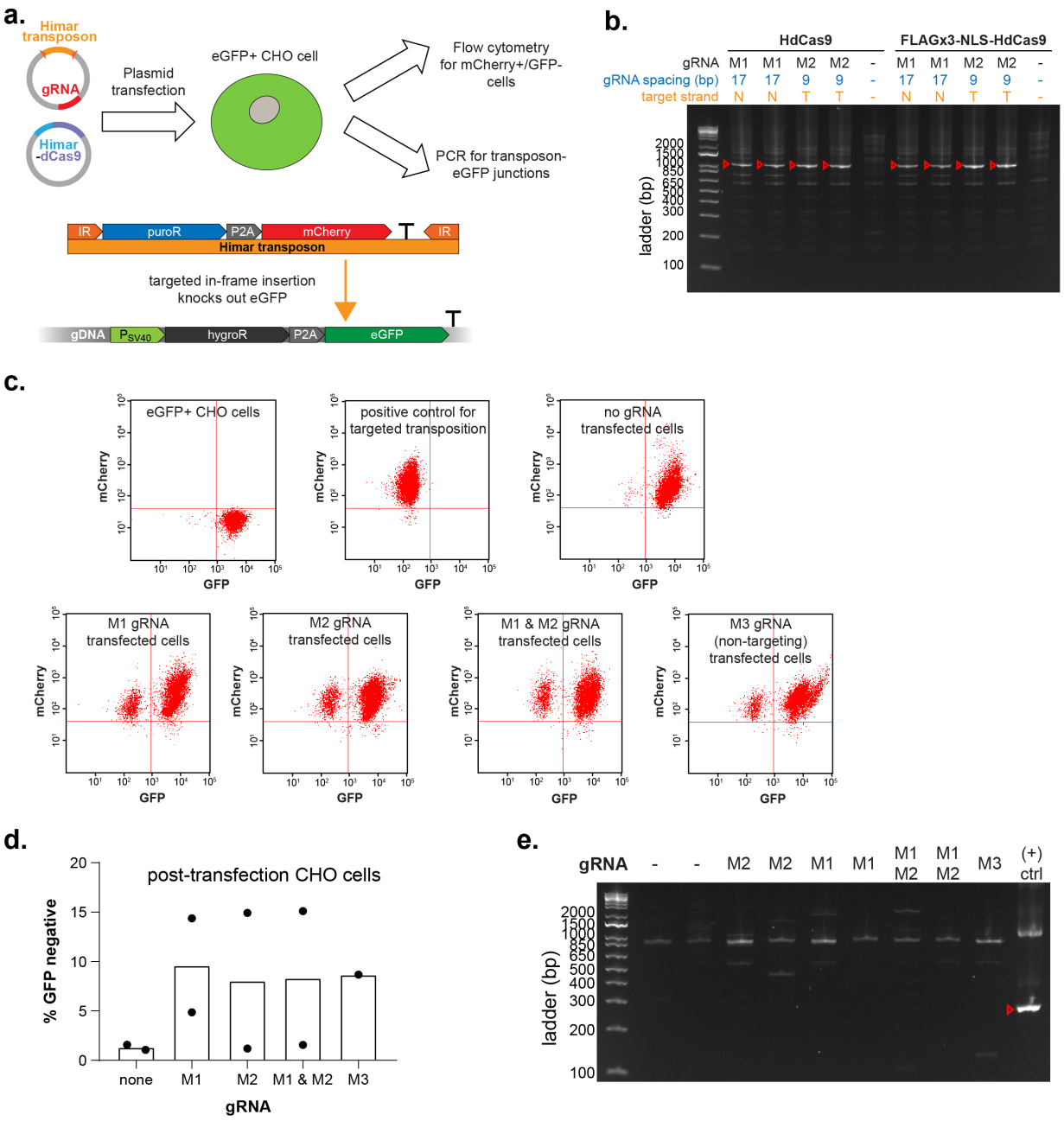
Supplementary Figure S2.3. PCR assay shows no evidence of *in vitro* targeted transposition using a Tn5-dCas9 fusion protein. Purified Tn5-dCas9-gRNA complexes (30 nM) with hyperactive E54K and L372P mutations were incubated with a Tn5 transposon donor plasmid (2.27 nM) and recipient vector pGT-B1 (2.27 nM) for 3 hours at 37C. Red arrows indicate expected PCR products for site-specific transposon insertions.



Supplementary Figure S2.4. ***In vitro*** **assay to analyze transposition by HdCas9 with 2 gRNAs. (a)** *In vitro* reactions containing 2 gRNAs were set up in 2 configurations to determine whether paired HdCas9 proteins bound at the same TA site would improve Himar dimerization and transposition activity, compared with HdCas9 proteins all bound individually to target plasmids. HdCas9 was first incubated with either gRNA A (red) or gRNA B (blue), and then the HdCas9-gRNA complexes were preloaded onto target plasmids as pairs (left) or as single complexes (right). Preloaded target plasmid-HdCas9-gRNA complexes were then mixed with transposon donor plasmids. The total final concentration of each protein-gRNA complex was 2.5 nM, and final concentrations of donor and target DNAs were 5 nM. **(b)** PCR analysis of transposition by HdCas9 with a single gRNA (left) or HdCas9 with 2 gRNAs (right), preloaded in separated (S) or paired configurations (P). Red arrows indicate expected PCR products for each reaction. **(c)** qPCR analysis of transposition by HdCas9 with a single gRNA, HdCas9 with 2 gRNAs (in a separated configuration), and HdCas9 with 2 gRNAs (in a paired configuration). n = 2-6 reactions per condition.



Supplementary Figure S2.5. HdCas9 performs *in vitro* site-specific transposition in the presence of background DNA. (a) PCR analysis of transposition reactions (n=3-6) with varying levels of background *E. coli* genomic DNA. Reactions were performed for 3 hours at 30C with 1 nM of target plasmid DNA, 1 nM of donor plasmid DNA, and 10 nM HdCas-gRNA 4 complex. Ratios of background to target plasmid DNA are by mass. **(b)** PCR analysis of transposition reactions (n=3) performed for different lengths of time, in the presence or absence of background non-specific DNA. Reactions were performed at 37C with 1 nM of recipient plasmid DNA, 1 nM of donor plasmid DNA, and 10 nM HdCas-gRNA 4 complex. Background *E. coli* genomic DNA was present at 10x the mass of recipient plasmid DNA. **(c)** qPCR measurement of transposition efficiency in reactions shown in panel (b). n=3 for each reaction condition.



Supplementary Figure S2.6. HdCas9 was not observed to target transposon insertions into a genomic locus in Chinese hamster ovary (CHO) cells. (a) eGFP+ CHO cells were transfected with an expression vector for HdCas9 and a transposon donor vector with expression constructs for gRNAs targeting the eGFP gene. The transposon contained a promoterless puromycin resistance gene and mCherry gene, which would both be expressed if the transposon integrated into the correct target site on eGFP. Puromycin-resistant cells resulting from transfection were analyzed by flow cytometry and PCR for transposon-target junctions. **(b)** PCR assay of *in vitro* transposition reactions with HdCas9 and eGFP-targeting gRNAs, using donor plasmid pHimar6 and recipient plasmid pZE41-eGFP. 2.27 nM of donor and recipient plasmids

along with 30 nM HdCas9-gRNA complex were incubated for 3 hours at 37C. Expected PCR products of targeted insertions are shown with red arrows. gRNAs M1 and M2 target the same insertion site. **(c)** Representative flow cytometry dot plots for transfected cells after 13 days of puromycin selection. A transposase-free control transfection did not produce viable cells and was not analyzed by flow cytometry. **(d)** Upon flow cytometry, 5-15% of cells in some transfections were GFP-negative. **(e)** PCR for eGFP-transposon junctions in genomic DNA resulting from *in vivo* transposition did not show evidence of site-specific transposition. The positive control PCR used a plasmid with the transposon cloned into the target site of eGFP as template. The red arrow indicates the expected size of the targeted transposition product, which is the same for gRNAs M1, M2, and M1 + M2.

Table 2.1. Plasmids used in this study.

Plasmid	Origin of replication	Size (bp)	Selection	Features	Purpose
pET-Himar-dCas9	ROP	10864	carb	6xHis tag, T7 promoter	HdCas9 protein purification
pGT-B1	pBBR1	6235	carb	constitutive sfGFP gene	target plasmid for in vitro assays
pHimar6	R6K	3394	kan	Himar transposon with chlor resistance cassette, RP4 oriT	Himar transposon donor plasmid for in vitro and E. coli in vivo assays
pTarget	ColE1	3237	spec	constitutive sfGFP gene	target plasmid for E. coli in vivo assays
pZE41-eGFP	ColE1	3154	spec	eGFP on pL-tetO promoter	target plasmid for in vitro testing of mammalian gRNAs
pHdCas9-gRNA1	p15A	8200	carb	HdCas9 on tet-inducible promoter, constitutively expressed gRNA 1	bacterial expression vector for HdCas9 and gRNA 1
pHdCas9-gRNA1-gRNA2	p15A	8392	carb	HdCas9 on tet-inducible promoter, constitutively expressed gRNA 1 and gRNA 2	bacterial expression vector for HdCas9, gRNA1, and gRNA 2
pHdCas9-gRNA4	p15A	8200	carb	HdCas9 on tet-inducible promoter, constitutively expressed gRNA 4	bacterial expression vector for HdCas9 and gRNA 4
pHdCas9-gRNA5	p15A	8200	carb	HdCas9 on tet-inducible promoter, constitutively expressed gRNA 5	bacterial expression vector for HdCas9 and gRNA 5
pHdCas9	p15A	7738	carb	HdCas9 on tet-inducible promoter	bacterial expression vector for HdCas9
pdCas9-carb	p15A	6847	carb	dCas9 on tet-inducible promoter	bacterial expression vector for HdCas9

pHP-M1	pBR322	4161	carb	Himar transposon with promoterless puroR-mCherry cassette; gRNA M1 expressed from U6 promoter	Himar transposon donor plasmid for mammalian cells; expression vector for gRNA M1
pHP-M2	pBR322	4161	carb	Himar transposon with promoterless puroR-mCherry cassette; gRNA M2 expressed from U6 promoter	Himar transposon donor plasmid for mammalian cells; expression vector for gRNA M2
pHP-M1-M2	pBR322	4540	carb	Himar transposon with promoterless puroR-mCherry cassette; gRNAs M1 and M2 expressed from U6 promoter	Himar transposon donor plasmid for mammalian cells; expression vector for gRNAs M1 and M2
pHP-M3	pBR322	4159	carb	Himar transposon with promoterless puroR-mCherry casette; Esp3I/BsmBI cloning site for gRNAs can also be expressed as a gRNA from U6 promoter	Himar transposon donor plasmid for mammalian cells; expression vector for non-targeting gRNA M3; Golden Gate cloning vector for gRNAs
pHP	pBR322	3742	carb	Himar transposon with promoterless puroR-mCherry cassette	Himar transposon donor plasmid for mammalian cells
pHP-on	pBR322	5340	carb	Himar transposon with puroR-mCherry cassette on CMV promoter; gRNAs M1 and M2 expressed from U6 promoter	mCherry expression plamid for transfection positive control
pHdCas9-mammalian	pBR322	8727	carb	HdCas9 gene with N-terminal 3xFLAG, N-terminal SV40 NLS, and C-terminal 6xHis, expressed from CMV promoter	HdCas9 expression vector for mammalian cells

pcDNA5/FRT/Hyg-eGFP	pBR322	4795	carb	FRT site with downstream hygromycin resistance-eGFP cassette	insert HygroR-eGFP into FLP-in cell genome to create transposon insertion target site
pcDNA5/FRT/Hyg-Himar	pBR322	6601	carb	FRT site with downstream hygromycin resistance-eGFP cassette; eGFP is interrupted by puroR-mCherry transposon	create positive control FLP-in cell line containing site-specific integration of Himar transposon

Table 2.2. gRNA sequences used in this study.

gRNA name	Sequence	Target gene	Target strand (T/N)	Spacing to TA site (bp)
1	GTCGTTACCAGAGTCGGCCA	sfGFP	N	8
2	TCAGTGCTTTGCTCGTTATC	sfGFP	T	7
3	CGTTCCTGCACATAGCCTTC	sfGFP	N	13
4	CGGCACGTACAAAACGCGTG	sfGFP	T	8
5	GTCGGCGGGGTGCTTCACGT	mCherry	N	10
7	ACCAGAGTCGGCCAAGGTAC	sfGFP	N	14
8	CTGCACATAGCCTTCCGGCA	sfGFP	N	18
9	CAATGCCTTTCAGCTCAATG	sfGFP	N	5
10	CAGCTCAATGCGGTTTACCA	sfGFP	N	15
11	GTAAACCGCATTGAGCTGAA	sfGFP	T	6
12	CAATATCCTGGGCCATAAGC	sfGFP	T	11
13	AGAACAGGACCATCACCGAT	sfGFP	N	17
14	GTGCTCAGATAGTGATTGTC	sfGFP	N	16
15	GAACTGGATGGTGATGTCAA	sfGFP	T	9
16	CCTTCCCCGAGGGCTTCAAG	mCherry	T	12
18	ACGCGATCACATGGTTCTGC	sfGFP	T	17
M1	GACCAGGATGGGCACCACCC	eGFP	N	17
M2	CAAGTTCAGCGTGTCCGGCG	eGFP	T	9
M3	GAGACGATTAATGCGTCTC	-	-	-

T indicates that the gRNA is complementary to the Template strand of the gene, while N indicates that the gRNA complements the Non-template strand. gRNAs that target the same TA insertion site are labeled with the same color. gRNAs 11, 13, and 15 all target different sites uniquely. M3 is a synthetic random non-targeting gRNA.

Table 2.3. Oligonucleotides used in this study.

Name	Sequence (5' -> 3')	Target	Tm (C)	Function
p433	CGCTTACAATTTC CATTCGCCATTC	pGT-B1	67	qPCR for Himar transposon-pGT-B1 junction
p415	CCCTGCAAAGCC CCTCTTTACG	pHimar6 transposon	71	qPCR for Himar transposon-pGT-B1 junction
p828	CTGCGCAACCCA AGTGCTAC	pGT-B1	70	Control qPCR for pGT-B1
p829	CAGTCCAGAGAA ATCGGCATTCA	pGT-B1	67	Control qPCR for pGT-B1
p923	Biotin/GCCATAAA CTGCCAGGCATC AA	pHimar6 transposon	68	In vitro transposon sequencing library preparation
p922	CCTTCTTGCGCA TCTCACG	pGT-B1	67	In vitro transposon sequencing library preparation
Adapter_T	Phosphate/AGATC GGAAGAGCACAC GTCTGAACTCCA GTCAC			Anneal to make Y-shapted adapter for Tn-seq library prep
Adapter_B	GTCTCGTGGGCT CGGGCTCTTCCG ATCT*N*N			Anneal to make Y-shapted adapter for Tn-seq library prep
p790	AATGATACGGCG ACCACCGAGATC TacacTAGATCGC CGCCagaccgggga cttatcatccaacctgt	Himar transposon IR	73	Add barcode & P5 sequence to Himar transposon ends for Illumina sequencing
p791	AATGATACGGCG ACCACCGAGATC TacacCTCTCTATC GCCagaccggggactt atcatccaacctgt	Himar transposon IR	73	Add barcode & P5 sequence to Himar transposon ends for Illumina sequencing
p792	AATGATACGGCG ACCACCGAGATC TacacTATCCTCTC GCCagaccggggactt atcatccaacctgt	Himar transposon IR	73	Add barcode & P5 sequence to Himar transposon ends for Illumina sequencing
p793	AATGATACGGCG ACCACCGAGATC TacacAGAGTAGA CGCCagaccgggga cttatcatccaacctgt	Himar transposon IR	73	Add barcode & P5 sequence to Himar transposon ends for Illumina sequencing
p794	AATGATACGGCG ACCACCGAGATC TacacGTAAGGAG CGCCagaccgggga cttatcatccaacctgt	Himar transposon IR	73	Add barcode & P5 sequence to Himar transposon ends for Illumina sequencing

p795	AATGATACGGCG ACCACCGAGATC TacacACTGCATA CGCCagaccgggga cttatcatccaacctgt	Himar transposon IR	73	Add barcode & P5 sequence to Himar transposon ends for Illumina sequencing
p712	CGCCagaccgggga cttatcatccaacctgt	Himar transposon IR	67	Read 1 primer for Illumina sequencing
p713	CGGAAGAGCCCG AGCCCACGAGAC	Himar sequencing library	67	Index 1 primer for Illumina sequencing
p898	TTTGAGTGAGCT GATACCGCTC	ColE1 oriR	67	qPCR for Himar transposon-plasmid junctions in pTarget-sfGFP and pZE41-eGFP plasmids
p899	GAGCGGTATCAG CTCACTCAAA	ColE1 oriR	67	Control qPCR for pTarget-sfGFP and pZE41-eGFP
p900	TCCCTTAACGTG AGTTTTCGTTCC	ColE1 oriR	67	Control qPCR for pTarget-sfGFP and pZE41-eGFP
p931	GGATAACATGGC CATCATCAAGGA G	mCherry	67	Control qPCR for Himar transposons in mammalian gDNA
p932	CCTCGTTGTGGG AGGTGATG	mCherry	68	Control qPCR for Himar transposons in mammalian gDNA
p933	AGGTCCTGTGAG CAAGGG	HygroR/eGFP cassette	67	qPCR for Himar transposon-eGFP junctions in mammalian gDNA
p946	AGAGTTCTTGCA GCTCGGTG	mammalian Himar transposon	68	qPCR for Himar transposon-eGFP junctions in mammalian gDNA
p858	CGACTACGCACT AGCCAAC	Tn5 transposon	66	qPCR for Tn5 transposon-pGT-B1 junction
p859	CAGGCTACGATA CGGAAGGTT	pGT-B1	66	qPCR for Tn5 transposon-pGT-B1 junction

Chapter 3: Metagenomic engineering of the mammalian gut microbiome *in situ*

Carlotta Ronda[1,*], Sway P. Chen[1,2,*], Vitor Cabral[1,*], Stephanie J. Yaung[3], Harris H. Wang[1,4,#]

4. Department of Systems Biology, Columbia University Medical Center, New York, USA
5. Integrated Program in Cellular, Molecular and Biomedical Studies, Columbia University Medical Center, New York, USA
6. Roche Sequencing Solutions, Pleasanton, CA, USA
7. Department of Pathology and Cell Biology, Columbia University Medical Center, New York, USA

* These authors have contributed equally

Correspondences should be addressed to H.H.W. (hw2429@columbia.edu).

3.1 ABSTRACT

Engineering microbial communities in open environments remains challenging. Here, we describe a platform to identify and modify genetically tractable mammalian microbiota by engineering community-wide horizontal gene transfer events *in situ*. With this approach, we demonstrate that diverse taxa in the murine gut microbiome can be modified directly with a desired genetic payload. *In situ* microbiome engineering in living animals enables introduction of novel capabilities into established communities in their native milieu.

3.2 MAIN TEXT

In nature, microbes live in open, dynamic and complex habitats that can be difficult, if not impossible, to recapitulate in a laboratory setting. Recent advances in deep sequencing have shed light on the vast microbial diversity that exists in many environments, including the mammalian gut. However, our ability to genetically access and alter these microbiomes remains limited despite advances in culturomics and synthetic biology[81, 99, 133, 134]. While genetic intractability is often attributed to host immunity such as restriction-methylation[135] or CRISPR-Cas processes[136], a myriad of other factors (e.g., DNA transformation, growth state, fitness burden) may also influence gene transfer potential[137]. The inability to genetically alter a bacterium greatly limits our basic understanding of the organism and its biotechnological potential. To overcome these challenges, we devised an approach, Metagenomic Alteration of Gut microbiome by In situ Conjugation (MAGIC), to genetically modify gut microbiota in their native habitat by engineering the mobilome—the repertoire of mobile genetic elements that permeate the gut microbiome. Implementing MAGIC directly on a complex gut microbiome in its natural habitat enables genetic modification of diverse microbes and development of new modifiable microbial chasses for synthetic applications (**Figure 3.1a**).

We sought to utilize the mammalian gut as a testbed for MAGIC because it harbors a diverse microbial community that also plays a key functional role in host physiology[4]. We constructed a laboratory *Escherichia coli* donor strain that can deliver a genetic payload into target recipients by broad host-range bacterial conjugation (**Figure 3.1b**). The IncPα-family RP4 conjugation system[138], which can efficiently conjugate into both Gram-positive and Gram-negative cells, was integrated into the *EcGT1* donor genome along with a constitutively expressing mCherry-specR cassette (ΔgalK::mCherry-specR). To strengthen biocontainment of the donor and to facilitate *in vitro* selection of recipients, an alternative strain *EcGT2* (Δasd::mCherry-specR) was generated to be

auxotrophic for the essential cell wall component diaminopimelic acid (DAP), thus requiring DAP supplementation in the growth media[139].

We developed a modular suite of mobile plasmids (*pGT*) that featured replicative origins with narrow to broad host-ranges, a RP4 transfer origin, a selectable marker, and the desired genetic payload (**Suppl. Tables 3.1-3.3, Suppl. Figure S3.1**). A broad host-range Himar transposon system was also utilized for delivering integrative payloads. As a demonstration of the system, we used a dual-reporter payload harboring a green fluorescent protein (GFP) and an antibiotic resistance gene (AbR). The use of Fluorescence Activated Cell Sorting (FACS) combined with 16S metagenomic analysis enables identification of successfully modified recipients or transconjugants, which can then be readily isolated on antibiotic selective plates. This multi-pronged strategy can increase the diversity of genetically tractable microbiota that can be captured. We first validated and optimized MAGIC protocols *in vitro* by assessing the gating stringency of FACS with control spike-ins of GFP-tagged bacteria into a complex sample community (**Suppl. Figure S3.2**). Subsequently, *in vitro* conjugations with defined recipient species (**Suppl. Figure S3.3**) and live bacterial communities extracted from mouse feces (**Suppl. Figure S3.4**) demonstrated the transfer of the payload from donors to recipients to yield GFP+ transconjugants that could be enriched by FACS (**Suppl. Figure S3.5**), which were confirmed by fluorescence microscopy (**Suppl. Figure S3.6**). 16S rRNA sequencing of FACS-enriched transconjugant populations revealed a diverse range of recipient bacteria **(Suppl. Figure S3.7**).

Since the gut microbiome can vary widely between hosts and can change due to dietary or other environmental variables, we explored the possibility of implementing MAGIC *in vivo*, directly in the native gut microbiome of an animal. We hypothesized that different groups of microbiota may be modified by using a library of pGT vectors that exhibit a range of gene expression levels and plasmid replication elements suitable for different gut bacteria. Libraries of pGT vectors (pGT-L1 to pGT-L6) were generated by modularly permuting pGT parts, including regulatory sequences of

varying activity, payload selectable genes (*bla*, *catP*, *tetQ*), transposon elements (Himar), and plasmid origins (RSF1010, pBBR1, p15A) (**Suppl. Tables 3.1-3.2**). We carried out 4 separate *in vivo* studies where EcGT2 donors containing pGT libraries were orally gavaged into conventionally raised C57BL6/J mice obtained from commercial vendors. To assess the transfer capacity of individual pGT replicative or integrative designs (pBBR1, p15A-Himar, and RSF1010), we introduced each pGT library (pGT-L1, pGT-L2, or pGT-L3, respectively) separately into a mice cohort from Taconic (**Suppl. Figure S3.8**). We tested larger combinatorial libraries (pGT-L3 to pGT-L6) in two independent mice cohorts to assess variability across cohorts (**Figure 3.2**, **Suppl. Figure S3.9**). To compare *in situ* transfer in different gut communities, we tested the pGT-L6 library in mice from a different source (Charles River) (**Suppl. Figure S3.10**).

We performed FACS enrichment and 16S metagenomic analysis on fecal material from all mice studies collected over time after oral gavage of pGT libraries (**Figure 3.2a**). Across *in situ* studies, up to 5% of resulting bacteria appeared to be successful transconjugants (i.e., $GFP^+/mCherry^-$) six hours post-gavage, compared to control groups (mice gavaged with PBS or EcGT2 carrying a non-transferrable vector pGT-NT) (**Figure 3.2b, Suppl. Figures S3.8a, S3.9a, S3.10a**). These $GFP^+/mCherry^-$ transconjugants persisted for up to 72 hours post-gavage (**Figure 3.2c, Supp. Fig S3.9b**). 16S metagenomic sequencing of these transconjugant populations revealed a wide phylogenetic breadth (**Figure 3.2d, Suppl. Figures S3.8b, S3.9c, S3.10b**). Importantly, we observed significant reproducible enrichment of Proteobacteria and Firmicutes, especially Clostridiales and Bacillales, amongst successful transconjugants across multiple independent experiments. Using the same pGT-L6 library in mice from different vendors, which harbored distinct microbiomes (**Suppl. Figure S3.10c**), yielded shared and distinct transconjugants (**Suppl. Figure S3.10d**). In parallel to FACS-metagenomic studies, we isolated individual transconjugants from these fecal samples by selective plating for the payload antibiotic resistance gene and confirmed the presence of the GFP-AbR payload by PCR (**Figure 3.2e**).

Across all experiments, we isolated and validated over 297 transconjugants belonging to 19 genera across 4 phyla (**Suppl. Figure S3.11**, **Suppl. Table 3.4**), validating the capacity of MAGIC to broadly transfer genetic material *in situ* into diverse recipients in the mammalian gut. In contrast, only 7 genera could be isolated from *in vitro* conjugation experiments using the same pGT vectors despite comparable diversity of transconjugants detected by FACS-metagenomics (**Suppl. Figure S3.7**). This difference may be due to *in vitro* conditions that sub-optimally support growth of diverse species during conjugation reactions, which underscores the value of implementing MAGIC *in situ* in an established complex microbiome.

Since transconjugants were no longer detected by 72 hours *in situ* (**Figure 3.2c**, **Suppl. Figure S3.9b**), we speculated that the genetic payload (GFP-AbR) on pGT vectors might be unstable or toxic, thus causing its negative selection in transconjugants. This hypothesis was tested *in vitro* by 20-30 serial passages of two transconjugant isolates of *Escherichia fergusonii* that contained the GFP-carbR payload either on a pGT-B1 (replicative pBBR1 origin) or a pGT-Ah1 (integrative Himar transposon) plasmid (**Suppl. Figure S3.12**). For the pGT-B1 population, we observed a significant increase in the fraction of GFP(-) cells (**Suppl. Figure S3.12a-c**). PCR assay of the origin of replication indicated that the pGT-B1 plasmid was no longer present in these GFP(-) cells (**Suppl. Figure S3.12d**). In contrast, cells in the pGT-Ah1 population remained GFP(+) despite a detectable loss of the plasmid in parts of the population over time (**Suppl. Figure S3.12 e-g**), which suggests a more stable maintenance of the GFP-CarbR payload as a integrative transposon within the host genome. Together, these results highlight the challenges of maintaining long-term *in vivo* stability of engineered genetic constructs in complex microbial communities, and suggest design considerations for more precise tuning of payload life-span and for improving payload biocontainment.

Microbes in the wild are known to possess a variety of natural mobile DNA elements and conjugation systems. Whole genome sequencing of three transconjugant strains of *Proteus*

mirabilis and *Escherichia fergusonii* from our studies (designated as Modifiable Gut Bacteria MGB3, MGB4, and MBG9) revealed the presence of putative endogenous DNA mobilization systems (**Suppl. Figure S3.13a-c**). We wondered whether these native mobilization systems could interface with our engineered pGT vectors and thus performed *in vitro* conjugations of the MGB strains with laboratory *E. coli* recipients. Surprisingly, we discovered that MGB4 and MGB9 (both *E. fergusonii*) were able to mobilize pGT vectors into recipients, although at a lower efficiency than our engineered EcGT2 donor (**Figure 3.3a, Suppl. Figure S3.13d**). These results suggest that some native gut bacteria can promote secondary transfer of engineered payloads using their endogenous conjugation machinery, which may improve payload transfer *in situ*.

In general, non-gut adapted bacteria, including common probiotics, do not colonize an established gut microbiome. Infiltration of foreign species usually requires drastic perturbations, such as use of broad-spectrum antibiotics to suppress the natural flora. Even then, exogenous species do not persist upon discontinuation of antibiotic suppression[140]. Since our donor strains did not readily colonize the murine gut and transconjugants were lost soon after (**Figure 3.2c, Suppl. Figures S3.9b, S3.14a**), we reasoned that using a colonizing donor strain may extend the persistence of payload constructs *in situ*. To explore this possibility, we first tested whether a mixed population of MGB strains (MGB3, MGB4, MGB9) could stably recolonize the native murine gut after a single oral dose without any antibiotic co-administration (**Figure 3.3b**). In sharp contrast to the rapid loss of a non-gut-adapted strain (EcGT1) within 48 hours, MGB strains (especially MGB4) recolonized the murine gut and stably persisted for at least 15 days (**Figure 3.3c, Suppl. Figure S3.15a**), populating along the entire gastrointestinal tract (**Suppl. Figure S3.15b**). FACS enrichment and 16S sequencing of GFP-expressing bacteria in feces from these mice revealed transconjugants resulting from *in situ* transfer of the pGT payload from MGB strains to the native microbiome after 6 hours (**Figure 3.3d**) and even 11 days post-gavage (**Suppl. Figure S3.15c**). These transconjugant populations had similar phylogeny although less diversity than those from

prior *in situ* experiments using the non-colonizing EcGT2 donor (**Figure 3.2d, Suppl. Figure 3.9c**). Together these results highlight the utility of MAGIC to isolate host-derived engineerable strains that can be modified and then used to stably recolonize the native community and mediate further transfer of engineered functions *in situ*.

In summary, MAGIC enables metagenomic infiltration of genetic payloads into a native microbiome and isolation of genetically modifiable strains from diverse communities. These modifiable native strains can then be reintroduced into their original community to maintain engineered functions via sustained vertical and horizontal transmission *in situ*. Future improvements to the system, such as optimization of vector stability and donor strain dosage (**Suppl. Figure S3.14b**), could enable better quantitative and temporal control of retention of genetic payloads *in situ*, which may be useful in applications requiring short-term or long-term actuation of engineered functions[32, 141, 142]. Designing genetic programs based on recipient-specific properties should enhance targeted execution of desired functions only in a defined subset of species in a community[143, 144]. Beyond the gut microbiome, MAGIC and complementary strategies to engineer the horizontal gene pool can facilitate programmable execution of genetic circuits in other microbial communities[64, 95, 145, 146]. Isolation of genetically tractable representatives from diverse microbiomes will expand the repertoire of new microbial chasses for emerging applications in synthetic biology and microbial ecology.

METHODS

Methods and any associated references are available in the online version of the paper.

DATA AVAILABILITY

The raw data from this study are available from the corresponding author upon request.

MATERIALS AVAILABILITY

All modular vector part sequences are listed in **Supplementary Table 3.3**. Full plasmid maps, vectors and strains used in this study are available from the corresponding author upon request.

ANIMAL ETHICS STATEMENT

All animal experiments were performed in compliance with Columbia University Medical Center IACUC protocols AC-AAAU6464 and AC-AAAL2503.

ACKNOWLEDGEMENTS

We thank members of the Wang lab for helpful scientific discussions and feedback, especially Ravi Sheth for helpful insights on NGS library preparation. H.H.W. acknowledges funding from DARPA (W911NF-15-2-0065), NIH (1DP5OD009172, 5R01AI132403, 1R01DK118044), NSF (MCB-1453219), ONR (N00014-15-1-2704), and Burroughs Wellcome PATH (1016691). C.R. is supported by a Junior Fellowship of the Simons Society of Fellows (#527896). S.P.C. is supported by a NIDDK F30 fellowship (F30 DK111145-01A1) and a NIH MSTP training grant (NIH

T32GM007367). We also thank Amir Figueroa at the Columbia University Medical Center Flow Cytometry Core for assistance with FACS studies.

AUTHOR CONTRIBUTIONS

C.R., V.C, S.P.C., S. Y. and H.H.W. designed the study. C.R., S.P.C. and V.C performed the experiments. C.R., S.P.C., V.C and H.H.W. analyzed the data and wrote the manuscript with input from all authors.

COMPETING FINANCIAL INTERESTS

A provisional patent application has been filed by The Trustees of Columbia University in the City of New York based on this work. The authors declare no additional competing financial interests.

3.3 METHODS

Media, chemicals and reagents. *E. coli, S. enterica, V. cholera,* and *P. aeruginosa* strains were grown in rich LB-Lennox media (BD) buffered to pH 7.45 with NaOH in aerobic conditions at 37°C, while *L. reuteri* was grown in MRS media (BD). *B. thetaiotaomicron* and *E. faecalis* were grown anaerobically at 37°C in Gifu Anaerobic Modified Medium (GAM) (Nissui Pharmaceutical) or BHI media (BD) supplemented with cysteine (1 g/L), hemin (5 mg/L), resazurin (1 mg/L), and Vitamin K (1 µL/L). All gut bacteria used in the study were grown in LB-Lennox or Gifu Anaerobic Modified Medium (GAM). Antibiotics were used at the following concentrations to select for *E. coli*: chloramphenicol (chlor) at 20 µg/ml, carbenicillin (carb) at 50 µg/ml, spectinomycin (spec) at 250 µg/ml, kanamycin (kan) at 50 µg/ml, tetracycline (tet) at 25 µg/ml, and erythromycin (erm) at 25 µg/ml. Antibiotics were used at the following ranges of concentrations to select for transconjugant gut bacteria: chloramphenicol (chlor) at 5-20 µg/ml, carbenicillin (carb) at 10-50 µg/ml, tetracycline (tet) at 5-25 µg/ml. Diaminopimelic acid (DAP) was supplemented at 50 µM as needed.

Isolation of live murine gut bacteria. Fresh fecal pellets were harvested from mice, and live gut bacteria were isolated by mechanical homogenization. Briefly, 250 µL of PBS was added to previously weighed pellets in a microcentrifuge tube. Pellets were thoroughly mechanically disrupted using a motorized pellet pestle before adding 750 µL of PBS. The disrupted pellets in PBS were then subjected to four iterations of vortex mixing for 15 sec at medium speed, centrifugation at 1,000 rpm for 30 sec at room temperature, recovery of 750 µL of supernatant into a new tube, and replacement of that volume of PBS before the next iteration. The resulting 3 ml of isolated cells were pelleted by centrifugation at 4,000xg for 5 min at room temperature, the supernatant was discarded, and cells were re-suspended in 0.5-1.0 ml of PBS. All gut bacteria isolations were performed in an anaerobic chamber (Coy Labs).

Donor strain construction. Donor strains EcGT1 and EcGT2 were derived from the S17 λpir *E. coli* strain[147] by generating modifications Δ*galK::mCherry-specR* and Δ*asd::mCherry-specR*, respectively, with λ-red recombineering using the pKD46 system[148]. Synthetic cassettes containing constitutively active mCherry and spectinomycin resistance genes were constructed with ~40 bp of homology on both ends to *galK* or *asd* flanking regions on the *E. coli* genome. 100 ng of mCherry-specR cassette DNA were electroporated into recombineering-competent S17-pKD46 cells. Cells were allowed to recover in 3 mL LB+carb at 30ºC for 3 hours prior to plating on LB+spec. Spectinomycin-resistant colonies were genotyped by PCR for validation of mutations. The pKD46 recombineering plasmid was cured out of validated recombinants by growth at 37ºC in the absence of carbenicillin to yield the EcGT1 and EcGT2 strains used throughout the study. When generating the EcGT2 strain, the growth media was supplemented with DAP at all stages of the protocol.

Plasmid construction. pGT vectors were designed to have modular components (e.g., selectable markers, regulatory elements, replication origins) that are interchangeable by isothermal assembly (ITA) or Golden Gate Assembly. Vector selection markers for *E. coli* were constitutively expressed, while the deliverable cargo or transposase cassettes were expressed using different regulatory elements to enable broad-host or narrow-host range gene expression. Regulatory elements used in this study exhibit a range of activity (**Suppl. Table 3.1**). Vector libraries used in this study are detailed in **Suppl. Table 3.2**. Full vector component sequences are listed in **Suppl. Table 3.3**. The non-transferrable vector pGT-NT used as a negative control was a minimal p15A cloning vector with no origin of transfer, containing a constitutively expressed sfGFP gene.

All plasmids were constructed by isothermal assembly (ITA) with NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs). Component parts were made by high-fidelity PCR with Q5 (NEB) or KAPA Hifi (Kapa Biosystems) polymerases, using existing vectors or gBlocks (Integrated DNA Technologies) as PCR templates. PCR products were digested with DpnI (NEB) and purified with the QIAquick PCR purification kit (Qiagen) prior to ITA and transformation into *E. coli*. All assembled plasmids were Sanger sequence-verified.

In vitro MAGIC studies on synthetic recipient community. Donor strains harboring pGT vectors and representative recipients (*E. coli* MG1655, *S. enterica* ATCC 700931, *V. cholera* C9503, *P. aeruginosa* PA01, *E. faecalis* ATCC 29200, *L. reuteri* ATCC 23272, *B. thetaiotaomicron* ATCC 29148) were grown overnight in appropriate media and cultivation conditions, and a 1:1000 dilution culture was re-grown for 14 hours at 37ºC prior to conjugation studies. To prepare cells for *in vitro* conjugation, donor and recipient populations were washed twice in PBS and cells were quantified by OD_{600} or flow cytometry using SYTO9 staining (Thermo Fisher). 10^8 donor cells and 10^8 recipient cells were mixed together, pelleted by centrifugation, and re-suspended in 10µL PBS. Donor and recipient mixes were spotted on an agar plate and incubated for 5 hours at 30ºC or 37ºC for conjugation. *In vitro* conjugations were performed on LB-Lennox (*E. coli*, *S. enterica*, *V. cholera*, *P. aeruginosa, E. faecalis*), MRS (*L. reuteri*), or supplemented BHI agar (*B. thetaiotaomicron*). Post-conjugation, cells were scraped from the plate into 1 mL PBS, and 100 uL was plated on appropriate antibiotics and incubated overnight at 30ºC or 37ºC to determine the number of colony forming units (CFU) of transconjugants.

In vitro MAGIC studies on natural recipient community. Donor strains harboring pGT vectors were streaked onto LB-Lennox agar plates with appropriate antibiotics and supplements, grown at 37ºC overnight, and then grown from a single colony in 2 mL liquid media for 10 hours at 37ºC

prior to conjugation. The recipient community was isolated anaerobically from fresh murine feces as described above, immediately before conjugation. Donor cells were washed twice in PBS and quantified by OD_{600}, while recipient cells were quantified by flow cytometry using SYTO9 staining. 10^8 donor cells and 10^9 recipient cells were mixed, pelleted by centrifugation at 5000 x g, and resuspended in 25 µL PBS. The mixes were spotted on PBS + 1.5% agar plates and incubated at 37ºC either aerobically or anaerobically overnight (9-10 hours). Post-conjugation, cells were scraped from the plate into 1 mL of PBS and subjected to antibiotic selection on GAM media, FACS enrichment, and metagenomic 16S analysis (see below).

In vitro assessment of pGT vectors horizontal gene transfer mediated by natural isolates. MGB natural isolates harboring pGT vectors (MGB3, MGB9, MBG4) were conjugated with a recipient *E. coli* strain harboring a kanamycin resistance plasmid compatible with pGT vectors. Prior to conjugations, all strains were streaked onto GAM agar plates with appropriate antibiotics, grown at 37ºC overnight, and then grown from a single colony in 5 mL liquid GAM for 10 hours at 37ºC prior to conjugation. MGB donor and recipient cells were washed twice in PBS and quantified by OD_{600}. 10^9 cells each of MGB and recipient strains were mixed, pelleted by centrifugation at 5000 x g, and resuspended in 15 µL PBS. The mixtures were spotted on GAM agar plates and incubated at 37ºC aerobically for 6 hours. Post-conjugation, cells were scraped from the plate into 1 mL of PBS and plated on selective and non-selective GAM media. Conjugation efficiency was calculated as $\frac{t}{n}$, where *t* is the number of *E. coli* transconjugant CFUs and *n* is the total number of *E. coli* CFUs.

Measurement of GFP expression in MGB strains. MGB isolates harboring pGT vectors (MGB3, MGB9, MBG4) were streaked onto GAM agar plates with appropriate antibiotics, grown

at 37ºC overnight, and then diluted to OD_{600} 0.001 in liquid GAM into a 96 well plate. The plate was incubated in a Synergy H1 (BioTek) microplate reader for 24 hours at 37ºC with orbital shaking. Measurements of OD_{600} and GFP expression (excitation 488 nm, emission 510 nm) were taken using Gen5 software (BioTek) at the end of 24 hours.

In vivo MAGIC studies in mice. Conventionally raised C57BL/6 female mice (Taconic Biosciences or Charles River Laboratories) were used throughout the study. Two control groups of 4 mice each were gavaged with PBS and EcGT2 containing a non-transferable GFP vector (pGT-NT). Three to four mice were used in each group gavaged with a pGT donor mix or with MGB strains. To equilibrate the murine gut microbiome ahead of time, mice from multiple litters were mixed, co-housed for at least 1 week prior to all experiments, and randomly allocated into groups. Mice were gavaged with 10^9 donor cells (EcGT2 or MGB strains) in 300 uL of PBS at 8-10 weeks old. Control mice were gavaged with 300 uL of PBS. Fecal matter was collected immediately before gavage and periodically after gavage to analyze the resulting microbiome populations by FACS, metagenomic 16S sequencing, and plating. Upon completion of the study, mice were euthanized and small and large intestinal tissues were extracted. Luminal contents were washed from each tissue sample with PBS and bacteria were extracted by homogenization of the luminal contents for plating and final CFU determination.

Flow cytometry and fluorescence-activated cell sorting (FACS) measurements. Gut bacteria isolated from fresh fecal pellets were analyzed for evidence of successful conjugation on a flow cytometer (Guava easyCyte HT) using red (642 nm) and blue (488 nm) lasers with Red2 and Green photodiodes to detect mCherry (587/610 nm) and sfGFP (485/510 nm) fluorescence, respectively. Bacteria at 100x and 1,000x dilutions in PBS were used for optimal detection of donor (GFP^+/mCh^+), gut microbes without a transferred vector (GFP^-/mCh^-), and transconjugants

(GFP^+/mCh^-). Data were collected and analyzed usingInCyte 3.1 software. For FACS enrichment studies, a BD FACS Aria II cell sorter operated with BD FACSDiva software was used to gate for sfGFP (FITC filter 515/10nm) and mCherry (mCherry filter 616/26nm). A double gating on GFP and mCherry channels was used to select for cells with GFP^+/mCh^- fluorescence. In addition, background events were also taken into account by using the GFP^+/mCh^- fluorescence detected in the fecal sample prior to gavage as baseline signal. An increase over the baseline signifies an enrichment of transconjugants. Population density (cells/gram fecal matter) was calculated based on number of cells sorted over the mass of the sorted fecal sample. Additional plating and direct colony counting were used to validate flow cytometric measurements. FACS plots were formatted using FCS Express 6.

Fluorescence microscopy of fecal bacteria. Bacteria were suspended in PBS and centrifuged at 5000xg to concentrate into a smaller volume, which varied depending on the concentration of bacteria. The bacteria were resuspended by pipetting, and a volume of 15 uL was dropped onto a Superfrost Plus microscope slide (Thermo Shandon) and covered with a glass cover slip. Slides were air-dried until the PBS receded from the edges of the cover slip and then sealed with clear nail polish. Bacteria were imaged at 40x magnification on a Nikon Eclipse Ti2 microscope on bright field, RFP, and GFP channels using NIS-Elements-AR software.

Validation of pGT vectors in transconjugants. Transconjugant validation was performed by colony PCR of the GFP-antibiotic resistance payload and/or the pGT vector backbone. PCR products with the expected size were further verified by Sanger sequencing. Taxonomy assignment of isolated colonies was based on 16S rRNA PCR amplification and Sanger sequencing. All transconjugant strains validated in the study are listed in **Suppl. Table 3.4**.

In vitro evolution of transconjugant gut bacteria. *Escherichia fergusonii* transconjugants MGB4 and MGB9 were serially passaged in LB media for 11-15 days. Starting from a single colony, the strains were inoculated into LB and grown at 37C with shaking. Every 12 hours the liquid culture was diluted 1:1000 into fresh LB media. At selected time points an aliquot of the saturated culture was plated on selective (50 μg/mL carbenicillin) and non-selective plates to quantify the percentage of cells expressing the payload antibiotic resistance and GFP genes. MGB9 cultures were also plated on selective plates with 20 μg/mL chloramphenicol to check for maintenance of the plasmid backbone.

Metagenomic 16S sequencing. Genomic DNA was extracted from isolated bacteria populations using the MasterPure Gram Positive DNA Purification Kit (Epicentre). PCR amplification of the 16S rRNA V4 region and multiplexed barcoding of samples were performed based on previous protocols[149]. The V4 region of the 16S rRNA gene was amplified using customized primers based on the method described in Kozich et al.[149] with the following modifications: (i) alteration of 16S primers to match updated EMP 505f and 806rB primers[150-152] and (ii) use of NexteraXT indices such that each index pair is separated by a Hamming distance of >2 and Illumina low-plex pooling guidelines can be used. Sequencing was performed using the Illumina MiSeq system (500V2 kit).

Analysis of 16S next-generation sequencing (NGS) data. Bacteria from fecal samples taken right before gavage (T0) and 6 hours post-gavage (T6) and were sorted by FACS to enrich for transconjugants. The compositions of the sorted transconjugant and total populations for each sample were determined from 16S sequencing data using the UPARSE pipeline[153] (USEARCH version 10.0.240) to generate Operational Taxonomic Unit (OTU) tables and abundances and the

RDP classifier[154] to assign the taxonomy. Phylogenetic associations were analyzed at the genus level with at least 90% confidence for 16S assignment. In all MiSeq runs, two blank controls with sterile water as input material were included to check for contaminants in the reagents and to filter out contaminant OTUs if present. Reads mapping to non-bacterial DNA (e.g., mitochondria, plastids, or other eukaryotic DNA) were also excluded from analysis. Only OTUs with more than 10 reads were considered in downstream analysis.

Relative abundances of OTUs in unsorted total fecal populations were calculated as the normalized number of reads in a sample. Relative abundances of OTUs in T0 FACS-enriched populations were used to measure false positive background fluorescence, which was subtracted from the T6 transconjugant populations. The corrected relative abundance of each OTU in a T6 FACS-enriched population is given by the formula:

$$RA_{6,i,sorted} = \frac{A_{6,i} * N_6 - A_{0,i} * N_0}{\sum_i (A_{6,i} * N_6 - A_{0,i} * N_0)}$$

where $RA_{t,i,sorted}$ is the corrected relative abundance of OTU i at time t, $A_{t,i}$ is the normalized number of reads of OTU i at time t in the FACS-sorted sample, and N_t is the fraction of mCherry-/GFP+ FACS-sorted events at time t. OTUs for which $RA_{6,i,sorted}$ is negative are eliminated from subsequent analysis, and all remaining $RA_{6,i,sorted}$ values are renormalized.

The *fold enrichment* of each OTU in the FACS-sorted population is defined as its relative abundance in the FACS-sorted population divided by its relative abundance in the unsorted total population at T6. To overcome the problem of detection limits (i.e., OTU i appears in the sorted population but is below the detection limit in the total population), we added a pseudo-count of p to all relative abundances when calculating fold enrichments. p is given by

$$p = 10^{\lfloor -\log_{10} n \rfloor}$$

where n is the total number of reads in the FACS-sorted sample and $\lfloor -\log_{10} n \rfloor$ is the floor function - the greatest integer less than or equal to - of $-\log_{10} n$. The fold enrichment of OTU i with the pseudo-count correction is calculated as

$$F_i = \frac{RA_{6,i,sorted} + p}{RA_{6,i,unsorted} + p}$$

If the relative abundance of OTU i in the unsorted population is below the detection limit, then the fold enrichment is calculable as $\frac{RA_{6,i,sorted} + p}{p}$, instead of $\frac{RA_{6,i,sorted}}{0}$.

The pseudo-count-corrected fold enrichment F_i overestimates the true fold enrichment ($\frac{RA_{6,i,sorted}}{RA_{6,i,unsorted}}$) by at most 10%, while possibly underestimating it. Because $0 < p \leq \frac{1}{n}$ and $RA_{6,i,sorted} \geq \frac{10}{n}$,

$$F_i = \frac{RA_{6,i,sorted} + p}{RA_{6,i,unsorted} + p} \leq \frac{RA_{6,i,sorted} + p}{RA_{6,i,unsorted}} \leq \frac{1.1 * RA_{6,i,sorted}}{RA_{6,i,unsorted}}$$

In all heat maps showing fold enrichment versus relative abundance, only OTUs with $F_i > 10$ are displayed to show more stringent and high confidence results. R code for this analysis is available upon request.

Whole genome sequencing of engineered mouse gut bacteria (MGB) isolates. To sequence MGB isolates, we prepared a sequencing library using the Nextera kit (Illumina) and utilized the Illumina HiSeq 2500 platform for 100 bp single-end reads. The SPAdes single cell assembler pipeline (version 3.9.1)[155] was employed to generate whole genome contigs. BLAST and PlasmidFinder (version 1.3)[156] were used to analyze the sequences and identify native mobilization systems. Geneious (version 7.1.5) was used to visualize contig alignments to genomes and plasmids.

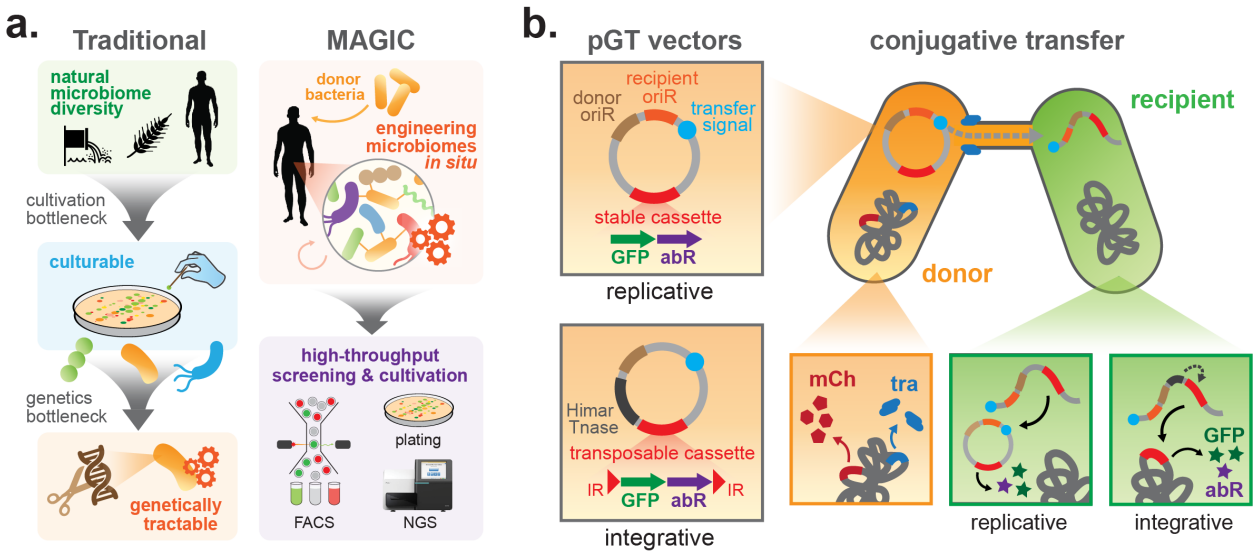


Figure 3.1. Overview of Metagenomic Alteration of Gut microbiome by In situ Conjugation (MAGIC). (a) In contrast to traditional approaches to cultivate microbes first and then test for genetic accessibility, MAGIC harnesses horizontal gene transfer in the native environment to genetically modify bacteria *in situ*. Transconjugant bacteria can be detected by FACS or antibiotic selection and further manipulated. **(b)** MAGIC implementation to transfer replicative or integrative pGT vectors from an engineered donor strain into amenable recipients in a complex microbiome. Replicative vectors feature a broad-host range origin of replication (oriR), while integrative vectors contain a transposable Himar cassette and transposase. The donor *E. coli* strain contains genomically integrated conjugative transfer genes (tra) and a mCherry gene. Transconjugant bacteria are detectable based on expression of an engineered payload that includes GFP and an antibiotic resistance gene (abR).

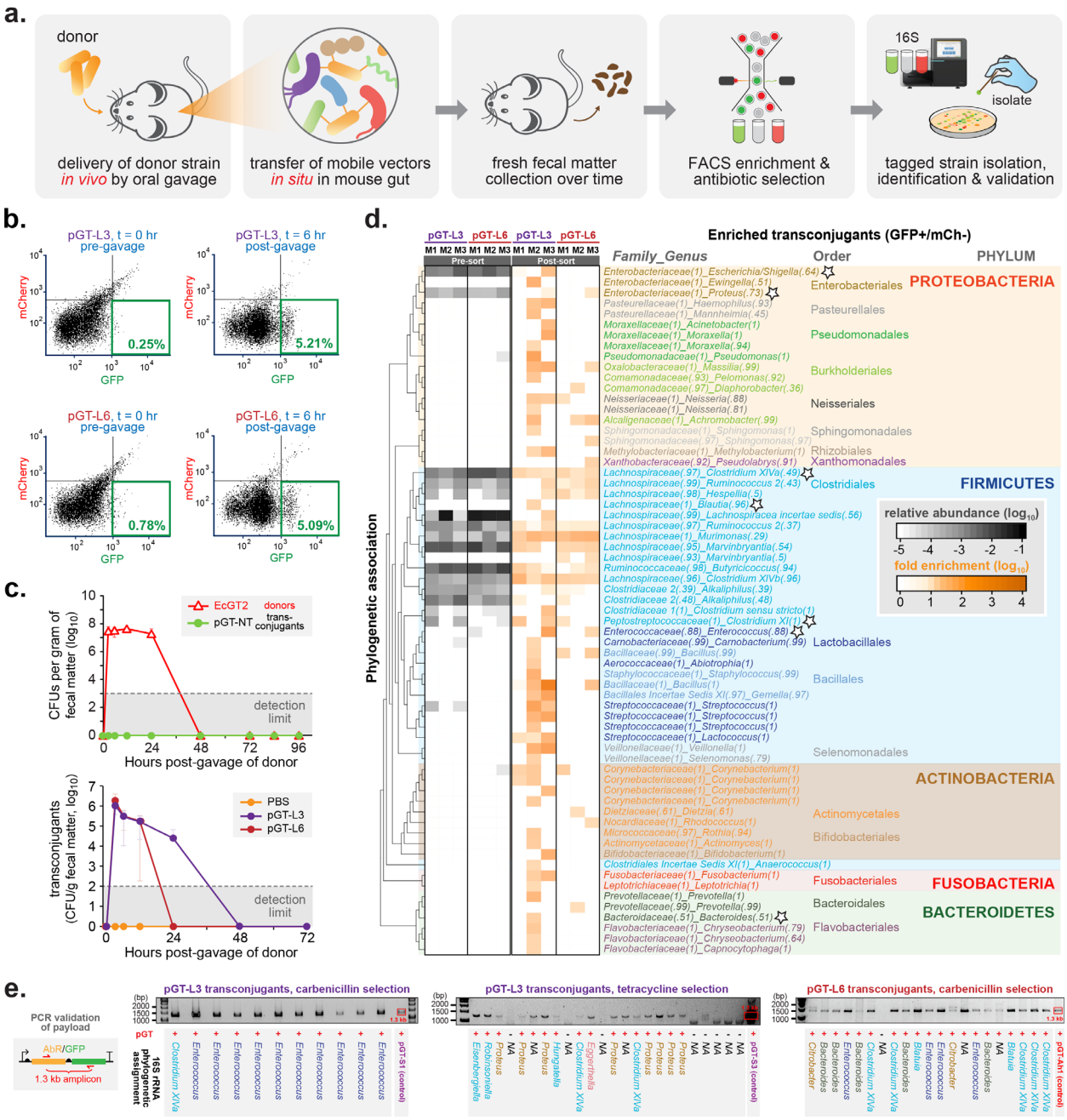


Figure 3.2. Identification and isolation of genetically tractable bacteria from the murine gut using MAGIC. (a) Implementation of MAGIC in a murine model with fecal bacterial analysis by FACS, antibiotic selection, and sequencing. **(b)** FACS dot plots of fecal bacteria, pre- and post-gavage of EcGT2 donors containing pGT-L3 or pGT-L6 vector libraries. Green boxes define the sorted GFP+/mCherry- transconjugant populations. For each vector library, fecal samples from 3 co-housed mice were independently evaluated by flow cytometry with similar results. **(c)** Longitudinal analysis of fecal microbiome by flow cytometry for presence of EcGT2 pGT-NT donor cells (red triangles, n= 4 mice) and transconjugants of vector libraries pGT-L3 (purple circles, n=3 mice), pGT-L6 (maroon circles, n=3 mice), pGT-NT control (green circles, n= 4 mice), or PBS (no donor) control (orange circles, n=2 mice). Donor cells and transconjugants were lost within 48

hours. The dotted line shows the detection limit. **(d)** 16S taxonomic classification of transconjugants (GFP^+/mCh^-) enriched by FACS of pGT-L3 and pGT-L6 recipient groups. Each column represents transconjugants from one mouse. Each OTU's relative abundance in the total bacterial population is shown in the grayscale heat-map, while each OTU's fold enrichment among transconjugants is shown in the orange heat-map.. Bracketed values indicate confidence of taxonomic assignment by RDP classifier. Genera with successfully cultivated isolates are denoted by white stars. **(e)** PCR confirmed the presence of the antibiotic resistance/GFP payload cassette from pGT-L3 and pGT-L6 vectors in diverse isolates that were engineered in the murine gut and isolated by selective plating with carbenicillin or tetracycline. NA indicates 16S sequences that were not available.

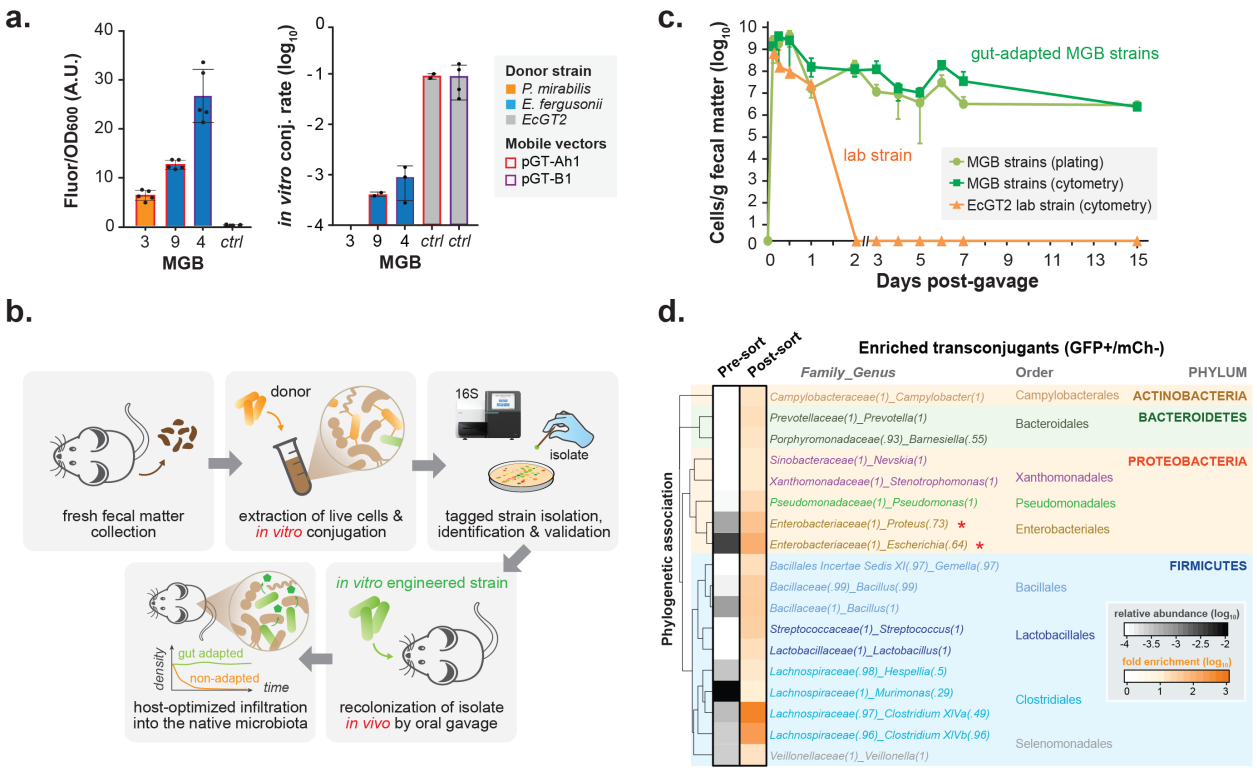
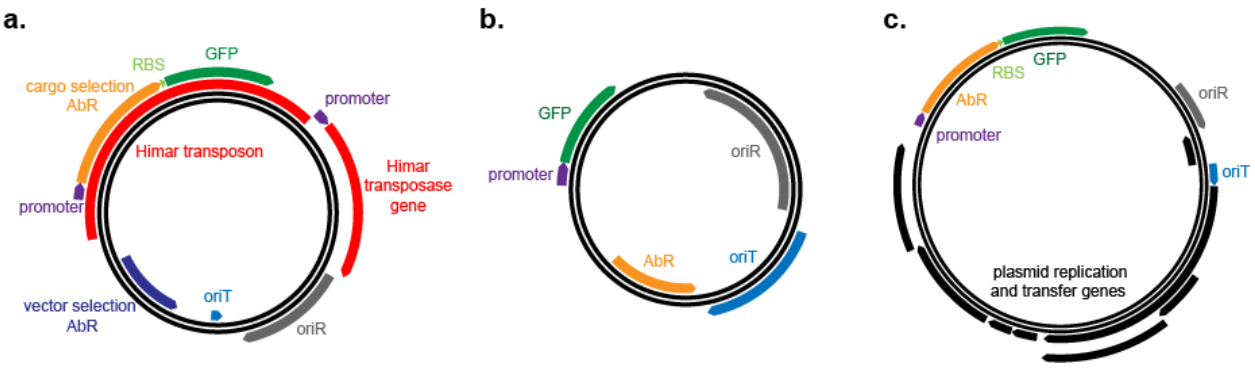
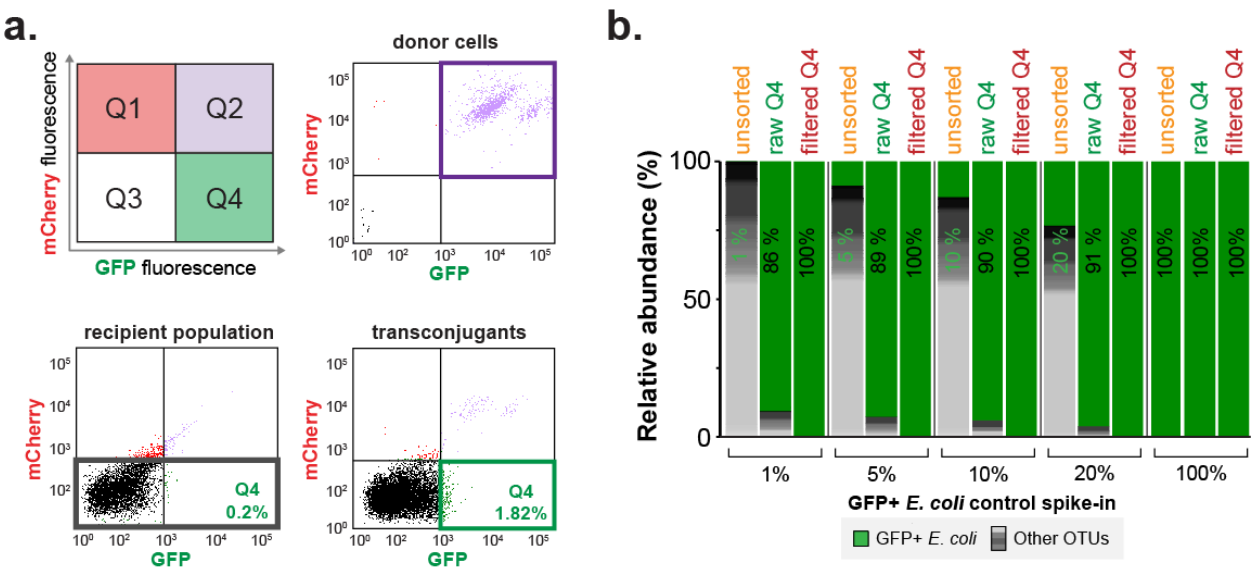


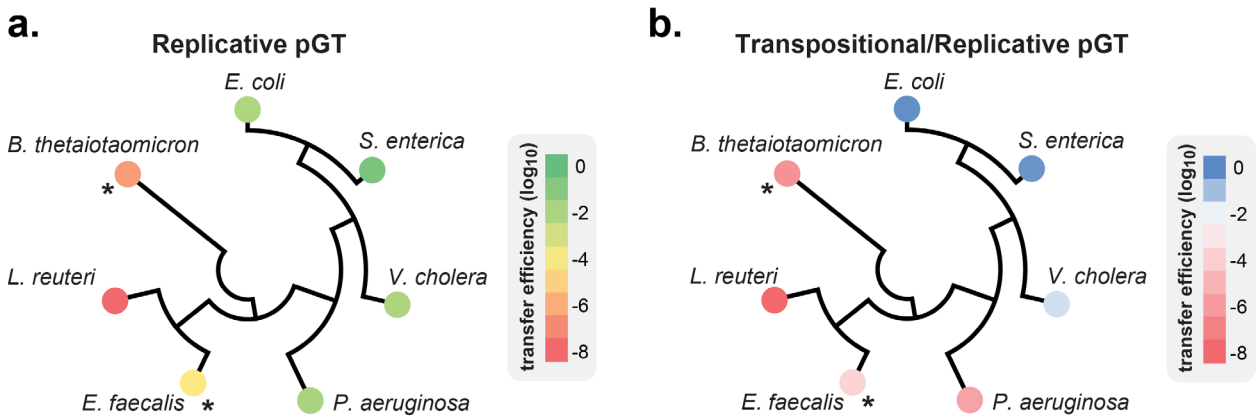
Figure 3.3. Transconjugant native gut bacteria recolonize the gut and mediate secondary transfer of engineered genetic payloads. (a) Left panel: GFP expression profiles of three isolates (MGB3, MGB4, MGB9; n=5 for each) versus control strain (*E. coli* MG1655, n=5). MGB isolates were *P. mirabilis* (orange bar) and *E. fergusonii* (blue bars) containing either vector pGT-Ah1 (red border) or vector pGT-B1 (purple border). *E. fergusonii* strains were genetically identical, but received two different vectors. Right panel: efficiency of *in vitro* conjugation of pGT vectors from MGB strains to *E. coli* MG1655 recipients. EcGT2 donors were used as positive controls (gray bars). Sample sizes are n=2-4. Bars indicate means; error bars indicate standard deviation. **(b)** Schematic diagram of experiment: genetically tractable gut microbiota were isolated from the murine microbiome *in vitro* and then orally gavaged to recolonize the gut. **(c)** Colonization of MGB strains and EcGT2 lab strain in mice (n=6, n=4 respectively) over time, after initial oral gavage. Cell densities were determined by both plating (light green) and flow cytometry (dark green) of fecal bacteria, and by flow cytometry for *E. coli* (orange). Error bars indicate standard deviation. **(d)** FACS enrichment and 16S taxonomic classification of top *in vivo* transconjugants at 6 hours post-gavage with MGB strains. Fecal samples from 6 mice were combined for analysis. Each OTU's relative abundance in the total bacterial population is shown in the grayscale heat-map, while each OTU's fold enrichment among transconjugants is shown in the orange heat-map. Bracketed values indicate confidence of taxonomic assignment by RDP classifier. Red asterisks denote OTUs that share the same genus as MGB donors.



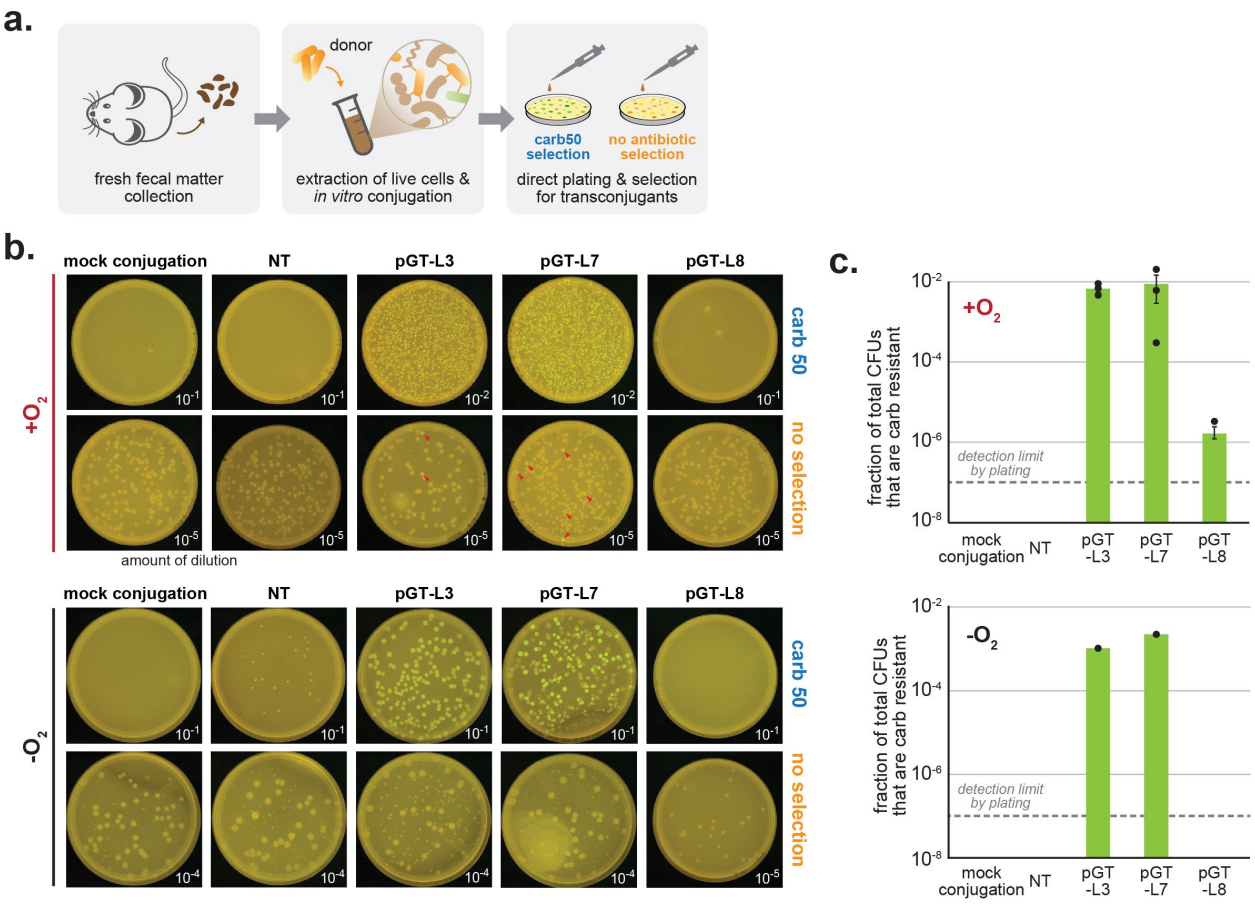
Supplementary Figure S3.1. Plasmid maps of vectors used in this study. (a) Map of Himar transposon integrative vectors (pGT-Ah and pGT-Kh variants found in libraries L2, L4, L5, L6, L7 and L8). **(b)** Map of replicative vectors with pBBR1 origin of replication (pGT-B variants found in libraries L1, L4, and L6). **(c)** Map of replicative vectors with RSF1010 origin of replication (pGT-S variants found in library L3). Although this vector backbone contain genes involved in conjugation (black), these vectors are not self-transmissible[157, 158].



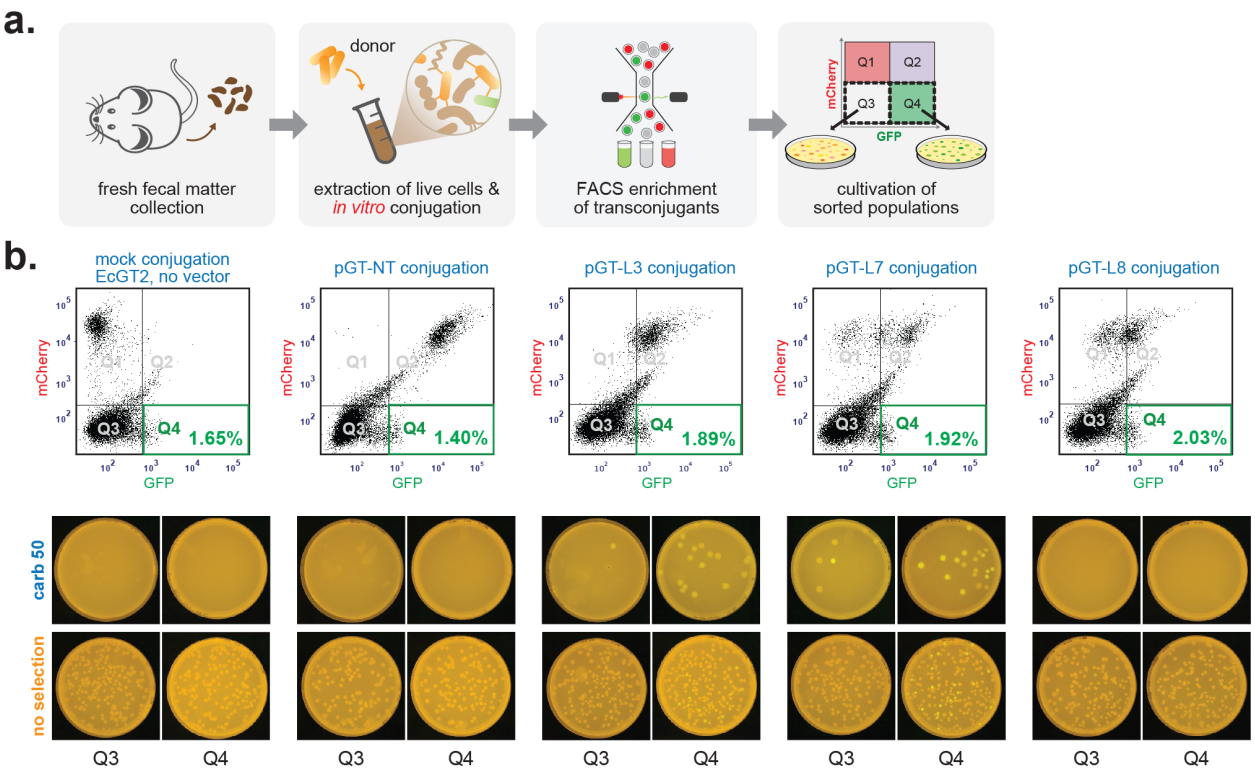
Supplementary Figure S3.2. FACS gating methodology for isolation of transconjugant bacteria. (a) Illustration of FACS enrichment method to isolate transconjugant cells from complex recipient populations. GFP and mCherry fluorescence are used to gate cell populations consisting of *E. coli* donors and diverse recipients. Quadrants Q1 and Q2 correspond to donor cells (mCh^+), while un-manipulated recipients are in quadrant Q3. Quadrant Q4 contains transconjugants that received the GFP gene cargo and are not naturally mCherry fluorescent (GFP^+, mCh^-). Q4 cells are isolated and further analyzed. This gating was used to analyze fecal samples from each individual mouse in each *in situ* experiment, as well as every *in vitro* conjugation in this study by flow cytometry. **(b)** To validate the FACS enrichment method, GFP^+ *E. coli* were mixed with a natural murine fecal bacterial community at given levels (1-100% of population) and retrieved by FACS. 16S sequencing of the samples showed that the fluorescent *E. coli* were efficiently and specifically enriched by FACS. Although the raw Q4 population contained some autofluorescent cells, the only remaining OTU in Q4 after applying an enrichment filter (see Methods) was *E. coli.*



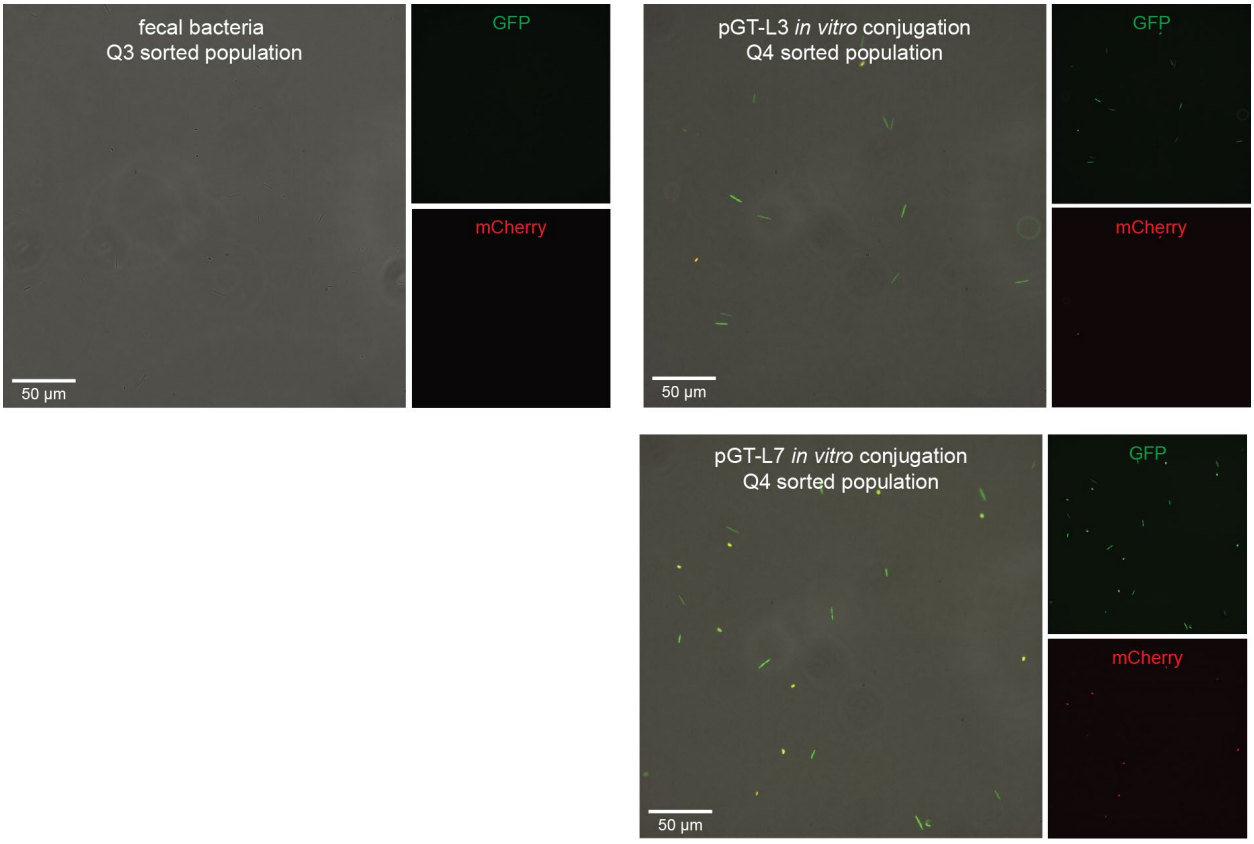
Supplementary Figure S3.3. pGT vectors were transferred from *E. coli* donors to representative recipient species during *in vitro* conjugations. (a) *In vitro* conjugation efficiency of replicative vector pGT-B1 from *E. coli* donor to various recipients, which are plotted by phylogenetic relationships. **(b)** *In vitro* conjugation efficiency of vector pG-Ah1 between *E. coli* donor and various recipients. This vector is replicative only in Proteobacteria (*E. coli, S. enterica, V. cholera, P. aeruginosa*) but delivered genetic cargo by transposition into a broader array of bacteria. Asterisks indicate cultures grown in anaerobic conditions, while all other cultures were grown aerobically. Conjugation efficiencies were calculated from 2 independent conjugations.



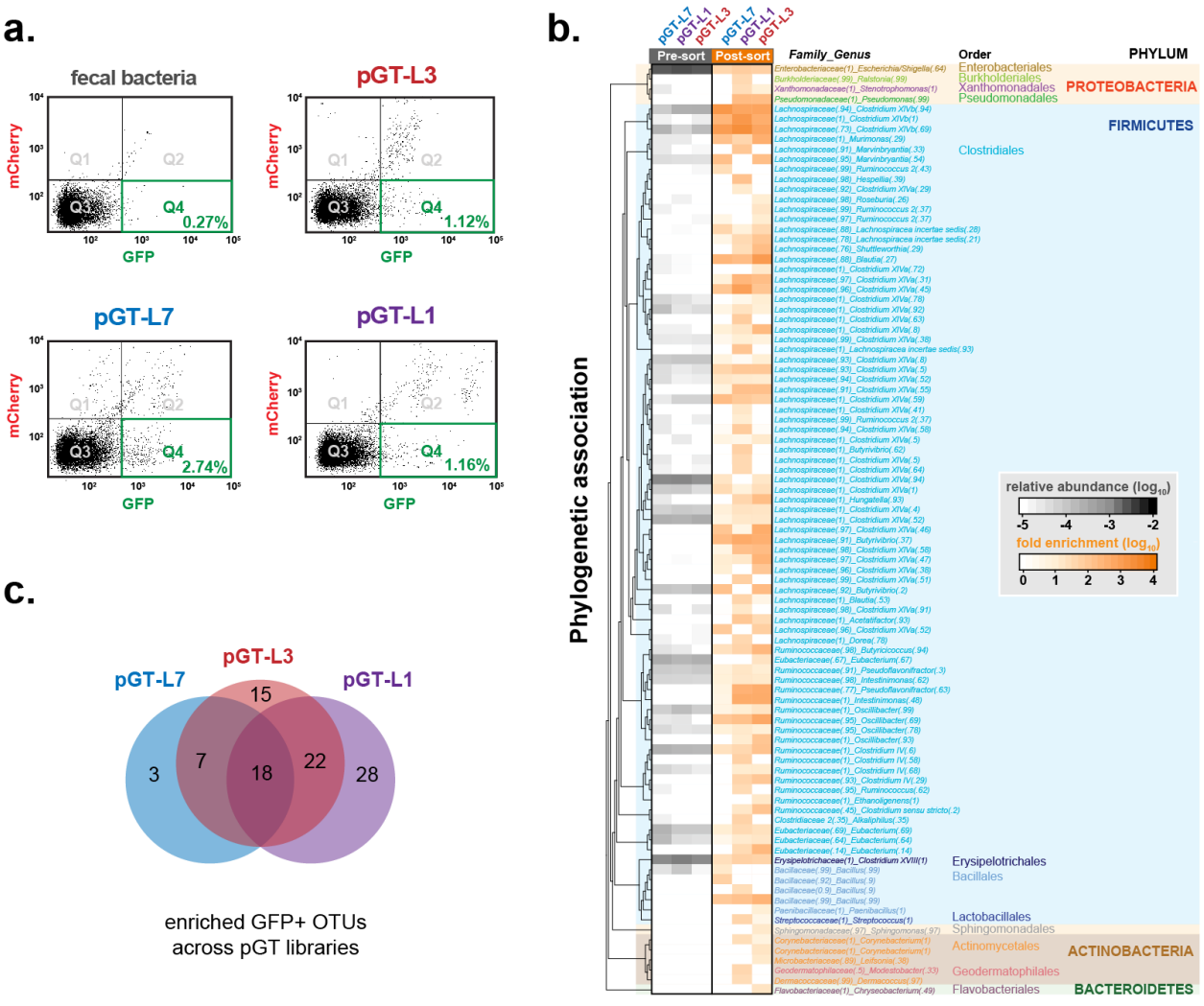
Supplementary Figure S3.4. pGT vectors were transferred from E. coli donors to murine fecal bacteria during *in vitro* conjugations. (a) *In vitro* conjugation of pGT vectors from EcGT2 donor strain into fecal bacteria extracted from murine feces. **(b)** Aerobic (top) and anaerobic (bottom) conjugations were performed using EcGT2 strains containing no vector (mock conjugation), a nontransferable vector (pGT-NT), pGT-L3, pGT-L7, and pGT-L8. Aerobic conjugations were plated on selective and non-selective media and grown aerobically at 37C for 24 hours. Anaerobic conjugations were plated on selective and non-selective media, grown anaerobically at 37C for 48 hours, and exposed to oxygen at room temperature for 48 hours. Red arrows indicate GFP+ CFUs on nonselective plates. **(c)** Efficiencies of aerobic (top) and anaerobic (bottom) conjugations. Aerobic conjugation efficiencies were calculated from 3 independent conjugations; anaerobic conjugation efficiencies were calculated from 1 conjugation.



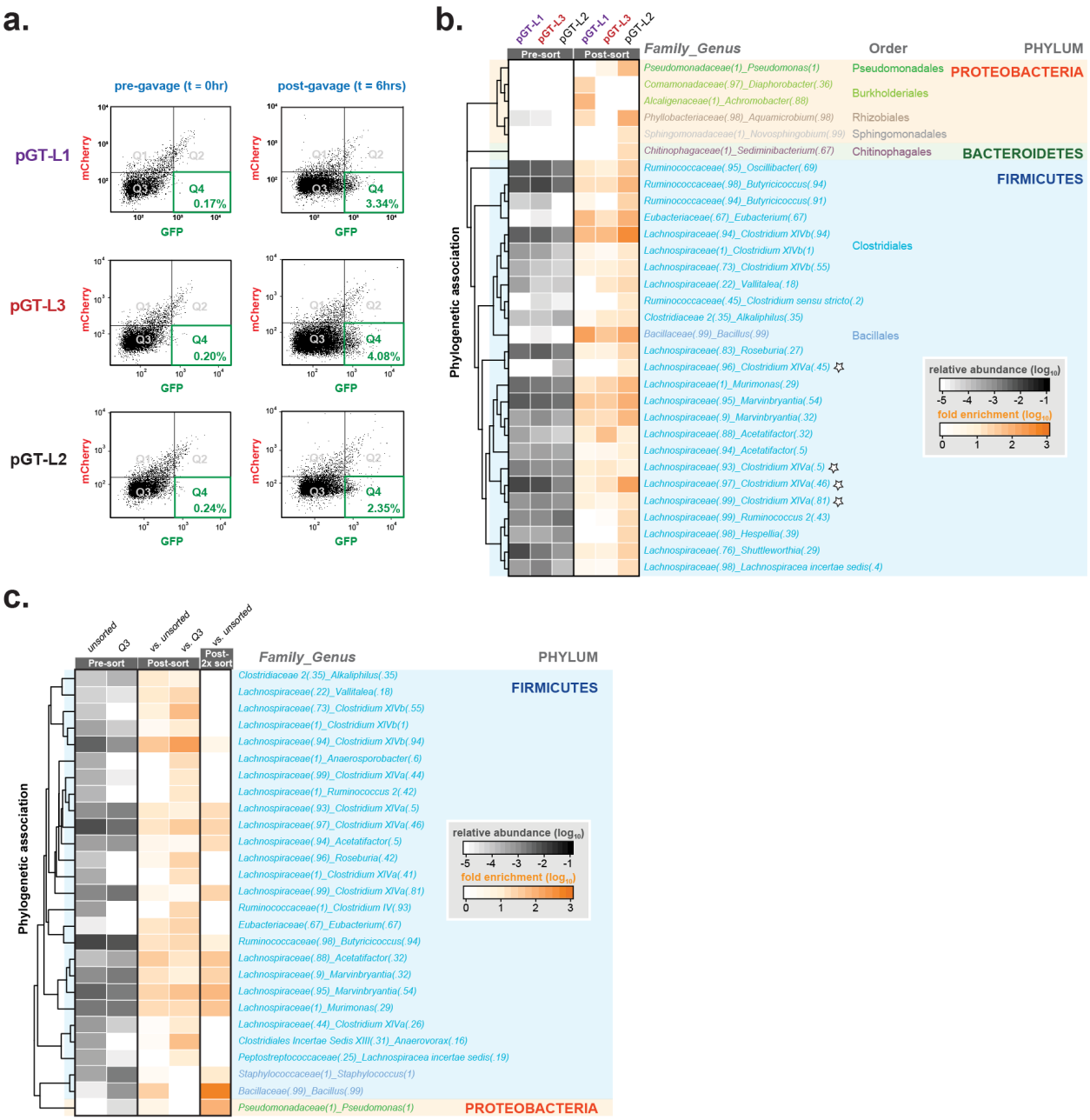
Supplementary Figure S3.5. FACS enriches for GFP+, antibiotic-resistant transconjugant gut bacteria arising from *in vitro* conjugations. (a) Implementation of FACS enrichment of *in vitro* conjugations. **(b)** Conjugations between EcGT2 harboring vector libraries pGT-L3, pGT-L7, and pGT-L8 and murine fecal bacteria were performed aerobically overnight. A mock conjugation using EcGT2 with no vector and a negative control conjugation using the pGT-NT non-transferable vector were also performed. 20,000 FACS sorted events from Q3 (mCherry-/GFP-) and Q4 (mCherry-/GFP+) populations were plated on selective and non-selective media and grown aerobically to select for transconjugants. Cultivable aerobic transconjugants of pGT-L3 and pGT-L7 vectors were successfully enriched by FACS, although GFP+ CFUs may appear dim against the autofluorescent media. This experiment was performed independently twice with similar results.



Supplementary Figure S3.6. FACS enriches for GFP+, antibiotic-resistant transconjugant gut bacteria arising from *in vitro* conjugations. (a) Implementation of FACS enrichment of *in vitro* conjugations. **(b)** Conjugations between EcGT2 harboring vector libraries pGT-L3, pGT-L7, and pGT-L8 and murine fecal bacteria were performed aerobically overnight. A mock conjugation using EcGT2 with no vector and a negative control conjugation using the pGT-NT non-transferable vector were also performed. 20,000 FACS sorted events from Q3 (mCherry-/GFP-) and Q4 (mCherry-/GFP+) populations were plated on selective and non-selective media and grown aerobically to select for transconjugants. Cultivable aerobic transconjugants of pGT-L3 and pGT-L7 vectors were successfully enriched by FACS, although GFP+ CFUs may appear dim against the autofluorescent media. This experiment was performed independently twice with similar results.

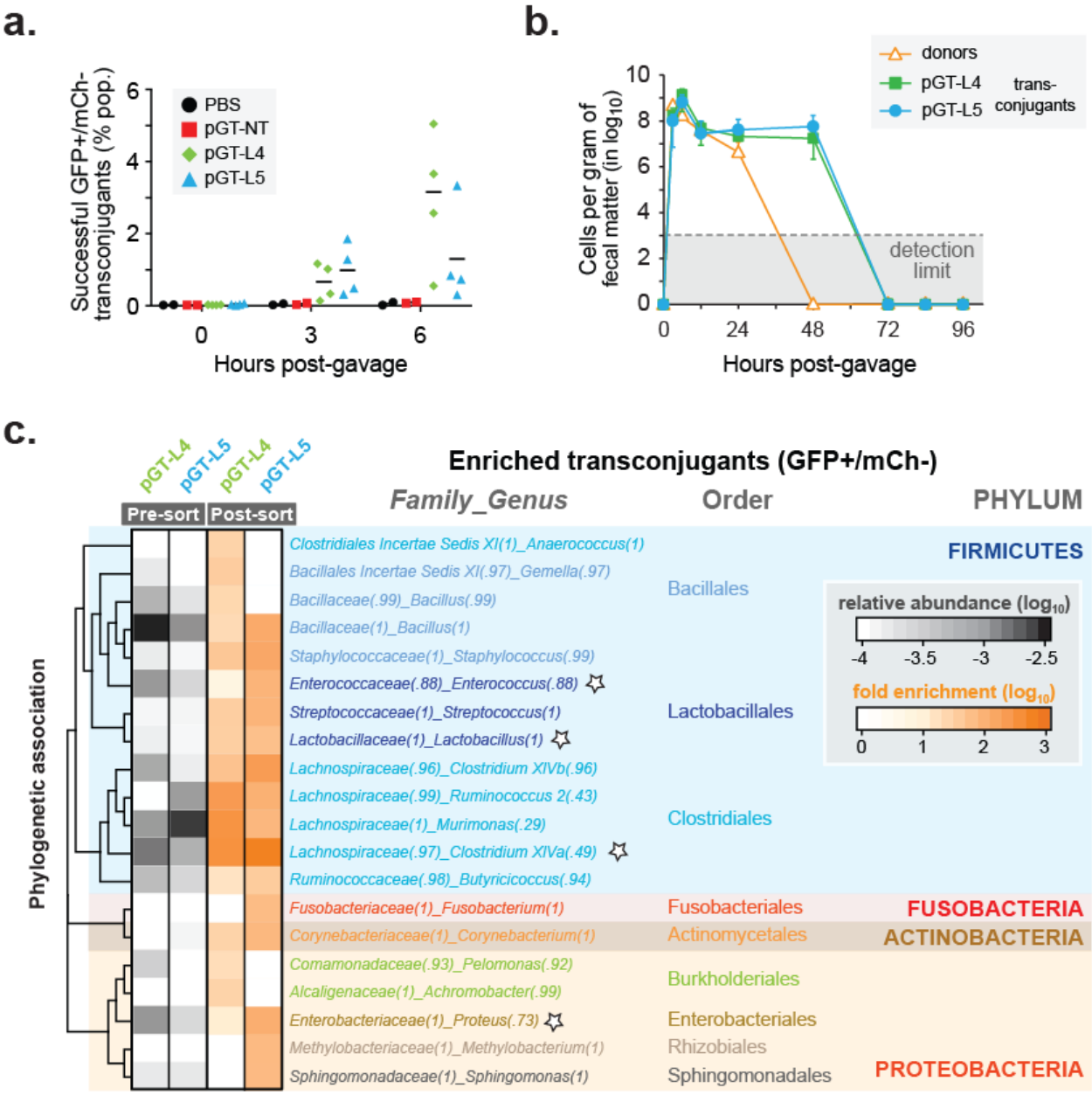


Supplementary Figure S3.7. Identification of FACS-enriched *in vitro* transconjugants by 16S sequencing. (a) FACS dot plots of *in vitro* conjugations of murine gut bacteria and EcGT2 donors with vector libraries pGT-L1, L3, and L7. This experiment was performed 3 times with similar results. Green boxes define the sorted GFP$^+$/mCherry$^-$ transconjugant populations. **(b)** 16S taxonomic classification of *in vitro* GFP$^+$/mCherry$^-$ transconjugants of pGT-L1, L3, and L7 enriched by FACS. Relative abundance of each OTU in the unsorted population is shown in the grayscale heat-map, while fold enrichment for transconjugants of each OTU is shown in the orange heat-map with annotated taxonomic identities. Bracketed values indicate confidence of taxonomic assignment by RDP classifier. Genera with successfully cultivated isolates are denoted by stars. Each column represents FACS-enriched transconjugants from one conjugation. **(c)** Comparison of OTUs shared between transconjugants arising from each vector library during *in vitro* conjugations. 18 OTUs were shared between all 3 libraries, with a total of 47 OTUs being shared between at least 2 libraries.



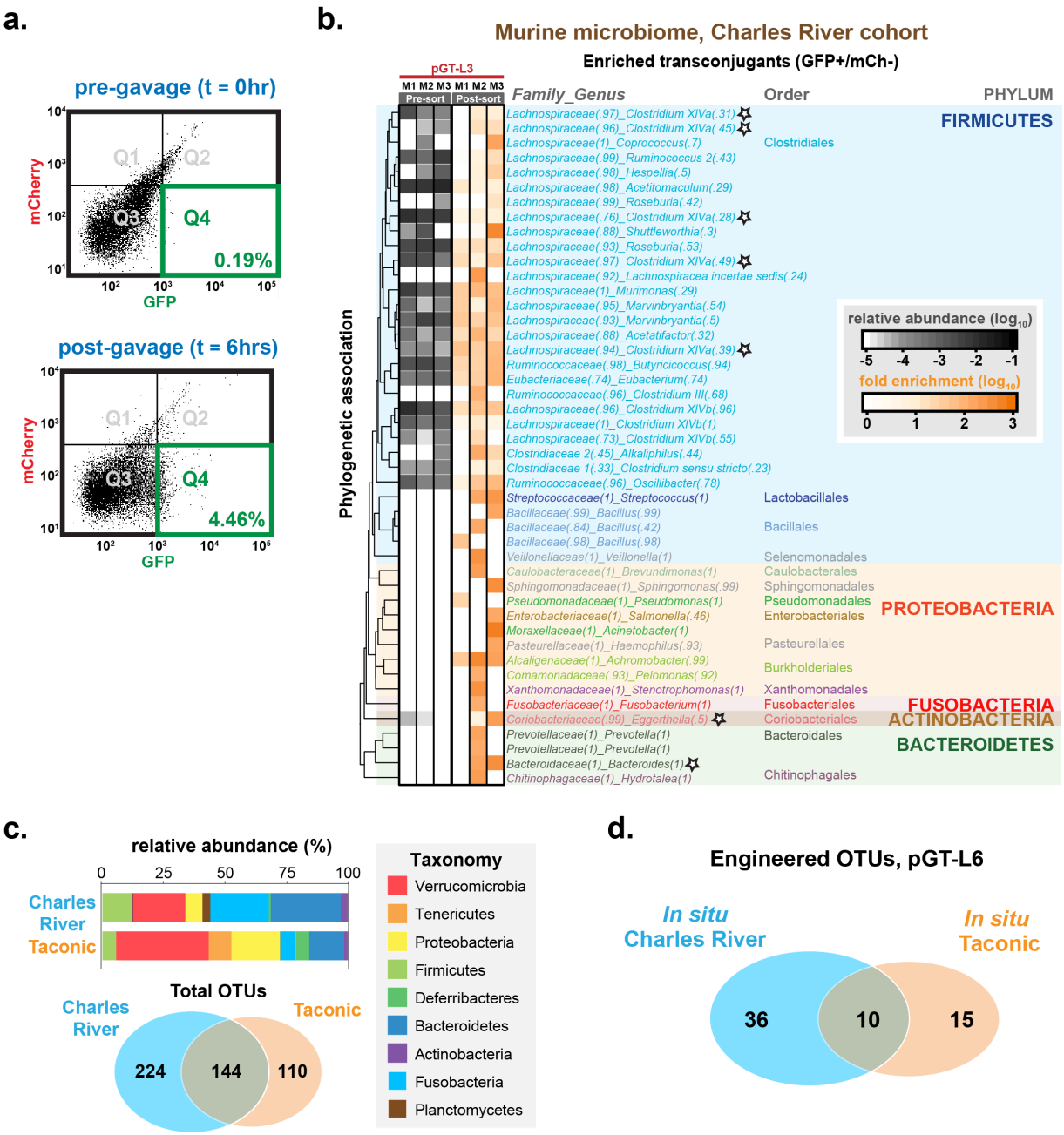
Supplementary Figure S3.8. Identification of FACS-enriched *in situ* transconjugants by 16S sequencing. (a) FACS dot plots of *in situ* conjugations using EcGT2 donors with vector libraries pGT-L1, L2, and L3. Green boxes define the sorted GFP^{+}/mCherry^{-} transconjugant populations. Each plot shows fluorescence expression of bacteria from the combined fecal samples of 3 co-housed mice. The experiment was run 3 independent times with similar results. **(b)** 16S taxonomic classification of FACS-enriched transconjugants from *in situ* mouse experiments using vector libraries pGT-L1, L2, and L3. Relative abundance of each OTU in the unsorted population is shown in the grayscale heat-map, while fold enrichment for transconjugants of each OTU is shown in the orange heat-map with annotated taxonomic identities. Bracketed values indicate confidence of taxonomic assignment by RDP classifier. Each column represents data from a

separately housed cohort of 3 mice whose fecal samples were combined for analysis. Genera with successfully cultivated isolates are denoted by stars. **(c)** The pGT-L3 transconjugant population from (b) was further analyzed by comparing Q4 enriched OTUs against Q3 OTUs, which represent a sample of the GFP- native bacteria population, and by performing enrichment analysis of Q4 samples that were sorted again for Q4. Enriched GFP+ transconjugants were robust whether compared against the total fecal population or against Q3. 7 out of 11 OTUs enriched in Q4 were present in the double-sorted Q4 population, indicating that Q4 sorting is robust. The OTUs lost upon double-sorting were obligate anaerobes and likely sensitive to prolonged aerobic conditions during double-sorting.



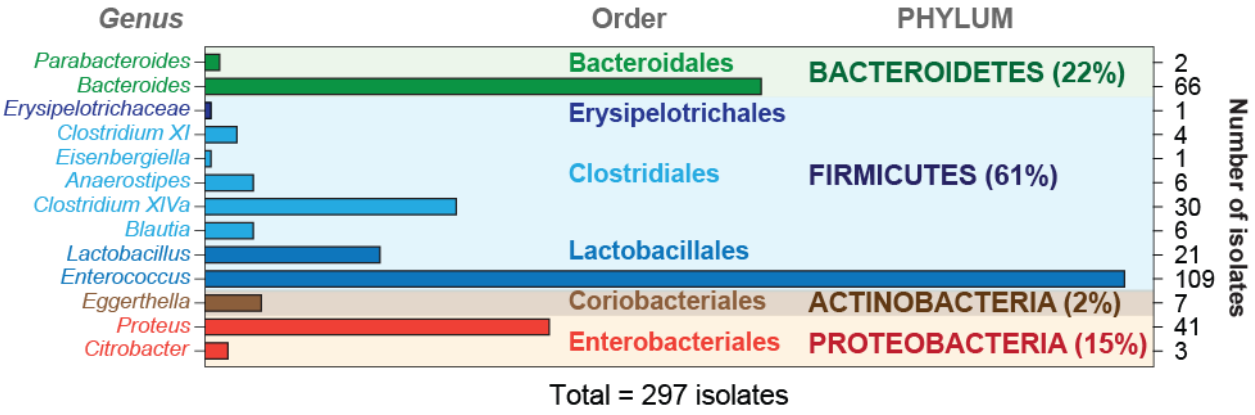
Supplementary Figure S3.9. Identification of FACS-enriched *in situ* transconjugants of multi-vector libraries. (a) Flow cytometric quantification of *in situ* transconjugants in the total bacterial population, post-gavage of EcGT2 donors containing pGT-L4 (green, n=4 mice) or pGT-L5 (blue, n=4 mice) vector libraries. Control groups gavaged with PBS (black, n=2 mice) or donors containing a non-transferrable pGT-NT vector (red, n=2 mice) produced no detectable transconjugants. Black bars indicate means. **(b)** Longitudinal analysis of murine fecal microbiome by flow cytometry for presence of transconjugants post-gavage of EcGT2 donors containing pGT-L4 (green, n=6 mice), or pGT-L5 (blue, n=6 mice). Donor cells of these libraries (orange, n=12 mice) were lost within 48 hours, while transconjugants were observed up to 72 hours post-gavage. The dotted line indicates the detection limit of flow cytometry. Error bars indicate standard

deviation. **(c)** 16S taxonomic classification of transconjugants (GFP^+/mCh^-) enriched by FACS of pGT-L4 and pGT-L5 recipient groups. Relative abundance of each OTU in the unsorted population is shown in the grayscale heat-map on the left, while fold enrichment for transconjugants of each OTU is shown in the orange heat-map on the right with annotated taxonomic identities. Bracketed values indicate confidence of taxonomic assignment by RDP classifier. Each column represents data from 6 mice from 2 independent cohorts whose fecal samples were combined for analysis. Genera with successfully cultivated isolates are denoted with stars.

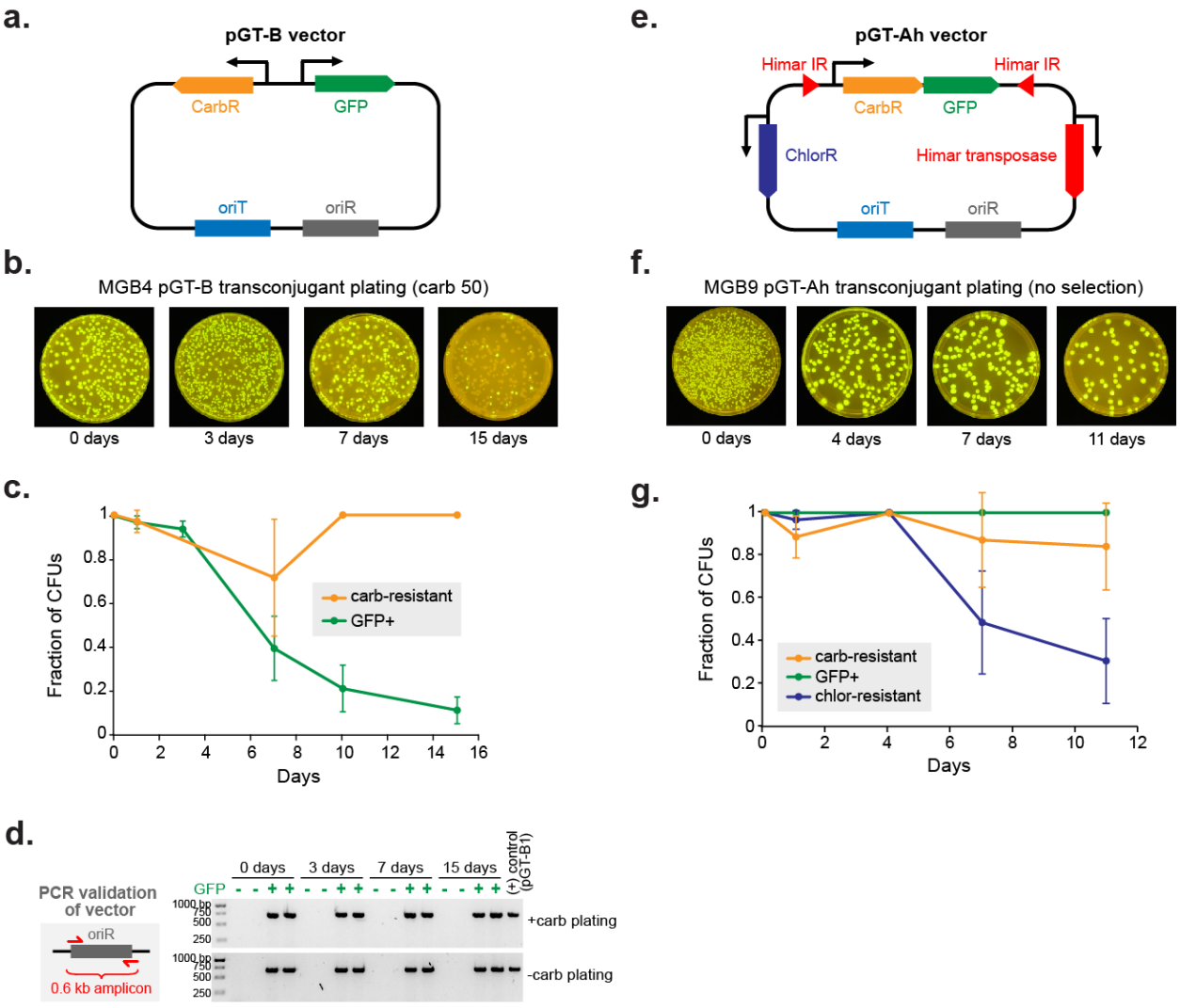


Supplementary Figure S3.10. Identification of FACS-enriched *in situ* transconjugants in mice from a different commercial vendor. (a) FACS dot plots of *in situ* conjugations using EcGT2 pGT-L3 donors in a cohort of mice from a different vendor (Charles River Laboratories). Green boxes define the sorted GFP$^+$/mCherry$^-$ transconjugant populations. Flow cytometry was performed 3 times, on fecal samples from individual co-housed mice, with similar results. **(b)** 16S taxonomic classification of FACS-enriched GFP$^+$/mCherry$^-$ transconjugants of pGT-L3. Relative abundance of each OTU in the unsorted population is shown in the grayscale heat-map, while fold enrichment for transconjugants of each OTU is shown in the orange heat-map with annotated taxonomic identities. Bracketed values indicate confidence of taxonomic assignment by RDP

classifier. Each column represents bacteria from one mouse. Genera with successfully cultivated isolates are denoted by stars. **(c)** Metagenomic 16S rRNA sequencing of mouse fecal samples shows that mice from different vendors have divergent gut microbiomes, with some shared OTUs. **(d)** In *in situ* experiments using the same vector library (pGT-L6) in cohorts of 3 mice each from different vendors, 10 transconjugant OTUs were shared between cohorts.



Supplementary Figure S3.11. PCR-validated transconjugant isolates from *in situ* mouse experiments. 297 PCR-validated isolates from *in situ* experiments using vector libraries pGT-L3 and pGT-L6 were identified by 16S Sanger sequencing and assigned to a genus using RDP classifier with assignment confidence >0.89.



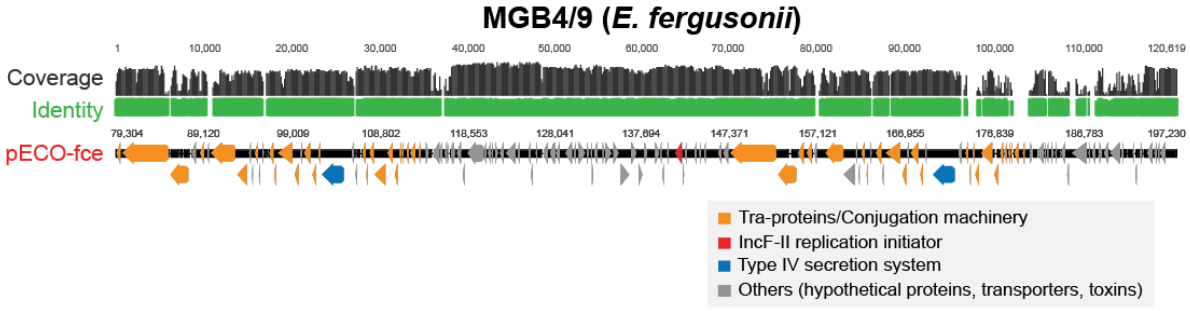
Supplementary Figure S3.12. Comparison of vector and payload stability in two transconjugant isolates. (a) Vector map of pGT-B1. GFP and beta-lactamase genes are expressed from separate promoters on a replicative pBBR1 origin plasmid. **(b)** MGB4, an *Escherichia fergusonii* isolate containing pGT-B1, lost GFP expression over time when serially passaged without selection for 15 days. Plating was performed for 3 independent serial passages. **(c)** Quantification of carb-resistant and GFP+ CFUs of MGB4 over time; all CFUs remained carb-resistant as the population lost GFP expression. Center values are the means of 3 serial passages; error bars represent standard deviation. **(d)** Colony PCR for the pGT-B1 backbone showed that the plasmid was absent in GFP- CFUs at all time points surveyed. Each lane shows the PCR product for one colony. This PCR was performed once. **(e)** Vector map of pGT-Ah1, which contains GFP and beta-lactamase genes on a transposable cassette. The plasmid backbone contains a chloramphenicol resistance gene for selection. **(f)** MGB9, an *Escherichia fergusonii* isolate containing pGT-Ah1, remained 100% GFP+ during serial passaging without selection over 11 days. Plating was performed for 3 independent serial passages. **(g)** Over time the proportion of MGB9 CFUs expressing the genes on the transposable cassette (GFP+ and carb-resistant) remained at 100%, while the chloramphenicol resistance conferred by the pGT-

Ah1 backbone was lost in some of the population. Center values are the means of 3 serial passages; error bars represent standard deviation.

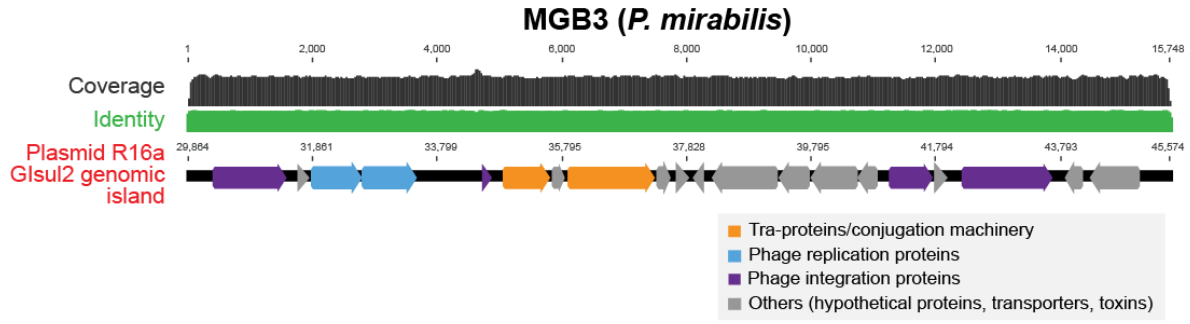
a.

Isolate	Strain	Vector	N50 (bp)	# of contigs	Total length (bp)
MGB3	*P. mirabilis*	pGT-Ah1	153,603	67	3,843,607
MGB9	*E. fergusonii*	pGT-Ah1	92,709	153	4,938,839
MGB4	*E. fergusonii*	pGT-B1	103,260	148	4,917,568

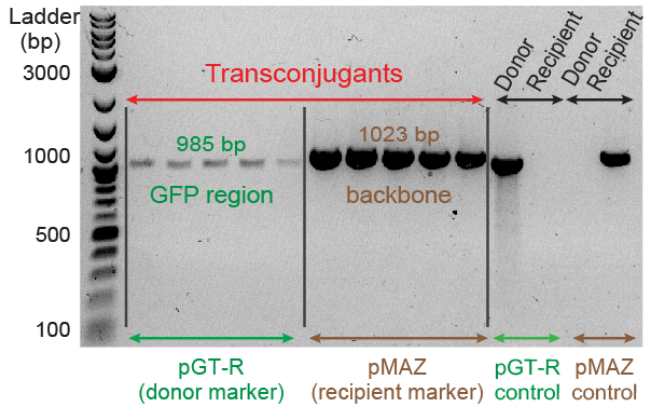
b.



c.

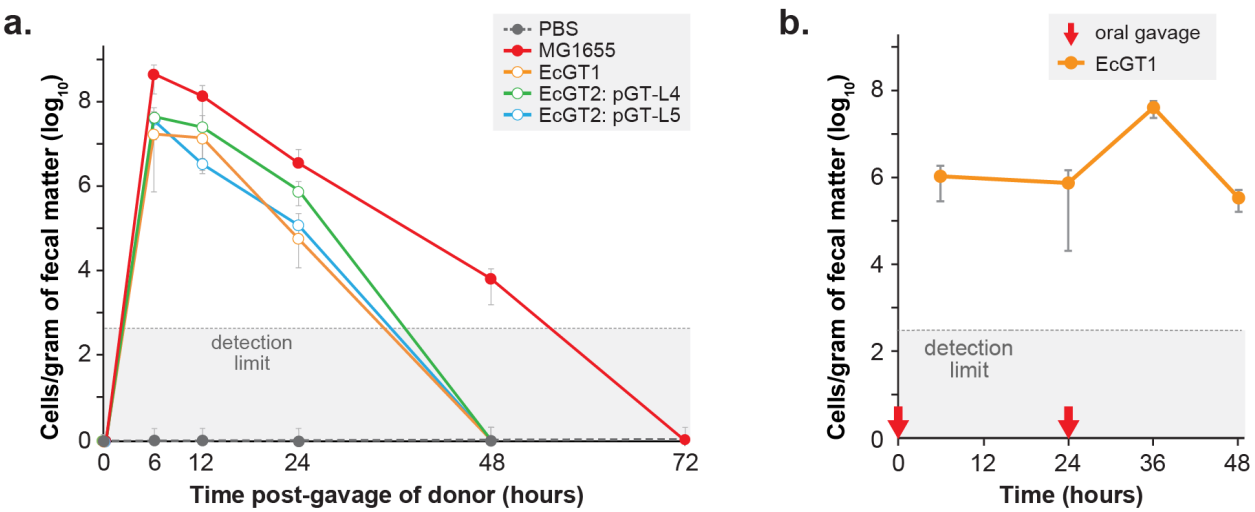


d.

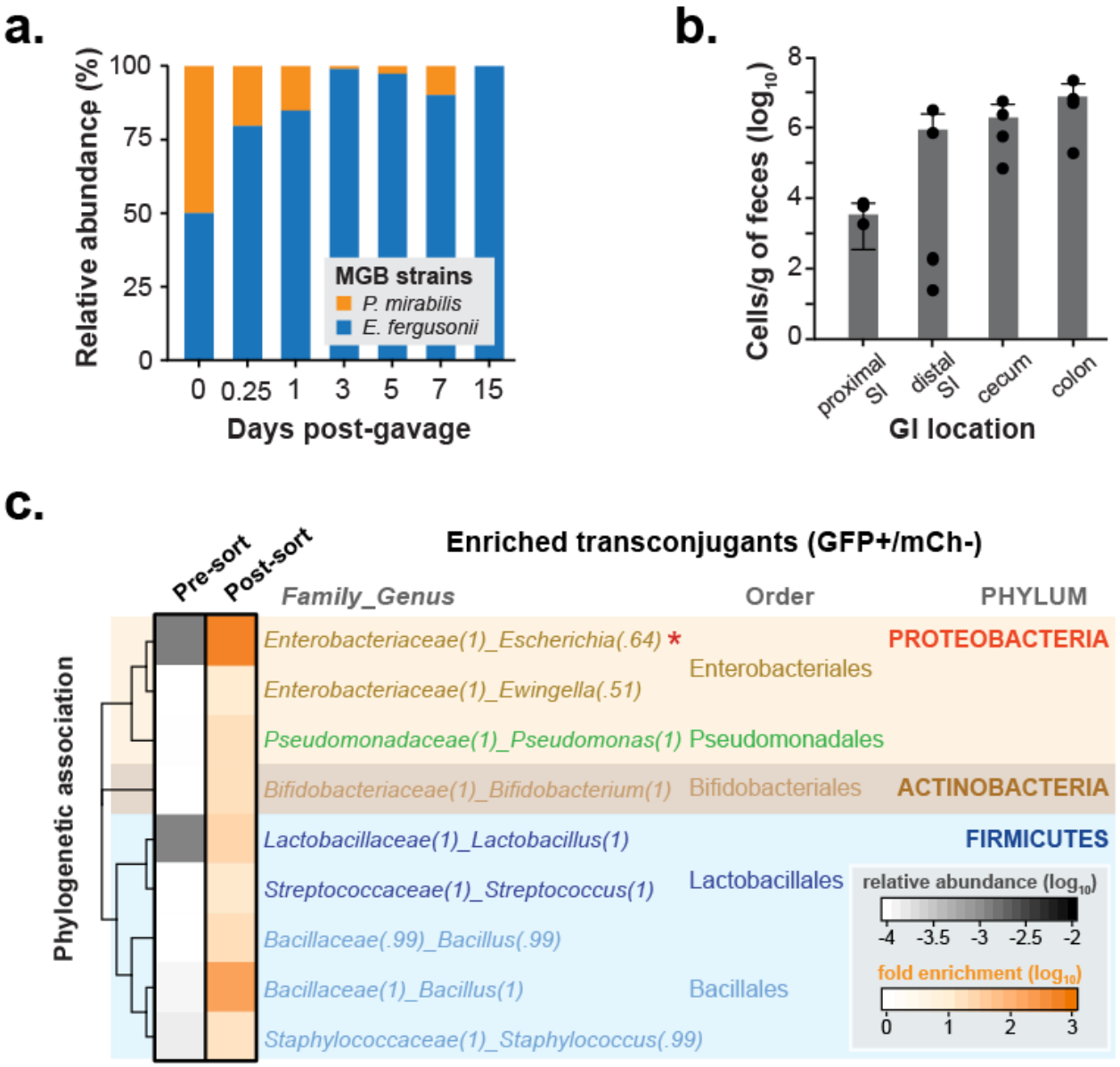


Supplementary Figure S3.13. Characterization of 3 Modifiable Gut Bacteria (MGB) strains by whole-genome sequencing and *in vitro* conjugation. (a) Three distinct MGB strains, isolated from *in vitro* conjugations between *E. coli* pGT donors and murine fecal bacteria, were analyzed by whole-genome sequencing. MGB4 and MGB9 appear to be the same strain isolated from separate experiments with different pGT vectors transferred. Sequencing of **(b)** MGB4/9 and **(c)** MGB3 revealed the presence of genes involved in conjugation and genetic transfer. However, only MGB4/9 strains that shared homology with the pECO-fce plasmid were observed to transfer

their pGT vectors to *E. coli* during *in vitro* conjugations. **(d)** PCR confirmation of pGT vector transfer from MGB4 to an *E. coli* recipient following *in vitro* conjugation. The conjugation was performed 3 times with similar results; 5 individual transconjugants were assessed by colony PCR.



Supplementary Figure S3.14. Longevity of donor *E. coli* strains in the murine gut following oral gavage. (a) *In vivo* gut colonization profiles of MAGIC donors EcGT1 (S17, *galK::mCherry),* EcGT2 (S17, *asd::mCherry*), and control *E. coli* MG1655 in C57BL/6 mice measured by flow cytometry of fecal bacteria after a single gavage of 10^9 cells. Mean values were calculated using feces from 2 gavaged mice; error bars indicate standard deviation. **(b)** Two orally gavaged doses of 10^9 EcGT1 cells resulted in a longer persistence of this donor in the gut. Mean values were calculated using feces from 2 gavaged mice; error bars indicate SEM.



Supplementary Figure S3.15. Characterization of MGB recolonization of the murine gut. (a) MGB3, MGB4, and MGB9 strains orally gavaged into mice (n=4) as a mixture recolonized the GI tract without any antibiotic treatment. MGBs were detectable in fecal samples for at least 15 days post-gavage. **(b)** MGB strains (namely MGB4) were present in all sampled locations along the GI tract when the mice (n=4) were euthanized 15 days post-gavage. Error bars represent standard deviation. **(c)** Phylogenetic tree of FACS-sorted GFP^{+}/$mCherry^{-}$.transconjugants in fecal samples from mice after 11 days post-gavage of MGB strains. Fecal samples from 4 mice were combined for analysis. Relative abundance of each OTU in the unsorted population is shown in the grayscale heat-map, while fold enrichment for transconjugants of each OTU is shown in the orange heat-map. Bracketed values indicate confidence of taxonomic assignment by RDP classifier. The red asterisk denotes the *Escherichia/Shigella* OTU that shares a genus with the MGB4/9 donors.

Supplementary Table 3.1. List of vectors and vector components.

Origins of replication (oriR):

Origin	Copy #	Host range	Code
R6K	10-20	Narrow (Proteobacteria)	**K**
p15A	14-16	Narrow (Enterobacteria)	**A**
oriV	4-7	Broad (Gram- and Gram+)	**V**
pBBR1	15-40	Broad (preferably Gram-)	**B**
RSF1010	12	Broad (Gram- and Gram+)	**S**
RCR	250-350	Broad (Eubacteria)	**W**

Integrative elements:

Transposase	Transposon inverted repeat	Host range	Code
none	-	-	-
Himar	ACAGGTTGGATGATAAGTCCCCGGTCT	Broad	h
Tn5	CTGTCTCTTATACACATCT	Broad	t

Vector selection genes:

Resistance gene	Antibiotic selection	[Ab] in *E. coli*
Beta-lactamase	Carbenicillin	50 µg/ml
Chlor	Chloramphenicol	20 µg/ml
Tet	Tet	25 µg/ml
Spec	Spec	200 µg/ml
Kan	Kan	50 µg/ml

Cargo selection cassettes:

Resistance cassette	Antibiotic selection	[Ab] in *E. coli*
GFP-Beta-lactamase	Carb	50 µg/ml
GFP-CatP	Chlor	20 µg/ml
GFP-Tet	Tet	25 µg/ml
GFP-Spec	Spec	250 µg/ml
GFP-Kan	Kan	50 µg/ml
GFP-ErmG	Erm	-

Regulation sequences:

Promoter/UTR	Expression in *E. coli*	Origin of sequence	Code
GATTGCATTAGGTTTTAGTTTCTTGTATAATGCTTAATGTTG GTCACTGACAGGCTACGATACGGAAGGTTGCTCACGCCCGGC CCCTTTGCCATGGCTAGTGTGTGGAAATTTCCGAGGAGCAAG TCTATTTCCAAAAATGGGCGAAAAAGGAGGTAATACA	+++	*Bacillus cellulosilyticus*	1
GGGAGAGCTTCAACGGCGCTTCTACCCATTTGCTTGGAAAGG ATGAGGAGCAGGAAGAAATTCCGTCCCCAATGCGACGGCCCT TTACATCCATGTTGTTTGATAGTATAATGGATACGGATTGAC CAAATTGTTCATTTAGTCAGTTTGAAGGATGAGGAGT	++	*Geobacillus sp.*	2
GTGAAGGATACGGCTGCGGCACTTCGACATCGCCCCATGTGG CGGCTTTGAACTGGGCTTATGAAACGCGTTCACAACCTTTTT TGACCATCGGCGCGAACGTGGTATCATGCGTTCAGCTTTTGC CCATACATACTACGTGCTCAATCTAGGAGGATTTCATAC	+++	*Eggerthella lenta*	3
CTCTAGAGTAGTAGATTATTTTAGGAATTTAGATGTTTTGTA TGAAATAGATGCTTCGTATGGAATTAATGAAATTTTTAGTCA GGTAAAAAAGGTAATAGGAGAATATT	+++	*Segmented filamentous bacteria*	4
GTTTTAAATGATGAAAAGAAATATTTAGGGAAGATTGTTTCG ACGCGAATTGTTGATCTGGAAAATGATCACCTTATCGGACAA GCTTTAAAATAGGAGGATATAAAAAT	++	*Segmented filamentous bacteria*	5
ATAAGGATTCTTTAAAGAGAGATATAGTTATGTCAAAGACTG TAGAATTTTTAGTAAATCAAAATAAAAAAAGAGGTATTAAAT AGAGTGTATTTTAAAGGAGGAGACTT	+++	*Segmented filamentous bacteria*	6
AAACACCAATAAAATTAGAATATTTAGGAGCGACTTTAAAAA AGTTTAATAAGAATTGTTTATGAGATATTTTTATTATATTTA AACTCAATTTAAAGTAGGGAGAATAG	+	*Segmented filamentous bacteria*	7
GCAAGTGTTCAAGAAGTTATTAAGTCGGGAGTGCAGTCGAAG TGGGCAAGTTGAAAAATTCACAAAAATGTGGTATAATATCTT TGTTCATTAGAGCGATAAACTTGAATTTGAGAGGGAACTTAG	+	*Clostridium perfringens*	8

Vector list:

Vector name	Cargo selection	Cargo promoter	Vector selection	Transposase promoter
pGT-Ah1	GFP-Beta-lactamase	4	Chlor	4
pGT-Ah2	GFP-Beta-lactamase	5	Chlor	5
pGT-Ah3	GFP-Beta-lactamase	6	Chlor	6
pGT-Ah4	GFP-Beta-lactamase	7	Chlor	7
pGT-Ah5	GFP-CatP	8	Kan	4
pGT-Ah6	GFP-CatP	8	Kan	5
pGT-Ah7	GFP-CatP	8	Kan	6
pGT-Ah8	GFP-CatP	8	Kan	7
pGT-Ah9	GFP-Tet	4	Chlor	4
pGT-Ah10	GFP-Tet	4	Chlor	5
pGT-Ah11	GFP-Tet	4	Chlor	6
pGT-B1	GFP	1	Beta-lactamase	-
pGT-B2	GFP	2	Beta-lactamase	-
pGT-B3	GFP	3	Beta-lactamase	-
pGT-S1	GFP-Beta-lactamase	4	Beta-lactamase	-
pGT-S2	GFP-Beta-lactamase	5	Beta-lactamase	-
pGT-S3	GFP-Tet	4	Tet	-
pGT-S4	GFP-Tet	5	Tet	-
pGT-Kh1	GFP-Beta-lactamase	4	Chlor	4
pGT-Kh2	GFP-Beta-lactamase	5	Chlor	5
pGT-Kh3	GFP-Beta-lactamase	7	Chlor	7

Supplementary Table 3.2. Vector libraries used in this study.

Library	Vectors
pGT-L1	B1, B2, B3
pGT-L2	Ah5, Ah6, Ah7, Ah8
pGT-L3	S1, S2, S3, S4
pGT-L4	Ah1, Ah3, B1, B2, B3
pGT-L5	Ah5, Ah6, Ah7, Ah8, Ah9, Ah10, Ah11
pGT-L6	Ah1, Ah3, Ah5, Ah6, Ah7, Ah8, Ah9, Ah10, Ah11, B1, B2, B3
pGT-L7	Ah1, Ah2, Ah3, Ah4
pGT-L8	Kh1, Kh2, Kh3

Supplementary Table 3.3: Full sequences of pGT vector parts.

Cargo selection genes	**Sequence**	**Notes**
Beta-lactamase (carbenicillin/ ampicillin resistance)	ATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTT CCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTG GGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGT TTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGC GCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTAT TCTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGC ATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCC AACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAAC ATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATA CCAAACGACGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAA CTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATG GAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTT ATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTG GGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCA ACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCAT TGGTAA	
CatP (chloramphenicol resistance)	ATGGTATTTGAAAAAATTGATAAAAATAGTTGGAACAGAAAAGAGTATTTTGACCAC TACTTTGCAAGTGTACCTTGTACATACAGCATGACCGTTAAAGTGGATATCACACAA ATAAAGGAAAAGGGAATGAAACTATATCCTGCAATGCTTTATTATATTGCAATGATT GTAAACCGCCATTCAGAGTTTAGGACGGCAATCAATCAAGATGGTGAATTGGGGATA TATGATGAGATGATACCAAGCTATACAATATTTCACAATGATACTGAAACATTTTCC AGCCTTTGGACTGAGTGTAAGTCTGACTTTAAATCATTTTTAGCAGATTATGAAAGT GATACGCAACGGTATGGAAACAATCATAGAATGGAAGGAAAGCCAAATGCTCCGGAA AACATTTTTAATGTATCTATGATACCGTGGTCAACCTTCGATGGCTTTAATCTGAAT TTGCAGAAAGGATATGATTATTTGATTCCTATTTTTACTATGGGGAAATATTATAAA GAAGATAACAAAATTATACTTCCTTTGGCAATTCAAGTTCATCACGCAGTATGTGAC GGATTTCACATTTGCCGTTTTGTAAACGAATTGCAGGAATTGATAAATAGTTAA	Derived from pJIR750 plasmid from *Clostridium perfringens*
Tet (tetracycline resistance)	ATGAATATTATAAATTTAGGAATTCTTGCTCACATTGATGCAGGAAAAACTTCCGTA ACCGAGAATCTGCTGTTTGCCAGTGGAGCAACGGAAAAGTGCGGCTGTGTGGATAAT GGTGACACCATAACGGACTCTATGGATATAGAGAAACGTAGAGGAATTACTGTTCGG GCTTCTACGACATCTATTATCTGGAATGGTGTGAAATGCAATATCATTGACACTCCG GGACACATGGATTTTATTGCGGAAGTGGAGCGGACATTCAAAATGCTTGATGGAGCA GTCCTCATCTTATCCGCAAAGGAAGGCATACAAGCGCAGACAAAGTTGCTGTTCAAT ACTTTACAGAAGCTGCAAATCCCGACAATTATATTTATCAATAAGATTGACCGAGCC GGTGTGAATTTGGAGCGTTTGTATCTGGATATAAAAGCAAATCTGTCTCAAGATGTC CTGTTTATGCAAAATGTTGTCGATGGATCGGTTTATCCGGTTTGCTCCCAAACATAT ATAAAGGAAGAATACAAAGAATTTGTATGCAACCATGACGACAATATATTAGAACGA TATTTGGCGGATAGCGAAATTTCACCGGCTGATTATTGGAATACGATAATCGCTCTT GTGGCAAAAGCCAAAGTCTATCCGGTGCTACATGGATCAGCAATGTTCAATATCGGT ATCAATGAGTTGTTGGACGCCATCACTTCTTTTATACTTCCTCCGGCATCGGTCTCA AACAGACTTTCATCTTATCTTTATAAGATAGAGCATGACCCCAAAGGACATAAAAGA AGTTTTCTAAAAATAATTGACGGAAGTCTGAGACTTCGAGACGTTGTAAGAATCAAC GATTCGGAAAAATTCATCAAGATTAAAAATCTAAAAACTATCAATCAGGGCAGAGAG ATAAATGTTGATGAAGTGGGCGCCAATGATATCGCGATTGTAGAGGATATGGATGAT TTTCGAATCGGAAATTATTTAGGTGCTGAACCTTGTTTGATTCAAGGATTATCGCAT CAGCATCCCGCTCTCAAATCCTCCGTCCGGCCAGACAGGCCCGAAGAGAGAAGCAAG GTGATATCCGCTCTGAATACATTGTGGATTGAAGACCCGTCTTTGTCCTTTTCCATA AACTCATATAGTGATGAATTGGAAATCTCGTTATATGGTTTAACCCAAAAGGAAATC ATACAGACATTGCTGGAAGAACGATTTTCCGTAAAGGTCCATTTTGATGAGATCAAG ACTATATACAAAGAACGACCTGTAAAAAAGGTCAATAAGATTATTCAGATCGAAGTG CCGCCCAACCCTTATTGGGCCACAATAGGGCTGACTCTTGAACCCTTACCGTTAGGG ACAGGGTTGCAAATCGAAAGTGACATCTCCTATGGTTATCTGAACCATTCTTTTCAA AATGCCGTTTTTGAAGGGATTCGTATGTCTTGCCAATCCGGGTTACATGGATGGGAA GTGACTGATCTGAAAGTAACTTTTACTCAAGCCGAGTATTATAGCCCGGTAAGTACA CCTGCTGATTTCAGACAGCTGACCCCTTATGTCTTCAGGCTGGCCTTGCAACAGTCA GGTGTGGACATTCTCGAACCGATGCTCTATTTTGAGTTGCAGATACCCCAAGCGGCA AGTTCCAAAGCTATTACAGATTTGCAAAAAATGATGTCTGAGATTGAAGACATCAGT TGCAATAATGAGTGGTGTCATATTAAAGGGAAAGTTCCATTAAATACAAGTAAAGAC TATGCATCAGAAGTAAGTTCATACACTAAGGGCTTAGGCATTTTTATGGTTAAGCCA TGCGGGTATCAAATAACAAAAGGCGGTTATTCTGATAATATCCGCATGAACGAAAAA GATAAACTTTTATTCATGTTCCAAAAATCAATGTCATCAAAATAA	

Spec (spectinomycin resistance)	ATGCGCTCACGCAACTGGTCCAGAACCTTGACCGAACGCAGCGGTGGTAACGGCGCA GTGGCGGTTTTCATGGCTTGTTATGACTGTTTTTTTGGGGTACAGTCTATGCCTCGG GCATCCAAGCAGCAAGCGCGTTACGCCGTGGGTCGATGTTTGATGTTATGGAGCAGC AACGATGTTACGCAGCAGGGCAGTCGCCCTAAAACAAAGTTAAACATCATGAGGGAA GCGGTGATCGCCGAAGTATCGACTCAACTATCAGAGGTAGTTGGCGTCATCGAGCGC CATCTCGAACCGACGTTGCTGGCCGTACATTTGTACGGCTCCGCAGTGGATGGCGGC CTGAAGCCACACAGTGATATTGATTTGCTGGTTACGGTGACCGTAAGGCTTGATGAA ACAACGCGGCGAGCTTTGATCAACGACCTTTTGGAAACTTCGGCTTCCCCTGGAGAG AGCGAGATTCTCCGCGCTGTAGAAGTCACCATTGTTGTGCACGACGACATCATTCCG TGGCGTTATCCAGCTAAGCGCGAACTGCAATTTGGAGAATGGCAGCGCAATGACATT CTTGCAGGTATCTTCGAGCCAGCCACGATCGACATTGATCTGGCTATCTTGCTGACA AAAGCAAGAGAACATAGCGTTGCCTTGGTAGGTCCAGCGGCGGAGGAACTCTTTGAT CCGGTTCCTGAACAGGATCTATTTGAGGCGCTAAATGAAACCTTAACGCTATGGAAC TCGCCGCCCGACTGGGCTGGCGATGAGCGAAATGTAGTGCTTACGTTGTCCCGCATT TGGTACAGCGCAGTAACCGGCAAAATCGCGCCGAAGGATGTCGCTGCCGACTGGGCA ATGGAGCGCCTGCCGGCCCAGTATCAGCCCGTCATACTTGAAGCTAGACAGGCTTAT CTTGGACAAGAAGAAGATCGCTTGGCCTCGCGCGCAGATCAGTTGGAAGAATTTGTC CACTACGTGAAAGGCGAGATCACCAAGGTAGTCGGCAAATAA	
Kan (kanamycin resistance)	ATGAGCCATATTCAACGGGAAACGTCGAGGCCGCGATTAAATTCCAACATGGATGCT GATTTATATGGGTATAAATGGGCTCGCGATAATGTCGGGCAATCAGGTGCGACAATC TATCGCTTGTATGGGAAGCCCGATGCGCCAGAGTTGTTTCTGAAACATGGCAAAGGT AGCGTTGCCAATGATGTTACAGATGAGATGGTCAGACTAAACTGGCTGACGGAATTT ATGCCTCTTCCGACCATCAAGCATTTTATCCGTACTCCTGATGATGCATGGTTACTC ACCACTGCGATCCCCGGAAAAACAGCATTCCAGGTATTAGAAGAATATCCTGATTCA GGTGAAAATATTGTTGATGCGCTGGCAGTGTTCCTGCGCCGGTTGCATTCGATTCCT GTTTGTAATTGTCCTTTTAACAGCGATCGCGTATTTCGTCTCGCTCAGGCGCAATCA CGAATGAATAACGGTTTGGTTGATGCGAGTGATTTTGATGACGAGCGTAATGGCTGG CCTGTTGAACAAGTCTGGAAAGAAATGCATAAACTTTTGCCATTCTCACCGGATTCA GTCGTCACTCATGGTGATTTCTCACTTGATAACCTTATTTTTGACGAGGGGAAATTA ATAGGTTGTATTGATGTTGGACGAGTCGGAATCGCAGACCGATACCAGGATCTTGCC ATCCTATGGAACTGCCTCGGTGAGTTTTCTCCTTCATTACAGAAACGGCTTTTTCAA AAATATGGTATTGATAATCCTGATATGAATAAATTGCAGTTTCATTTGATGCTCGAT GAGTTTTTCTAA	
ermG (erythromycin resistance)	ATGAACAAAGTAAATATAAAAGATAGTCAAAATTTTATTACTTCAAAATATCACATA GAAAAAATAATGAATTGCATAAGTTTAGATGAAAAAGATAACATCTTTGAAATAGGT GCAGGGAAAGGTCATTTTACTGCTGGATTGGTAAAGAGATGTAATTTTGTAACGGCG ATAGAAATTGATTCTAAATTATGTGAGGTAACTCGTAATAAGCTCTTAAATTATCCT AACTATCAAATAGTAAATGATGATATACTGAAATTTACATTTCCTAGCCACAATCCA TATAAAATATTTGGCAGCATACCTTACAACATAAGCACAAATATAATTCGAAAAATT GTTTTTGAAAGTTCAGCCACAATAAGTTATTTAATAGTGGAATATGGTTTTGCTAAA ATGTTATTAGATACAAACAGATCACTAGCATTGCTGTTAATGGCAGAGGTAGATATT TCTATATTAGCAAAAATTCCTAGGTATTATTTCCATCCAAAACCTAAAGTGGATAGC ACATTAATTGTATTAAAAAGAAAGCCAGCAAAAATGGCATTTAAAGAGAGAAAAAAA TATGAAACTTTTGTAATGAAATGGGTTAACAAAGAGTACGAAAAACTGTTTACAAAA AATCAATTTAATAAAGCTTTAAAACATGCGAGAATATATGATATAAACAATATTAGT TTCGAACAATTTGTATCGCTATTTAATAGTTATAAAATATTTAACGGCTAA	
sfGFP	ATGCGTAAAGGCGAAGAGCTGTTCACTGGTTTCGTCACTATTCTGGTGGAACTGGAT GGTGATGTCAACGGTCATAAGTTTTCCGTGCGTGGCGAGGGTGAAGGTGACGCAACT AATGGTAAACTGACGCTGAAGTTCATCTGTACTACTGGTAAACTGCCGGTACCTTGG CCGACTCTGGTAACGACGCTGACTTATGGTGTTCAGTGCTTTGCTCGTTATCCGGAC CACATGAAGCAGCATGACTTCTTCAAGTCCGCCATGCCGGAAGGCTATGTGCAGGAA CGCACGATTTCCTTTAAGGATGACGGCACGTACAAAACGCGTGCGGAAGTGAAATTT GAAGGCGATACCCTGGTAAACCGCATTGAGCTGAAAGGCATTGACTTTAAAGAAGAC GGCAATATCCTGGGCCATAAGCTGGAATACAATTTTAACAGCCACAATGTTTACATC ACCGCCGATAAACAAAAAAATGGCATTAAAGCGAATTTTAAAATTCGCCACAACGTG GAGGATGGCAGCGTGCAGCTGGCTGATCACTACCAGCAAAACACTCCAATCGGTGAT GGTCCTGTTCTGCTGCCAGACAATCACTATCTGAGCACGCAAAGCGTTCTGTCTAAA GATCCGAACGAGAAACGCGATCACATGGTTCTGCTGGAGTTCGTAACCGCAGCGGGC ATCACGCATGGTATGGATGAACTGTACAAATAA	

Vector selection genes	**Sequence**	**Notes**
Chlor (chloramphenicol resistance)	ATGGAGAAAAAAATCACTGGATATACCACCGTTGATATATCCCAATGGCATCGTAAA GAACATTTTGAGGCATTTCAGTCAGTTGCTCAATGTACCTATAACCAGACCGTTCAG CTGGATATTACGGCCTTTTTAAAGACCGTAAAGAAAAATAAGCACAAGTTTTATCCG GCCTTTATTCACATTCTTGCCCGCCTGATGAATGCTCATCCGGAATTACGTATGGCA ATGAAAGACGGTGAGCTGGTGATATGGGATAGTGTTCACCCTTGTTACACCGTTTTC CATGAGCAAACTGAAACGTTTTCATCGCTCTGGAGTGAATACCACGACGATTTCCGG CAGTTTCTACACATATATTCGCAAGATGTGGCGTGTTACGGTGAAAACCTGGCCTAT TTCCCTAAAGGGTTTATTGAGAATATGTTTTTCGTCTCAGCCAATCCCTGGGTGAGT TTCACCAGTTTTGATTTAAACGTGGCCAATATGGACAACTTCTTCGCCCCCGTTTTC ACCATGGGCAAATATTATACGCAAGGCGACAAGGTGCTGATGCCGCTGGCGATTCAG GTTCATCATGCCGTTTGTGATGGCTTCCATGTCGGCAGAATGCTTAATGAATTACAA CAGTACTGCGATGAGTGGCAGGGCGGGGCGTAA	All vector selection markers not listed here are the same as the ones in the "Cargo selection genes" section.

Origins of replication/ Plasmid backbones	**Sequence**	**Notes**
R6K origin of replication	ATCCCTGGCTTGTTGTCCACAACCGTTAAACCTTAAAAGCTTTAAAAGCCTTATATATTCTTT TTTTTCTTATAAAACTTAAAACCTTAGAGGCTATTTAAGTTGCTGATTTATATTAATTTTATT GTTCAAACATGAGAGCTTAGTACGTGAAACATGAGAGCTTAGTACGTTAGCCATGAGAGCTTA GTACGTTAGCCATGAGGGTTTAGTTCGTTAAACATGAGAGCTTAGTACGTTAAACATGAGAGC TTAGTACGTGAAACATGAGAGCTTAGTACGTACTATCAACAGGTTGAACTGCTGATCTTC	Requires additional pir gene for replication
p15A origin of replication	AACAACTTATATCGTATGGGGCTGACTTCAGGTGCTACATTTGAAGAGATAAATTGCACTGAA ATCTAGAAATATTTTATCTGATTAATAAGATGATCTTCTTGAGATCGTTTTGGTCTGCGCGTA ATCTCTTGCTCTGAAAACGAAAAAACCGCCTTGCAGGGCGGTTTTTCGAAGGTTCTCTGAGCT ACCAACTCTTTGAACCGAGGTAACTGGCTTGGAGGAGCGCAGTCACCAAAACTTGTCCTTTCA GTTTAGCCTTAACCGGCGCATGACTTCAAGACTAACTCCTCTAAATCAATTACCAGTGGCTGC TGCCAGTGGTGCTTTTGCATGTCTTTCCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGC GCAGCGGTCGGACTGAACGGGGGGTTCGTGCATACAGTCCAGCTTGGAGCGAACTGCCTACCC GGAACTGAGTGTCAGGCGTGGAATGAGACAAACGCGGCCATAACAGCGGAATGACACCGGTAA ACCGAAAGGCAGGAACAGGAGAGCGCACGAGGGAGCCGCCAGGGGGAAACGCCTGGTATCTTT ATAGTCCTGTCGGGTTTCGCCACCACTGATTTGAGCGTCAGATTTCGTGATGCTTGTCAGGGG GGCGGAGCCTATGGAAAAACGGCTTTGCCGCGGCCCTCTCACTTCCCTGTTAAGTATCTTCCT GGCATCTTCCAGGAAATCTCCGCCCCGTTCGTAAGCCATTTCCGCTCGCCGCAGTCGAACGAC CGAGCGTAGCGAGTCAGTGAGCGAGGAAGCGGAATATATCC	
oriV	AGCGGGCCGGGAGGGTTCGAGAAGGGGGGGCACCCCCCTTCGGCGTGCGCGGTCACGCGCCAG GGCGCAGCCCTGGTTAAAAACAAGGTTTATAAATATTGGTTTAAAAGCAGGTTAAAAGACAGG TTAGCGGTGGCCGAAAAACGGGCGGAAACCCTTGCAAATGCTGGATTTTCTGCCTGTGGACAG CCCCTCAAATGTCAATAGGTGCGCCCCTCATCTGTCATCACTCTGCCCCTCAAGTGTCAAGGA TCGCGCCCCTCATCTGTCAGTAGTCGCGCCCCTCAAGTGTCAATACCGCAGGGCACTTATCCC CAGGCTTGTCCACATCATCTGTGGGAAACTCGCGTAAAATCAGGCGTTTTCGCCGATTTGCGA GGCTGGCCAGCTCCACGTCGCCGGCCGAAATCGAGCCTGCCCCTCATCTGTCAACGCCGCGCC GGGTGAGTCGGCCCCTCAAGTGTCAACGTCCGCCCCTCATCTGTCAGTGAGGGCCAAGTTTTC CGCGTGGTATCCACAACGCCGGCGGCCGCGGTGTCTCGCACACGGCTTCGACGGCGTTTCTGG CGCGTTTGCAGGGCCATAGACGGCCGCCAGCCCAGCGGCGAGGGCAACCAGCCCGGTGAGCGT CGGAAAGGCGCTGGAAGCCCCGTAGCGACGCGGAGAGGGGCGAGACAAGCCAAGGGCGCAGGC TCGATGCGCAGCACGACATAGCCGGTTCTCGCAAGGACGAGAATTTCCCTGCGGTGCCCCTCA AGTGTCAATGAAAGTTTCCAACGCGAGCCATTCGCGAGAGCCTTGAGTCCACGCTAGATCTAT CTCA	Requires trfA protein for replication

pBBR1 origin of replication	CTACGGGCTTGCTCTCCGGGCTTCGCCCTGCGCGGTCGCTGCGCTCCCTTGCCAGCCCGTGGA TATGTGGACGATGGCCGCGAGCGGCCACCGGCTGGCTCGCTTCGCTCGGCCCGTGGACAACCC TGCTGGACAAGCTGATGGACAGGCTGCGCCTGCCCACGAGCTTGACCACAGGGATTGCCCACC GGCTACCCAGCCTTCGACCACATACCCACCGGCTCCAACTGCGCGGCCTGCGGCCTTGCCCCA TCAATTTTTTTAATTTTCTCTGGGGAAAAGCCTCCGGCCTGCGGCCTGCGCGCTTCGCTTGCC GGTTGGACACCAAGTGGAAGGCGGGTCAAGGCTCGCGCAGCGACCGCGCAGCGGCTTGGCCTT GACGCGCCTGGAACGACCCAAGCCTATGCGAGTGGGGGCAGTCGAAGGCGAAGCCCGCCCGCC TGCCCCCCGAGCCTCACGGCGGCGAGTGCGGGGGTTCCAAGGGGGCAGCGCCACCTTGGGCAA GGCCGAAGGCCGCGCAGTCGATCAACAAGCCCCGGAGGGGCCACTTTTTGCCGGAGGGGGAGC CGCGCCGAAGGCGTGGGGGAACCCCGCAGGGGTGCCCTTCTTTGGGCACCAAAGAACTAGATA TAGGGCGAAATGCGAAAGACTTAAAAATCAACAACTTAAAAAAGGGGGGTACGCAACAGCTCA TTGCGGCACCCCCCGCAATAGCTCATTGCGTAGGTTAAAGAAAATCTGTAATTGACTGCCACT TTTACGCAACGCATAATTGTTGTCGCGCTGCCGAAAAGTTGCAGCTGATTGCGCATGGTGCCG CAACCGTGCGGCACCCTACCGCATGGAGATAAGCATGGCCACGCAGTCCAGAGAAATCGGCAT TCAAGCCAAGAACAAGCCCGGTCACTGGGTGCAAACGGAACGCAAAGCGCATGAGGCGTGGGC CGGGCTTATTGCGAGGAAACCCACGGCGGCAATGCTGCTGCATCACCTCGTGGCGCAGATGGG CCACCAGAACGCCGTGGTGGTCAGCCAGAAGACACTTTCCAAGCTCATCGGACGTTCTTTGCG GACGGTCCAATACGCAGTCAAGGACTTGGTGGCCGAGCGCTGGATCTCCGTCGTGAAGCTCAA CGGCCCCGGCACCGTGTCGGCCTACGTGGTCAATGACCGCGTGGCGTGGGGCCAGCCCCGCGA CCAGTTGCGCCTGTCGGTGTTCAGTGCCGCCGTGGTGGTTGATCACGACGACCAGGACGAATC GCTGTTGGGGCATGGCGACCTGCGCCGCATCCCGACCCTGTATCCGGGCGAGCAGCAACTACC GACCGGCCCCGGCGAGGAGCCGCCCAGCCAGCCCGGCATTCCGGGCATGGAACCAGACCTGCC AGCCTTGACCGAAACGGAGGAATGGGAACGGCGCGGGCAGCAGCGCCTGCCGATGCCCGATGA GCCGTGTTTTCTGGACGATGGCGAGCCGTTGGAGCCGCCGACACGGGTCACGCTGCCGCGCCG GTAG	Includes coding sequence of required replication protein
RSF1010 plasmid backbone (continued to next page)	GCTCGACCAGGCGTACGCTTATGGGTGCCTTTCCGCAGCTTGGAACGCGGATGGAGAAGAGGA GCAACGCGATCTAGCTATCGCGGCCGCGATCAAGCAGGTGCGACAGACGTCATACTAGATATC AAGCGACTTCTCCTATCCCCTGGGAACACATCAATCTCACCGGAGAATATCGCTGGCCAAAGC CTTAGCGTAGGATTCCGCCCCTTCCCGCAAACGACCCCAAACAGGAAACGCAGCTGAAACGGG AAGCTCAACACCCACTGACGCATGGGTTGTTCAGGCAGTACTTCATCAACCAGCAAGGCGGCA CTTTCGGCCATCCGCCGCGCCCCACAGCTCGGGCAGAAACCGCGACGCTTACAGCTGAAAGCG ACCAGGTGCTCGGCGTGGCAAGACTCGCAGCGAACCCGTAGAAAGCCATGCTCCAGCCGCCCG CATTGGAGAAATTCTTCAAATTCCCGTTGCACATAGCCCGGCAATTCCTTTCCCTGCTCTGCC ATAAGCGCAGCGAATGCCGGGTAATACTCGTCAACGATCTGATAGAGAAGGGTTTGCTCGGGT CGGTGGCTCTGGTAACGACCAGTATCCCGATCCCGGCTGGCCGTCCTGGCCGCCACATGAGGC ATGTTCCGCGTCCTTGCAATACTGTGTTTACATACAGTCTATCGCTTAGCGGAAAGTTCTTTT ACCCTCAGCCGAAATGCCTGCCGTTGCTAGACATTGCCAGCCAGTGCCCGTCACTCCCGTACT AACTGTCACGAACCCCTGCAATAACTGTCACGCCCCCCTGCAATAACTGTCACGAACCCCTGC AATAACTGTCACGCCCCCAAACCTGCAAACCCAGCAGGGGCGGGGGCTGGCGGGGTGTTGGAA AAATCCATCCATGATTATCTAAGAATAATCCACTAGGCGCGGTTATCAGCGCCCTTGTGGGGC GCTGCTGCCCTTGCCCAATATGCCCGGCCAGAGGCCGGATAGCTGGTCTATTCGCTGCGCTAG GCTACACACCGCCCCACCGCTGCGCGGCAGGGGGAAAGGCGGGCAAAGCCCGCTAAACCCCAC ACCAAACCCCGCAGAAATACGCTGGAGCGCTTTTAGCCGCTTTAGCGGCCTTTCCCCCTACCC GAAGGGTGGGGGCGCGTGTGCAGCCCCGCAGGGCCTGTCTCGGTCGATCATTCAGCCCGGCTC ATCCTTCTGGCGTGGCGGCAGACCGAACAAGGCGCGGTCGTGGTCGCGTTCAAGGTACGCATC CATTGCCGCCATGAGCCGATCCTCCGGCCACTCGCTGCTGTTCACCTTGGCCAAAATCATGGC CCCCACCAGCACCTTGCGCCTTGTTTCGTTCTTGCGCTCTTGCTGCTGTTCCCTTGCCCGCAC CCGCTGAATTTCGGCATTGATTCGCGCTCGTTGTTCTTCGAGCTTGGCCAGCCGATCCGCCGC CTTGTTGCTCCCCTTAACCATCTTGACACCCCATTGTTAATGTGCTGTCTCGTAGGCTATCAT GGAGGCACAGCGGCGGCAATCCCGACCCTACTTTGTAGGGGAGGGCGCACTTACCGGTTTCTC TTCGAGAAACTGGCCTAACGGCCACCCTTCGGGCGGTGCGCTCTCCGAGGGCCATTGCATGGA GCCGAAAAGCAAAAGCAACAGCGAGGCAGCATGGCGATTTATCACCTTACGGCGAAAACCGGC AGCAGGTCGGGCGGCCAATCGGCCAGGGCCAAGGCCGACTACATCCAGCGCGAAGGCAAGTAT GCCCGCGACATGGATGAAGTCTTGCACGCCGAATCCGGGCACATGCCGGAGTTCGTCGAGCGG CCCGCCGACTACTGGGATGCTGCCGACCTGTATGAACGCGCCAATGGGCGGCTGTTCAAGGAG GTCGAATTTGCCCTGCCGGTCGAGCTGACCCTCGACCAGCAGAAGGCGCTGGCGTCCGAGTTC GCCCAGCACCTGACCGGTGCCGAGCGCCTGCCGTATACGCTGGCCATCCATGCCGGTGGCGGC GAGAACCCGCACTGCCACCTGATGATCTCCGAGCGGATCAATGACGGCATCGAGCGGCCCGCC GCTCAGTGGTTCAAGCGGTACAACGGCAAGACCCCGGAGAAGGGCGGGGCACAGAAGACCGAA GCGCTCAAGCCCAAGGCATGGCTTGAGCAGACCCGCGAGGCATGGGCCGACCATGCCAACCGG GCATTAGAGCGGGCTGGCCACGACGCCCGCATTGACCACAGAACACTTGAGGCGCAGGGCATC GAGCGCCTGCCCGGTGTTCACCTGGGGCCGAACGTGGTGGAGATGGAAGGCCGGGGCATCCGC ACCGACCGGGCAGACGTGGCCCTGAACATCGACACCGCCAACGCCCAGATCATCGACTTACAG GAATACCGGGAGGCAATAGACCATGAACGCAATCGACAGAGTGAAGAAATCCAGAGGCATCAA CGAGTTAGCGGAGCAGATCGAACCGCTGGCCCAGAGCATGGCGACACTGGCCGACGAAGCCCG GCAGGTCATGAGCCAGACCCAGCAGGCCAGCGAGGCGCAGGCGGCGGAGTGGCTGAAAGCCCA GCGCCAGACAGGGGCGGCATGGGTGGAGCTGGCCAAAGAGTTGCGGGAGGTAGCCGCCGAGGT GAGCAGCGCCGCGCAGAGCGCCCGGAGCGCGTCGCGGGGGTGGCACTGGAAGCTATGGCTAAC CGTGATGCTGGCTTCCATGATGCCTACGGTGGTGCTGCTGATCGCATCGTTGCTCTTGCTCGA CCTGACGCCACTGACAACCGAGGACGGCTCGATCTGGCTGCGCTTGGTGGCCCGATGAAGAAC	Includes genes for mobilization proteins A, B, C and replication proteins A, B, C

GACAGGACTTTGCAGGCCATAGGCCGACAGCTCAAGGCCATGGGCTGTGAGCGCTTCGATATC
GGCGTCAGGGACGCCACCACCGGCCAGATGATGAACCGGGAATGGTCAGCCGCCGAAGTGCTC
CAGAACACGCCATGGCTCAAGCGGATGAATGCCCAGGGCAATGACGTGTATATCAGGCCCGCC
GAGCAGGAGCGGCATGGTCTGGTGCTGGTGGACGACCTCAGCGAGTTTGACCTGGATGACATG
AAAGCCGAGGGCCGGGAGCCTGCCCTGGTAGTGGAAACCAGCCCGAAGAACTATCAGGCATGG
GTCAAGGTGGCCGACGCCGCAGGCGGTGAACTTCGGGGGCAGATTGCCCGGACGCTGGCCAGC
GAGTACGACGCCGACCCGGCCAGCGCCGACAGCCGCCACTATGGCCGCTTGGCGGGCTTCACC
AACCGCAAGGACAAGCACACCACCCGCGCCGGTTATCAGCCGTGGGTGCTGCTGCGTGAATCC
AAGGGCAAGACCGCCACCGCTGGCCCGGCGCTGGTGCAGCAGGCTGGCCAGCAGATCGAGCAG
GCCCAGCGGCAGCAGGAGAAGGCCCGCAGGCTGGCCAGCCTCGAACTGCCCGAGCGGCAGCTT
AGCCGCCACCGGCGCACGGCGCTGGACGAGTACCGCAGCGAGATGGCCGGGCTGGTCAAGCGC
TTCGGTGATGACCTCAGCAAGTGCGACTTTATCGCCGCGCAGAAGCTGGCCAGCCGGGGCCGC
AGTGCCGAGGAAATCGGCAAGGCCATGGCCGAGGCCAGCCCAGCGCTGGCAGAGCGCAAGCCC
GGCCACGAAGCGGATTACATCGAGCGCACCGTCAGCAAGGTCATGGGTCTGCCCAGCGTCCAG
CTTGCGCGGGCCGAGCTGGCACGGGCACCGGCACCCCGCCAGCGAGGCATGGACAGGGGCGGG
CCAGATTTCAGCATGTAGTGCTTGCGTTGGTACTCACGCCTGTTATACTATGAGTACTCACGC
ACAGAAGGGGGTTTTATGGAATACGAAAAAAGCGCTTCAGGGTCGGTCTACCTGATCAAAAGT
GACAAGGGCTATTGGTTGCCCGGTGGCTTTGGTTATACGTCAAACAAGGCCGAGGCTGGCCGC
TTTTCAGTCGCTGATATGGCCAGCCTTAACCTTGACGGCTGCACCTTGTCCTTGTTCCGCGAA
GACAAGCCTTTCGGCCCCGGCAAGTTTCTCGGTGACTGATATGAAAGACCAAAAGGACAAGCA
GACCGGCGACCTGCTGGCCAGCCCTGACGCTGTACGCCAAGCGCGATATGCCGAGCGCATGAA
GGCCAAAGGGATGCGTCAGCGCAAGTTCTGGCTGACCGACGACGAATACGAGGCGCTGCGCGA
GTGCCTGGAAGAACTCAGAGCGGCGCAGGGCGGGGGTAGTGACCCCGCCAGCGCCTAACCACC
AACTGCCTGCAAAGGAGGCAATCAATGGCTACCCATAAGCCTATCAATATTCTGGAGGCGTTC
GCAGCAGCGCCGCCACCGCTGGACTACGTTTTGCCCAACATGGTGGCCGGTACGGTCGGGGCG
CTGGTGTCGCCCGGTGGTGCCGGTAAATCCATGCTGGCCCTGCAACTGGCCGCACAGATTGCA
GGCGGGCCGGATCTGCTGGAGGTGGGCGAACTGCCCACCGGCCCGGTGATCTACCTGCCCGCC
GAAGACCCGCCCACCGCCATTCATCACCGCCTGCACGCCCTTGGGGCGCACCTCAGCGCCGAG
GAACGGCAAGCCGTGGCTGACGGCCTGCTGATCCAGCCGCTGATCGGCAGCCTGCCCAACATC
ATGGCCCCGGAGTGGTTCGACGGCCTCAAGCGCGCCGCCGAGGGCCGCCGCCTGATGGTGCTG
GACACGCTGCGCCGGTTCCACATCGAGGAAGAAAACGCCAGCGGCCCCATGGCCCAGGTCATC
GGTCGCATGGAGGCCATCGCCGCCGATACCGGGTGCTCTATCGTGTTCCTGCACCATGCCAGC
AAGGGCGCGGCCATGATGGGCGCAGGCGACCAGCAGCAGGCCAGCCGGGGCAGCTCGGTACTG
GTCGATAACATCCGCTGGCAGTCCTACCTGTCGAGCATGACCAGCGCCGAGGCCGAGGAATGG
GGTGTGGACGACGACCAGCGCCGGTTCTTCGTCCGCTTCGGTGTGAGCAAGGCCAACTATGGC
GCACCGTTCGCTGATCGGTGGTTCAGGCGGCATGACGGCGGGGTGCTCAAGCCCGCCGTGCTG
GAGAGGCAGCGCAAGAGCAAGGGGGTGCCCCGTGGTGAAGCCTAAGAACAAGCACAGCCTCAG
CCACGTCCGGCACGACCCGGCGCACTGTCTGGCCCCCGGCCTGTTCCGTGCCCTCAAGCGGGG
CGAGCGCAAGCGCAGCAAGCTGGACGTGACGTATGACTACGGCGACGGCAAGCGGATCGAGTT
CAGCGGCCCGGAGCCGCTGGGCGCTGATGATCTGCGCATCCTGCAAGGGCTGGTGGCCATGGC
TGGGCCTAATGGCCTAGTGCTTGGCCCGGAACCCAAGACCGAAGGCGGACGGCAGCTCCGGCT
GTTCCTGGAACCCAAGTGGGAGGCCGTCACCGCTGAATGCCATGTGGTCAAAGGTAGCTATCG
GGCGCTGGCAAAGGAAATCGGGGCAGAGGTCGATAGTGGTGGGGCGCTCAAGCACATACAGGA
CTGCATCGAGCGCCTTTGGAAGGTATCCATCATCGCCCAGAATGGCCGCAAGCGGCAGGGGTT
TCGGCTGCTGTCGGAGTACGCCAGCGACGAGGCGGACGGGCGCCTGTACGTGGCCCTGAACCC
CTTGATCGCGCAGGCCGTCATGGGTGGCGGCCAGCATGTGCGCATCAGCATGGACGAGGTGCG
GGCGCTGGACAGCGAAACCGCCCGCCTGCTGCACCAGCGGCTGTGTGGCTGGATCGACCCCGG
CAAAACCGGCAAGGCTTCCATAGATACCTTGTGCGGCTATGTCTGGCCGTCAGAGGCCAGTGG
TTCGACCATGCGCAAGCGCCGCCAGCGGGTGCGCGAGGCGTTGCCGGAGCTGGTCGCGCTGGG
CTGGACGGTAACCGAGTTCGCGGCGGGCAAGTACGACATCACCCGGCCCAAGGCGGCAGGCTG
ACCCCCCCCACTCTATTGTAAACAAGACATTTTTATCTTTTATATTCAATGGCTTATTTTCCT
GCTAATTGGTAATACCATGAAAAATACCATGCTCAGAAAAGGCTTAACAATATTTTGAAAAAT
TGCCTACTGAGCGCTGCCGCACAGCTCCATAGGCC

RCR	TCCGCCGCCCTAGACCTAGTGTCATTTTATTTCCCCCGTTTCAGCATCAAGAACCTTTGCATA ACTTGCTCTATATCCACACTGATAATTGCCCTCAAACCATAATCTAAAGGCGCTAGAGTTTGT TGAAACAATATCTTTTACATCATTCGTATTTAAAATTCCAAACTCCGCTCCCCTAAGGCGAAT AAAAGCCATTAAATCTTTTGTATTTACCAAATTATAGTCATCCACTATATCTAAGAGTAAATT CTTCAATTCTCTTTTTTGGCTTTCATCAAGTGTTATATAGCGGTCAATATCAAAATCATTAAT GTTCAAAATATCTTTTTTGTCGTATATATGTTTATTCTTAGCAATAGCGTCCTTTGATTCATG AGTCAAATATTCATATGAACCTTTGATATAATCAAGTATCTCAACATGAGCAACTGAACTATT CCCCAATTTTCGCTTAATCTTGTTCCTAACGCTTTCTATTGTTACAGGATTTCGTGCAATATA TATAACGTGATAGTGTGGTTTTTTATAGTGCTTTCCATTTCGTATAACATCACTACTATTCCA TGTATCTTTATCTTTTTTTTCGTCCATATCGTGTAAAGGACTGACAGCCATAGATACGCCCAA ACTCTCTAATTTTTCCTTCCAATCATTAGGAATTGAGTCAGGATATAATAAAAATCCAAAATT TCTAGCTTTAGTATTTTTAATAGCCATGATATAATTACCTTATCAAAAACAAGTAGCGAAAAC TCGTATCCTTCTAAAAACGCGAGCTTTCGCTTATTTTTTTTGTTCTGATTCCTTTCTTGCATA TTCTTCTATAGCTAACGCCGCAACCGCAGATTTTGAAAAACCTTTTTGTTTCGCCATATCTGT TAATTTTTTATCTTGCTCTTTTGTCAGAGAAATCATAACTCTTTTTTTCGATTCTGAAATCAC CATTTAAAAAACTCCAATCAAATAATTTTATAAAGTTAGTGTATCACTTTGTAATCATAAAAA CAACAATAAAGCTACTTAAATATAGATTTATAAAAAACGTTGGCGAAAACGTTGGCGATTCGT TGGCGATTGAAAAACCCCTTAAACCCTTGAGCCAGTTGGGATAGAGCGTTTTTGGCACAAAAA TTGGCACTCGGCACTTAATGGGGGGTCGTAGTACGGAAGCAAAATTCGCTTCCTTTCCCCCCA TTTTTTTCCAAATTCCAAATTTTTTTCAAAAATTTTCCAGCGCTACCGCTCGGCAAAATTGCA AGCAATTTTTAAAATCAAACCCATGAGGGAATTTCATTCCCTCATACTCCCTTGAGCCTCCTC CAACCGAAATAGAAGGGCGCTGCGCTTATTATTTCATTCAGTCATCGGCTTTCATAATCTAAC AGACAACATCTTCGCTGCAAAGCCACGCTACGCTCAAGGGCTTTTACGCTACGATAACGCCTG TTTTAACGATTATGCCGATAACTAAACGAAATAAACGCTAAAACGTCTCAGAAACGATTTTGA GACGTTTTAATAAAAAATCGCCTAGTGC	

Transposon inverted repeat sequences	**Sequence**
Himar	ACAGGTTGGATGATAAGTCCCCGGTCT
Tn5	CTGTCTCTTATACACATCT

Regulatory sequences (5' UTRs, incl. promoter and RBS)	**Sequence**	**Notes**
1	GATTGCATTAGGTTTTAGTTTCTTGTATAATGCTTAAT GTTGGTCACTGACAGGCTACGATACGGAAGGTTGCTCA CGCCCGGCCCCTTTGCCATGGCTAGTGTGTGGAAATTT CCGAGGAGCAAGTCTATTTCCAAAAATGGGCGAAAAAG GAGGTAATACA	From *Bacillus cellulosilyticus*
2	GGGAGAGCTTCAACGGCGCTTCTACCCATTTGCTTGGA AAGGATGAGGAGCAGGAAGAAATTCCGTCCCCAATGCG ACGGCCCTTTACATCCATGTTGTTTGATAGTATAATGG ATACGGATTGACCAAATTGTTCATTTAGTCAGTTTGAA GGATGAGGAGT	From *Geobacillus sp.*
3	GTGAAGGATACGGCTGCGGCACTTCGACATCGCCCCAT GTGGCGGCTTTGAACTGGGCTTATGAAACGCGTTCACA ACCTTTTTTGACCATCGGCGCGAACGTGGTATCATGCG TTCAGCTTTTGCCCATACATACTACGTGCTCAATCTAG GAGGATTTCATAC	From *Eggerthella lenta*
4	CTCTAGAGTAGTAGATTATTTTAGGAATTTAGATGTTT TGTATGAAATAGATGCTTCGTATGGAATTAATGAAATT TTTAGTCAGGTAAAAAAGGTAATAGGAGAATATT	From Segmented Filamentous Bacteria (SFB)
5	GTTTTAAATGATGAAAAGAAATATTTAGGGAAGATTGT TTCGACGCGAATTGTTGATCTGGAAAATGATCACCTTA TCGGACAAGCTTTAAAATAGGAGGATATAAAAAT	From Segmented Filamentous Bacteria (SFB)
6	ATAAGGATTCTTTAAAGAGAGATATAGTTATGTCAAAG ACTGTAGAATTTTTAGTAAATCAAAATAAAAAAAGAGG TATTAAATAGAGTGTATTTTAAAGGAGGAGACTT	From Segmented Filamentous Bacteria (SFB)
7	AAACACCAATAAAATTAGAATATTTAGGAGCGACTTTA AAAAAGTTTAATAAGAATTGTTTATGAGATATTTTTAT TATATTTAAACTCAATTTAAAGTAGGGAGAATAG	From Segmented Filamentous Bacteria (SFB)
8	GCAAGTGTTCAAGAAGTTATTAAGTCGGGAGTGCAGTC GAAGTGGGCAAGTTGAAAAATTCACAAAAATGTGGTAT AATATCTTTGTTCATTAGAGCGATAAACTTGAATTTGA GAGGGAACTTAG	From *Clostridium perfringens*

Primers for PCR validation of transconjugants:

16S forward	AGAGTTTGATCATGGCTCAG
16S reverse	CGGTTACCTTGTTACGACTT
GFP validation primer forward	ATGCGTAAAGGCGAAGAGC
GFP validation primer reverse	TTATTTGTACAGTTCATCCATACCATG
Beta-lactamase validation primer forward	ATGAGTATTCAACATTTCCGTGTC
Beta-lactamase validation primer reverse	TTACCAATGCTTAATCAGTGAGGC
pGT-B backbone validation primer forward	CTGCGCAACCCAAGTGCTAC
pGT-B backbone validation primer reverse	CAGTCCAGAGAAATCGGCATTCA
pGT-Ah backbone validation primer forward	ATGGAAAAAAAGGAATTTCGTGTTTTG
pGT-Ah backbone validation primer reverse	TTATTCAACATAGTTCCCTTCAAGAGC
CarbR internal forward primer	CCGAAGAACGTTTTCCAATGATGAG
GFP internal reverse primer	TGATTGTCTGGCAGCAGAAC
catP (chlor resistance) validation primer forward	GCAAGTGTTCAAGAAGTTATTAAGTC
catP (chlor resistance) validation primer reverse	TTAACTATTTATCAATTCCTGCAATTCG
tetQ (tet resistance) internal forward primer	TGGAAGAACGATTTTCCGTAAAGGT

Supplementary Table 3.4. List of isolated transconjugant strains. Strains are grouped by the mouse cohort they were isolated from and the vector library used in the study. All family-level assignments were made using the RDP classifier with confidence >0.89.

Taconic mice *in situ* conjugations

Vector library	Family	Genus	Genus-level assignment confidence	Vector received	Antibiotic resistance
pGT-L6	Erysipelotrichaceae (Clostridium XVIII)	*Erysipelotrichaceae incertae sedis*	1	pGT-Ah	carb
	Bacteroidaceae	*Bacteroides*	1	pGT-Ah	carb
	Enterobacteriaceae	*Proteus*	1	pGT-Ah	carb
	Enterobacteriaceae	*Citrobacter*	1	pGT-Ah	carb
	Enterococcaceae	*Enterococcus*	1	pGT-Ah	carb
	Lachnospiraceae	*Hungatella*	0.72	pGT-Ah	carb
	Lachnospiraceae	*Clostridium XlVa*	1	pGT-Ah	carb
	Lachnospiraceae	*Anaerostipes*	1	pGT-Ah	carb
	Lachnospiraceae	*Moryella*	0.19	pGT-Ah	carb
	Lachnospiraceae	*Blautia*	1	pGT-Ah	carb
	Lactobacillaceae	*Lactobacillus*	1	pGT-Ah	carb
	Peptostreptococcaceae	*Clostridium XI*	1	pGT-Ah	carb
pGT-L3	Coriobacteriaceae	*Eggerthella*	1	pGT-S	tet
	Enterobacteriaceae	*Cosenzaea*	0.73	pGT-S	tet
	Enterobacteriaceae	*Proteus*	1	pGT-S	tet
	Enterococcaceae	*Enterococcus*	1	pGT-S	carb
	Lachnospiraceae	*Lactonifactor*	0.7	pGT-S	tet
	Lachnospiraceae	*Clostridium XlVa*	1	pGT-S	carb
	Lachnospiraceae	*Hungatella*	0.71	pGT-S	tet
	Lachnospiraceae	*Clostridium XlVa*	1	pGT-S	tet
	Lachnospiraceae	*Blautia*	1	pGT-S	tet
	Lachnospiraceae	*Robinsoniella*	0.42	pGT-S	tet
	Lachnospiraceae	*Eisenbergiella*	0.99	pGT-S	tet
	Lactobacillaceae	*Lactobacillus*	0.89	pGT-S	tet

Charles River mice *in situ* conjugations

Vector library	Family	Genus	Genus-level assignment confidence	Vector received	Antibiotic resistance
pGT-L6	Bacteroidaceae	*Bacteroides*	1	pGT-Ah	carb
	Enterococcaceae	*Enterococcus*	1	pGT-Ah	carb
	Lactobacillaceae	*Lactobacillus*	1	pGT-Ah	carb
	Porphyromonadaceae	*Parabacteroides*	1	pGT-Ah	carb

***In vitro* conjugations**

Vector library	Family	Genus	Genus-level assignment confidence	Vector received	Antibiotic resistance
pGT-L7	Enterobacteriaceae	*Proteus*	1	pGT-Ah	carb
	Enterococcaceae	*Enterococcus*	1	pGT-Ah	carb
	Enterobacteriaceae	*Escherichia*	1	pGT-Ah	carb
	Lactobacillaceae	*Lactobacillus*	1	pGT-Ah	carb
	Bacillaceae	*Bacillus*	1	pGT-Ah	carb
pGT-L3	Enterobacteriaceae	*Escherichia*	1	pGT-S	carb
	Enterococcaceae	*Enterococcus*	1	pGT-S	carb
	Enterobacteriaceae	*Proteus*	1	pGT-S	carb
pGT-L5	Enterobacteriaceae	*Cosenzae*	0.89	pGT-Ah	chl
	Enterobacteriaceae	*Proteus*	1	pGT-Ah	chl
	Burkholderiaceae	*Cupriavidus*	1	pGT-Ah	chl
pGT-L4	Enterobacteriaceae	*Escherichia*	1	pGT-Ah	carb
	Enterobacteriaceae	*Proteus*	1	pGT-Ah	carb
	Enterobacteriaceae	*Escherichia*	1	pGT-B	carb

Chapter 4: Genetic engineering of Segmented Filamentous Bacteria using MAGIC

4.1. Introduction

Mammalian gut bacteria are adapted to live in a highly complex environment teeming with diverse microbes and a wide range of chemicals arising from food digestion, host metabolism, and microbial metabolism. Within this environment, some bacteria occupy niches which are difficult or impossible to recapitulate *in vitro*, making these bacteria recalcitrant to cultivation. Although advances in "culturomics" have enabled cultivation of increasing numbers of gut commensals, over 20% of known human gut microbes remain uncultivable and difficult to study genetically[81]. The Metagenomic Alteration of Gut microbiome by *In situ* Conjugation (MAGIC) system described in Chapter 3 circumvents the need for cultivation, potentially enabling genetic manipulation of undomesticated gut bacteria *in situ* and greatly expanding the scope of functional genetic studies of host-microbiome interactions that can be done.

One gut bacterium of great interest in the context of host-microbial interactions that remains difficult to cultivate is Segmented Filamentous Bacteria (SFB), also known as *Candidatus* Arthromitus or *Candidatus* Savagella. SFB is a spore-forming, Gram positive member of the *Clostridiaceae* family. The first study of SFB's habitat, morphology, and life cycle was published in 1974, describing segmented, filamentous-shaped bacteria found in the ilea of weaned mice and rats that grew with one filament end tightly attached to gut epithelial cells[159]. Bacteria with this morphology were subsequently identified in a wide range of animals, including amphibians, fish, birds, and mammals, including monkeys and humans[160, 161]. In nearly all species examined, SFB-like bacteria colonize the ileum and attach directly to the epithelium, suggesting that these bacteria occupy similar niches and interact closely with their hosts[162].

The SFB of mice and rats are the most closely studied out of this panoply of morphologically similar bacteria. Early studies in the 1970s used electron microscopy to visualize SFB and characterize its life cycle, which involves attachment of a cell to the gut epithelium by a holdfast, growth of the filament by division of segments, and differentiation and release of new holdfasts from the distal end of the filament[163]. The use of gnotobiotic mice and rats mono-associated with SFB has enabled *in vivo* cultivation of SFB monocultures[164], which have been used to sequence the SFB genome[165, 166]. Sequencing studies have shown that SFB is an obligate anaerobe with a small genome size (1.6 Mbp) and low genomic GC content (28%) and lacks genes for the biosynthesis of most amino acids, vitamins/cofactors and nucleotides[167, 168]. These findings suggest that SFB is an obligate or facultative mammalian symbiont and relies on the host for many essential nutrients.

Mouse studies have shown that SFB directly affects host physiology and disease pathogenesis. Colonization with SFB is sufficient to induce differentiation of Th17 cells in the small-intestinal lamina propria of germ-free mice and protect the mice against infection by *Citrobacter rodentium*, an intestinal pathogen[21]. In conventional mice that harbor SFB along with other intestinal bacteria, SFB antigens stimulate of over 60% of intestinal Th17 cells, indicating that SFB is a dominant microbial signal for Th17 induction[96]. Furthermore, Th17 induction requires SFB colonization and adhesion to the intestine and cannot be caused by presence of SFB antigens alone, indicating that that SFB has organismal properties that modulate the host immune system[96, 169]. SFB is host-specific: although rats and mice both have phylogenetically similar SFB, each species' SFB only adheres and induces an immune response specifically in its native host[169]. SFB also induces intestinal IgA secretion and activates intraepithelial lymphocytes[170]. As a consequence of its role as an immune system regulator, SFB modulates the incidence and progression of several autoimmune diseases in mouse models. SFB causes experimental autoimmune encephalomyelitis[171], a model for multiple sclerosis, as well as autoimmune

arthritis[172], two diseases dependent on Th17 cell activation. Interestingly, SFB protects against the development of type 1 diabetes in female, but not male, mice[173].

Given the key role SFB plays in immune regulation and its relevance to autoimmune diseases, the biology of SFB warrants further study. The difficulty of cultivating and propagating SFB *in vitro*[174, 175] has hindered the development of genetic tools for SFB; SFB has yet to be transformed or otherwise genetically manipulated. Because SFB's unique relationship to the host is dependent on the organism itself, functional genetic studies would be invaluable for elucidating SFB's life cycle and symbiosis with the host. In the work described in this chapter, we tested the hypothesis that SFB could be genetically manipulated for the first time using the MAGIC system. Specifically, we wanted to demonstrate that that SFB could receive engineered DNA by horizontal gene transfer from an *E. coli* donor within the murine gut, and that SFB is capable of expressing heterologous genes. The work presented in this chapter describes the design principles for optimizing *in situ* gene transfer and gene expression in an organism with no prior available genetic tools, and the experimental principles for measuring trans-gene expression in an uncultivable gut commensal. These preliminary studies lay the foundation for future optimization of genetic tools for SFB and serve as a proof of principle for the utility of MAGIC in genetically manipulating undomesticated microbes.

4.2. Results

Identification of SFB genetic parts and design of MAGIC vectors.

We aimed to implement MAGIC technology in SFB by *in situ* conjugative delivery of two novel genetically encoded traits, GFP and an antibiotic resistance gene. These two genetic functions would enable assessment of trans-gene expression by two orthogonal methods—measurement of GFP fluorescence by microscopy or FACS, and antibiotic selection (**Figure**

4.1a). The genetic payload was delivered on a conjugative Himar transposon, with an associated hyperactive Himar transposase (**Figure 4.1b**). Because no known replicative plasmids were available for SFB, the host-independent, broadly active Himar transposon was used as an integrative vector to enable long-term, stable maintenance of the delivered payload in SFB. An additional advantage of using an integrative Himar transposon to deliver the payload is that the resulting transposon mutant library could potentially be analyzed by transposon sequencing to determine genes in SFB that affect fitness (**Figure 4.1a**).

Because no toolbox of genetic parts was readily available for SFB, we curated a library of antibiotic resistance genes, promoters, and ribosomal binding sites (RBS) that were likely active in SFB in order to build the MAGIC vectors. We chose specific antibiotic resistance genes for SFB MAGIC vectors based on available Clostridia-specific resistance markers and on experimental measurement of SFB's antibiotic susceptibilities in the mouse gut. To maximize the likelihood of conferring a new antibiotic resistance to SFB, we searched the literature for resistance genes originally identified in closely related Clostridia, which were more likely to have biological activity in SFB. Resistance genes for chloramphenicol (*catP*) and erythromycin (*ermBP*) had been identified in *Clostridium perfringens* and were subsequently shown to confer resistance to *E. coli* when cloned into plasmids[176-178], suggesting that these genes were broadly active across bacteria and likely to also function in SFB. We confirmed that SFB is susceptible to these two drugs by treating C57BL/6 mice with antibiotics in drinking water and measuring fecal SFB levels relative to all bacteria over time by qPCR (**Figure 4.1c**). During treatment with either chloramphenicol or erythromycin for 4 days, SFB levels dropped significantly and stayed low after antibiotic treatment ended. SFB prevalence was not affected by neomycin administration. These results indicate that SFB is susceptible to chloramphenicol and erythromycin, but not neomycin, at dosages commonly used in veterinary medicine and that these antibiotics may be used for selection of antibiotic-resistant SFB.

We also sought to optimize expression of the selectable payload and the Himar transposase by identifying regulatory elements (promoters and ribosomal binding sites) likely to be active in SFB. The native *C. perfringens* promoter/RBS sequence from plasmid pJIR750[178] was used to drive expression of the payload, a two-gene operon consisting of *C. perfringens* chloramphenicol resistance gene catP and a superfolding GFP gene (sfGFP). To drive expression of the Himar transposase, we used four promoter sequences from the SFB genome that were likely to be constitutively active. These promoters were all 110 bp 5' untranslated regions (UTRs) of putative constitutively expressed housekeeping genes in SFB. We identified these genes based on homology to the most highly constitutively expressed genes in the *Clostridium beijerinckii* genome[179], hypothesizing that the SFB homologs performed similar functions and were expressed similarly. Each UTR also contained a bacterial RBS[180]; the last 36 bp of SFB promoter 4, containing the RBS, was used to drive translation of sfGFP in the catP/sfGFP operon (**Figure 4.1b**). We validated that all 5' UTRs were capable of driving GFP expression in *E. coli* (**Table 4.1**), suggesting that these UTRs are sufficient to initiate transcription and translation in a broad range of bacteria. Putative SFB genetic parts are summarized in **Table 4.1**, and the MAGIC vectors for SFB (pSFB plasmids) constructed using these parts are summarized in **Table 4.2**.

Design of MAGIC donor strains for optimized gene transfer into SFB.

As in Chapter 3, we utilized an *E. coli* donor strain with a genomically integrated IncPα-family RP4 conjugation system[138] to conjugate genetic vectors into SFB *in situ.* Donor strain *EcGT2* (Δasd::mCherry-specR) was additionally engineered to be auxotrophic for the essential cell wall component diaminopimelic acid (DAP)[139]. Spectinomycin resistance and DAP auxotrophy provide simple selection and counter-selection for the donor strain, while the constitutively expressed, genomically encoded mCherry allows for identification of donor cells in FACS or microscopy applications.

In general, restriction-methylation systems are barriers to horizontal gene transfer between species, as these systems recognize and cleave transferred DNA with foreign methylation patterns[135, 181]. We anticipated restriction-methylation encoded in the SFB genome to limit the transfer of MAGIC vectors from *E. coli*. To circumvent this potential problem, we expressed native SFB DNA methyltransferases in the *E. coli* donor to methylate MAGIC vectors in an SFB-like pattern prior to conjugation. We created two derivative strains of EcGT2, EcGT3 and EcGT4, each of which contains 2 annotated SFB methyltransferases on a pSC101 origin plasmid (**Figure 4.2a**). In EcGT4, one of the methyltransferases belongs to a Type I restriction-modification system and was cloned with its associated specificity subunit into the plasmid[182]. The presence of additional methyltransferases in the *E. coli* donor strain resulted in slightly slower growth *in vitro* (**Figure 4.2b**).

Preliminary studies of *in situ* gene delivery in SFB-monoassociated mice.

With the help of our collaborators in Dr. Kenya Honda's lab at RIKEN Center for Integrative Medical Sciences, we tested the MAGIC system for genetic engineering of SFB *in situ* using SFB-monoassociated (SFB-mono) mice in a gnotobiotic facility. EcGT3 and EcGT4 strains carrying one of four pSFB MAGIC vectors were gavaged as a mix of 8 strains into young adult SFB-mono mice over 2 days (**Figure 4.3a**, groups A-C). In a separate gnotobiotic isolator (group D), mice were gavaged with EcGT3 and EcGT4 strains without MAGIC vectors as a control to examine the length of *E. coli* colonization in the SFB-mono mouse gut and serve as a negative control for flow cytometry. Fecal samples taken over time post-gavage were analyzed by microscopy of fecal smears and by flow cytometry of fecal bacteria to look for GFP+ SFB. The presence of *E. coli* in the gut over time was monitored by plating fecal contents *in vitro* and counting CFUs as well as by microscopy of fecal smears. Additional fecal samples were taken at each time point post-gavage and frozen for future analysis. Tissue samples and luminal contents were also collected from each mouse post-euthanasia for further analysis.

The EcGT3 and EcGT4 donors transiently colonized the GI tract of SFB-mono mice post-oral gavage. In a separate preliminary experiment, mice gavaged with a single bolus of 10^9 *E. coli* cells had no detectable *E. coli* in their feces after 72 hours (data not shown). After being orally gavaged to SFB-mono mice as two boluses of 10^10 cells, the EcGT3 and EcGT4 donors transiently colonized the gut for at least 4 days but less than 7 days, being undetectable by plating and microscopy by 7 days (**Figure 4.3b**). Flow cytometry for SFB on day 4 showed the presence of GFP-expressing, mCherry-negative bacteria in feces from the mice treated with MAGIC vector donors, but not the mice treated with donor controls or untreated mice (**Figure 4.4a-b**). As plating and microscopy showed that the feces was free of contaminating bacteria, the GFP+/mCherry- bacteria were likely to be GFP-expressing SFB that had received the pSFB vector.

The mice that received the pSFB vectors were given antibiotics in drinking water starting at day 4 or day 8 to select for pSFB recipients. Mice in groups A and B were given chloramphenicol to select for SFB with the pSFB genetic payload and neomycin to select against *E. coli* donors, while mice in group C were given only neomycin. Feces of group 1A mice were examined at day 7 and 8, after 3 and 4 days of antibiotic treatment, respectively. SFB filaments were present in feces on both days, although in lower numbers than before. Filaments on day 8 appeared fainter and shorter than those at day 7 (**Figure 4.3b**), suggesting that SFB were being selected against by chloramphenicol. Flow cytometry of fecal bacteria collected on day 8 showed no GFP+ SFB in groups A, B, or D (**Figure 4.4c**). This result indicates that GFP+ SFB did not maintain GFP expression 4 days after being detected, in the presence or absence of antibiotic selection for the payload. After antibiotic selection was completed, mice were euthanized for collection of intestinal tissue samples.

Preliminary study of *in situ* gene delivery in NSG mice.

NSG mice colonized with SFB-mono mouse feces were used as a specific pathogen-free (SPF) animal testbed for genetically manipulating SFB within a conventional microbiome. NSG mice orally gavaged with feces from SFB-mono mice have higher titers of SFB (personal communication with Ivo Ivanov) than wild-type mice, potentially increasing the likelihood of detecting transconjugant SFB after MAGIC. We gavaged these mice with EcGT3 and EcGT4 donors harboring pSFB vectors (n = 4 mice) or with empty EcGT3/EcGT4 donors as a negative control (n=1) over three days (**Figure 4.5**). We collected fecal samples over 6 days following the first gavage for FACS enrichment of GFP+ SFB and tissue samples at the end of the experiment. Mice were not treated with antibiotics, so any gene transfer into SFB would occur in the context of a native microbiome.

We attempted to isolate GFP-expressing SFB directly by FACS, following previously published protocols for FACS isolation of SFB from SFB-mono mouse feces[183]. The NycoPrep enrichment of SFB described by Farkas et al. failed to enrich SFB to sortable levels in the NSG mouse feces, so we omitted that step and sorted fecal bacteria directly extracted from feces, stained with SYTO40, and filtered through a cell strainer. However, we were unable to detect SFB based on SSC-W and SSC-A profiles, likely due to the low abundance of SFB compared with other gut bacteria. Fecal samples were frozen for future analysis.

4.3. Discussion

In this chapter, we rationally curated the first set of genetic parts and selection markers for SFB and attempted pilot experiments to genetically engineer SFB *in situ* in gnotobiotic SFB-mono mice and in SPF NSG mice. The SFB-mono mouse experiment provided useful information on the survival of *E. coli* MAGIC donors in the guts of gnotobiotic mice and promising FACS data showing potential GFP-expressing SFB prior to antibiotic treatment. In the NSG mouse experiment, SFB

was not detectable by FACS among the high abundance of other bacteria in a conventional mouse microbiome, making it difficult to assess whether horizontal gene transfer into SFB occurred. These studies require additional follow-up work to confirm and replicate the observed results and to optimize protocols for genetically manipulating SFB in functional genetic studies.

In the SFB-mono experiment, mice treated with *E. coli* donors of pSFB conjugative transposon plasmids appeared to have GFP-expressing, mCherry-negative bacteria in their feces 4 days later, which were presumed, but not confirmed, to be genetically modified SFB. This experiment should be repeated, with FACS-enrichment for the presumed GFP+ SFB and 16S sequencing on these cells to verify that they were in fact SFB and not incorrectly gated *E. coli*. Microscopy on the FACS-enriched population would enable distinguishing between GFP+ SFB vs. *E. coli*, which have very different morphologies. Using a GFP+ negative control donor strain, rather than the mCherry+/GFP- empty donor strain, would also provide a more directly comparable GFP+ background in flow cytometry studies.

The antibiotic selection protocol used in the pilot SFB-mono experiment requires additional optimization for future experiments. We performed antibiotic selection using chloramphenicol and neomycin to select for SFB carrying the genetic payload in one group of pSFB-treated mice (Group A, **Figure 4.2a**), but failed to see expansion of the GFP+ SFB population after antibiotics, only a decrease in the number and size of filaments (**Figure 4.2b**). We did not follow up on the effects of antibiotic treatment on the other groups, which were dosed on different schedules, nor on group A after day 8, when antibiotics were discontinued. PCR quantification of SFB prevalence in fecal samples collected throughout the experiment will elucidate whether some SFB survived the antibiotic regimen and expanded after antibiotic selection. If all SFB were killed, the chloramphenicol concentration we used may have been too high. We had previously validated that this level of chloramphenicol was sufficient to eliminate SFB from the gut (**Figure 4.1c**), but had not identified the minimum inhibitory concentration (MIC) of chloramphenicol on SFB. The MIC can be

determined by giving cohorts of SFB-harboring mice different dosages of chloramphenicol (or other antibiotics) and measuring SFB prevalence in feces by qPCR. In future iterations of the mouse experiment, selecting for payload-carrying SFB using the MIC of chloramphenicol may result in better retention of those SFB. Additionally, the co-administration of neomycin to kill off *E. coli* donor cells may not have been optimal. It is possible that administering chloramphenicol only or the two antibiotics in succession might have resulted in more effective selection for payload-containing SFB; further experiments testing different antibiotic dosing schedules are required.

Optimization of the donor strain and donor dosage is also likely to influence the rate of HGT into SFB. We attempted to optimize donor strains EcGT3/4 for HGT into SFB by expressing SFB DNA methyltransferase systems in these strains. However, we have not yet confirmed that the presence of these additional methyltransferases actually results in new DNA methylation patterns in these cells, which can be done by single-molecule real-time (SMRT) sequencing[184]. It is also unclear to what extent the native *E. coli* methyltransferases present in EcGT3/4 inhibit HGT into SFB. If DNA methylation is a significant barrier to gene transfer, the methylation-free *E. coli* strain ER3413[185] may serve as the basis for engineering an optimized donor strain harboring only SFB DNA methyltransferases. The timing and dosage of donor cells also needs to be further studied. In SFB-mono mice, the donor strain persisted for 4-7 days after being gavaged as 2 boluses of 10^10 cells, but even at day 4, there were few donor cells present in the gut. The life cycle of SFB takes about 3 days, and it is unclear at what stage of its life cycle SFB is most receptive to HGT. The donor gavage schedule may need to be altered to maintain a high titer of donor cells for a longer time to facilitate HGT events.

If the MAGIC platform can be optimized to enable efficient gene delivery and expression in SFB, it can enable functional genetic studies of SFB for the first time, without the need for *in vitro* cultivation. Introducing a random transposon and associated transposase (such as Himar) into SFB would allow for the generation of a transposon knockout library and screening for genes that are

involved in epithelial adhesion, life cycle control, and immune response stimulation. Novel genes (for example, genes from rat SFB) could be introduced into SFB to determine which genes are responsible for host specificity[169]. With genetic engineering tools available, the biology of SFB could be greatly elucidated by functional screens, and SFB could potentially be developed as a synthetic biological chassis for delivering molecules to the intestinal epithelium in therapeutic applications.

4.4. Materials and methods

Construction of pSFB vectors. Plasmids were constructed by isothermal assembly (ITA) using existing plasmid backbones (p15A), resistance markers (catP from pJIR750, *E. coli* beta-lactamase), and Himar1 transposon and transposase sequences from pSAM-BT. Novel promoters identified from the SFB genome were ordered as a gBlock of 4 concatenated promoters (Integrated DNA Technologies) and subsequently PCR amplified and used in ITA reactions. PCR reactions were performed using Kapa Hifi PCR Master Mix (KAPA Biosystems). ITA reactions were performed using the Hifi DNA Assembly kit (New England Biolabs).

Construction of SFB methyltransferase plasmids. Five SFB DNA methylation systems (SFBM_0801 and SFBM_0802, SFBM_0805, SFBM_0018 and SFBM_0020, SFBM_1080, SFBM_0619) were inferred from the annotated SFB-mouse-Japan genome (NCBI accession number NC_015913). These methylation systems were amplified from SFB genomic DNA using the primers listed in Table 3, using Kapa Hifi PCR Master Mix. To make a plasmid containing all 5 methylation systems, the 5 PCR products and a pSC101 vector backbone were mixed in equimolar ratios in an ITA reaction using the Hifi DNA Assembly kit (New England Biolabs). Among the reaction products, we did not produce the intended plasmid, but only plasmids containing at most 2 out of 5 methylation systems, likely due to the large DNA fragment sizes and low GC content.

Plasmids pSFBMT1 and pSFBMT2 were confirmed by PCR and Sanger sequencing to contain 2 methylation systems each and were transformed into *E. coli* strain EcGT2 to make strains EcGT3 and EcGT4.

Mouse care. Gnotobiotic mice were housed in germ-free isolators and fed a standard germ-free diet. NSG mice were purchased from Jackson Laboratories and raised in a SPF animal facility, until 2 days prior to gavage, when they were moved to sterile cages within a sterile laminar flow hood. NSG mice were fed a standard autoclaved diet. Female young adults (7-8 weeks old) were used for all experiments. When applicable, drinking water was supplemented with DAP (500μM), chloramphenicol (0.5 g/L), neomycin (1 g/L), and/or erythromycin (0.1 g/L). Water containing antibiotics was also supplemented with sterile filtered 1% sucrose to encourage water intake. Water was changed every 2 days to prevent antibiotic degradation.

Culturing E. coli donor strains for gavage. EcGT3 and EcGT4 strains were grown in LB + DAP 50μM + tet 15 μg/mL. EcGT3/EcGT4 strains harboring pSFB plasmids were grown in LB + DAP 50μM + tet 15 μg/mL + carb 50 μg/mL. *E. coli* were grown in a 37C shaking incubator for 12-14 hours from a frozen glycerol stock; 200 μL of the saturated overnight culture was diluted into 4 mL of fresh media and grown at 37C with shaking for 6 hours. Cell density was quantified by OD600 using a Nanodrop. Cells were spun down at maximum speed on a tabletop centrifuge at room temperature and resuspended in PBS with 50μM DAP twice; 10^9 -10^{10} cells were then suspended in a final volume of 300 μL PBS with 50μM DAP for gavage.

Analysis of SFB-mono fecal bacteria. Fecal samples collected from SFB-mono mice were analyzed by microscopy, plating, and flow cytometry. Fecal smears on microscope slides were Gram stained and imaged by bright field microscopy. Fecal pellets were weighed and

homogenized in 1 mL final volume of PBS per pellet, as described in Chapter 3. Bacterial extracts were plated on LB + 50 μM DAP or GAM + 50 μM DAP agar plates in dilutions to count *E. coli* CFUs and check for contamination.

To prepare bacteria for flow cytometry, bacterial extracts from 300 mg of feces were combined and diluted 1:10 in PBS and exposed to air for 30 minutes. The bacteria were then stained with 1.5 μL/mL SYTO40 (ThermoFisher) for 30 minutes protected from light and filtered through a 40 μm mesh strainer and analyzed on a BD FACSAria III or BD LSR Fortessa. Gating for GFP+ SFB was accomplished by gating out mCherry+ donor cells on a PE-A (yellow autofluorescence) vs. mCherry-A scatter plot and gating for SYTO40 stained cells on a PE-A vs. Pacific Blue-A scatter plot. The SYTO40 stained cells (presumably SFB) were then gated into GFP+ and GFP- groups on the GFP Alexa 488-A channel.

FACS analysis of NSG mouse fecal bacteria. Fecal pellets were homogenized in 1 mL final volume of PBS per pellet, as described in Chapter 3, and bacterial extracts were combined by cage. Bacterial extracts were exposed to air for 30 minutes and then stained with 5 μM SYTO 40 for 30 minutes, protected from light, and filtered through a 40 μm mesh strainer. The stained bacteria were diluted 1:10 and 1:50 in PBS and analyzed on a BD FACSAria II. SYTO40 stained cells were gated on the Pacific Blue fluorescence channel, and GFP (FITC channel) and SSC-W gates were set using bacteria from cage 2 (negative control) mice. The gate for GFP-expressing SFB was SYTO40+, high SSC-W, GFP+, mCherry-, but we were unable to detect these cells.

16S qPCR for SFB in fecal samples. Metagenomic DNA was extracted from fecal bacteria samples using the MasterPure Gram Positive DNA purification kit. qPCR reactions were performed with primers 5′-GACGCTGAGGCATGAGAGCAT-3′ and 5′-GACGGCACGGATTGTTATTCA-3′ for SFB, and 5′-ACTCCTACGGGAGGCAGCAGT-3′ and 5′-ATTACCGCGGCTGCTGCG-3′ for Eubacteria (universal)[183]. qPCRs was performed using the

Kapa SYBR Fast qPCR Master Mix on a BioRad CFX96 Touch thermocycler. Reaction conditions were as follows: 3 minutes initial denaturation at 95C followed by 40 cycles of 3 second denaturation at 95C and 20 second annealing/amplification at 62C. A known SFB-containing fecal bacteria sample was used as a positive PCR control, and *E. coli* gDNA was used as a negative control.

Acknowledgements

Experiments in gnotobiotic mice, microscopy, and FACS analysis were performed by our collaborators Seiko Narushima, Koji Atarashi, and Kenya Honda at RIKEN Center for Integrative Medical Sciences. Ivo Ivanov, Casandra Panea, and Carolyn Lee assisted with preliminary mouse experiments at Columbia. Plasmid pJIR750 was obtained from ATCC (#87015). Antibiotic icon in figure 4.1 made by mavadee from www.flaticon.com.

Ethics Statement

All experiments were performed in accordance with Columbia University Medical Center IACUC protocol #. All gnotobiotic mice experiments were performed in accordance with RIKEN Center for Integrative Medical Sciences IACUC guidelines.

Table 4.1. SFB genetic parts. 5' UTR sequences from SFB and *C. perfringens* have Shine-Dalgarno sequences (putative RBSs) underlined.

Promoter sequence	Expression in *E. coli*	Origin of sequence	Code
CTCTAGAGTAGTAGATTATTTTAGGAATTTAGATGTTTTGTATGAAATAGAT GCTTCGTATGGAATTAATGAAATTTTTAGTCAGGTAAAAAAGGTAATAGGAG AATATT	**+++**	SFBM_1374 (methionine aminopeptidase) 5' UTR	1
GTTTTAAATGATGAAAAGAAATATTTAGGGAAGATTGTTTCGACGCGAATT GTTGATCTGGAAAATGATCACCTTATCGGACAAGCTTTAAAATAGGAGGAT ATAAAAAT	**++**	SFBM_0930 (histidine triad family protein) 5' UTR	2
ATAAGGATTCTTTAAAGAGAGATATAGTTATGTCAAAGACTGTAGAATTTT TAGTAAATCAAAATAAAAAAAGAGGTATTAAATAGAGTGTATTTTAAAGGA GGAGACTT	**+++**	SFBM_0302 (clpP protease) 5' UTR	3
AAACACCAATAAAATTAGAATATTTAGGAGCGACTTTAAAAAAGTTTAATA AGAATTGTTTATGAGATATTTTTATTATATTTAAACTCAATTTAAAGTAGG GAGAATAG	**+**	SFBM_1493 (murA) 5' UTR	4
GCAAGTGTTCAAGAAGTTATTAAGTCGGGAGTGCAGTCGAAGTGGGCAAGT TGAAAAATTCACAAAAATGTGGTATAATATCTTTGTTCATTAGAGCGATAA ACTTGAATTTGAGAGGGAACTTAG	**+**	*Clostridium perfringens* catP 5' UTR	5

Antibiotic resistance gene	Sequence	Origin of sequence	Selection in *C. perfringens*
catP	ATGGTATTTGAAAAAATTGATAAAAATAGTTGGAACAGAAAAGAGTAT TTTGACCACTACTTTGCAAGTGTACCTTGTACATACAGCATGACCGTT AAAGTGGATATCACACAAATAAAGGAAAAGGGAATGAAACTATATCCT GCAATGCTTTATTATATTGCAATGATTGTAAACCGCCATTCAGAGTTT AGGACGGCAATCAATCAAGATGGTGAATTGGGGATATATGATGAGATG ATACCAAGCTATACAATATTTCACAATGATACTGAAACATTTTCCAGC CTTTGGACTGAGTGTAAGTCTGACTTTAAATCATTTTTAGCAGATTAT GAAAGTGATACGCAACGGTATGGAAACAATCATAGAATGGAAGGAAAG CCAAATGCTCCGGAAAACATTTTTAATGTATCTATGATACCGTGGTCA ACCTTCGATGGCTTTAATCTGAATTTGCAGAAAGGATATGATTATTTG ATTCCTATTTTTACTATGGGGAAATATTATAAAGAAGATAACAAAATT ATACTTCCTTTGGCAATTCAAGTTCATCACGCAGTATGTGACGGATTT CACATTTGCCGTTTTGTAAACGAATTGCAGGAATTGATAAATAGTTAA	*Clostridium perfringens*	Chloramphenicol 10-20 µg/mL[186, 187]
ermBP	ATGAACAAAAATATAAAATATTCTCAAAACTTTTTAACGAGTGAAAAA GTACTCAACCAAATAATAAAACAATTGAATTTAAAAGAAACCGATACC GTTTACGAAATTGGAACAGGTAAAGGGCATTTAACGACGAAACTGGCT AAAATAAGTAAACAGGTAACGTCTATTGAATTAGACAGTCATCTATTC AACTTATCGTCAGAAAAATTAAAACTGAATACTCGTGTCACTTTAATT CACCAAGATATTCTACAGTTTCAATTCCCTAACAAACAGAGGTATAAA ATTGTTGGGAGTATTCCTTACCATTTAAGCACACAAATTATTAAAAAA GTGGTTTTTGAAAGCCATGCGTCTGACATCTATCTGATTGTTGAAGAA GGATTCTACAAGCGTACCTTGGATATTCACCGAACACTAGGGTTGCTC TTGCACACTCAAGTCTCGATTCAGCAATTGCTTAAGCTGCCAGCGGAA TGCTTTCATCCTAAACCAAAAGTAAACAGTGTCTTAATAAAACTTACC CGCCATACCACAGATGTTCCAGATAAATATTGGAAGCTATATACGTAC TTTGTTTCAAAATGGGTCAATCGAGAATATCGTCAACTGTTTACTAAA AATCAGTTTCATCAAGCAATGAAACACGCCAAAGTAAACAATTTAAGT ACCGTTACTTATGAGCAAGTATTGTCTATTTTTAATAGTTATCTATTA TTTAACGGGAGGAAATAA	*Clostridium perfringens*	Erythromycin 30 µg/mL[188]

Table 4.2. pSFB vectors for MAGIC implementation in SFB.

Vector	SFB selection payload	Payload promoter	Himar transposase promoter	Origin of replication	E. coli selection
pSFB 1	catP-sfGFP	5	1	p15A	Beta-lactamase
pSFB 2	catP-sfGFP	5	2	p15A	Beta-lactamase
pSFB 3	catP-sfGFP	5	3	p15A	Beta-lactamase
pSFB 4	catP-sfGFP	5	4	p15A	Beta-lactamase

Table 4.3. Primers used to construct SFB methyltransferase vectors. Primers contain 5' overhangs to allow for isothermal assembly.

Amplicon	Forward Primer	Reverse Primer	Amplicon size	Annealing temperature (˚C)
pSC101 + tetA vector backbone	TATATGTCATACACAGAAGCGTTACATTGTCGATCTGTTCATGGT	TTTTACGCATTATTATATTTATTGGAGGGGAAATGTGCGCGGAAC	3.4 kb	66
SFBM_0619	TCAATTGAATGAGAGTTAACATCGGATTTACTTAATACAGATGAA	GTAAATTTACTATGTTTTTAGTTGATATCAATGTGCCCAAGAGTTATT	760 bp	55
SFBM_1080	GGCACATTGATATCAACTAAAAACATAGTAAATTTACCAAAATTTAG	ATCAGAAACCCAAGCGTATGAGTTGAAAGATAAGCTTAATGG	3.6 kb	60
SFBM_0018, SFBM_0019, SFBM_0020	AGCTTATCTTTCAACTCATACGCTTGGGTTTCTGATAAATTTGGG	AACAGATCGACAATGTAACGCTTCTGTGTATGACATATATTAACC	3.1 kb	62
SFBM_0805	CGCGCACATTTCCCCTCCAATAAATATAATAATGCGTAAAATTCA	AAATAATTTTTATAGGGGATGGTTTAATAGTTTTGTTAAATTTATAAATTGTAAAAAAG	1.1 kb	58
SFBM_0801, SFBM_0802	CAATTTATAAATTTAACAAAACTATTAAACCATCCCCTATAAAAATTATTTATTTCCT	ATCTGTATTAAGTAAATCCGATGTTAACTCTCATTCAATTGATGAT	3.2 kb	62

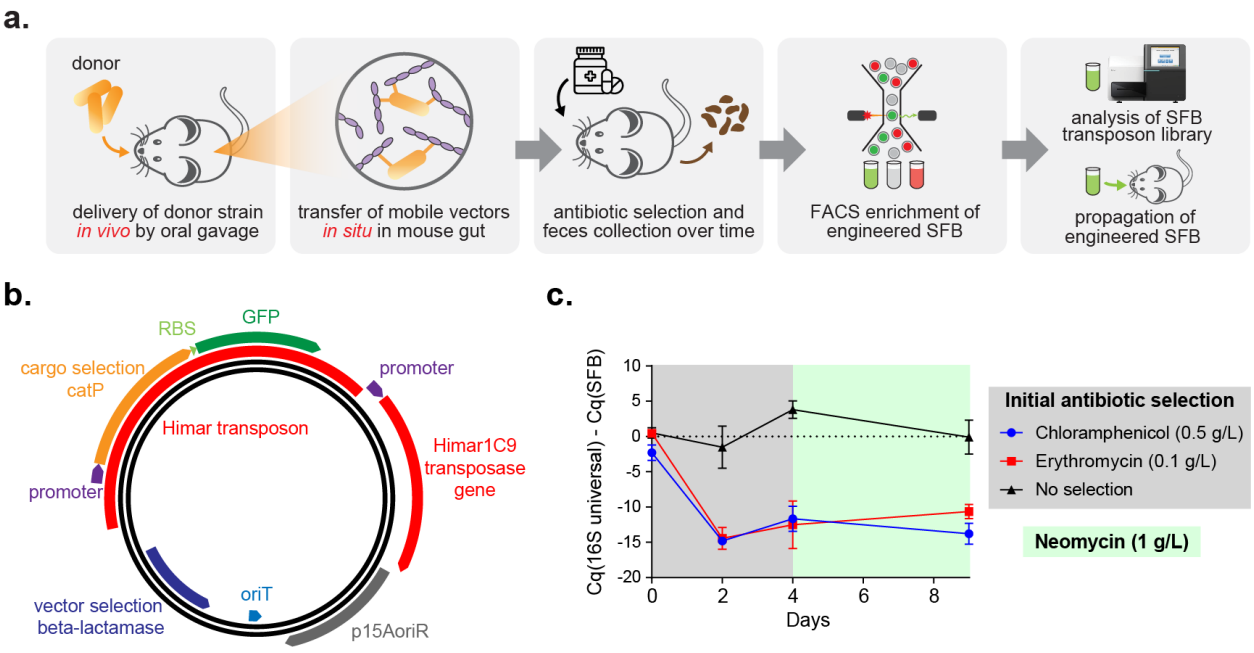


Figure 4.1. Implementation of MAGIC to genetically engineer Segmented Filamentous Bacteria (SFB). (a) Schematic of *in situ* horizontal transfer of selectable genetic payloads into SFB. **(b)** Plasmid map of pSFB vectors containing a payload with the catP chloramphenicol resistance gene and a GFP gene on a Himar transposon. The associated Himar1C9 hyperactive transposase is delivered with the transposon to stably integrate the payload into the SFB genome. Promoter and RBS sequences were obtained from the SFB genome and from a *C. perfringens* plasmid. **(c)** Measurement of SFB antibiotic susceptibility in conventional B6 mice. Cages of 2 mice each were treated with 4 days of chloramphenicol, erythromycin, or no antibiotic in drinking water, then switched to neomycin in drinking water. SFB prevalence in fecal samples collected over time was measured by qPCR.

a.

Strain	Plasmids	Genotype	Growth media
EcGT2	-	*RP4-2(Km::Tn7,Tc::Mu-1), pro-82, LAMpir, recA1, endA1, thiE1, hsdR17, creC510, asd::mCherry-specR*	LB DAP 50 µM
EcGT3	pSFB-MT1 (contains SFBM_0805 and SFBM_0619; pSC101 oriR; resistance gene tetA)	*RP4-2(Km::Tn7,Tc::Mu-1), pro-82, LAMpir, recA1, endA1, thiE1, hsdR17, creC510, asd::mCherry-specR*	LB DAP 50 µM tetracycline 15 µg/mL
EcGT4	pSFB-MT2 (contains SFBM_1080, SFBM_0018, SFBM_0019, and SFBM_0020; pSC101 oriR; resistance gene tetA)	*RP4-2(Km::Tn7,Tc::Mu-1), pro-82, LAMpir, recA1, endA1, thiE1, hsdR17, creC510, asd::mCherry-specR*	LB DAP 50 µM tetracycline 15 µg/mL

b.

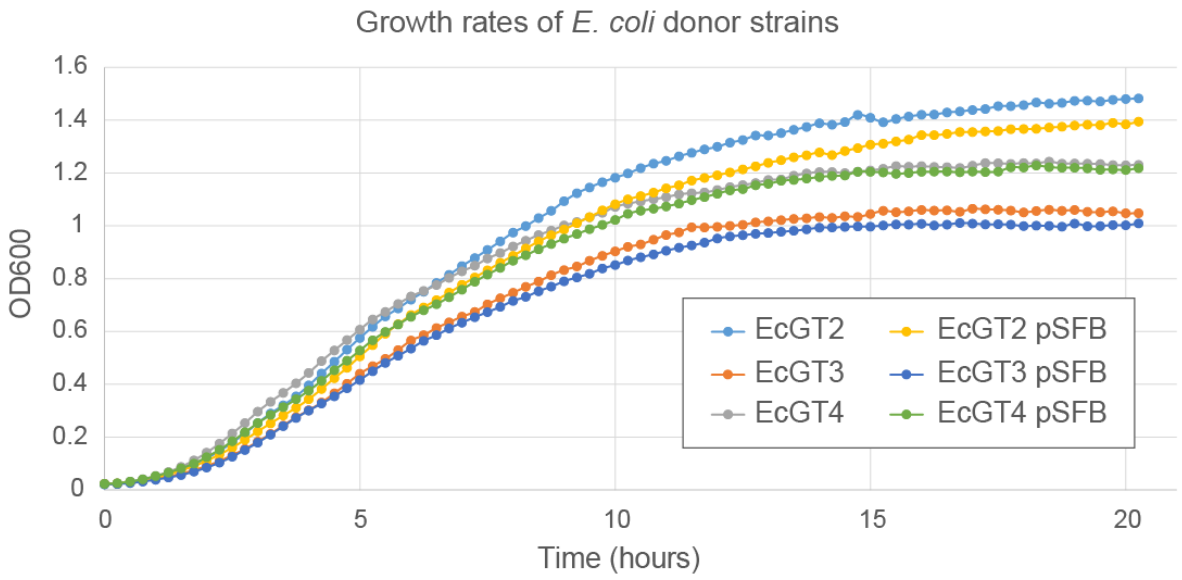


Figure 4.2. Characterization of SFB-optimized donor *E. coli* strains. (a) Donor strains EcGT3 and EcGT4 were generated from parent strain EcGT2 with the addition of SFB DNA methyltransferase systems on a pSC101 plasmid. **(b)** EcGT3 and EcGT4 strains grow slightly slower than the EcGT2 strain, both with and without an additional pSFB1 plasmid, and reach a lower concentration at saturation. OD600 was measured during growth at 37C on a BioTek plate reader; each curve represents the average of 3 replicates.

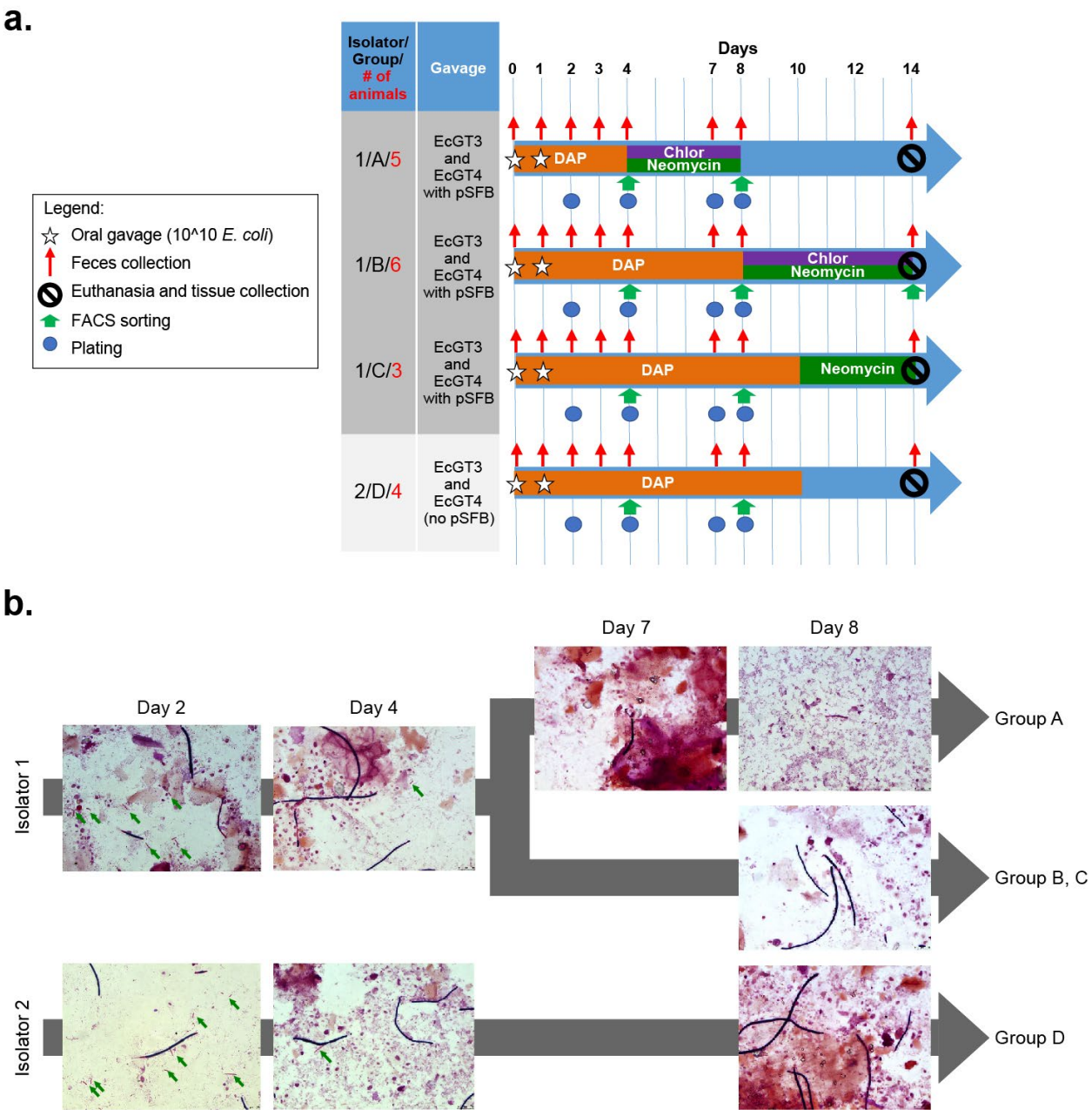


Figure 4.3. Implementation of MAGIC to genetically engineer SFB in SFB-monoassociated mice. (a) Experimental setup. Groups A-C were each housed in their own cage within the same gnotobiotic isolator, while Group D was housed in a cage in a separate isolator. Mice were gavaged with 2 boluses of 10^10 *E. coli* donor cells mixed in equal proportions. The colored bars on the timeline show the duration of antibiotic treatment or DAP supplementation in drinking water. Fecal samples were collected over time for microscopy, FACS analysis, and plating to quantify *E. coli* presence and identify contaminants. Tissue samples were collected at the end of the experiment. **(b)** Micrographs of fecal smears over time. *E. coli* (green arrows) were visible in samples collected up to day 4, but not at day 7 or after. Group A (treated with antibiotics from day 4 to day 8) showed decreased numbers of filaments and smaller, fainter filaments at days 7 and

8, while the other groups' SFB filaments remained relatively unchanged post-gavage (but pre-antibiotics).

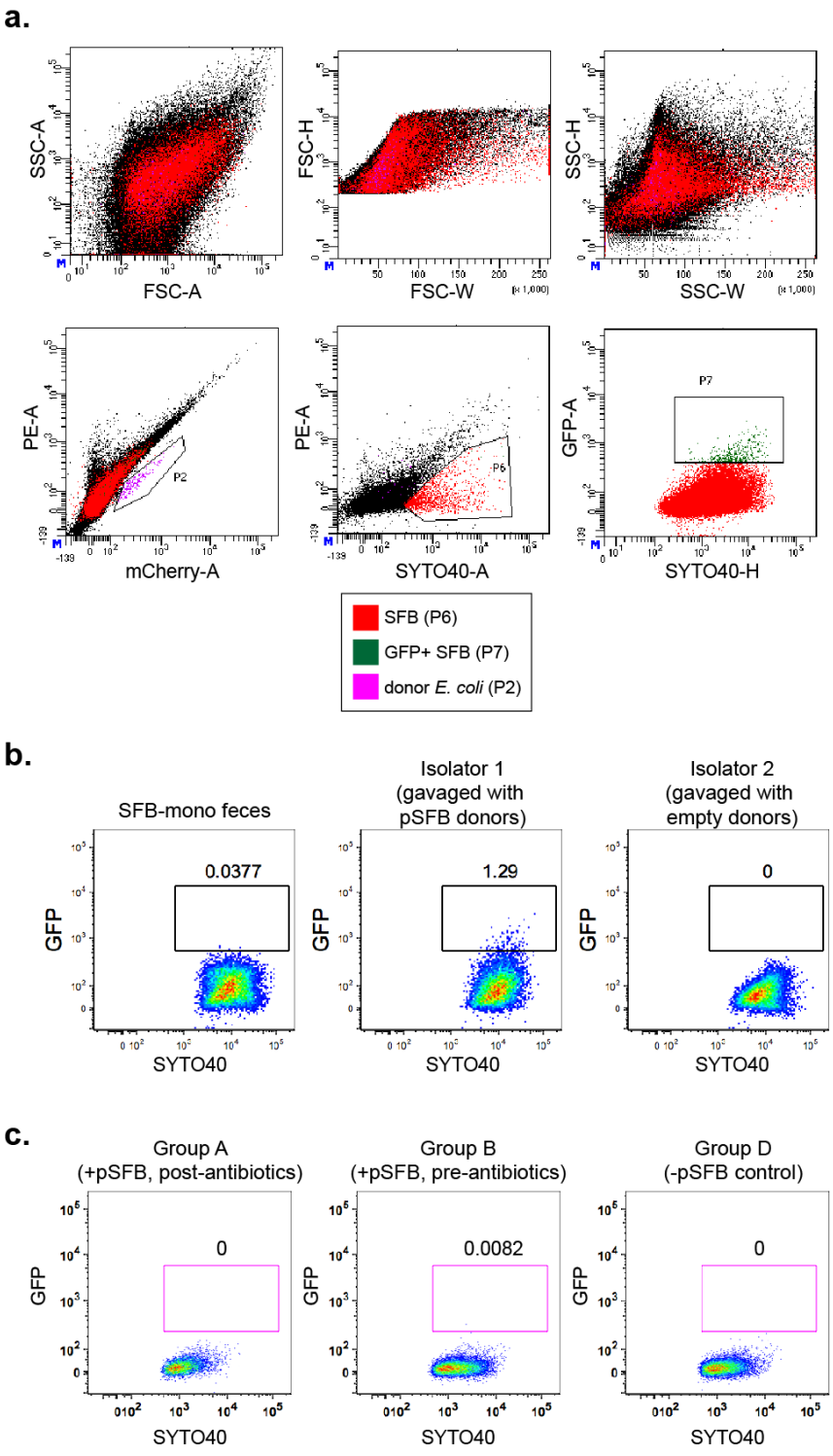


Figure 4.4. Flow cytometric analysis of fecal bacteria from SFB-monoassociated mice after gavage with MAGIC donors. (a) Gates used to identify GFP+ SFB. mCherry+ donor *E. coli* (P2, magenta) were gated out on the basis of having mCherry fluorescence greater than PE autofluorescence. SFB stained with SYTO40 (P6, red) were gated on the basis of having blue fluorescence greater than PE autofluorescence. P7 (green) represents the population of GFP+

SFB. **(b)** Flow cytometry of fecal bacteria on day 4 of the experiment. Mice from Isolator 1 had a small fraction of GFP+ SFB present in feces, while mice in isolator 2 (gavaged with the empty donor strains) and untreated mice did not. **(c)** Flow cytometry of fecal bacteria on day 8 of the experiment. Mice in Group A had undergone 4 days of antibiotic treatment to select for chloramphenicol-resistant SFB, while Groups B and D had not undergone antibiotic treatment. In all groups, GFP+ SFB were not apparent at this time.

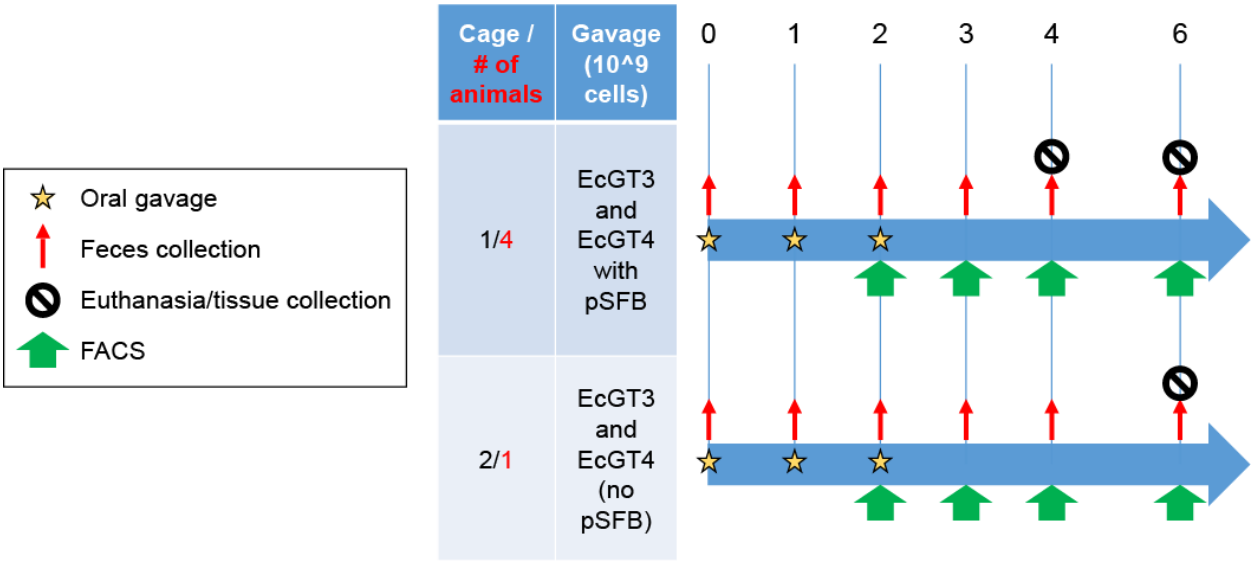


Figure 4.5. Implementation of MAGIC to engineer SFB in specific pathogen free (SPF) NSG mice harboring SFB. Mice were gavaged with 3 boluses of 10^9 donor cells (mixed in equal proportions) over 3 days. Fecal samples were collected over time for FACS to enrich for GFP+ SFB. In Group 1, two mice were euthanized at day 4 and two were euthanized at day 6.

Conclusions and Future Perspectives

As the field of microbiome research progresses from characterization of microbiomes using -omics approaches to experimentally testing hypotheses about microbial ecology and host-microbiome interactions, tools for making functional genetic modifications to the microbial community will be invaluable. The Cas-Transposon and MAGIC platforms presented here aim to address the need for technologies that enable genome alterations across a wide phylogenetic range, including microbes that are uncultivable or not currently genetically tractable, because many of them are involved in important physiological processes. The Cas-Transposon system enables host-independent targeted genome editing using Himar-dCas9 (HdCas9) fusion transposases as an alternative to CRISPR-Cas genome editing. Because the Himar transposase operates independently of host factors, it can efficiently insert larger pieces of DNA than homologous recombination machinery, potentially enabling insertion of operon-sized (multi-kb) pieces of DNA into a target locus. In Chapter 2, we demonstrated that HdCas9 transposition activity is targeted to a gRNA-defined locus in both cell-free *in vitro* reactions and in plasmid-based assays in *E. coli*. The MAGIC platform complements Cas-Transposons, serving as a delivery tool to introduce DNA into a wide range of organisms within the gut microbiome without the need for cultivation. Using MAGIC, we showed that diverse bacteria spanning multiple phyla were amenable to genetic manipulation by *in situ* conjugation, expanding the range of gut bacteria that can potentially be used as targets or chasses for insertion of synthetic DNA (Chapter 3). Furthermore, native gut bacteria could be engineered by *in vitro* conjugation and then deployed back into their host animal to recolonize the environment and promote long-term infiltration of engineered gene constructs into the metagenome. We piloted the use of MAGIC to genetically modify Segmented Filamentous Bacteria, an uncultivable gut commensal, with promising preliminary results (Chapter 4).

Optimization of platform technologies

At present, both technologies described in this work remain in an early phase of development and can be optimized further. In the case of the Cas-Transposon system, we demonstrated *in vitro* and *in vivo* gRNA-targeted transpositions into a plasmid using the HdCas9 protein, but targeting transposition to a single-copy genomic locus within a genome is a greater challenge, requiring high efficiency of transposition at the target locus and low efficiency at all off-target sites. We propose several approaches to overcome this challenge. First, protein engineering may produce HdCas9 proteins with both higher transposition specificity and efficiency; a directed evolution approach might be used to evolve transposases that insert preferentially at gRNA-targeted sites. Rational protein design, such as making alterations to the allosteric inhibition domain of the Himar protein while also mutating the dimerization interface, could achieve the same effect by increasing the efficiency of transposition when transposases are stably bound to the target locus, but decreasing the probability that transposases dimerize at low local concentrations elsewhere. Further optimization of delivery/expression of the system components (protein, gRNA, and transposon DNA) into cells is also likely necessary for *in vivo* function. We can also engineer different varieties of targeted transposases, containing different transposase domains and DNA-binding catalytically dead nucleases[189, 190], with different targeting behaviors, which may be useful in organisms with different genomic distributions of insertion sites and protospacer adjacent motifs.

Future improvements to the MAGIC system would address the need for targeted gene delivery and actuation of engineered functions in specific microbes or subsets of microbes. The specificity of the MAGIC system can be tailored to particular microbes in several ways. First, MAGIC vectors can be engineered with specific or broad host-range origins of replication and with regulatory elements that are selectively recognized by certain microbes[143, 191]. Conjugation into selected strains can also be enhanced by using a donor strain that localizes in a specific area of

the gut, longitudinally (proximal vs. distal) or radially (in the lumen vs. attached to the mucosa), to increase the likelihood of conjugation into the more abundant microbes at that site[192]. Horizontal gene transfer (HGT) into targeted microbes can also be enhanced by designing vectors to avoid motifs that are recognized by the recipient's defense mechanisms against foreign DNA, such as restriction-modification systems[135] and CRISPR/Cas systems[91], or by expressing recipient DNA methyltransferases in the donor strain, as we did for SFB in Chapter 4. As we discover and characterize new anti-DNA defense systems in bacteria[193], strategies to enhance HGT by avoiding these systems may additionally come into play.

Greater temporal control of genetic payload persistence in the gut can be achieved by optimizing the lifetime of the donor strain. As we observed in Chapter 3, a single gavage of a lab-adapted donor strain resulted in a transient (<72 hour) perturbation of the mouse gut metagenome, while a single gavage of an engineered donor strain resulted in stable recolonization and engraftment of the payload genes into the gut metagenome. By altering the donor strain (gut adapted vs. non-adapted) and gavage conditions (single vs. multiple doses, number of bacteria per dose) we may be able to fine-tune the concentration of donor cells and temporal duration of the metagenomic perturbation. To control for the number of cells surviving passage through the stomach and entering the intestines, it may also be useful to encapsulate the donor bacteria in a protective coating during delivery through the stomach; this encapsulation strategy may reduce variations between individual mice in the level of HGT and metagenomic perturbation.

Future research directions

The technologies described here enable multiple types of studies of naturally occurring microbiomes. The Cas-Transposon and MAGIC technologies can be combined, with MAGIC

serving as a delivery platform for Cas-Transposon components, to enable engraftment of a novel gene function into a microbiome and/or removal of an existing gene from the metagenome without altering the microbial composition of the community. Genes within the gut microbiome that may be candidates for functional manipulation in host-microbiome interaction studies include bile salt metabolism genes, which encode microbial proteins that alter the balance of bile acids present in the body and modulate hormonal signaling pathways that use bile acids as signal molecules[194]. The abundance of particular bile salt metabolism genes within mammalian microbiomes has been shown to affect resistance to pathogen invasion and host metabolism[40, 195], making the population of bile salt hydrolase and dehydrogenase genes a potential target for gene therapy aimed at improving host health. Other microbial genetic functions that could be deleted to improve host health include genes for antibiotic resistance and virulence factors in pathogens, as well as genes in commensal bacteria that are involved in the production of harmful compounds such as trimethylamine, which promotes atherosclerosis and has been linked to colorectal cancer[25, 196].

In the setting of other microbiomes, *in situ* genetic engineering may be utilized for bioremediation applications. For example, soil and marine environments are frequently contaminated with pollutants such as plastic particles or toxins from industrial/agricultural activity. It would be useful to introduce novel gene pathways into the bacteria that occupy these environments for sensing and metabolism of pollutants, although safety and biocontainment of genes delivered into open environments would be paramount concerns. Antibiotic resistance genes are also commonly distributed among bacteria in agricultural settings, as a result of antibiotic use in farm animals, and these resistant bacteria pose health hazards to both humans and animals. Removal of antibiotic resistance genes from these environmental bacteria would be another useful bioremediation application.

At present, engraftment of new genetic functions is more feasible than removal of a gene from a microbial population. We observed in Chapter 3 that up to 5% of mouse gut bacteria could

be identified as transconjugants of our libraries of conjugative vectors, showing that new genes can be disseminated and expressed from a significant fraction of the native microbiome. However, removing a gene from a population of cells is more difficult, since the probability of on-target homologous recombination (if using CRISPR-Cas9 genome editing) or transposition (if using Cas-Transposons) in each cell is small, preserving the gene function in most cells. To improve the efficacy of removing a gene from a microbiome, the Cas-Transposon system could be adapted to serve as a gene drive, in which genes for the targeted transposase and gRNAs complementing the unwanted gene are propagated through a population on a conjugative transposon. In such a system, targeted gene knockout events are not limited only to microbes that directly receive the Cas-Transposon system, but also recipients of secondary conjugative transfers (and beyond). Mariner transposases such as Himar operate in a cut-and-paste fashion, but leave a 2-nucleotide duplication upon insertion and excision; if Himar-dCas9 were to be used as a gene drive, an on-target transposition event followed by excision and conjugation of the transposon would result in a frameshift mutation of the target gene.

MAGIC can also be used to characterize HGT pathways within a microbiome in detail. In our studies in Chapter 3, we identified the recipient ranges for several vectors by looking for expression of one or more genetic markers present on the transferred vector. However, these recipients represent only a subset of microbes that actually received the vector. It is likely that some microbes received a vector via conjugation but did not express any of the genetic markers, due to either incompatibility of protein function or transcriptional/translational regulatory elements. For example, GFP requires oxygen to fold and may not be expressed well in gut anaerobes that are particularly sensitive to oxygen, while antibiotic resistance proteins may not fold properly, bind their target, or be transported to the correct cellular location in some species, impeding their function. To understand HGT, it is necessary to differentiate between microbes that received a genetic vector but did not express the genes on the vector and those that were non-recipients.

Using a DNA detection system, such as fluorescence *in situ* hybridization (FISH), in conjunction with MAGIC, it would be possible to detect bacteria that were successful recipients of a given vector, regardless of gene expression. Correlating recipients identified by gene expression versus by direct DNA detection would yield information about the ability of each recipient to express the transferred genes. We could also correlate the relative abundance of each transconjugant containing transferred DNA with its abundance in the whole population to determine which bacteria are more likely to receive the vector. Bacteria resistant to conjugation may have defense systems that recognize and eliminate the vector, contain native plasmids which are incompatible with the vector, or be unable to produce necessary replication/integration factors. Further analysis of transconjugants using the MAGIC approach will elucidate the landscape of HGT within naturally occurring microbiomes.

Results in Chapter 3 also highlight the possibility of developing engineered personalized probiotics tailored to individual hosts. Because every individual has their own stable microbiome, exogenous bacteria such as commercial probiotics do not readily infiltrate the microbiome and are only transiently present and active after consumption. To confer longer-lasting effects on the microbiome, native strains from the host's microbiome could be engineered with novel genetic functions and then redeployed into the gut. This finding can be generalized for other hosts, potentially including humans; native gut bacteria from individuals can be engineered with useful functions to serve as host-adapted probiotics that persist for a longer period of time after consumption. New functions that may be interesting to confer to native bacteria include sensing systems for small molecules, metabolism genes to manipulate energy harvest within the gut, and pathways for the production of beneficial small molecules. However, these bacteria would also be need to engineered with safety mechanisms to prevent the indefinite persistence of the engineered functions once they are no longer desired, and to prevent the spread of the genes into other members of the microbiome.

Biocontainment of synthetic DNA in microbial communities

The potential deployment of synthetic gene constructs by *in situ* conjugation into open environments necessitates the discussion of biocontainment strategies. In Chapter 3, we observed that genes transferred into native gut bacteria could have different lengths of persistence in the transconjugant population without selection, depending on the gene function and the payload design. A strain of *E. fergusonii* containing a replicative vector with beta-lactamase and GFP genes lost the GFP gene from most of the population within several days, but 100% of the population remained beta-lactam resistant during that time. When beta-lactamase and GFP genes were introduced into *E. fergusonii* as a single operon on an integrative transposon, the two genes persisted together in 100% of the population, even as the chloramphenicol resistance gene on the vector backbone was lost over time. The spread and longevity of antibiotic resistance genes in microbial communities is a global public health hazard, as it renders antibiotics ineffective in the treatment of disease. Our studies in Chapter 3 highlight the potential danger of letting a synthetic gene construct loose in the wild, as that gene construct may persist and propagate indefinitely.

Several biocontainment strategies may be employed to limit the conjugative spread of synthetic DNA. The first is to utilize narrow host range vectors for gene delivery, such as plasmids that require specific replication factors that are not widely distributed (e.g., R6K origin of replication, which needs the pir protein to replicate). Bacteriophages, which only infect specific hosts, are another way to deliver DNA into a narrow host range of recipients. Genetic payloads can also be linked to a toxin-antitoxin system, so that the payload is always co-expressed with a toxin and leads to cell death in any off-target transconjugants that do not contain the antitoxin gene. The toxin gene could be built into the same operon as the payload, or encoded into a payload gene in a different reading frame. Finally, a Cas-Transposon-based gene drive against

an unwanted genetic payload could be used to eliminate that payload from a given population. Development of novel biocontainment strategies will improve our control over HGT of engineered DNA, which is necessary to prevent the unintended spread of potentially dangerous genes in nature.

Conclusion

With continued advances in the field of synthetic biology, the array of genetic engineering tools available to biologists will enable increasingly fine-tunable and programmable modifications of a wider range of organisms. The technologies presented in this thesis are two of among many developments ongoing in this field, and can be used synergistically with many other genome editing and gene delivery approaches. Using novel genetic engineering techniques, we will be able to better understand the mechanisms driving ecology in natural microbiomes, through functional genetic studies of complex populations, and also be able to engineer microbiomes in useful ways to further our well-being.

References

1. Hamady, M. & Knight, R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome research* **19**, 1141-1152 (2009).
2. Integrative, H.M.P.R.N.C. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* **16**, 276-289 (2014).
3. Kolmeder, C.A. & de Vos, W.M. Metaproteomics of our microbiome - developing insight in function and activity in man and model systems. *J Proteomics* **97**, 3-16 (2014).
4. Human Microbiome Project, C. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214 (2012).
5. Duvallet, C., Gibbons, S.M., Gurry, T., Irizarry, R.A. & Alm, E.J. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature communications* **8**, 1784 (2017).
6. Mazidi, M., Rezaie, P., Kengne, A.P., Mobarhan, M.G. & Ferns, G.A. Gut microbiome and metabolic syndrome. *Diabetes Metab Syndr* **10**, S150-157 (2016).
7. Backhed, F., Ley, R.E., Sonnenburg, J.L., Peterson, D.A. & Gordon, J.I. Host-bacterial mutualism in the human intestine. *Science* **307**, 1915-1920 (2005).
8. Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65 (2010).
9. Arumugam, M. et al. Enterotypes of the human gut microbiome. *Nature* **473**, 174-180 (2011).
10. Tap, J. et al. Towards the human intestinal microbiota phylogenetic core. *Environ Microbiol* **11**, 2574-2584 (2009).
11. Lozupone, C.A., Stombaugh, J.I., Gordon, J.I., Jansson, J.K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220-230 (2012).
12. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222-227 (2012).
13. Turnbaugh, P.J. et al. A core gut microbiome in obese and lean twins. *Nature* **457**, 480-484 (2009).
14. den Besten, G. et al. The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *J Lipid Res* **54**, 2325-2340 (2013).
15. Nicholson, J.K. et al. Host-gut microbiota metabolic interactions. *Science* **336**, 1262-1267 (2012).
16. Ridaura, V.K. et al. Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* **341**, 1241214 (2013).
17. Smith, M.I. et al. Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science* **339**, 548-554 (2013).
18. Haiser, H.J. et al. Predicting and manipulating cardiac drug inactivation by the human gut bacterium Eggerthella lenta. *Science* **341**, 295-298 (2013).
19. Wallace, B.D. et al. Alleviating cancer drug toxicity by inhibiting a bacterial enzyme. *Science* **330**, 831-835 (2010).
20. Mazmanian, S.K., Liu, C.H., Tzianabos, A.O. & Kasper, D.L. An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell* **122**, 107-118 (2005).
21. Ivanov, II et al. Induction of intestinal Th17 cells by segmented filamentous bacteria. *Cell* **139**, 485-498 (2009).
22. Diaz Heijtz, R. et al. Normal gut microbiota modulates brain development and behavior. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 3047-3052 (2011).
23. Neufeld, K.M., Kang, N., Bienenstock, J. & Foster, J.A. Reduced anxiety-like behavior and central neurochemical change in germ-free mice. *Neurogastroenterol Motil* **23**, 255-264, e119 (2011).

24. Gibson, G.R. et al. Expert consensus document: The International Scientific Association for Probiotics and Prebiotics (ISAPP) consensus statement on the definition and scope of prebiotics. *Nat Rev Gastroenterol Hepatol* **14**, 491-502 (2017).
25. Wang, Z. et al. Non-lethal Inhibition of Gut Microbial Trimethylamine Production for the Treatment of Atherosclerosis. *Cell* **163**, 1585-1595 (2015).
26. O'Toole, P.W. & Cooney, J.C. Probiotic bacteria influence the composition and function of the intestinal microbiota. *Interdiscip Perspect Infect Dis* **2008**, 175285 (2008).
27. Bravo, J.A. et al. Ingestion of Lactobacillus strain regulates emotional behavior and central GABA receptor expression in a mouse via the vagus nerve. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 16050-16055 (2011).
28. Turroni, F. et al. Deciphering bifidobacterial-mediated metabolic interactions and their impact on gut microbiota by a multi-omics approach. *ISME J* **10**, 1656-1668 (2016).
29. Rund, S.A., Rohde, H., Sonnenborn, U. & Oelschlaeger, T.A. Antagonistic effects of probiotic Escherichia coli Nissle 1917 on EHEC strains of serotype O104:H4 and O157:H7. *International journal of medical microbiology : IJMM* **303**, 1-8 (2013).
30. Schierack, P. et al. E. coli Nissle 1917 Affects Salmonella adhesion to porcine intestinal epithelial cells. *PloS one* **6**, e14712 (2011).
31. Deriu, E. et al. Probiotic bacteria reduce salmonella typhimurium intestinal colonization by competing for iron. *Cell Host Microbe* **14**, 26-37 (2013).
32. Steidler, L. et al. Treatment of murine colitis by Lactococcus lactis secreting interleukin-10. *Science* **289**, 1352-1355 (2000).
33. Bermudez-Humaran, L.G. et al. Serine protease inhibitors protect better than IL-10 and TGF-beta anti-inflammatory cytokines against mouse colitis when delivered by recombinant lactococci. *Microb Cell Fact* **14**, 26 (2015).
34. Vandenbroucke, K. et al. Orally administered L. lactis secreting an anti-TNF Nanobody demonstrate efficacy in chronic colitis. *Mucosal Immunol* **3**, 49-56 (2010).
35. Zmora, N. et al. Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is Associated with Unique Host and Microbiome Features. *Cell* **174**, 1388-1405 e1321 (2018).
36. Kristensen, N.B. et al. Alterations in fecal microbiota composition by probiotic supplementation in healthy adults: a systematic review of randomized controlled trials. *Genome Med* **8**, 52 (2016).
37. McNulty, N.P. et al. The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins. *Sci Transl Med* **3**, 106ra106 (2011).
38. Gough, E., Shaikh, H. & Manges, A.R. Systematic review of intestinal microbiota transplantation (fecal bacteriotherapy) for recurrent Clostridium difficile infection. *Clin Infect Dis* **53**, 994-1002 (2011).
39. Seekatz, A.M. et al. Fecal Microbiota Transplantation Eliminates Clostridium difficile in a Murine Model of Relapsing Disease. *Infection and immunity* **83**, 3838-3846 (2015).
40. Buffie, C.G. et al. Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. *Nature* **517**, 205-208 (2015).
41. Dodd, D. et al. A gut bacterial pathway metabolizes aromatic amino acids into nine circulating metabolites. *Nature* **551**, 648-652 (2017).
42. Lawley, T.D. & Walker, A.W. Intestinal colonization resistance. *Immunology* **138**, 1-11 (2013).
43. Mai, V., Ukhanova, M., Visone, L., Abuladze, T. & Sulakvelidze, A. Bacteriophage Administration Reduces the Concentration of Listeria monocytogenes in the Gastrointestinal Tract and Its Translocation to Spleen and Liver in Experimentally Infected Mice. *Int J Microbiol* **2010**, 624234 (2010).

44. Mai, V., Ukhanova, M., Reinhard, M.K., Li, M. & Sulakvelidze, A. Bacteriophage administration significantly reduces Shigella colonization and shedding by Shigella-challenged mice without deleterious side effects and distortions in the gut microbiota. *Bacteriophage* **5**, e1088124 (2015).
45. Lu, T.K. & Collins, J.J. Engineered bacteriophage targeting gene networks as adjuvants for antibiotic therapy. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 4629-4634 (2009).
46. Manrique, P., Dills, M. & Young, M.J. The Human Gut Phage Community and Its Implications for Health and Disease. *Viruses* **9** (2017).
47. Esvelt, K.M. & Wang, H.H. Genome-scale engineering for systems and synthetic biology. *Molecular systems biology* **9**, 641 (2013).
48. Urnov, F.D., Rebar, E.J., Holmes, M.C., Zhang, H.S. & Gregory, P.D. Genome editing with engineered zinc finger nucleases. *Nature reviews. Genetics* **11**, 636-646 (2010).
49. Joung, J.K. & Sander, J.D. TALENs: a widely applicable technology for targeted genome editing. *Nat Rev Mol Cell Biol* **14**, 49-55 (2013).
50. Horvath, P. & Barrangou, R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167-170 (2010).
51. Wiedenheft, B., Sternberg, S.H. & Doudna, J.A. RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482**, 331-338 (2012).
52. Deveau, H., Garneau, J.E. & Moineau, S. CRISPR/Cas system and its role in phage-bacteria interactions. *Annu. Rev. Microbiol.* **64**, 475-493 (2010).
53. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012).
54. Sander, J.D. & Joung, J.K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature biotechnology* **32**, 347-355 (2014).
55. Jacobs, J.Z., Ciccaglione, K.M., Tournier, V. & Zaratiegui, M. Implementation of the CRISPR-Cas9 system in fission yeast. *Nature communications* **5**, 5344 (2014).
56. Bassett, A.R., Tibbit, C., Ponting, C.P. & Liu, J.L. Highly efficient targeted mutagenesis of Drosophila with the CRISPR/Cas9 system. *Cell reports* **4**, 220-228 (2013).
57. Hwang, W.Y. et al. Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature biotechnology* **31**, 227-229 (2013).
58. Wang, H. et al. One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910-918 (2013).
59. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823 (2013).
60. Mali, P. et al. RNA-Guided Human Genome Engineering via Cas9. *Science* **339**, 823-826 (2013).
61. Shuman, S. & Glickman, M.S. Bacterial DNA repair by non-homologous end joining. *Nature reviews. Microbiology* **5**, 852-861 (2007).
62. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L.A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* **31**, 233-239 (2013).
63. Citorik, R.J., Mimee, M. & Lu, T.K. Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. *Nature biotechnology* **32**, 1141-1145 (2014).
64. Bikard, D. et al. Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. *Nature biotechnology* **32**, 1146-1150 (2014).
65. Yu, D. et al. An efficient recombination system for chromosome engineering in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 5978-5983 (2000).

66. Wang, H.H. et al. Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894-898 (2009).
67. Datta, S., Costantino, N., Zhou, X. & Court, D.L. Identification and analysis of recombineering functions from Gram-negative and Gram-positive bacteria and their phages. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 1626-1631 (2008).
68. Andrews, B.J., Proteau, G.A., Beatty, L.G. & Sadowski, P.D. The FLP recombinase of the 2 micron circle DNA of yeast: interaction with its target sequences. *Cell* **40**, 795-803 (1985).
69. Abremski, K. & Hoess, R. Bacteriophage P1 site-specific recombination. Purification and properties of the Cre recombinase protein. *The Journal of biological chemistry* **259**, 1509-1514 (1984).
70. Bolusani, S. et al. Evolution of variants of yeast site-specific recombinase Flp that utilize native genomic sequences as recombination target sites. *Nucleic acids research* **34**, 5259-5269 (2006).
71. Buchholz, F. & Stewart, A.F. Alteration of Cre recombinase site specificity by substrate-linked protein evolution. *Nature biotechnology* **19**, 1047-1052 (2001).
72. Karberg, M. et al. Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. *Nature biotechnology* **19**, 1162-1167 (2001).
73. Yao, J. & Lambowitz, A.M. Gene targeting in gram-negative bacteria by use of a mobile group II intron ("Targetron") expressed from a broad-host-range vector. *Applied and environmental microbiology* **73**, 2735-2743 (2007).
74. Heap, J.T. et al. The ClosTron: Mutagenesis in Clostridium refined and streamlined. *Journal of microbiological methods* **80**, 49-55 (2010).
75. Gaudelli, N.M. et al. Programmable base editing of A*T to G*C in genomic DNA without DNA cleavage. *Nature* **551**, 464-471 (2017).
76. Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A. & Liu, D.R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420-424 (2016).
77. Lampe, D.J., Churchill, M.E. & Robertson, H.M. A purified mariner transposase is sufficient to mediate transposition in vitro. *The EMBO journal* **15**, 5470-5479 (1996).
78. van Opijnen, T. & Camilli, A. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nature reviews. Microbiology* **11**, 435-442 (2013).
79. Johnston, C., Martin, B., Fichant, G., Polard, P. & Claverys, J.P. Bacterial transformation: distribution, shared mechanisms and divergent control. *Nature reviews. Microbiology* **12**, 181-196 (2014).
80. Rettedal, E.A., Gumpert, H. & Sommer, M.O. Cultivation-based multiplex phenotyping of human gut microbiota allows targeted recovery of previously uncultured bacteria. *Nature communications* **5**, 4714 (2014).
81. Lagier, J.C. et al. Culture of previously uncultured members of the human gut microbiota by culturomics. *Nat Microbiol* **1**, 16203 (2016).
82. Cascales, E. & Christie, P.J. The versatile bacterial type IV secretion systems. *Nature reviews. Microbiology* **1**, 137-149 (2003).
83. Gogarten, J.P. & Townsend, J.P. Horizontal gene transfer, genome innovation and evolution. *Nature reviews. Microbiology* **3**, 679-687 (2005).
84. Smillie, C.S. et al. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241-244 (2011).
85. Lozupone, C.A. et al. The convergence of carbohydrate active gene repertoires in human gut microbes. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 15076-15081 (2008).
86. Hehemann, J.H. et al. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**, 908-912 (2010).

87. Salyers, A.A., Gupta, A. & Wang, Y. Human intestinal bacteria as reservoirs for antibiotic resistance genes. *Trends Microbiol* **12**, 412-416 (2004).
88. Stecher, B. et al. Gut inflammation can boost horizontal gene transfer between pathogenic and commensal Enterobacteriaceae. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 1269-1274 (2012).
89. Brito, I.L. et al. Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **535**, 435-439 (2016).
90. Frost, L.S., Leplae, R., Summers, A.O. & Toussaint, A. Mobile genetic elements: the agents of open source evolution. *Nature reviews. Microbiology* **3**, 722-732 (2005).
91. Marraffini, L.A. & Sontheimer, E.J. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843-1845 (2008).
92. Thomas, C.M. & Nielsen, K.M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews. Microbiology* **3**, 711-721 (2005).
93. Gormley, E.P. & Davies, J. Transfer of plasmid RSF1010 by conjugation from Escherichia coli to Streptomyces lividans and Mycobacterium smegmatis. *Journal of bacteriology* **173**, 6705-6708 (1991).
94. Fang, F.C. & Helinski, D.R. Broad-host-range properties of plasmid RK2: importance of overlapping genes encoding the plasmid replication initiation protein TrfA. *Journal of bacteriology* **173**, 5861-5868 (1991).
95. Klumper, U. et al. Broad host range plasmids can invade an unexpectedly diverse fraction of a soil bacterial community. *ISME J* **9**, 934-945 (2015).
96. Yang, Y. et al. Focused specificity of intestinal TH17 cells towards commensal bacterial antigens. *Nature* **510**, 152-156 (2014).
97. Riglar, D.T. et al. Engineered bacteria can function in the mammalian gut long-term as live diagnostics of inflammation. *Nature biotechnology* **35**, 653-658 (2017).
98. Daeffler, K.N. et al. Engineering bacterial thiosulfate and tetrathionate sensors for detecting gut inflammation. *Molecular systems biology* **13**, 923 (2017).
99. Mimee, M., Tucker, A.C., Voigt, C.A. & Lu, T.K. Programming a Human Commensal Bacterium, Bacteroides thetaiotaomicron, to Sense and Respond to Stimuli in the Murine Gut Microbiota. *Cell Syst* **1**, 62-71 (2015).
100. Dou, J. & Bennett, M.R. Synthetic Biology and the Gut Microbiome. *Biotechnology journal* **13**, e1700159 (2018).
101. Munoz-Lopez, M. & Garcia-Perez, J.L. DNA transposons: nature and applications in genomics. *Curr Genomics* **11**, 115-128 (2010).
102. Curcio, M.J. & Derbyshire, K.M. The outs and ins of transposition: from mu to kangaroo. *Nat Rev Mol Cell Biol* **4**, 865-877 (2003).
103. Richardson, J.M. et al. Mechanism of Mos1 transposition: insights from structural analysis. *The EMBO journal* **25**, 1324-1334 (2006).
104. Richardson, J.M., Colloms, S.D., Finnegan, D.J. & Walkinshaw, M.D. Molecular architecture of the Mos1 paired-end complex: the structural basis of DNA transposition in a eukaryote. *Cell* **138**, 1096-1108 (2009).
105. Claeys Bouuaert, C., Lipkow, K., Andrews, S.S., Liu, D. & Chalmers, R. The autoregulation of a eukaryotic DNA transposon. *eLife* **2**, e00668 (2013).
106. Zhang, L., Sankar, U., Lampe, D.J., Robertson, H.M. & Graham, F.L. The Himar1 mariner transposase cloned in a recombinant adenovirus vector is functional in mammalian cells. *Nucleic acids research* **26**, 3687-3693 (1998).
107. Lampe, D.J., Grant, T.E. & Robertson, H.M. Factors affecting transposition of the Himar1 mariner transposon in vitro. *Genetics* **149**, 179-187 (1998).

108. Lampe, D.J., Akerley, B.J., Rubin, E.J., Mekalanos, J.J. & Robertson, H.M. Hyperactive transposase mutants of the Himar1 mariner transposon. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 11428-11433 (1999).
109. Goodman, A.L. et al. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe* **6**, 279-289 (2009).
110. van Opijnen, T., Bodi, K.L. & Camilli, A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nature methods* **6**, 767-772 (2009).
111. Zhang, J.K., Pritchett, M.A., Lampe, D.J., Robertson, H.M. & Metcalf, W.W. In vivo transposon mutagenesis of the methanogenic archaeon Methanosarcina acetivorans C2A using a modified version of the insect mariner-family transposable element Himar1. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 9665-9670 (2000).
112. Maragathavally, K.J., Kaminski, J.M. & Coates, C.J. Chimeric Mos1 and piggyBac transposases result in site-directed integration. *FASEB J* **20**, 1880-1882 (2006).
113. Owens, J.B. et al. Chimeric piggyBac transposases for genomic targeting in human cells. *Nucleic acids research* **40**, 6978-6991 (2012).
114. Owens, J.B. et al. Transcription activator like effector (TALE)-directed piggyBac transposition in human cells. *Nucleic acids research* **41**, 9197-9207 (2013).
115. Luo, W. et al. Comparative analysis of chimeric ZFP-, TALE- and Cas9-piggyBac transposases for integration into a single locus in human cells. *Nucleic acids research* **45**, 8411-8422 (2017).
116. Feng, X., Bednarz, A.L. & Colloms, S.D. Precise targeted integration by a chimaeric transposase zinc-finger fusion protein. *Nucleic acids research* **38**, 1204-1216 (2010).
117. Qi, L.S. et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173-1183 (2013).
118. Gilbert, L.A. et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442-451 (2013).
119. Guilinger, J.P., Thompson, D.B. & Liu, D.R. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nature biotechnology* **32**, 577-582 (2014).
120. Tsai, S.Q. et al. Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nature biotechnology* **32**, 569-576 (2014).
121. Chaikind, B., Bessen, J.L., Thompson, D.B., Hu, J.H. & Liu, D.R. A programmable Cas9-serine recombinase fusion protein that operates on DNA sequences in mammalian cells. *Nucleic acids research* **44**, 9758-9770 (2016).
122. Pickens, L.B., Tang, Y. & Chooi, Y.H. Metabolic engineering for the production of natural products. *Annu Rev Chem Biomol Eng* **2**, 211-236 (2011).
123. Esvelt, K.M., Smidler, A.L., Catteruccia, F. & Church, G.M. Concerning RNA-guided gene drives for the alteration of wild populations. *eLife* **3** (2014).
124. Goryshin, I.Y., Miller, J.A., Kil, Y.V., Lanzov, V.A. & Reznikoff, W.S. Tn5/IS50 target recognition. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 10716-10721 (1998).
125. Vigdal, T.J., Kaufman, C.D., Izsvak, Z., Voytas, D.F. & Ivics, Z. Common physical properties of DNA affecting target site selection of sleeping beauty and other Tc1/mariner transposable elements. *Journal of molecular biology* **323**, 441-452 (2002).
126. Trubitsyna, M., Morris, E.R., Finnegan, D.J. & Richardson, J.M. Biochemical characterization and comparison of two closely related active mariner transposases. *Biochemistry* **53**, 682-689 (2014).
127. Lampe, D.J. Bacterial genetic methods to explore the biology of mariner transposons. *Genetica* **138**, 499-508 (2010).
128. Warming, S., Costantino, N., Court, D.L., Jenkins, N.A. & Copeland, N.G. Simple and highly efficient BAC recombineering using galK selection. *Nucleic acids research* **33**, e36 (2005).

129. Li, X.T., Thomason, L.C., Sawitzke, J.A., Costantino, N. & Court, D.L. Positive and negative selection using the tetA-sacB cassette: recombineering and P1 transduction in Escherichia coli. *Nucleic acids research* **41**, e204 (2013).
130. Liu, D. & Chalmers, R. Hyperactive mariner transposons are created by mutations that disrupt allosterism and increase the rate of transposon end synapsis. *Nucleic acids research* **42**, 2637-2645 (2014).
131. Goryshin, I.Y. & Reznikoff, W.S. Tn5 in vitro transposition. *The Journal of biological chemistry* **273**, 7367-7374 (1998).
132. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012).
133. Yaung, S.J., Church, G.M. & Wang, H.H. Recent progress in engineering human-associated microbiomes. *Methods in molecular biology* **1151**, 3-25 (2014).
134. Cuiv, P.O. et al. Isolation of Genetically Tractable Most-Wanted Bacteria by Metaparental Mating. *Sci Rep* **5**, 13282 (2015).
135. Tock, M.R. & Dryden, D.T. The biology of restriction and anti-restriction. *Curr Opin Microbiol* **8**, 466-472 (2005).
136. Marraffini, L.A. CRISPR-Cas immunity in prokaryotes. *Nature* **526**, 55-61 (2015).
137. Thomas, C.M. & Nielsen, K.M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* **3**, 711-721 (2005).
138. Pansegrau, W. et al. Complete nucleotide sequence of Birmingham IncP alpha plasmids. Compilation and comparative analysis. *J Mol Biol* **239**, 623-663 (1994).
139. Hapfelmeier, S. et al. Reversible microbial colonization of germ-free mice reveals the dynamics of IgA immune responses. *Science* **328**, 1705-1709 (2010).
140. Myhal, M.L., Laux, D.C. & Cohen, P.S. Relative colonizing abilities of human fecal and K 12 strains of Escherichia coli in the large intestines of streptomycin-treated mice. *Eur J Clin Microbiol* **1**, 186-192 (1982).
141. Kommineni, S. et al. Bacteriocin production augments niche competition by enterococci in the mammalian gastrointestinal tract. *Nature* **526**, 719-722 (2015).
142. Saeidi, N. et al. Engineering microbes to sense and eradicate Pseudomonas aeruginosa, a human pathogen. *Mol Syst Biol* **7**, 521 (2011).
143. Wegmann, U., Horn, N. & Carding, S.R. Defining the bacteroides ribosomal binding site. *Applied and environmental microbiology* **79**, 1980-1989 (2013).
144. Sheth, R.U., Cabral, V., Chen, S.P. & Wang, H.H. Manipulating Bacterial Communities by in situ Microbiome Engineering. *Trends Genet* **32**, 189-200 (2016).
145. Dahlberg, C., Bergstrom, M. & Hermansson, M. In Situ Detection of High Levels of Horizontal Plasmid Transfer in Marine Bacterial Communities. *Appl Environ Microbiol* **64**, 2670-2675 (1998).
146. Brophy, J.A.N. et al. Engineered integrative and conjugative elements for efficient and inducible DNA transfer to undomesticated bacteria. *Nat Microbiol* **3**, 1043-1053 (2018).
147. Simon, R., Priefer, U. & Puhler, A. A Broad Host Range Mobilization System for In Vivo Genetic Engineering: Transposon Mutagenesis in Gram Negative Bacteria. *Nat Biotech* **1**, 784-791 (1983).
148. Datsenko, K.A. & Wanner, B.L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *P Natl Acad Sci USA* **97**, 6640-6645 (2000).
149. Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K. & Schloss, P.D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and environmental microbiology* **79**, 5112-5120 (2013).

150. Caporaso, J.G. et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America* **108 Suppl 1**, 4516-4522 (2011).
151. Parada, A.E., Needham, D.M. & Fuhrman, J.A. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol* **18**, 1403-1414 (2016).
152. Apprill, A., McNally, S., Parsons, R. & Weber, L. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquatic Microbial Ecology* **75**, 129-137 (2015).
153. Edgar, R.C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* **10**, 996-998 (2013).
154. 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 Microb* **73**, 5261-5267 (2007).
155. Nurk, S. et al. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J Comput Biol* **20**, 714-737 (2013).
156. Carattoli, A. et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* **58**, 3895-3903 (2014).
157. Guerry, P., van Embden, J. & Falkow, S. Molecular nature of two nonconjugative plasmids carrying drug resistance genes. *Journal of bacteriology* **117**, 619-630 (1974).
158. Scholz, P. et al. Complete nucleotide sequence and gene organization of the broad-host-range plasmid RSF1010. *Gene* **75**, 271-288 (1989).
159. Davis, C.P. & Savage, D.C. Habitat, succession, attachment, and morphology of segmented, filamentous microbes indigenous to the murine gastrointestinal tract. *Infection and immunity* **10**, 948-956 (1974).
160. Klaasen, H.L. et al. Intestinal, segmented, filamentous bacteria in a wide range of vertebrate species. *Lab Anim* **27**, 141-150 (1993).
161. Yin, Y. et al. Comparative analysis of the distribution of segmented filamentous bacteria in humans, mice and chickens. *ISME J* **7**, 615-621 (2013).
162. Ericsson, A.C., Hagan, C.E., Davis, D.J. & Franklin, C.L. Segmented filamentous bacteria: commensal microbes with potential effects on research. *Comp Med* **64**, 90-98 (2014).
163. Chase, D.G. & Erlandsen, S.L. Evidence for a complex life cycle and endospore formation in the attached, filamentous, segmented bacterium from murine ileum. *Journal of bacteriology* **127**, 572-583 (1976).
164. Klaasen, H.L., Koopman, J.P., Van den Brink, M.E., Van Wezel, H.P. & Beynen, A.C. Mono-association of mice with non-cultivable, intestinal, segmented, filamentous bacteria. *Archives of microbiology* **156**, 148-151 (1991).
165. Sczesnak, A. et al. The genome of th17 cell-inducing segmented filamentous bacteria reveals extensive auxotrophy and adaptations to the intestinal environment. *Cell Host Microbe* **10**, 260-272 (2011).
166. Prakash, T. et al. Complete genome sequences of rat and mouse segmented filamentous bacteria, a potent inducer of th17 cell differentiation. *Cell Host Microbe* **10**, 273-284 (2011).
167. Pamp, S.J., Harrington, E.D., Quake, S.R., Relman, D.A. & Blainey, P.C. Single-cell sequencing provides clues about the host interactions of segmented filamentous bacteria (SFB). *Genome research* **22**, 1107-1119 (2012).
168. Kuwahara, T. et al. The lifestyle of the segmented filamentous bacterium: a non-culturable gut-associated immunostimulating microbe inferred by whole-genome sequencing. *DNA Res* **18**, 291-303 (2011).

169. Atarashi, K. et al. Th17 Cell Induction by Adhesion of Microbes to Intestinal Epithelial Cells. *Cell* **163**, 367-380 (2015).
170. Schnupf, P., Gaboriau-Routhiau, V. & Cerf-Bensussan, N. Host interactions with Segmented Filamentous Bacteria: an unusual trade-off that drives the post-natal maturation of the gut immune system. *Semin Immunol* **25**, 342-351 (2013).
171. Lee, Y.K., Menezes, J.S., Umesaki, Y. & Mazmanian, S.K. Proinflammatory T-cell responses to gut microbiota promote experimental autoimmune encephalomyelitis. *Proceedings of the National Academy of Sciences of the United States of America* **108 Suppl 1**, 4615-4622 (2011).
172. Wu, H.J. et al. Gut-residing segmented filamentous bacteria drive autoimmune arthritis via T helper 17 cells. *Immunity* **32**, 815-827 (2010).
173. Kriegel, M.A. et al. Naturally transmitted segmented filamentous bacteria segregate with diabetes protection in nonobese diabetic mice. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 11548-11553 (2011).
174. Ericsson, A.C. et al. Isolation of segmented filamentous bacteria from complex gut microbiota. *BioTechniques* **59**, 94-98 (2015).
175. Schnupf, P. et al. Growth and host interaction of mouse segmented filamentous bacteria in vitro. *Nature* **520**, 99-103 (2015).
176. Bannam, T.L. & Rood, J.I. Relationship between the Clostridium perfringens catQ gene product and chloramphenicol acetyltransferases from other bacteria. *Antimicrobial agents and chemotherapy* **35**, 471-476 (1991).
177. Berryman, D.I. & Rood, J.I. Cloning and hybridization analysis of ermP, a macrolide-lincosamide-streptogramin B resistance determinant from Clostridium perfringens. *Antimicrobial agents and chemotherapy* **33**, 1346-1353 (1989).
178. Bannam, T.L. & Rood, J.I. Clostridium perfringens-Escherichia coli shuttle vectors that carry single antibiotic resistance determinants. *Plasmid* **29**, 233-235 (1993).
179. Wang, Y., Li, X., Mao, Y. & Blaschek, H.P. Single-nucleotide resolution analysis of the transcriptome structure of Clostridium beijerinckii NCIMB 8052 using RNA-Seq. *BMC genomics* **12**, 479 (2011).
180. Salis, H.M. The ribosome binding site calculator. *Methods in enzymology* **498**, 19-42 (2011).
181. Corvaglia, A.R. et al. A type III-like restriction endonuclease functions as a major barrier to horizontal gene transfer in clinical Staphylococcus aureus strains. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 11954-11958 (2010).
182. Murray, N.E. Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). *Microbiology and molecular biology reviews : MMBR* **64**, 412-434 (2000).
183. Farkas, A.M. et al. Induction of Th17 cells by segmented filamentous bacteria in the murine intestine. *J Immunol Methods* **421**, 104-111 (2015).
184. Fang, G. et al. Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nature biotechnology* **30**, 1232-1239 (2012).
185. Anton, B.P. et al. Complete Genome Sequence of ER2796, a DNA Methyltransferase-Deficient Strain of Escherichia coli K-12. *PloS one* **10**, e0127446 (2015).
186. Melville, S.B., Labbe, R. & Sonenshein, A.L. Expression from the Clostridium perfringens cpe promoter in C. perfringens and Bacillus subtilis. *Infection and immunity* **62**, 5550-5558 (1994).
187. Bannam, T.L., Teng, W.L., Bulach, D., Lyras, D. & Rood, J.I. Functional identification of conjugation and replication regions of the tetracycline resistance plasmid pCW3 from Clostridium perfringens. *Journal of bacteriology* **188**, 4942-4951 (2006).
188. Li, J. & McClane, B.A. A novel small acid soluble protein variant is important for spore resistance of most Clostridium perfringens food poisoning isolates. *PLoS Pathog* **4**, e1000056 (2008).

189. Murugan, K., Babu, K., Sundaresan, R., Rajan, R. & Sashital, D.G. The Revolution Continues: Newly Discovered Systems Expand the CRISPR-Cas Toolkit. *Mol Cell* **68**, 15-25 (2017).
190. Luo, M.L., Mullis, A.S., Leenay, R.T. & Beisel, C.L. Repurposing endogenous type I CRISPR-Cas systems for programmable gene repression. *Nucleic acids research* **43**, 674-681 (2015).
191. Johns, N.I. et al. Metagenomic mining of regulatory elements enables programmable species-selective gene expression. *Nature methods* **15**, 323-329 (2018).
192. Tropini, C., Earle, K.A., Huang, K.C. & Sonnenburg, J.L. The Gut Microbiome: Connecting Spatial Organization to Function. *Cell Host Microbe* **21**, 433-442 (2017).
193. Doron, S. et al. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359** (2018).
194. Ridlon, J.M., Kang, D.J. & Hylemon, P.B. Bile salt biotransformations by human intestinal bacteria. *J Lipid Res* **47**, 241-259 (2006).
195. Joyce, S.A. et al. Regulation of host weight gain and lipid metabolism by bacterial bile acid modification in the gut. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 7421-7426 (2014).
196. Xu, R., Wang, Q. & Li, L. A genome-wide systems analysis reveals strong link between colorectal cancer and trimethylamine N-oxide (TMAO), a gut microbial metabolite of dietary meat and fat. *BMC genomics* **16 Suppl 7**, S4 (2015).

Appendix 1

Trends in Genetics

CellPress

Opinion

Manipulating Bacterial Communities by *in situ* Microbiome Engineering

Ravi U. Sheth,[1,2] Vitor Cabral,[1] Sway P. Chen,[1,2] and Harris H. Wang[1,3,*]

Microbial communities inhabit our entire planet and have a crucial role in biogeochemical processes, agriculture, biotechnology, and human health. Here, we argue that '*in situ* microbiome engineering' represents a new paradigm of community-scale genetic and microbial engineering. We discuss contemporary applications of this approach to directly add, remove, or modify specific sets of functions and alter community-level properties in terrestrial, aquatic, and host-associated microbial communities. Specifically, we highlight emerging *in situ* genome engineering approaches as tractable techniques to manipulate microbial communities with high specificity and efficacy. Finally, we describe opportunities for technological innovation and ways to bridge existing knowledge gaps to accelerate the development of *in situ* approaches for microbiome manipulations.

Trends

High-throughput sequencing advances have provided a detailed survey of microbial community composition and prevalence, but a functional and mechanistic understanding of microbial ecology is lacking.

Manipulating microbiome composition and function is of great interest for basic science and engineering applications; contemporary methods for manipulating microbial communities *in situ* yield perturbations limited to particular specificities and magnitudes.

Emerging *in situ* genome engineering tools can precisely alter the metagenomic content of microbial communities over a range of specificities and magnitudes.

Advances and Roadblocks in Microbiome Research

Over the past decade, breakthroughs in high-throughput sequencing to read DNA from genomes have vastly outpaced our capabilities to edit genetic information. Developments in metagenomic and transcriptomic sequencing of mixed cell populations have enabled large-scale quantification of microbial community composition, function, and dynamics in a culture-independent manner [1]. However, we still lack a basic mechanistic understanding of the individual genetic factors that drive overall function and emergent ecological principles in these communities. To understand these complex communities and engineer them in useful ways, we must address critical technological and knowledge gaps in systems-level genetic manipulation [2].

Most microbial communities in nature exist in complex, dynamic consortia with highly interconnected networks of metabolic and ecological interactions that have yet to be unraveled. Consequently, the natural environments of most microbial communities are difficult, if not impossible, to recreate experimentally. In fact, a large majority of microbes have not been cultivated in the laboratory [3]. These microbes are not accessible for genetic studies, and their function and properties within their communities remain unknown. Even for culturable microbes, only a few genetic systems have been utilized in specific model strains [4]. Development of new genetic systems requires significant time and effort, and resulting tools are not necessarily transferrable between different microbes. Furthermore, studies of single model strains in laboratory conditions may not necessarily reflect behaviors relevant to natural environments. No general methods are available to genetically engineer consortia of different organisms *in situ*. These challenges greatly limit the functional analysis and forward engineering of polymicrobial communities of any level of complexity.

[1]Department of Systems Biology, Columbia University Medical Center, New York, NY, USA
[2]Integrated Program in Cellular, Molecular and Biomedical Studies, Columbia University Medical Center, New York, NY, USA
[3]Department of Pathology and Cell Biology, Columbia University Medical Center, New York, NY, USA

*Correspondence: hw2429@columbia.edu (H.H. Wang).

Engineering Microbial Communities *in situ*

***In situ* microbiome engineering** (see Glossary) methods allow for the manipulation and study of microbial communities in their native context without the need for individual laboratory domestication. These approaches can be classified by several characteristics: the magnitude of perturbation to community composition and function; specificity of perturbation to community members or processes; and degree of engineerability (Figure 1). While some approaches have low specificity and can lead to large-scale changes (e.g., **microbiota** transplants), others can be more easily designed to affect specific members while minimizing the overall impact on the community (e.g., engineered probiotics). Here, we discuss contemporary *in situ* microbiome engineering approaches and offer *in situ* **genome engineering** as a new paradigm to directly manipulate communities with greater control of magnitude and specificity.

Contemporary Methods for *in situ* Microbiome Engineering

Chemical, cellular, and phage-based methods can be used to alter microbial communities *in situ*. Common examples of each approach are outlined in Table 1 and discussed in detail below.

Chemical Modifiers of Microbiomes

Biochemical availability can predictably affect microbiome composition and function [5]. **Prebiotics** are naturally occurring chemicals that selectively promote growth or activity in a community. Human-associated prebiotics are often nondigestible dietary polysaccharides that stimulate growth of commensal bacteria in the gut [6,7]. Prebiotics can be used in other settings,

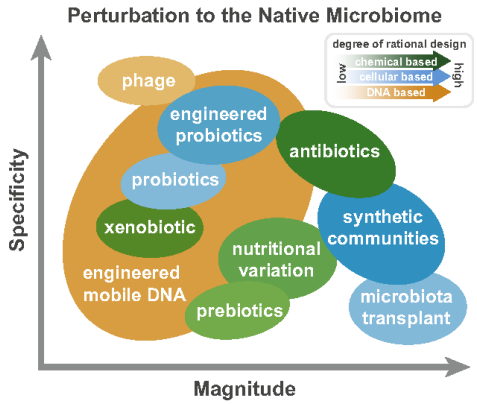


Figure 1. *In situ* Microbiome Perturbations Vary in Magnitude, Specificity, and Degree of Rational Design Required. A variety of approaches, based on chemical (green), cellular (blue), and DNA (orange) methods, can be applied to manipulate microbial communities in their native context. Each method can vary in its magnitude of perturbation to the native microbiome, shown increasing on the horizontal axis; its specificity of targeting to particular community members, shown increasing on the vertical axis; and its degree of required rational design, shown with increasing shading density. Particular combinations of magnitudes and specificities may be desirable for given target applications. Chemical-based approaches, such as xenobiotics, prebiotics, and nutritional variation, yield relatively broad-spectrum changes, with varying magnitudes. Antibiotics, a class of xenobiotics, can produce larger magnitude changes with higher specificity. Cellular-based techniques, such as probiotics and engineered probiotics, can yield low-magnitude, specific perturbations, while large-scale microbiota transplants or synthetic communities can lead to larger, but less-specific changes. Finally, DNA-based methods, such as phages, can yield highly specific, albeit low-magnitude perturbations, while engineered mobile DNA can yield perturbations over a large range of magnitudes and specificity. This flexible control of magnitude and specificity implies that engineered mobile DNA may be a desirable and tractable method for manipulating microbial communities compared to other methods.

Glossary

Bacteriophage/phage: a virus that infects and hijacks the machinery of a bacterium to reproduce; may integrate stably into the bacterial genome.
Conjugation: a mechanism of genetic material transfer via direct cell–cell contact.
Genome engineering: technologies or approaches to alter genetically inheritable information in a targeted or specific manner.
***In situ* microbiome engineering:** manipulation of microbial communities in their native environment.
Metagenome: the collection of genes, genomes, and inheritable information present in a given environment.
Microbiome: encompassing term referring to the microbiota, metagenome, and surrounding environment of a microbial community.
Microbiota: the set of microorganisms present in a given environment.
Mobile genetic elements: genetic information that can be transferred between cells; includes conjugative plasmids, transposons, and bacteriophages.
Prebiotics: naturally occurring chemicals that can promote growth or activity in a community in a selective manner.
Probiotics: bacteria that can confer a benefit to a particular host environment.
Xenobiotics: biochemical compounds unnatural to an environment that can promote or limit the growth or function of specific microbial community members.

Table 1. Commonly Utilized *in situ* Microbiome-Engineering Methods

Method	Class (Common Formulations)	Predominant Targets	Mechanism of Action
Prebiotics (chemical based)	Dietary fibers (inulin), polysaccharides (oligosaccharides)	*Lactobacillus*, *Bifidobacteria*	Promote bacterial growth, mechanism generally unknown
Antibiotics (chemical based)	β-lactams (cephalosporins, carbapenems)	*Clostridium*, *Staphylococcus*, *Streptococcus*	Block cell wall synthesis
	Aminoglycosides (kanamycin, streptomycin)	*Klebsiella*, *Pseudomonas*	Block protein synthesis
	Macrolides (erythromycin, azithromycin)	*Chlamydia*, *Legionella*, *Mycoplasma*	Block protein synthesis
	Glycopeptides (vancomycin)	*Enterococcus*, *Clostridium*, *Staphylococcus*	Block peptidoglycan synthesis
	Quinolones (ciprofloxacin, levofloxacin)	*Neisseria*, *Pseudomonas*, *Streptococcus*	Block DNA replication
	Metronidazole	*Bacteroides*, *Clostridium*	Block DNA/RNA synthesis
Probiotics (cellular based)	Firmicutes (*Lactobacillus*), Actinobacteria (*Bifidobacteria*), Proteobacteria	Variable, broadly targeting	Compete for nutrients, produce antimicrobials, modulate environment
Microbiota transplants (cellular based)	Fecal microbiota transplants	Variable, broadly targeting	Replace native community, mechanism generally unknown
Bacteriophages (phage based)	Specific phage strains or cocktails of phages	Variable, but strain specific	Cell lysis (lytic), genomic integration (lysogenic)

such as the food industry, where the polysaccharide β-glucan has been commercially sold to improve fish health and resistance to infection [8]. However, a major limitation of prebiotics is the inability to rationally predict or change the specificity of their manipulations. To address this shortcoming, approaches such as functional metagenomics, and transcriptome and transposon sequencing have been used to identify gene-level fitness determinants of individual microbes in specific metabolic niches [9,10]. These measurements have yielded more nuanced nutritional variation approaches that can be exploited to modulate the microbiome with higher precision [11]. For example, transposon sequencing of a model gut microbiome community revealed a specific dietary metabolite that could modulate the abundance of a single strain in the community [12]. Furthermore, transcriptome sequencing of the gut bacterium *Eggerthella lenta* revealed specific genes that inactivate the cardiac drug digoxin and their transcriptional regulation, enabling the design of a dietary modification to reduce inactivation *in vivo* [13].

Targeted biochemical modulation of microbiota can also utilize **xenobiotics**, compounds that are foreign to an environment designed to modulate microbial function or growth. For example, β-glucuronidase inhibitors have been used to reduce the toxicity of a chemotherapeutic by inactivating bacterial enzymes that reactivate the drug [14]. Asparaginase can be delivered *in vivo* to protect against infection by degrading asparagine, a regulator of group A *Streptococcus* proliferation [15]. Finally, a small-molecule structural analog of choline can inhibit TMA production by microbes, reducing levels of TMAO, a metabolite associated with cardiac disease [16].

Antibiotics are a widely used class of xenobiotics that modulate microbial growth by inhibiting essential cellular machinery. Most antibiotics target membrane integrity, protein synthesis, or

replication processes [17]. However, due to the widely conserved function of mechanisms targeted by antibiotics, they tend to affect a broad spectrum of bacteria, causing large and potentially undesirable changes in the community [18]. Antibiotic use can lead to persistent alterations in microbiota composition over time, and can further select for antibiotic-resistance genes [19]. With the emergence of many antibiotic-resistant pathogens, there is a critical need for new classes of antibiotics that selectivssely target undesirable strains without promoting the spread of broad-spectrum resistance. Antimicrobial peptides and secondary metabolites, such as bacteriocins, are outstanding candidates for such novel antibiotics, as because they display selective elimination of particular strains and a diverse set of these compounds is present in natural communities [20,21]. A better mechanistic understanding of specific biochemical processes and strain-level genetic factors will improve strategies to modulate microbiome growth and function with high specificity.

Cellular Modifiers of Microbiomes

Beyond biochemical approaches, live bacterial strains or communities can be used to manipulate microbial ecosystems. In contrast to molecular modulators, these cellular approaches can yield more nuanced interaction and function over space and time. **Probiotics** are bacteria that can confer a benefit to a particular host environment [22]. For example, probiotic *Lactobacilli* have been used in livestock to decrease the incidence of pathogenic infections [23]. In humans, probiotic bacteria alter the gut microbiota by competing for nutrients, producing antimicrobial compounds, or modulating host immunity [24,25]. However, our fledgling understanding of probiotic mechanistic function has limited their value as a tool for predictive microbiome manipulation.

Genetically engineered probiotics have potential for more targeted community manipulations. For example, in the poultry industry, genetically modified *Salmonella* lacking virulence factors have been used to vaccinate chickens against infection [26]. Synthetic biology tools can be applied to engineer probiotics with precise and novel functions. For example, bacteria have been modified to modulate microbiota or host physiology by secreting chemicals or proteins, including human interleukin-10 to reduce inflammation [27], NAPEs to reduce food intake and obesity [28], and bacterial quorum signals to modulate microbiota composition [29]. Wholly new functions can also be engineered: Danino *et al.* engineered an *Escherichia coli* probiotic as an orally administered diagnostic of liver metastasis in mice through production of a detectable signal in urine [30]. More complex synthetic biology circuits, such as combinatorial logic [31] or memory [32] circuits, can be layered upon these simple designs to improve the precision, specificity, and controllability of desired perturbations.

Mixtures of bacteria can also be utilized to manipulate microbiomes. Microbiota transplantation is the beneficial transfer of live bacteria from one environment into another. This approach has been recently popularized through successful clinical trials of fecal microbiota transplants to treat recurrent *Clostridium difficile* infections [33]. It is hypothesized that transplantation may replace the existing microbial community with a more infection-resistant community from a healthy donor, but the mechanisms of this process remain largely unclear [34]. A more refined approach to transplantation is the transfer of synthetic communities, which could replicate the functions of complex consortia, but contain defined members that are amenable to detailed genetic and biochemical studies [35–37]. We envision that novel experimental tools to delineate interspecies and host–microbe interactions, and improved metabolic and ecological modeling [38], combined with functional studies in gnotobiotic animals [39], will enable better design of microbial consortia to precisely manipulate microbial communities.

Phage-Based Modifiers of Microbiomes

Bacteriophages (phages) are the most abundant, diverse, and rapidly replicating life forms on Earth [40]. Phages infect a host microbe, hijack its replication machinery to reproduce, and then

replicate via stable genomic integration (lysogenic cycle) or lysis of the host and dissemination (lytic cycle). This life cycle makes them ideal genetic-engineering candidates to selectively eliminate strains or transfer specific genes in microbial populations. Indeed, natural phages have been used to limit the growth of undesirable or pathogenic bacteria in humans [41], agriculture [42], food processing [43], and aquaculture [44]. Phages have been further genetically engineered to deliver specific DNA payloads or to alter host specificity. For example, phages have been designed to deliver biofilm dispersal enzymes [45] or genes that increase antibiotic sensitivity [46]. By delivering the CRISPR-Cas RNA-guided nuclease system with phages, designated strains can be selectively eliminated based on their genetic content [47]. Furthermore, phage host ranges can be modularly engineered by swapping phage tail components [48]. Altering phage populations represents another avenue for microbiome modulation; exposing the gut microbiome to antibiotics alters its associated virome and ecological networks [49]. Quantitative characterization of phage ecology, combined with advances in forward genetic engineering of phage function, will allow for more complex manipulations of microbial communities *in situ*.

In situ Genome Engineering: An Emerging Frontier in Microbiome Modulation

Despite much progress, contemporary methods for modifying microbiomes have not seen widespread success in achieving desired manipulations. We attribute these shortcomings to two major knowledge and technical barriers. First, we lack a fine-scale understanding of how individual microbial species function in the context of their natural environments, and a subsequent large-scale understanding of emergent ecosystem function. Second, we lack techniques to efficiently manipulate microbial communities over a large range of magnitudes and specificities. These barriers have limited the effective design and physical implementation of manipulations.

Current methods face a variety of outstanding engineering challenges. For example, chemical manipulations are specific to particular microbial strains and biochemical processes and cannot be broadly applied. Cellular approaches require colonization of a foreign strain into an ecologically competitive environment, which may be difficult to engineer or lead to unwanted consequences.

Rather than targeting specific strains or functions, one could instead directly modify the metagenomic content of a community to achieve a desired manipulation. While genomes vary greatly between microbes, the **metagenome** of a community is more constant [50], and governs its biochemical and cellular function. Compared with currently available techniques, direct genomic manipulation could enable perturbations with magnitude and specificity tunable over a greater range. For example, a metabolic pathway could be added directly to the genome of a native microbe, rather than introducing a foreign strain containing the pathway, thus reducing off-target effects and achieving high-specificity manipulation. Alternatively, the same pathway could be targeted to a range of native organisms to achieve a large-magnitude manipulation. Advances in materials science to manipulate and polymerize chemical building blocks enabled the proliferative use of plastics during the 20th century; we analogously envision that direct, tunable manipulations of the genetic building blocks of microbial communities will enable novel bioengineering applications. Here, we advocate for the development of *in situ* genome engineering approaches, or techniques to directly manipulate genetic information and engineer new functions in complex microbial communities (Figure 2).

Given that complex communities are difficult to recapitulate in the laboratory, new approaches are needed for genome engineering *in situ*. In nature, microbial genomes are in constant flux as a result of abundant horizontal gene transfer events mediated by **mobile genetic elements** [51]. These horizontal gene transfer events have been increasingly recognized for the important roles they have in the evolution of individual genomes [52] and of entire microbial communities [49]. Furthermore, these events occur rapidly, on timescales of less than a week amongst bacterial

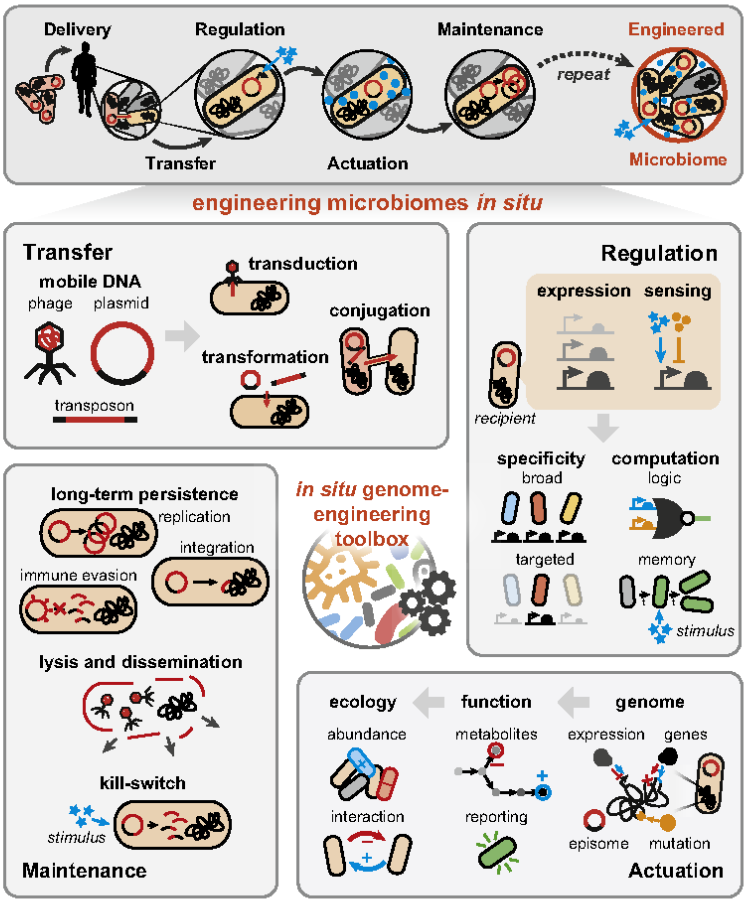


Figure 2. The *in situ* Genome Engineering Toolbox. The genomic content of native microbial communities can be directly engineered via *in situ* genome engineering. As we depict in the top panel, mobile genetic elements can be delivered and transferred to an endogenous microbiome, where they elicit desired functions via a combination of regulation and actuation strategies; these elements can then maintain themselves over time to achieve their long term desired functions. A toolbox of existing and novel genetic tools will enable engineering of mobile genetic elements for *in situ* genome engineering methods. Transfer methods, such as phages, plasmids, and transposons, can be used to deliver and circulate engineered DNA sequences to microbial communities, via processes such as transduction, transformation, and conjugation. Regulatory parts, including transcription and translation parts and sensors of endogenous and exogenous chemical ligands, will enable the construction of more complex genetic devices to tune host range and endow higher order functions, such as logic and memory. Actuation of genomes, including addition and removal of genes, modulation of expression, targeted mutations, and episomal modifications, will allow for changes to underlying community metabolic function, or the introduction of wholly new functions, such as reporting on the state of an environment. These functional manipulations can further alter communities at the ecological level by altering the abundances of specific strains or introducing competitive or cooperative interactions. Maintenance of these mobile elements allows for dynamic and long term control of engineered genetic content; with replication, integration, and optimized immune evasion, vectors can be stably propagated over time. Alternatively, lysis can quickly disseminate phages across a community, or engineered circuits, such as kill switches, could be used as a safeguard.

species in the gut microbiome [53]. Thus, natural horizontal gene transfer processes, including **conjugation**, natural transformation, and phage delivery, are viable entry points for *in situ* genome engineering.

Conjugation via plasmids or transposable elements enables transfer of genetic material through direct cell–cell contact. An *in situ* conjugative system could utilize a transiently introduced donor cell carrying an engineered mobile plasmid or transposable element that would be transferred to native microbiota. Secondary conjugations could be engineered to occur following the first transfer to promote further propagation. Bacteria can also uptake extracellular DNA through natural transformation machinery [54] or chemical and physical transformation processes, such as abrasion with mineral particles [55], which creates small membrane pores that allow DNA into the cytoplasm [56]. Phages can also be utilized to transfer and insert genes into bacterial genomes.

Building an Expansive *in situ* Genome Engineering Toolbox

We envision a suite of genetic tools that will significantly expand our ability to activate stable and controllable synthetic gene circuits in complex natural microbiomes. First, natural horizontal gene transfer vectors could be engineered with tunable host ranges and dynamics. These vectors could then be augmented with existing and emerging synthetic biology tools, such as transcriptional and translational regulatory parts, logic gates, and genome editing tools to add, remove, or modify particular functions and ecologies. Finally, maintenance of these engineered vectors could be precisely controlled, resulting in propagation across specific members of a community or destruction via controlled kill switches. Recent studies suggest that such mobile genetic element-mediated transfer is a tractable approach to manipulate diverse communities. Plasmids can be broadly mobilized into naturally occurring soil bacteria, transferring to bacteria from 11 different phyla [57] with efficiencies of up to 1 in 10 000 cells within a few days [58,59]. Furthermore, viral tagging experiments have revealed that, in nature, a single bacterial host can harbor dozens of different phage populations, suggesting that viruses can be isolated and engineered for a broad range of hosts [60]. Here, we highlight key components that need to be developed to form the foundation for this new *in situ* genome engineering toolbox.

New replicative or integrative plasmids are needed for the stable propagation of exogenous DNA in the microbiome. Recent sequencing efforts have demonstrated that plasmids are prevalent in microbial communities [61]. However, the host ranges of most plasmids remain unknown, and their associated proteins and modes of regulation are poorly characterized [62]. Few, if any, new plasmids are being developed, as almost all current vectors are based on plasmids isolated during the pre-genomic era. Characterizing more natural plasmids with various host ranges will elucidate the principles underlying their host specificity, transfer dynamics, and stability, and will aid in the construction of new synthetic vectors [63].

Tunable gene regulation systems with varying strengths and host specificities are needed to better control the activity of synthetic circuits across microbial communities. A repository of characterized regulatory parts, such as promoters and ribosome-binding sites, for diverse microbes does not yet exist and needs to be developed. Furthermore, basic measurement techniques to assess genetic circuit function across many species in parallel have not been developed. New strategies will be needed to engineer genetic circuits with broad and defined host ranges [64]. Existing and novel genetic parts, combined with community-level measurement strategies, will yield design principles for building genetic circuits with predictable performance in different hosts.

Recent advances in genome editing have enabled programmable modification of microbial genomes. RNA-guided CRISPR-Cas9 systems have been used to site-specifically edit bacterial genomes [65], while LI.LtrB group II introns (retrotransposons that undergo RNA intermediary

steps) have been successfully repurposed for targeted gene editing of multiple bacterial species [66]. Although these techniques enable powerful interrogation of the genome, new tools to enable strain- and site-specific genomic integration of large synthetic constructs are still required, because current approaches are often inefficient or difficult to target to desired genomic sites. Additionally, increased knowledge and functional annotation of microbial genomes, coupled with advances in modeling techniques to predict the effects of particular genomic modifications, will be necessary to maximize the utility of existing genome-editing tools.

More sophisticated genetic devices are also needed to perform higher function processes. Programmable transcriptional and post-transcriptional regulators [67,68], combined with new chemical sensing pathways, will enable sophisticated genetic circuitry, such as memory devices and kill switches, to control engineered functions in complex communities [69]. Other strategies, such as a gene drive to propagate engineered function in higher-order organisms [70], conceptually similar to bacterial *in situ* genome engineering, may be further developed for microbial populations.

A better understanding of the function and dynamics of natural mobile genetic elements will enable strategies for long-term persistence of mobile genetic elements. A major challenge in utilizing these natural elements is ensuring efficient delivery, transfer, and stability of the system over time. Delivery of mobile genetic elements and subsequent transfer between endogenous microbiota *in situ* could minimize perturbations to the overall structure of a given community. Mobile elements found in the wild use elegant strategies to ensure their long-term presence; for example, an *Enterococcus faecalis* conjugative plasmid expressing a bacteriocin enhances niche colonization of its host, facilitating transfer to other *E. faecalis* strains in the mammalian gut [53]. Similar strategies that couple a niche or metabolic advantage to a mobile element could be used to engineer selection and stability over time.

Finally, many bacterial immune systems, such as CRISPRs and restriction endonucleases, prevent foreign DNA from infiltrating the cell. Active immune evasion is required to enable efficient transfer and propagation of engineered DNA, which may require sequence recoding to avoid restriction enzyme digestion or modification of DNA methylation patterns to match that of the recipient cells. In fact, methylation matching has been shown to increase gene transfer rates by several orders of magnitude [71]. Experimental and design tools for bacterial immune evasion are needed to predictably manipulate gene transfer efficiencies *in situ*.

Design Principles, Knowledge Gaps and Applications of *in situ* Microbiome Engineering

Successful *in situ* microbiome engineering will require an expanded understanding of basic ecological principles; new systems, measurement methods, and genetic parts; and the application of existing and new quantitative modeling frameworks (Figure 3, Key Figure). These three major knowledge areas directly inform and influence each other; for example, underlying ecology provides a starting point for determining new genetic parts for systems engineering. The performance of these parts can then be used to parameterize quantitative models of complex microbial systems, which could ultimately reveal hidden or underlying ecological interactions. However, as outlined in Figure 3, key knowledge gaps remain to be addressed.

The application of *in situ* microbiome engineering to basic science questions will enable a new class of experiments to elucidate the determinants of individual and community function. In individual strains, these new tools will allow for genetic studies of unculturable microbes and the mechanistic basis of microbial fitness and function. In communities, these tools will enable large-scale perturbations of community composition and interactions at unprecedented resolution and scale. Furthermore, *in situ* microbiome engineering has numerous applications to the

Key Figure

Principles and Knowledge Gaps in *in situ* Genome Engineering

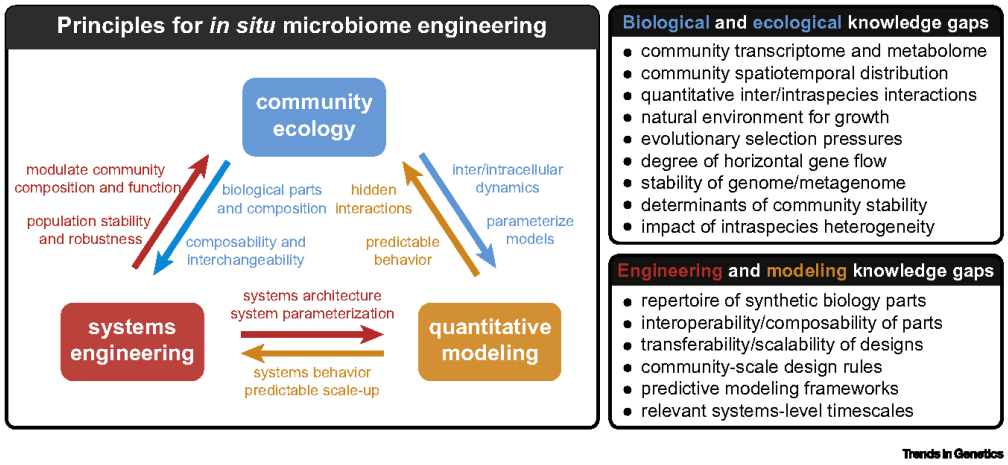


Figure 3. Key principles for *in situ* microbiome engineering include community ecology (blue), systems engineering (red), and quantitative modeling (orange). On the left, we illustrate the direct interplay between each of the three principles; these directional interactions and knowledge that inform one another are denoted via color-coded arrows and text. On the right, we detail specific biological, ecological, engineering, and modeling knowledge gaps.

addition, deletion, or modification of community-level functions and properties (Figure 4). In human-associated communities, microbiome manipulations could improve health and nutrition via removal of detrimental host interactions (e.g., chronic inflammation) and addition of beneficial processes (e.g., producing essential nutrients). Altering community-level properties, such as resilience to infection, could yield novel ecology-based treatments for infectious diseases. In an agricultural setting, targeted manipulations could increase crop yield by accelerating nutritional absorption or enhance bioremediation by removing toxins. In human-made environments, such as buildings, pipelines, and ship hulls, engineered communities could enable exclusion of strains with undesirable properties (e.g., pathogenic or biofouling) and augmentation of materials with desirable 'smart' properties, such as self-healing, chemical production, and recording exposure to biochemical compounds. Finally, synthetic communities with defined functions could replace natural communities in certain settings to enable predictability and control over particular biochemical processes. Such communities could be used to colonize environments lacking endogenous microbiota, such as other planets, to improve habitability for humans.

Safety and Regulation

The manipulation and engineering of microbial ecosystems in natural environments will require significant advances in our ability to reliably predict engineering outcomes and safeguard against undesirable events. Current manipulations of microbial ecosystems, intended or not, are widespread and subject to a complex litany of regulatory policies with varying stringency. New policies and regulatory frameworks will be required to appropriately evaluate the safety and implementation of emerging approaches, such as genetically engineered probiotics and mobile vectors. Analogous to conversations around the use of gene drives to cause forced

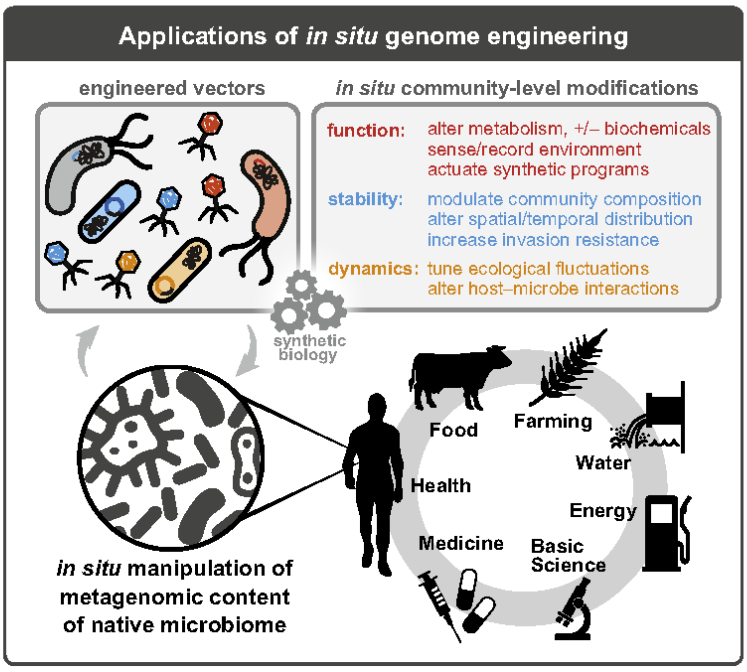


Figure 4. Applications of *in situ* Genome Engineering. *In situ* genome engineering approaches could be utilized to address fundamental basic science questions and key applications in medicine, health, food, farming, water, and energy. Engineered vectors introduced into native environments could allow for the manipulation of metagenomic content, and would propagate within the community over time. These engineered vectors could then endow these communities with desirable alterations to community level function, stability, and dynamics, detailed in the upper right panel.

inheritance of engineered traits in a population [72], or CRISPR-Cas9 for gene-editing purposes [73], the societal implications of microbiome-engineering technologies and their regulation must be carefully considered, evaluated, and communicated to the public for these open-environment engineering approaches to be rationally evaluated and safely adopted.

Concluding Remarks

In situ microbiome and genome engineering offer exciting opportunities at the frontier of population and ecological engineering with applications in basic science, human health, agriculture, and beyond. As we move from understanding and engineering individual organisms to entire ecosystems, we envision that these emerging techniques will reveal a vast diversity and elegance underlying natural microbial ecosystems, and will correspondingly suggest wholly new strategies to manipulate microbial communities (see Outstanding Questions).

Outstanding Questions

How do specific biological factors, such as genetic content or biochemical interactions, lead to emergent ecological properties, including spatial variation, dynamics, and stability?

What is the extent of horizontal gene transfer in microbiomes, and how does this impact intraspecies heterogeneity, genomic stability, evolutionary selection, and community function?

How can standardized synthetic biology genetic parts for genetic transfer, regulation, and genome editing be characterized and designed across many different bacteria?

What systems engineering principles can be leveraged to increase transferability and composability of these parts into more complex genetic devices across many different hosts?

What are appropriate quantitative modeling frameworks to predict specific microbiome manipulations?

Acknowledgments

We thank members of the Wang lab for helpful discussions and feedback that help shape the content of this work. H.H.W. acknowledges funding support from the NIH (1DP5OD009172 02, 1U01GM110714 01), NSF (MCB 1453219), Sloan Foundation (FR 2015 65795), DARPA (W911NF 15 2 0065), and ONR (N00014 15 1 2704). R.U.S. is supported by a Fannie and John Hertz Foundation Graduate Fellowship and an NSF Graduate Research Fellowship (DGE 11 44155). S.P.C. is supported by a NIH MSTP training grant (NIH T32GM007367).

References

1. Turnbaugh, P.J. *et al.* (2007) The human microbiome project. *Nature* 449, 804–810
2. Esvelt, K.M. and Wang, H.H. (2013) Genome-scale engineering for systems and synthetic biology. *Mol. Syst. Biol.* 9, 641
3. Stewart, E.J. (2012) Growing unculturable bacteria. *J. Bacteriol.* 194, 4151–4160
4. Yaung, S.J. *et al.* (2014) Recent progress in engineering human-associated microbiomes. *Methods Mol. Biol.* 1151, 3–25
5. Faith, J.J. *et al.* (2011) Predicting a human gut microbiota's response to diet in gnotobiotic mice. *Science* 333, 101–104
6. Bouhnik, Y. *et al.* (2004) The capacity of nondigestible carbohydrates to stimulate fecal bifidobacteria in healthy humans: a double-blind, randomized, placebo-controlled, parallel-group, dose-response relation study. *Am. J. Clin. Nutr.* 80, 1658–1664
7. Roberfroid, M. *et al.* (2010) Prebiotic effects: metabolic and health benefits. *Br. J. Nutr.* 104 (Suppl. 2), S1–S63
8. Song, S.K. *et al.* (2014) Prebiotics as immunostimulants in aquaculture: a review. *Fish Shellfish Immunol.* 40, 40–48
9. Yaung, S.J. *et al.* (2015) Improving microbial fitness in the mammalian gut by in vivo temporal functional metagenomics. *Mol. Syst. Biol.* 11, 788
10. van Opijnen, T. and Camilli, A. (2013) Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat. Rev. Microbiol.* 11, 435–442
11. David, L.A. *et al.* (2014) Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505, 559–563
12. Wu, M. *et al.* (2015) Genetic determinants of in vivo fitness and diet responsiveness in multiple human gut Bacteroides. *Science* 350, aac5992
13. Haiser, H.J. *et al.* (2013) Predicting and manipulating cardiac drug inactivation by the human gut bacterium Eggerthella lenta. *Science* 341, 295–298
14. Wallace, B.D. *et al.* (2010) Alleviating cancer drug toxicity by inhibiting a bacterial enzyme. *Science* 330, 831–835
15. Baruch, M. *et al.* (2014) An extracellular bacterial pathogen modulates host metabolism to regulate its own sensing and proliferation. *Cell* 156, 97–108
16. Wang, Z. *et al.* (2015) Non-lethal inhibition of gut microbial trimethylamine production for the treatment of atherosclerosis. *Cell* 163, 1585–1595
17. Kohanski, M.A. *et al.* (2010) How antibiotics kill bacteria: from targets to networks. *Nat. Rev. Microbiol.* 8, 423–435
18. Robinson, C.J. and Young, V.B. (2010) Antibiotic administration alters the community structure of the gastrointestinal micobiota. *Gut Microbes* 1, 279–284
19. Gibson, M.K. *et al.* (2015) Antibiotics and the developing infant gut microbiota and resistome. *Curr. Opin. Microbiol.* 27, 51–56
20. Brogden, K.A. (2005) Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? *Nat. Rev. Microbiol.* 3, 238–250
21. Cullen, T.W. *et al.* (2015) Gut microbiota. Antimicrobial peptide resistance mediates resilience of prominent gut commensals during inflammation. *Science* 347, 170–175
22. O'Toole, P.W. and Cooney, J.C. (2008) Probiotic bacteria influence the composition and function of the intestinal microbiota. *Interdiscip. Perspect. Infect. Dis.* 2008, 175285
23. Mappley, L.J. *et al.* (2013) Oral treatment of chickens with Lactobacillus reuteri LM1 reduces Brachyspira pilosicoli-induced pathology. *J. Med. Microbiol.* 62, 287–296
24. Spinler, J.K. *et al.* (2008) Human-derived probiotic Lactobacillus reuteri demonstrate antimicrobial activities targeting diverse enteric bacterial pathogens. *Anaerobe* 14, 166–171
25. Buffie, C.G. *et al.* (2015) Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. *Nature* 517, 205–208
26. Hassan, J.O. and Curtiss, R., 3rd (1994) Development and evaluation of an experimental vaccination program using a live avirulent Salmonella typhimurium strain to protect immunized chickens against challenge with homologous and heterologous Salmonella serotypes. *Infect. Immun.* 62, 5519–5527
27. Steidler, L. *et al.* (2000) Treatment of murine colitis by Lactococcus lactis secreting interleukin-10. *Science* 289, 1352–1355
28. Chen, Z. *et al.* (2014) Incorporation of therapeutically modified bacteria into gut microbiota inhibits obesity. *J. Clin. Invest.* 124, 3391–3406
29. Thompson, J.A. *et al.* (2015) Manipulation of the quorum sensing signal AI-2 affects the antibiotic-treated gut microbiota. *Cell Rep.* 10, 1861–1871
30. Danino, T. *et al.* (2015) Programmable probiotics for detection of cancer in urine. *Sci. Transl. Med.* 7, 289ra84
31. Mimee, M. *et al.* (2015) Programming a human commensal bacterium, Bacteroides thetaiotaomicron, to sense and respond to stimuli in the murine gut microbiota. *Cell Systems* 1, 62–71
32. Kotula, J.W. *et al.* (2014) Programmable bacteria detect and record an environmental signal in the mammalian gut. *Proc. Natl. Acad. Sci. U.S.A.* 111, 4838–4843
33. van Nood, E. *et al.* (2013) Duodenal infusion of donor feces for recurrent Clostridium difficile. *N. Engl. J. Med.* 368, 407–415
34. Borody, T.J. and Khoruts, A. (2012) Fecal microbiota transplantation and emerging applications. *Nat. Rev. Gastroenterol. Hepatol.* 9, 88–96
35. Faith, J.J. *et al.* (2010) Creating and characterizing communities of human gut microbes in gnotobiotic mice. *ISME J.* 4, 1094–1098
36. Tvede, M. and Rask-Madsen, J. (1989) Bacteriotherapy for chronic relapsing Clostridium difficile diarrhoea in six patients. *Lancet* 1, 1156–1160
37. Petrof, E.O. *et al.* (2013) Stool substitute transplant therapy for the eradication of Clostridium difficile infection: 'RePOOPulating' the gut. *Microbiome* 1, 3
38. Shoaie, S. *et al.* (2015) Quantifying diet-induced metabolic changes of the human gut microbiome. *Cell Metab.* 22, 320–331
39. Seedorf, H. *et al.* (2014) Bacteria from diverse habitats colonize and compete in the mouse gut. *Cell* 159, 253–266
40. Weinbauer, M.G. (2004) Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* 28, 127–181
41. Summers, W.C. (2001) Bacteriophage therapy. *Annu. Rev. Microbiol.* 55, 437–451
42. Jones, J.B. *et al.* (2012) Considerations for using bacteriophages for plant disease control. *Bacteriophage* 2, 208–214
43. Perera, M.N. *et al.* (2015) Bacteriophage cocktail significantly reduces or eliminates Listeria monocytogenes contamination on lettuce, apples, cheese, smoked salmon and frozen foods. *Food Microbiol.* 52, 42–48
44. Silva, Y.J. *et al.* (2014) Influence of environmental variables in the efficiency of phage therapy in aquaculture. *Microb. Biotechnol.* 7, 401–413
45. Lu, T.K. and Collins, J.J. (2007) Dispersing biofilms with engineered enzymatic bacteriophage. *Proc. Natl. Acad. Sci. U.S.A.* 104, 11197–11202
46. Lu, T.K. and Collins, J.J. (2009) Engineered bacteriophage targeting gene networks as adjuvants for antibiotic therapy. *Proc. Natl. Acad. Sci. U.S.A.* 106, 4629–4634
47. Bikard, D. *et al.* (2014) Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. *Nat. Biotechnol.* 32, 1146–1150
48. Ando, H. *et al.* (2015) Engineering modular viral scaffolds for targeted bacterial population editing. *Cell Syst.* 1, 187–196
49. Modi, S.R. *et al.* (2013) Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* 499, 219–222
50. Qin, J. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65
51. Smillie, C.S. *et al.* (2011) Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480, 241–244
52. Ochman, H. *et al.* (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304
53. Kommineni, S. *et al.* (2015) Bacteriocin production augments niche competition by enterococci in the mammalian gastrointestinal tract. *Nature* 526, 719–722

54. Chen, I. and Dubnau, D. (2004) DNA uptake during bacterial transformation. *Nat. Rev. Microbiol.* 2, 241–249
55. Rodriguez-Beltran, J. *et al.* (2013) The animal food supplement sepiolite promotes a direct horizontal transfer of antibiotic resistance plasmids between bacterial species. *Antimicrob. Agents Chemother.* 57, 2651–2653
56. Wilharm, G. *et al.* (2010) A simple and rapid method of bacterial transformation. *J. Microbiol. Methods* 80, 215–216
57. Klumper, U. *et al.* (2015) Broad host range plasmids can invade an unexpectedly diverse fraction of a soil bacterial community. *ISME J.* 9, 934–945
58. Musovic, S. *et al.* (2014) Long-term manure exposure increases soil bacterial community potential for plasmid uptake. *Environ. Microbiol. Rep.* 6, 125–130
59. Klumper, U. *et al.* (2014) Novel assay to measure the plasmid mobilizing potential of mixed microbial communities. *Front. Microbiol.* 5, 730
60. Deng, L. *et al.* (2014) Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. *Nature* 513, 242–245
61. Shintani, M. *et al.* (2015) Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. *Front. Microbiol.* 6, 242
62. Frost, L.S. *et al.* (2005) Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* 3, 722–732
63. Jones, B.V. and Marchesi, J.R. (2007) Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. *Nat. Methods* 4, 55–61
64. Kushwaha, M. and Salis, H.M. (2015) A portable expression resource for engineering cross-species genetic circuits and pathways. *Nat. Commun.* 6, 7832
65. Jiang, W. *et al.* (2013) RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* 31, 233–239
66. Karberg, M. *et al.* (2001) Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. *Nat. Biotechnol.* 19, 1162–1167
67. Na, D. *et al.* (2013) Metabolic engineering of Escherichia coli using synthetic small regulatory RNAs. *Nat. Biotechnol.* 31, 170–174
68. Nielsen, A.A. and Voigt, C.A. (2014) Multi-input CRISPR/Cas genetic circuits that interface host regulatory networks. *Mol. Syst. Biol.* 10, 763
69. Farzadfard, F. and Lu, T.K. (2014) Synthetic biology. Genomically encoded analog memory with precise in vivo DNA writing in living cell populations. *Science* 346, 1256272
70. Esvelt, K.M. *et al.* (2014) Concerning RNA-guided gene drives for the alteration of wild populations. *Elife* 2014, e03401
71. Butler, C.A. and Gotschlich, E.C. (1991) High-frequency mobilization of broad-host-range plasmids into Neisseria gonorrhoeae requires methylation in the donor. *J. Bacteriol.* 173, 5793–5799
72. Oye, K.A. *et al.* (2014) Biotechnology. Regulating gene drives. *Science* 345, 626–628
73. Baltimore, D. *et al.* (2015) Biotechnology. A prudent path forward for genomic engineering and germline gene modification. *Science* 348, 36–38

Appendix 2

Gastroenterology 2015;149:883–885

BRIEF REPORTS

Proton Pump Inhibitors Alter Specific Taxa in the Human Gastrointestinal Microbiome: A Crossover Trial

Daniel E. Freedberg,[1] Nora C. Toussaint,[2] Sway P. Chen,[3] Adam J. Ratner,[4] Susan Whittier,[5] Timothy C. Wang,[1] Harris H. Wang,[5,6,§] and Julian A. Abrams[1,§]

[1]*Department of Medicine, Division of Digestive and Liver Diseases, Columbia University Medical Center;* [2]*New York Genome Center;* [3]*College of Physicians and Surgeons, Columbia University Medical Center;* [4]*Department of Pediatrics, Division of Infectious Diseases, Columbia University Medical Center;* [5]*Department of Pathology and Cell Biology; and* [6]*Department of Systems Biology, Columbia University, New York, New York*

See Covering the Cover synopsis on page 827; see editorial on page 848.

We conducted an open-label crossover trial to test whether proton pump inhibitors (PPIs) affect the gastrointestinal microbiome to facilitate *Clostridium difficile* infection (CDI). Twelve healthy volunteers each donated 2 baseline fecal samples, 4 weeks apart (at weeks 0 and 4). They then took PPIs for 4 weeks (40 mg omeprazole, twice daily) and fecal samples were collected at week 8. Six individuals took the PPIs for an additional 4 weeks (from week 8 to 12) and fecal samples were collected from all subjects at week 12. Samples were analyzed by 16S ribosomal RNA gene sequencing. We found no significant within-individual difference in microbiome diversity when we compared changes during baseline vs changes on PPIs. There were, however, significant changes during PPI use in taxa associated with CDI (increased Enterococcaceae and Streptococcaceae, decreased Clostridiales) and taxa associated with gastrointestinal bacterial overgrowth (increased Micrococcaceae and Staphylococcaceae). In a functional analysis, there were no changes in bile acids on PPIs, but there was an increase in genes involved in bacterial invasion. These alterations could provide a mechanism by which PPIs predispose to CDI. ClinicalTrials.gov ID NCT01901276.

Keywords: *Clostridium difficile* Infection; Pharmacology; Gastroesophageal Reflux Disease; Acid Suppression.

Proton pump inhibitors (PPIs) have been associated with *Clostridium difficile* infection (CDI), but the mechanism linking PPIs and CDI is unknown. Broad-spectrum antibiotics are the most important risk factor for CDI and cause loss of diversity within the gastrointestinal microbiome.[1] There are also more specific changes within the microbiome that precede CDI. Increases in Enterococcaceae and decreases within key Clostridial taxa at the time of hospital admission are associated with increased risk for subsequent development of CDI.[2]

This study tested whether PPIs given in the absence of antibiotics alter the human colonic microbiome to predispose to CDI. Twelve healthy volunteers each donated a fecal sample at week 0 and week 4 of the study (see Supplementary Methods for complete description). They subsequently all took omeprazole 40 mg twice daily for 4 weeks and donated an additional sample (week 8). The subjects were then randomized 1:1 to stop PPIs or continue them for an additional 4 weeks, after which they donated a final sample (week 12) (Supplementary Figure 1). We excluded those who used antibiotics within 1 year or tested positive for the *C difficile* toxin B gene at week 0 (Supplementary Table 1 and Supplementary Figure 2). We used 16S ribosomal RNA gene sequencing to describe the fecal microbiome. Our a priori primary outcome was fecal microbial diversity, defined as the within-individual difference in Shannon's index of diversity comparing change during the 4-week baseline period to change during the 4-week period on PPIs. To focus on taxa predisposing to CDI, we prespecified taxa of interest referencing studies of lower gastrointestinal microbiome changes preceding CDI, and studies of upper gastrointestinal microbiome changes after PPIs (Supplementary Table 2).

We found no within-individual changes in diversity after 4 weeks of PPI treatment (Figure 1*A*). Two subjects received antibiotics between week 8 and week 12 for reasons unrelated to the study; for these subjects, the samples taken after antibiotics were excluded from the final analyses. In the remaining subjects that received 8 weeks of PPIs (n = 5), there was no difference between 8 weeks of PPIs compared with baseline ($P = .79$). On principal coordinates analyses, there was no distinct clustering of samples from before vs after PPI treatment (Figure 2). In sum, overall fecal microbial composition remained stable during use of PPIs.

However, PPI treatment for 4 weeks did induce significant within-individual increases in Enterococcaceae and Streptococcaceae, taxa that have been associated with exposure to antibiotics and increased risk for CDI (Figure 1*B*).[3–7] In a hospital-based study, patients who later

[§]Authors share co-corresponding authorship.

Abbreviations used in this paper: **CDI, *Clostridium difficile* infection; PPI, proton pump inhibitor.**

Most current article

0016-5085/$36.00
http://dx.doi.org/10.1053/j.gastro.2015.06.043

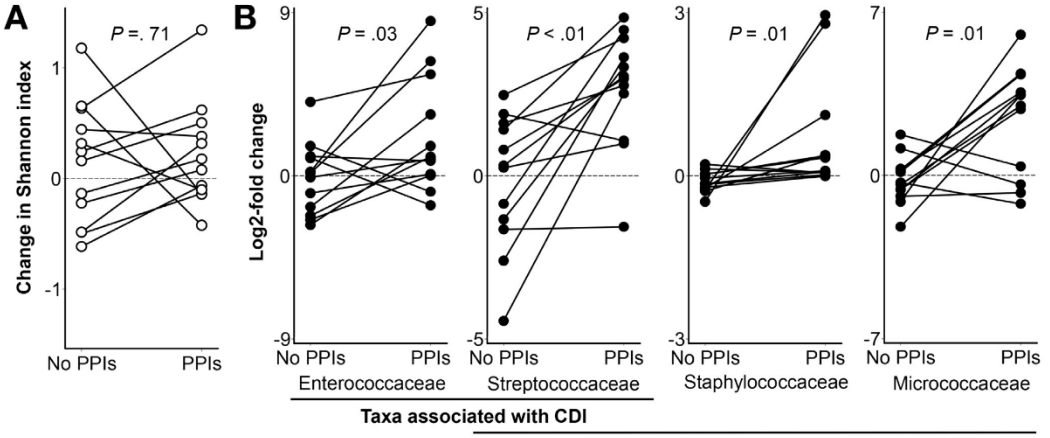


Figure 1. Changes in fecal microbiotal diversity and specific taxa throughout the study. (*A*) There was no significant change in overall diversity after 4 weeks of PPIs compared with 4 weeks at baseline. (*B*) Four weeks of PPIs induced within-individual changes in the relative abundance of prespecified taxa associated with CDI and with upper gastrointestinal bacterial overgrowth. Lines connect individuals.

developed CDI had low diversity and increased Enterococcaceae compared with control patients.[3] Enterococci are present in low abundance in human stool, but can rapidly expand after broad-spectrum antibiotics.[4] In mice, treatment with clindamycin is followed by proliferation of enterococci and CDI.[5] Dynamic modeling suggests that an increase in enterococci is a key step preceding *C difficile* colonization.[6] Streptococcaceae, which are predominantly upper gastrointestinal tract organisms, were increased >10-fold after PPIs. Gastric and small intestinal bacterial overgrowth with *Streptococcus* is an established consequence of PPIs, but the direct pH-raising effects of PPIs are attenuated by the distal duodenum.[8–10] Streptococcaceae are disrupted by broad-spectrum antibiotics and have been associated with CDI.[7] Our results are consistent with the hypothesis that PPI-induced hypochloryhydria causes increased gastric and fecal *Streptococcus*, leading to increased risk for CDI.

To identify additional changes caused by 4 weeks of PPIs, we looked across 97 bacterial families present in all samples and compared within-individual changes before

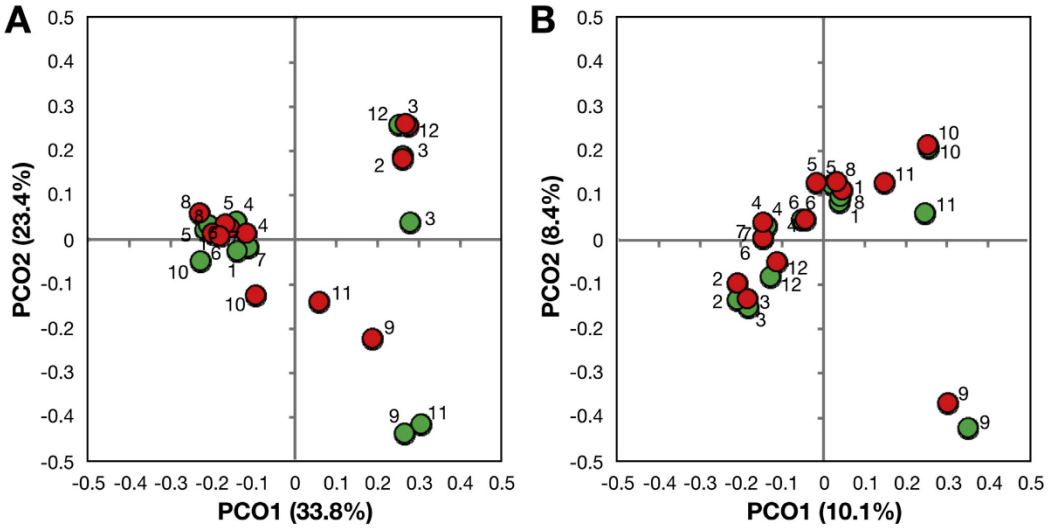


Figure 2. Principal coordinate analyses. Weighted (*A*) and unweighted (*B*) UniFrac analyses from immediately before and after 4 weeks of PPIs. *Circles* represent samples from before PPIs (*green*) or after PPIs (*red*); the corresponding subject's number is adjacent to each *circle*.

and after PPIs. We found a 44% median decrease in Clostridiaceae ($P = .03$). There were no further alterations in prespecified taxa in either arm of the study during the final 4 weeks of the study (Supplementary Figure 3). Stool polymerase chain reaction testing and culture for *C difficile* was performed on all samples. One subject had an equivocal toxin B test after 8 weeks of PPIs and also growth of *C difficile* in culture before and after PPIs.

Bacterial production of secondary bile salts can play an important role in CDI by inhibiting *C difficile* spore germination. In cirrhotics given PPIs, there were decreased levels of urinary dimethylamine, which is produced by the bacterial metabolism of bile salts; subjects also had increased fecal Streptococcaceae after PPIs, as was seen in our study.[11] Cluster XIVa Clostridia including *Clostridium scindens* actively produce secondary bile acids.[12] In our study, there was a decrease in Clostridiaceae after PPIs, but we were unable to directly map *C scindens* within our reference library. Instead, we performed real-time polymerase chain reaction for the *baiCD* gene, which encodes the rate-limiting enzyme in the production of secondary bile acids.[12] We found no change in *baiCD* gene copy number after PPIs ($P = .79$) and, to further investigate bile acids, we then used liquid chromatography and mass spectrometry to directly assess bile acid levels in our samples. There was no change after PPIs in any of 10 dominant human primary and secondary bile acids (Supplementary Figure 4).

Next, we used PICRUSt to impute the metagenome from our 16S sequencing results.[13] We found no changes in the KEGG pathways for bile acid biosynthesis after PPIs (Supplementary Figure 5*A*). We then performed an unbiased metagenomic analysis across all KEGG pathways, assessing for within-individual differences after PPIs compared with the baseline period. After 4 weeks of PPIs, there was a significant increase in the pathway corresponding to genes for *Staphylococcus aureus* infection, which includes genes for antimicrobial lectins (Supplementary Figure 5*B*). After 8 weeks of PPIs, there were significant increases in the pathways corresponding to genes for bacterial invasion of epithelial cells and for the renin-angiotensin system (Supplementary Figure 5*C*); these pathways include genes for antibacterial peptides and maintenance of epithelial integrity. Together, these results imply that PPIs do not increase risk for CDI by altering fecal levels of secondary bile acids, but rather that PPIs might be important after *C difficile* sporulation, by lowering colonization resistance.

Antibiotics cause CDI and reduce the diversity and overall size of the microbiome.[14] We did not find a reduction in fecal microbial diversity after PPIs, but loss of diversity may represent an epiphenomenon that often accompanies the key changes within specific taxa that are permissive for CDI. One prior human study found a small but statistically significant reduction in total bacterial operational taxonomic units after PPIs, but did not identify changes in Shannon diversity or within specific bacterial taxa.[15] In contrast, we did not find changes in operational taxonomic unit counts after PPIs ($P = .12$).

In conclusion, 4 weeks of high-dose PPIs did not change fecal microbial diversity beyond baseline variability, but significantly affected certain taxa including Streptococcaceae and Enterococcaceae. PPIs may increase risk for CDI by altering crucial taxa involved in colonization resistance to *C difficile*.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at http://dx.doi.org/10.1053/j.gastro.2015.06.043.

References

1. Chang JY, et al. J Infect Dis 2008;197:435–438.
2. Manges AR, et al. J Infect Dis 2010;202:1877–1884.
3. Vincent C, et al. Microbiome 2013;1:18.
4. Donskey CJ, et al. N Engl J Med 2000;343:1925–1932.
5. Lawley TD, et al. Infect Immun 2009;77:3661–3669.
6. **Stein RR, Bucci V**, et al. PLoS Comput Biol 2013;9:e1003388.
7. Perez-Cobas AE, et al. Gut 2013;62:1591–1601.
8. Rosen R, et al. JAMA Pediatr 2014;168:932–937.
9. Rosen R, et al. J Pediatr 2015;166:917–923.
10. Lo WK, et al. Clin Gastroenterol Hepatol 2013;11:483–490.
11. Bajaj JS, et al. Am J Physiol Gastrointest Liver Physiol 2014;307:G951–G957.
12. Buffie CG, et al. Nature 2015;517:205–208.
13. Langille MG, et al. Nat Biotechnol 2013;31:814–821.
14. Antonopoulos DA, et al. Infect Immun 2009;77:2367–2375.
15. Seto CT, et al. Microbiome 2014;2:42.

Author names in bold designate shared co-first authorship.

Received January 29, 2015. Accepted June 29, 2015.

Reprint requests
Address requests for reprints to: Harris H. Wang, PhD, Department of Pathology and Cell Biology, Columbia University, 701 West 168th Street, New York, New York 10032. e-mail: hw2429@cumc.columbia.edu; fax: (212) 851-5149; or Julian A. Abrams, MD, Division of Digestive and Liver Diseases, Columbia University Medical Center, 630 West 168th Street, New York, New York 10032. e-mail: ja660@cumc.columbia.edu; fax: (212) 305-5576.

Acknowledgments
The authors thank Justin Cross for assistance in measuring primary and secondary fecal bile acid levels.

Author contributions: DEF and JAA were involved in all aspects of the study. NCT was involved in acquisition, analysis, and interpretation of data, and revision of the manuscript. SPC and SW were involved in acquisition of the data and revision of the manuscript. AJR was involved in acquisition and analysis of the data and revision of the manuscript. TCW was involved in study concept, interpretation of the data, and revision of the manuscript. HHW was involved in study concept, acquisition, analysis, and interpretation of data, and revision of the manuscript.

Conflicts of interest
The authors disclose no conflicts.

Funding
Daniel E. Freedberg was funded by the National Center for Advancing Translational Sciences (NIH KL2 TR000081). Harris H. Wang was funded by the National Institutes of Health Director's Early Independence Award (1DP5OD009172-02) and the National Science Foundation CAREER Award (MCB-1453219). Sway P. Chen was funded by the Columbia University's Medical Scientist Training Program. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or other funding organizations.

Supplementary Materials and Methods

Subjects

Participants considered for the study were healthy volunteers 18 years or older lacking the following exclusion criteria: use of systemic antibiotics within 1 year, use of proton pump inhibitors (PPIs) or histamine-2 receptor antagonists within 1 year, new medications within 1 month, chronic gastrointestinal mucosal disease, abnormal bowel frequency (minimum of 3 bowel movements per week, maximum of 3 per day), use of medications with potential interactions with PPIs, pregnancy, and travel planned outside of the United States during the study period. Subjects were instructed to avoid probiotics and major dietary shifts during the study period.

Study Design

The study had a randomized, crossover design. At the initial study visit, information was gathered regarding medical history, diet, use of medications, and anthropomorphic data, including height, weight, and waist and hip circumference. Stool was tested for the *C difficile* toxin B gene by polymerase chain reaction (PCR) at all study visits and those testing positive at baseline were excluded. All subjects were observed for 4 weeks and subsequently block randomized to 4 vs 8 weeks of omeprazole 40 mg twice daily (Supplementary Figure 1). Study visits were scheduled 0, 4, 8, and 12 weeks after enrollment. At each study visit, subjects provided a stool sample, answered questions regarding interval history, and completed a food frequency questionnaire derived from the National Health Interview Survey and validated for assessment of fat and fiber intake over the preceding 4 weeks.[1,2] Omeprazole compliance was assessed via pill counts and by performing mass spectrometry for omeprazole and its metabolites, which were detected at week 8 in all subjects and at week 12 in all subjects randomized to 8 weeks of PPIs. All authors had access to the study data and approved the final manuscript. This study was approved by the Institutional Review Board of Columbia University and registered at ClinicalTrials.gov (ID NCT01901276).

Sample Preparation

Stool specimens were captured in standard collection containers and brought promptly (<1 hour) or temporarily frozen. At study visits, specimens were mixed and aliquoted in a sterile manner, and frozen at −80°C. At the end of the study, batched DNA extraction was performed using the PowerFecal DNA Isolation Kit (Mo Bio, Carlsbad, CA). Polymerase chain reaction was performed targeting the V4 hypervariable region of the 16S ribosomal RNA gene with primers derived from the human microbiome project.[3] Samples were pooled and purified with the QIAquick PCR kit (Qiagen, Valencia, CA) and library quantification performed using a KAPA Library Quantification Kit (Kapa Biosystems, Wilmington, MA).

16S Ribosomal RNA Gene Sequencing

Sequencing of the 16S ribosomal RNA gene V4 region was performed using the Illumina MiSeq 300PE platform (Illumina, San Diego, CA). Singleton reads were discarded and read pairs were merged, trimmed, and filtered for quality using *mothur*.[4] Subsequently singleton contigs were discarded to yield 24,187,741 total sequence contigs or an average 484,000 reads per sample. Greengenes[5] was used as a reference database with additional sequences of interest retrieved from the National Center for Biotechnology, including sequences corresponding to *C difficile*. Clustering of taxonomic units was made at 97% sequence similarity using USEARCH[6] and taxonomic assignments were made using *mothur*.[5] FastTree2[7] (version 2.1.7) was used to generate a phylogenetic tree of the contigs.[7,8] Using *mothur* and the phylogenetic tree, we calculated α- and β-diversity indices including weighted and unweighted UniFrac distances.

Outcomes

The a priori primary outcome was a change in fecal microbial diversity, defined as the within-individual difference in Shannon's index of diversity comparing the change during the 4-week period before PPIs to the change during the 4-week period on PPIs. To assess for a change in fecal microbial composition, we additionally compared Bray-Curtis indices corresponding to the 4-week period before PPIs and the 4-week period on PPIs. To focus on taxa predisposing to the development of CDI, we prespecified taxa of interest referencing 2 types of prior studies: studies of change within the lower gastrointestinal microbiome preceding CDI,[9,10] and studies of change within the upper gastrointestinal microbiome after PPIs[11,12] (Supplementary Table 1). A post-hoc analysis was performed among all family-level assignments to examine for significant changes, comparing the period after with the period before PPI exposure.

Clostridium difficile *Polymerase Chain Reaction and Culture*

Fresh aliquots of stool samples were tested via commercial PCR for the *C difficile* toxin B gene according to the manufacturer's protocol (BD GeneOhm, Sparks, MD). For anaerobic culture, aliquots of frozen stool specimens were thawed and inoculated under anaerobic conditions onto Brucella 5% sheep's blood agar containing hemin and vitamin K1 and agar with cefoxitin, cycloserine, and fructose (Remel, Lenexa, KS). Presumptive Clostridial isolates were identified using the RapID ANA II System (Remel).[13]

Reverse Transcription Polymerase Chain Reaction for baiCD *Gene*

To quantify the relative abundance of the *baiCD* gene in fecal samples, we measured *baiCD* and 16S ribosomal RNA by quantitative PCR using the CFX96 Real-Time System and C1000 Touch Thermal Cycler (Bio-Rad). *BaiCD* was amplified using primers *baiCD*_f (5′-GGWTTCAGCC CRCAGATCTTCTTTG-3′) and *baiCD*_r (5′-TGTGWGYGCATG GAATTCWACTGC-3′). These primers were designed to

amplify a 160-bp region of the aligned sequences of bile acid-inducible operons from *Clostridium hiranonis* TO-931 and *C scindens* VPI 12708. 16S quantitative PCR was performed using previously described primers, which amplify a 172-bp amplicon.[14] PCR reactions were performed using SsoAdvanced SYBR Green Supermix (Bio-Rad) with a template of 10 ng genomic DNA and 0.5 uM of each primer in a total volume of 20 uL. Genomic DNA from *Escherichia coli* K-12 MG1655 and *C scindens* (DSM 5676) were used as templates for negative and positive controls, respectively. The PCR program consisted of an initial step of 98°C for 2 minutes, followed by 40 cycles of 98°C denaturation for 5 seconds and 62°C annealing/extension for 30 seconds, and 65°C−95°C melting for 3 minutes.

Fecal Bile Acids

Methods used to quantify fecal bile acids were similar to those described in Buffie et al.[15] In brief, samples were homogenized, corrected to a final concentration of 0.5 mg/10 μL, and then sonicated. After adding an internal standard (d4-chenodeoxycholic acid), we performed 2 methanol extractions and transferred filtered samples to a mass spectrometry vial containing a reduced volume glass insert. Bile acids were then separated using an Agilent 1290 HPLC and Cogent C18 column (2.1 mm × 150 mm, 2.2 μm; MicroSolv, Eatontown, NJ). Mobile phase A was water + 0.05% formic acid; mobile phase B was acetone + 0.05% formic acid, run at a flow rate of 0.35 mL/min. The injection volume was 5 uL and the liquid chromatography gradient was 25%−70% B in 25 minutes. Bile acids were detected using an Agilent 6550 Q-TOF mass spectrometer with JetStream source, operating in negative ionization mode. Acquisition was from m/z 50 to 1100 at 1 spectra/s, gas temperature 275°C, drying gas 11 L/min, nebulizer 30 psig, sheath gas 325°C, sheath gas flow 10 L/min, VCap 4000 V, fragmentor 365 V, and Oct1 RF 750 V. Bile acids were confirmed by alignment to authentic standards (Steraloids Inc, Newport, RI or Sigma Aldrich, St Louis, MO). Final abundances and normalization for 10 dominant human bile acids (Supplementary Table 3) was performed using the ProFinder and Mass Profiler Professional software (Agilent Technologies, Santa Clara, CA).

Imputed Metagenomic Analysis

To impute the metagenome, we used PICRUSt, a freely available predictive strategy that uses 16S sequences as input data.[16] In brief, QIIME[17] (v1.8.0) was used to perform closed-reference OTU picking at 97% similarity of the preprocessed 16S reads within our dataset against the Greengenes[5] reference OTU database (May 2013, 99% OTU clustering). The resulting OTUs were used to predict functional composition with PICRUSt and the predicted functions were collapsed into KEGG pathways. Differential pathway abundance analysis was performed using phyloseq[18] combined with DESeq2.[19]

Statistical Analysis

Within-individual differences in Shannon indices were tested using paired t tests (normally distributed data). Changes in relative abundance of specific taxa were tested using paired t tests for normally distributed data or Wilcoxon signed rank tests. Data was analyzed using STATA 12 (StataCorp, College Station, TX). To determine our population size, we referred to studies that assessed the effects of antibiotics on diversity and powered our study to detect a difference of half that magnitude.[20,21] Significance tests for prespecified taxa of interest were not adjusted for multiple hypothesis testing (Supplementary Table 2); for other tests, the Benjamini-Hochberg method was used to adjust for multiple hypothesis testing. All significance tests were performed 2-sided and at the $\alpha = .05$ level.

Primer Sequences and Polymerase Chain Reaction Protocol

Primer	Sequence
515_f1	GAGTTCAGACGTGTGCTCTTCCGATCT GTGCCAGCMGCCGCGGTAA
806_r1_N3	CCTACACGACGCTCTTCCGATCT NNN GGACTACHVGGGTWTCTAAT
806_r1_N4	CCTACACGACGCTCTTCCGATCT NNNN GGACTACHVGGGTWTCTAAT
806_r1_N5	CCTACACGACGCTCTTCCGATCT NNNNN GGACTACHVGGGTWTCTAAT
806_r1_N6	CCTACACGACGCTCTTCCGATCT NNNNNN GGACTACHVGGGTWTCTAAT
P5_r2	AATGATACGGCGACCACCGAGATCT ACAC TCTTTC CCTACACGACGCTCTTCCGATCT
P7_bc01	CAAGCAGAAGACGGCATACGAGAT CAGGTT GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc02	CAAGCAGAAGACGGCATACGAGAT TCACAA GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc03	CAAGCAGAAGACGGCATACGAGAT ACATCA GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc04	CAAGCAGAAGACGGCATACGAGAT AGCGCA GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc05	CAAGCAGAAGACGGCATACGAGAT CATCAA GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc06	CAAGCAGAAGACGGCATACGAGAT GCTATT GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc07	CAAGCAGAAGACGGCATACGAGAT TAGATC GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc08	CAAGCAGAAGACGGCATACGAGAT CATGAC GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc09	CAAGCAGAAGACGGCATACGAGAT GAATCG GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc10	CAAGCAGAAGACGGCATACGAGAT TCTTCT GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc11	CAAGCAGAAGACGGCATACGAGAT ATTCCG GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc12	CAAGCAGAAGACGGCATACGAGAT GGAATT GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc13	CAAGCAGAAGACGGCATACGAGAT ACGGTG GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc14	CAAGCAGAAGACGGCATACGAGAT CTCAGC GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc15	CAAGCAGAAGACGGCATACGAGAT TCCGGT GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc16	CAAGCAGAAGACGGCATACGAGAT TGCAGT GTGACTG GAGTTCAGACGTGTGCTCTTC

Continued

Primer	Sequence
P7_bc17	CAAGCAGAAGACGGCATACGAGAT TTCATA GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc18	CAAGCAGAAGACGGCATACGAGAT ATACAC GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc19	CAAGCAGAAGACGGCATACGAGAT CGTTAT GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc20	CAAGCAGAAGACGGCATACGAGAT CTCGGA GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc21	CAAGCAGAAGACGGCATACGAGAT TGTGTG GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc22	CAAGCAGAAGACGGCATACGAGAT ACCGCG GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc23	CAAGCAGAAGACGGCATACGAGAT GATCGG GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc24	CAAGCAGAAGACGGCATACGAGAT TCACGG GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc25	CAAGCAGAAGACGGCATACGAGAT ATTACT GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc26	CAAGCAGAAGACGGCATACGAGAT CTTAGA GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc27	CAAGCAGAAGACGGCATACGAGAT GCAGCT GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc28	CAAGCAGAAGACGGCATACGAGAT TCCTCC GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc29	CAAGCAGAAGACGGCATACGAGAT GAACTA GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc30	CAAGCAGAAGACGGCATACGAGAT ACAACC GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc31	CAAGCAGAAGACGGCATACGAGAT GGTAAC GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc32	CAAGCAGAAGACGGCATACGAGAT GTGGTC GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc33	CAAGCAGAAGACGGCATACGAGAT CCGCGT GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc34	CAAGCAGAAGACGGCATACGAGAT CTGACA GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc35	CAAGCAGAAGACGGCATACGAGAT CCGAAT GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc36	CAAGCAGAAGACGGCATACGAGAT AGCCGC GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc37	CAAGCAGAAGACGGCATACGAGAT TAGCGC GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc38	CAAGCAGAAGACGGCATACGAGAT TGACCT GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc39	CAAGCAGAAGACGGCATACGAGAT CTTATC GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc40	CAAGCAGAAGACGGCATACGAGAT GTAGCC GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc41	CAAGCAGAAGACGGCATACGAGAT CCATAG GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc42	CAAGCAGAAGACGGCATACGAGAT GAGGCA GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc43	CAAGCAGAAGACGGCATACGAGAT AATTGA GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc44	CAAGCAGAAGACGGCATACGAGAT ACTCAC GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc45	CAAGCAGAAGACGGCATACGAGAT AAGTTG GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc46	CAAGCAGAAGACGGCATACGAGAT TACGAT GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc47	CAAGCAGAAGACGGCATACGAGAT CACCAC GTGACTG GAGTTCAGACGTGTGCTCTTC
P7_bc48	CAAGCAGAAGACGGCATACGAGAT GCATTC GTGACTG GAGTTCAGACGTGTGCTCTTC

Polymerase Chain Reaction Protocol

Low-cycle PCR was performed using KAPA SYBR FAST qPCR (Kapa Biosystems, Wilmington, MA) per the manufacturer's instruction. Nested PCR cycles were performed using the following protocol on a Bio-Rad CFX96 Touch real-time PCR instrument (Bio-Rad, Hercules, CA).

Polymerase Chain Reaction 1

Starting template: 20 ng DNA.
Primer sets: 515_f1 and 806_r1_N3-6.
Step 1 at 95°C for 3 minutes;
Step 2 at 95°C for 10 seconds;
Step 3 at 60°C for 30 seconds;
Repeat steps 2 × 3–30 cycles;
Step 5 at 68°C for 5 minutes;
Step 6 at 4°C on hold.

Polymerase Chain Reaction 2

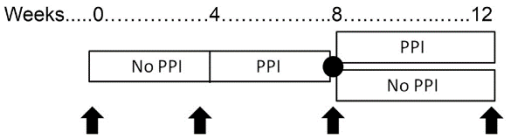
Starting template: 1 μL of PCR product.
Primer sets: P7_bc01-48 and P5_r2 with P5/P7 adaptors.
Step 1 at 95°C for 3 minutes;
Step 2 at 95°C for 10 seconds;
Step 3 at 60°C for 30 seconds;
Repeat steps 2 through 3 × 10 cycles;
Step 5 at 68°C for 5 minutes;
Step 6 at 4°C on hold.

References

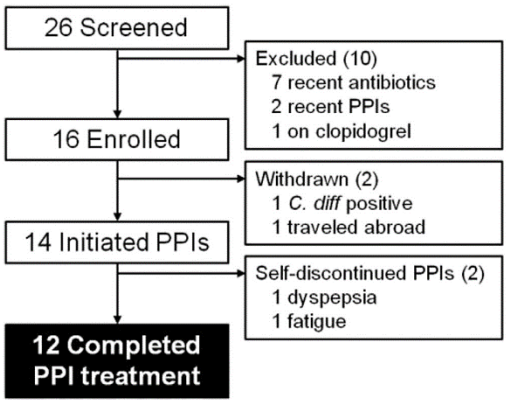
1. Thompson FE, et al. Public Health Nutr 2004; 7:1097–1105.
2. Thompson FE, et al. J Am Diet Assoc 2005;105:352–363. quiz 487.
3. Kuczynski J, et al. Nat Rev Genet 2012;13:47–58.
4. Schloss PD, et al. Appl Environ Microbiol 2009; 75:7537–7541.
5. McDonald D, et al. ISME J 2012;6:610–618.
6. Edgar RC. Bioinformatics 2010;26:2460–2461.
7. Price MN, et al. PLoS One 2010;5:e9490.
8. Lozupone C, et al. Appl Environ Microbiol 2005; 71:8228–8235.
9. Vincent C, et al. Microbiome 2013;1:18.
10. **Stein RR**, **Bucci V**, et al. PLoS Comput Biol 2013; 9:e1003388.

11. Pyleris E, et al. Dig Dis Sci 2012;57:1321–1329.
12. Rosen R, et al. JAMA Pediatr 2014;168:932–937.
13. Celig DM, et al. J Clin Microbiol 1991;29:457–462.
14. Salzman NH, et al. Nat Immunol 2010;11:76–83.
15. Buffie CG, et al. Nature 2015;517(7533):205–208.
16. Langille MG, et al. Nat Biotechnol 2013;31:814–821.
17. Caporaso JG, et al. Nat Methods 2010;7:335–336.
18. McMurdie PJ, et al. PLoS One 2013;8:e61217.
19. Love MI, et al. Genome Biol 2014;15:550.
20. De La Cochetiere MF, et al. J Clin Microbiol 2005; 43:5588–5592.
21. Jakobsson HE, et al. PLoS One 2010;5:e9836.

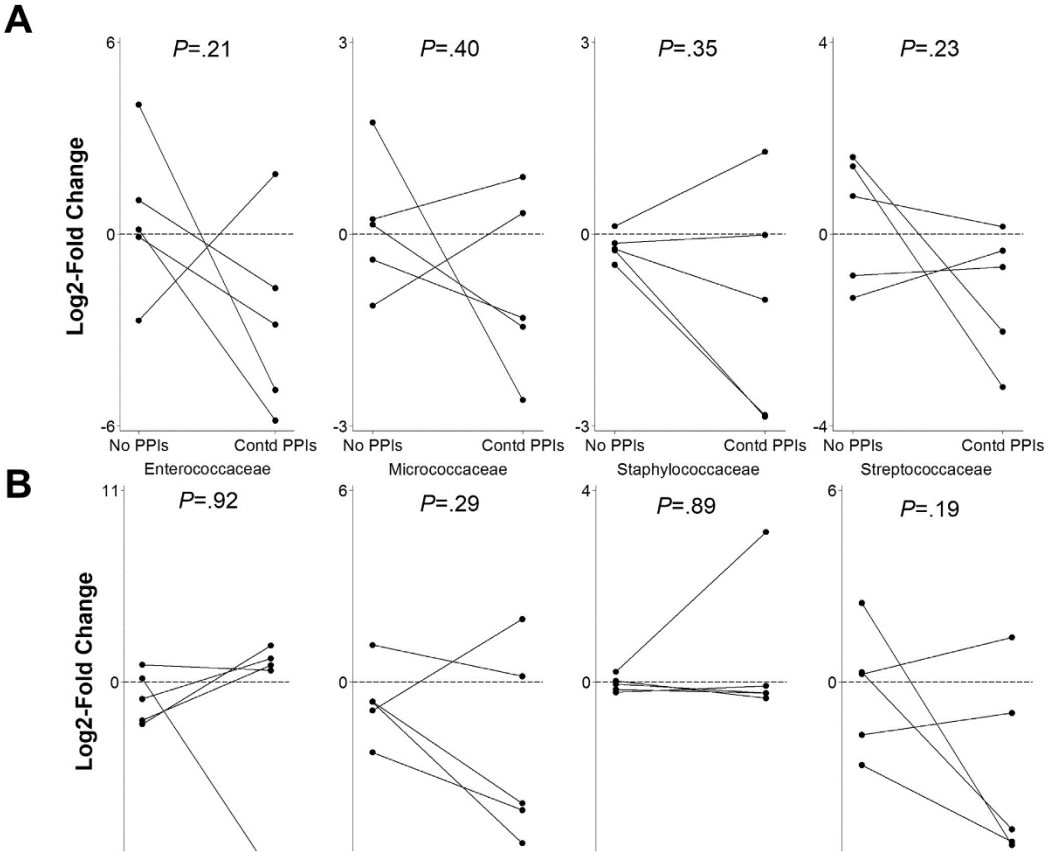
Author names in bold designate shared co-first authorship.



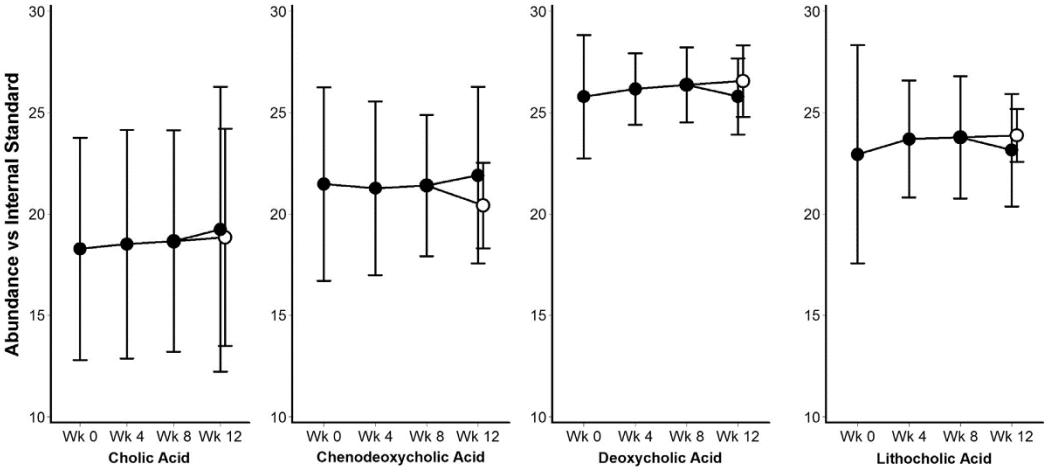
Supplementary Figure 1. Design of the study. Time in weeks is given at top. *Black circle,* randomization. *Black arrowheads*, study visits.



Supplementary Figure 2. Flow of patients into the study.



Supplementary Figure 3. Week 12 data. Log2-fold change in relative abundance, comparing change during the baseline period to change during the period corresponding to weeks 8 to 12 of the study. This is shown for subjects who continued PPIs during weeks 8 to 12 (*A*) and for subjects who completed 4 weeks of PPIs and then were off PPIs during Weeks 8 to 12 (*B*).

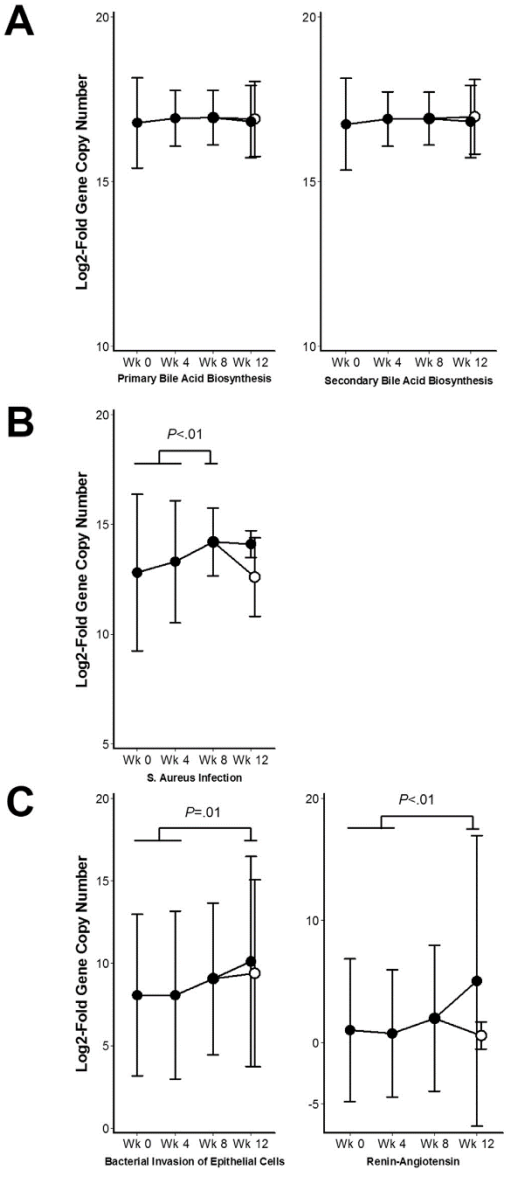


Supplementary Figure 4. Bile acid levels. Combined data for the 2 main primary bile acids (cholic acid and chenodeoxycholic acid) and for the 2 main secondary bile acids (deoxycholic acid and lithocholic acid). Comparing the baseline period (weeks 0 to 4) to the period on PPIs (weeks 4 to 8), there were no differences in levels of these bile acids or any of the other 6 dominant human primary and secondary bile acids. *Vertical lines* are 95% confidence intervals; subjects randomized to discontinue PPIs after the week 8 visit are shown as *open circles*.

Supplementary Table 1. Baseline Characteristics of Subjects

Variable	Median (IQR)
Age, *y*	39.5 (29.0–51.5)
Sex	
Male	3
Female	9
Anthropomorphics	
BMI	28.4 (21.3–34.5)
Waist: hip ratio	0.89 (0.83–0.97)
Diet	
Calories as fat, %	34.7 (32.1–38.2)
Daily grams fiber	14.5 (12.7–18.9)

BMI, body mass index; IQR, interquartile range.



Supplementary Figure 5. Imputed metagenomic changes. PICRUSt was used to perform an unbiased estimatation of within-individual metagenomic changes. There were no changes in the genes for the pathways corresponding to primary bile acid biosynthesis or to secondary bile acid biosynthesis (*A*). After adjusting for multiple hypothesis testing using the Benjamini-Hochberg method, there was a significant increase in genes within the KEGG pathways for *Staphylococcus aureus* infection after 4 weeks of PPIs (*B*) and for bacterial invasion of epithelial cells and renin-angiotensin after 8 weeks of PPIs (*C*). For all pathways, *vertical lines* show 95% confidence intervals. Subjects randomized to discontinue PPIs after the week 8 visit are shown as *open circles*.

Supplementary Table 2. Taxa of Interest

Taxon	Taxonomic level	Primary reference, first author
Associated with *Clostridium difficile* infection		
Bacteroidetes	Phylum	Vincent[9]
Enterococcaceae	Family	Vincent[9]
Tissiarellaceae	Family	Vincent[9]
Streptococcaceae	Family	Rosen[12]
Blautia	Genus	Stein[10]
Coprobacillus	Genus	Stein[10]
Akkermansia	Species	Stein[10]
Associated with bacterial overgrowth		
Enterobacteriaceae	Family	Pyleris[11]
Micrococcaceae	Family	Rosen[12]
Staphylococcaceae	Family	Rosen[12]
Streptococcaceae	Family	Rosen[12]
Veillonellaceae	Family	Rosen[12]
Serratia	Genus	Pyleris[11]

Supplementary Table 3. Bile Acids Tested

Name	Molecular formula	Molecular weight, *g/mol*	Retention time, *min*
Primary bile acids			
Cholic acid	$C_{24}H_{40}O_5$	408.57	15.6
Chenodeoxycholic acid	$C_{24}H_{40}O_4$	392.57	23.5
Glycocholic acid	$C_{26}NH_{43}O_6$	465.63	15.0
Glycochenodeoxycholic acid	$C_{26}NH_{43}O_5$	449.62	19.4
Taurocholic acid	$C_{26}H_{45}NO_7S$	515.70	12.4
Taurochenodeoxycholic acid	$C_{26}H_{45}NO_6S$	499.70	16.5
Secondary bile acids			
Lithocholic acid	$C_{24}H_{40}O_3$	376.57	28.7
Deoxycholic acid	$C_{24}H_{40}O_4$	392.57	24.0
Ursodeoxycholic acid	$C_{24}H_{40}O_4$	392.57	19.6
Hyodeoxycholic acid	$C_{24}H_{40}O_4$	392.57	20.0

Appendix 3: Technical notes on MAGIC experiments on mice

We performed MAGIC to alter the gut microbiota of mice over several independent experiments, which took place over 3 years and in 2 different mouse facilities, using multiple cohorts of mice. Here, we describe some technical considerations that are likely to affect the efficacy of *in situ* conjugation and gene expression in transconjugants, which may explain some of the variations seen between mice and between experiments.

1. Facilities: Experiments in Figures 3.2 and 3.3 and Suppl. Figures S3.9, S3.10, S3.11, S3.14, and S3.15 were performed in 2016-2017 in the Hammer Health Sciences Center (HHSC) mouse facility, which did not have an air shower at the entrance and where mice were kept in cages with individual food racks and water bottles. Experiments in Suppl. Figure S3.8 were performed in 2018 in the Irving Center for Cancer Research (ICRC) mouse facility, which had an air shower at every entrance and where mice were provided water from a central source (but food racks were placed in individual cages). The highest proportions of transconjugants and most diverse transconjugants were observed in experiments performed in HHSC, although we did not control for the starting microbiomes of the mouse cohorts involved. In conjugations of vector library pGT-L3 in each facility (Figure 3.2, Suppl. Figures S3.8 and S3.10) the proportion of transconjugants was similar, but the diversity of transconjugants was lower in the ICRC experiments.
2. Mice: For all experiments, we used conventionally raised C57BL/6 female mice from Taconic Biosciences or Charles River Laboratories. The mice were purchased at 6 weeks of age. Gavage experiments were performed when the mice were 8-10 weeks old. Mice were randomly assigned to cages of 3-4 each and co-housed for at least 1 week before the gavage to equilibrate their microbiomes, and then randomly allocated into treatment groups prior to gavage. We thus assumed that the starting microbiome of each mouse used in a single experiment was relatively similar, across treatment groups. We observed

similar phylogeny and diversity of transconjugants between mice from the two vendors, despite quite different starting microbiomes. (Taconic: Figure 3.2, Suppl. Figure S3.9. Charles River: Suppl. Figures S3.8, S3.10.) We did not test other strains, ages, or sexes of mice, which may have different microbiomes, nor did we control for the starting microbiomes of the mice across different experiments.

3. Donor strain:
    a. For all *in situ* conjugation experiments, we used EcGT2 as the donor strain; it remains unclear how other donor strains would perform in conjugation efficiency and ability to survive the gut.
    b. In all experiments, the donor strain was grown to saturation from a glycerol stock, and then rediluted and grown to saturation again at 37C. Donor strains in different growth phases (exponential phase, for example) were not tested and may have different survival and conjugation rates in the gut environment.
    c. To wash donor cells for gavage, cells were pelleted and resuspended 2x in PBS. 10^9 washed donor cells, quantified by OD600 and in equal ratios for each vector in the vector library, were diluted into 300 uL of PBS for gavage. We kept 3-4 mice in a cage for each treatment group and combined n+1 300 uL doses of donor strain in a syringe. The same syringe was used to gavage slightly more than 300 uL of donor strain per mouse, and the bulb-tipped needle was not switched between mice in each treatment group.
    d. Although EcGT2 is auxotrophic for DAP, we did not add DAP to the PBS during gavage, as DAP is produced by other gut bacteria and is present in non-limiting concentrations in a conventional gut microbiome. We do not expect that adding DAP to the gavage media would affect EcGT2 growth *in situ*, but altering the gavage media in other ways may promote donor survival and HGT.

4. Timing of experiments: Mice have circadian rhythms and the gut bacteria are likely to fluctuate based on the time of day; it is possible that the bacteria are more amenable to HGT at some times of day rather than others. We started all of the experiments with time 0 feces collection and gavage at 2-3 AM, followed by the 6 hour time point of fecal collection at 9 AM, as we were constrained by the logistics of booking the FACS machine in the core facility.
5. Preparation of fecal samples:
    a. We tested two conditions for storage of fecal samples between the time of collection and downstream analysis by FACS or antibiotic plating, which was usually several hours later the same day: (1) Fecal samples were collected in microcentrifuge tubes and put on ice in the 4C room, aerobically, and (2) Fecal samples were collected in microcentrifuge tubes and put into an anaerobic chamber at ambient temperature (25-30C). Fecal samples kept anaerobically at ambient temperature contained far more viable transconjugants upon antibiotic selection than the fecal samples on ice, likely because the more oxygen and temperature sensitive species died under condition (1).
    b. To improve efficiency of bacterial extraction from feces, we found that collecting 3 sets of fecal samples (or separating the fecal collection into 3 tubes) was useful. One set was immediately put into anaerobic conditions for downstream extraction in anaerobic conditions, followed by antibiotic selection. One set was frozen at -80C for future use/analysis. One set was extracted with aerobic PBS aerobically for FACS analysis, which was much faster than doing the extraction in an anaerobic chamber, and provided the advantage of exposing the bacteria to oxygen to allow GFP to fold.