(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
21 May 2015 (21.05.2015)

WIPO | PCT

(10) International Publication Number
WO 2015/074054 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/US2014/066173

(22) International Filing Date:
18 November 2014 (18.11.2014)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/905,530 18 November 2013 (18.11.2013) US

(71) Applicant: **THE TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK** [US/US]; West 116th Street and Broadway, New York, NY 10027 (US).

(72) Inventor: **WANG, Harris**; c/o Columbia University, West 116th Street and Broadway, New York, NY 10027 (US).

(74) Agents: **DAVITZ, Michael, A.** et al.; Ascenda Law Group PC, 84 West Santa Clara Street, Suite 550, San Jose, CA 95113 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

— *with sequence listing part of description (Rule 5.2(a))*

(54) Title: IMPROVING MICROBIAL FITNESS IN THE MAMMALIAN GUT

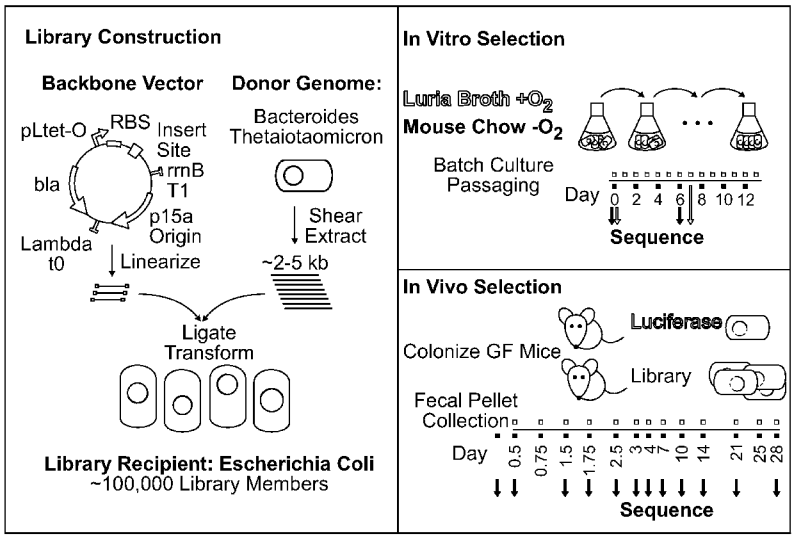

FIG. 1

(57) Abstract: The present invention provides for a powerful and systematic method of identifying genes that enhance bacterial fitness in the mammalian gastrointestinal (GI) tract. The identified genes can then be used to generate recombinant bacteria which may improve the health of a mammal, such as a human, or may be used to combat a gastrointestinal disorder.

WO 2015/074054 A1

IMPROVING MICROBIAL FITNESS IN THE MAMMALIAN GUT

Cross Reference to Related Application

This application claims priority to U.S. Provisional Application No. 61/905,530 filed on November 18, 2013, which is incorporated herein by reference in its entirety.

Government License Rights

This invention was made with government support under the NIH Director's Early Independence Award (Grant No. 1DP5OD009172-01) awarded by the National Institutes of Health. The government may have certain rights in the invention.

Field of the Invention

The present invention relates to methods and compositions for improving the health of a mammal by modulating the population of bacteria in the mammalian gastrointestinal (GI) tract. In particular, the present invention relates to promoting the growth of beneficial gut bacteria.

Background of the Invention

The human body is colonized by trillions of microbes, with the highest concentration along the gastrointestinal (GI) tract. There is an intimate relationship between gut microbes and their mammalian host during development and maintenance of homeostasis. The GI tract is a hostile environment for poorly adapted microbes. Nonetheless, diverse groups of microbes have evolved to prosper in the GI tract, in the setting of intense interspecies competition, physical and chemical stressors, and the host immune system (Ley *et al*, 2006; Dethlefsen *et al*, 2007). These microorganisms also support the normal homeostatic functions of the host by helping to extract nutrients, stimulate the immune system, and provide protection against colonization by pathogens (Ley *et al*, 2006; Gill *et al*, 2006; Bäckhed *et al*, 2005; Stappenbeck *et al*, 2002; Hooper, 2004). Only the most gut-adapted microbes are able to flourish, and thus the genes and pathways that they carry are enriched over time. It is now increasingly clear that the repertoires of genes encompassed in the microbiota, and not just simply the composition of the microbes themselves, drive the maintenance of microbiome health.

Next-generation sequencing has enabled systematic studies of the mammalian microbiota, and great strides have been made in characterizing the structure of bacterial communities and their genetic potential *in vivo*. For instance, the Human Microbiome Project (HMP) (Turnbaugh *et al*, 2007; Peterson *et al*, 2009; Huttenhower *et al*, 2012) and MetaHIT (Qin *et al*, 2010) have generated maps of bacterial species abundances throughout the human body, reference genomes, and catalogs of more than 100 million microbial genes assembled from shotgun sequencing of *in vivo* communities. Although these studies have generated vast amounts of descriptive data, the functions of most bacterial genes in these collections remain poorly characterized or wholly unknown. While much is being revealed about the microbial colonizers at the general population level, little is known about specific factors that promote gut colonization in each microbe at the genetic and molecular level.

Traditional methods to characterize the functions of microbial genes require the isolation, cultivation, and introduction of foreign DNA into a recipient organism. However, an estimated 60-80% of mammalian-associated microbiota species remain uncultivated (Walker *et al*, 2014). Even after successful culture and introduction of genetic material into a microorganism, the DNA must integrate into the microbial genome or be maintained episomally. This requires known compatible replication and restriction-modification systems, which may not be feasible for many microbes. If these barriers can be overcome, standard low-throughput methods for functional characterization of genes may be employed. Newer approaches, such as transposon mutagenesis coupled with next-generation sequencing, are also applicable. In this latter approach, random locations on the genome are disrupted with a transposon containing a selectable marker; the resulting library is subjected to selection conditions and deep-sequenced to determine enriched and depleted mutants (van Opijnen *et al*, 2009). A limitation of this approach is that essential genes or those that are important to cell fitness are difficult to study, since inactivation of these genes by transposon mutagenesis would be lethal to the organism under study. An additional constraint is that transposon mutagenesis may disrupt the expression of bystander genes that are near the relevant locus, thus causing confounding phenotypic effects.

Nonetheless, the presence of certain genes is known to impart significant advantage to microbes in the gut – the classic example being antibiotic resistance genes. Since it is now clear that horizontal gene acquisition is a key mechanism driving gene flow across the microbiome,

we need to identify the class and function of genes that directly alter microbial fitness in the gut – well beyond just reservoirs for antibiotic resistance.

Summary

The present invention provides for a method of identifying genes that enhance bacterial fitness in the gastrointestinal (GI) tract. The method may comprise the steps of: (a) constructing a genomic or metagenomic library comprising fragments of the genome of at least one donor bacterium; (b) introducing the library to recipient bacteria; (c) introducing the recipient bacteria carrying the library into the GI tract of a mammal; (d) taking stool samples of the mammal at different time points, e.g., T_1 through T_n (n is an integer); (e) isolating DNA (e.g., plasmids and/or genomic DNAs) from the stool samples of step (d); and (f) sequencing the DNA of step (e). The mammal may be a mouse, e.g., a mouse that is germ-free before introduction of the recipient bacteria. The mouse may be healthy, or have inflamed GI tract.

In certain embodiments, a gene is identified to enhance bacterial fitness in the mammalian GI tract, when the gene has an abundance at a time point of step (d) that is at least 10 fold, at least 20 fold, at least 50 fold, at least 100 fold, at least 200 fold, at least 500 fold, or at least 1000 fold, of its abundance in the library. The time points (e.g., T_1 through T_n) may range from about 0 day to about 30 days after introduction of the recipient bacteria carrying the library into the GI tract of a mammal (e.g., step (c) of the method).

The DNA may be fragmented by sonication or digestion by at least one restriction enzyme. The fragments of the donor bacterial genome can be under the control of a constitutive or inducible promoter in the recipient bacteria. The DNA may be sequenced by deep sequencing, or Sanger sequencing.

The donor bacterium may belong to genera *Bacteroides* or *Clostridium*. For example, the donor bacterium can be *Bacteroides thetaiotaomicron* or *Clostridium butyricum*.

The donor bacterium can also be from a natural gut microbiota. The natural gut microbiota may be from, e.g., the gut of a healthy mammal, or the gut of a mammal with inflamed GI tract. For example, the mammal with inflamed GI tract may have an inflammatory bowel disease (IBD, such as ulcerative colitis or Crohn's disease) or irritable bowel syndrome (IBS). The healthy mammal, or the mammal with inflamed GI tract, may be a human subject.

The recipient bacteria may belong to phyla Bacteroidetes, Firmicutes, Proteobacteria, or Actinobacteria. The recipient bacteria may belong to genera *Bacteroides*, *Clostridium, Escherichia, Bacillus, Lactobacillus*, or *Bifidobacterium*. Non-limiting examples of recipient

bacteria include *Escherichia coli (E. coli)*, *Bacillus subtilis (B. subtilis)*, *Lactobacillus plantarum (L. plantarum), Lactobacillus reuteri (L. reuteri), Lactobacillus rhamnosus (L. rhamnosus), Bifidobacterium longum (B. longum), Bacteroides thetaiotaomicron*, *Clostridium butyricum, Bacteroides fragilis, Bacteroides melaninogenicus, Bacteroides oralis, Bacteroides amylophilus, Clostridium butyricum, Clostridium perfringens, Clostridium tetani,* and *Clostridium septicum.*

Also encompassed by the present invention is a probiotic composition comprising recombinant bacteria comprising a gene encoding a protein such as a glycoside hydrolase, a galactokinase or a glucose/galactose transporter. The gene may be heterologous. The gene may be endogenous which has been engineered to overexpress the encoded protein. The gene may be under the control of a constitutive or inducible promoter. The protein can be wild-type or a mutant. The gene may be wild-type or mutated. For example, the gene may be truncated at the 5' end by from about 1 base pair (bp) to about 50 bp from the start codon. In one embodiment, the protein is *Bacteroides thetaiotaomicron* glycoside hydrolase whose encoding gene may be wild-type or may be truncated at the 5' end by about 4 bp from the start codon.

The recombinant bacteria may be capable of metabolizing sucrose, galactose, or both sucrose and galactose.

In the probiotic composition, the gene may be integrated into the bacterial chromosome or may be episomal.

The probiotic composition may be a food composition, a beverage composition, a pharmaceutical composition, or a feedstuff composition. For example, the probiotic composition can be a dairy product.

Brief Description of the Drawings

Figure 1. Experimental design.
(Left panel) Map of the library backbone vector. The vector was linearized and ligated to sheared fragments of donor genome to generate the heterologous insert library.
(Right panels) Passaging of the E. coli library in two liquid media conditions (top) and inoculation of the library or a control luciferase plasmid into germ-free (GF) mice (bottom). Small boxes across the time line denote sample collection time points. Arrows indicate deep-sequenced samples.

Figure 2. Input library characterization.
(A) Even coverage of the *Bacteroides thetaiotaomicron* genome. The solid and dashed wavy lines represent per-base coverage values for the chromosome and native *B. thetaiotaomicron* p5482 plasmid, respectively. The histogram (top right) shows the distribution of genes by their coverage (normalized to gene length).
(B) Insert size distribution of library.
(C) Plasmid retention calculated by comparing number of colonies on LB vs. LB+carbenicillin plates from *in vitro* passaging experiments in aerobic LB or anaerobic mouse chow (MC) filtrate.

Figure 3. *In vivo* selection experiments.
(A) Plasmid retention calculated by comparing number of colonies on LB vs. LB+carbenicillin plates from mouse fecal samples. n = 5 mice; error bars = standard deviation
(B) Effective positional coverage across the entire Bt genome for each mouse, begins with essentially even coverage of the Bt genome of ~6 Mb, but drops rapidly over the experimental time-course, representative of selection at specific loci.
(C) Representative longitudinal selection of Bt genes in a single mouse (Mouse 2). For each mouse and time point, ~10^9 sequenced bases were mapped to the *B. thetaiotaomicron* genome. Of those mapped bases, the percentage mapping to each gene is plotted. Genes with less than 0.2% are grouped together (dark gray bars).

Figure 4. BT_1759 glycoside hydrolase.

(A) Selection kinetics by Fragments Per Kilobase Mapped (FPKM) fold change and normalized effective coverage of genes BT_1757, BT_1758, and BT_1759. m1-m5 = Mouse 1- Mouse 5.
(B) Mapped reads to each base in the region with deep sequencing and Sanger sequencing of isolated clones (below, length of inserts are to scale to the gene map). Read values are the mean across five mice and were normalized to 1 billion mapped bases per run to compare across time points. Sanger sequencing was performed on ten clones per mouse at Day 7 and eight clones per mouse at Day 28. nt = nucleotides.
(C) Functional characterization in minimal media with sucrose as the sole carbon source. Three sets of strains were studied: 1) starting *E. coli* strains transformed with the CDS of each gene cloned into the backbone vector, 2) *E. coli* clones directly isolated from stool samples, and 3) starting *E. coli* strains re-transformed with individual plasmids isolated from stool samples. All clones isolated at Day 28 carried the BT_1759 locus. Lines represent the mean.

Figure 5. BT_0370 galactokinase and BT_371 glucose/galactose transporter.
(A) Selection kinetics by Fragments Per Kilobase Mapped (FPKM) fold change and normalized effective coverage of genes BT_0368, BT_0369, BT_0370, BT_0371, and BT_0372.
(B) Mapped reads to each base in the region with deep sequencing and Sanger sequencing of isolated clones (below). Read values are the mean across five mice and were normalized to 1 billion mapped bases per run to compare across time points. Isolation of individual clones allowed for insert size profiling at Day 7. Screened isolates from Day 28 did not reveal any galactokinase inserts.
(C) Functional characterization in minimal media with galactose as the sole carbon source. Three sets of strains were studied: 1) starting *E. coli* strains transformed with the CDS of each gene cloned into the backbone vector, 2) *E. coli* clones directly isolated from stool samples, and 3) starting *E. coli* strains re-transformed with individual plasmids isolated from stool samples. All clones isolated at Day 28 carried the BT_1759 locus. Lines represent the mean.
(D) Genotyping of the background *E. coli* genome at the galK locus. 20 clones were screened for each time point in the mice inoculated with the library, while 30 clones were screened at Day 28 in the lux control mice.

Figure 6. Distribution of mapped bases to each Bt gene by mouse. For each mouse and time point, ~10^9 sequenced bases were mapped to the *B. thetaiotaomicron* genome. Of those mapped bases, the percentage mapping to each gene is shown. Genes with < 0.2% are grouped together ("Other genes (each <0.2%) and intergenic regions"). Specific genes >= 0.2% that appeared in one mouse but not the others are indicated in smaller font and colored differently.

Figure 7. COG (cluster of orthologous group) functional categories of bases mapped to the entire Bt genome averaged across the five mice.

Figure 8. Growth characterization of clones with genomic single nucleotide variants (SNVs). Growth curves over 42 hours at 37°C in M9 with 0.2% galactose and carbenicillin of (A) mouse-isolated clones from Day 28 and (B) cloned BT_0369, BT_0370, BT_0371, BT_0372, and BT_0370-BT_0372. The mean of four replicates is plotted in filled circles; error bars represent the standard deviation. **(C)** Endpoint optical density after 96 hours of growth. Two mouse-isolated strains with the BT_0370 insert were compared to isogenic strains transformed with those plasmids (4.0 or 4.3 kb insert). The strain with the galR SNV is shown in circles filled with dots. Lines represent the mean.

Detailed Description of the Invention

The present invention provides for a powerful and systematic method of identifying genes that enhance bacterial fitness in the mammalian gastrointestinal (GI) tract. The identified genes can be used to generate recombinant bacteria to be included in a probiotic composition. The composition may improve the health of a mammal, such as a human, or may be used to combat a gastrointestinal disorder.

The method of identifying a bacterial fitness gene may contain the following steps: (a) constructing a genomic or metagenomic library comprising fragments of the genome of at least one donor bacterium; (b) introducing the library to recipient bacteria; (c) introducing the recipient bacteria carrying the library into the GI tract of a mammal (to initiate the *in vivo* selection process); (d) taking stool samples of the mammal at different time points, e.g., T_1 through T_n; (e) isolating DNA (e.g., plasmids and/or genomic DNAs) from the stool samples of step (d); and (f) sequencing the DNA of step (e).

When the abundance of a gene or a DNA segment during the *in vivo* selection (e.g., at a time point, or multiple time points, of step (d)) is greater than its abundance before the *in vivo* selection (e.g., its abundance in the original library), the gene may be identified to be able to enhance bacterial fitness in the mammalian GI tract. For example, the abundance of a gene during the *in vivo* selection (e.g., at a time point, or multiple time points such as 2, 3, 4, 5, 6, 7, 8, 9, 10 or more time points, of step (d), such as on day 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30; after 1, 2, 3 or 4 weeks; after 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 or 12 months; or after 1, 2, 3, 4, 5 years or longer, counting from the initiation of the *in vivo* selection) may be at least about 1.2 fold, at least about 1.4 fold, at least about 1.5 fold, at least about 1.8 fold, at least about 2 fold, at least about 3 fold, at least about 4 fold, at least about 5 fold, at least about 6 fold, at least about 7 fold, at least about 8 fold, at least about 10 fold, at least about 15 fold, at least about 20 fold, at least about 25 fold, at least about 30 fold, at least about 35 fold, at least about 40 fold, at least about 50 fold, at least about 60 fold, at least about 70 fold, at least about 80 fold, at least about 90 fold, at least about 100 fold, at least about 200 fold, at least about 250 fold, at least about 300 fold, at least about 400 fold, at least about 500 fold, at least about 600 fold, at least about 700 fold, at least about 800 fold, at least about 900 fold, at least about 1,000 fold, at least about 1,100 fold, at least about 1,200 fold, at least about

1,300 fold, at least about 1,400 fold, at least about 1,500 fold, at least about 1,600 fold, at least about 1,700 fold, at least about 1,800 fold, at least about 1,900 fold, or at least about 2,000 fold, of its abundance before the *in vivo* selection (e.g., its abundance in the library).

In one embodiment, a genomic library composed of 3-5 kb uniformly fragmented genomic DNA from donor *Bacteroides thetaiotaomicron* is first cloned and expressed in *E. coli*. *Bacteroides thetaiotaomicron* is an abundant Gram-negative commensal microbe with known important metabolic capabilities in the gut. A cell library containing ~10^5 unique clones will provide on average >50x coverage of the donor genome. This Gram-negative genomic library is then introduced by gavage to healthy, germ-free mice for passaging over a period of a month. Gnotobiotic murine experiments are conducted. *In vivo* selection of beneficial genes that improve gut colonization leads to a measureable enrichment of their relative abundance across the microbial population. The selected library from fecal pellets collected daily is extracted and deep-sequenced to identify and quantify genes that are enriched during the *in vivo* selection.

Also encompassed by the present invention are recombinant bacteria (e.g., in a probiotic composition) engineered with genes that can enhance bacterial fitness in the mammalian GI tract. The genes may encode a glycoside hydrolase, a galactokinase, a glucose/galactose transporter, or any other proteins involved in carbohydrate (e.g., sucrose, galactose, etc.) metabolism and/or transport. The gene may be under the control of a constitutive or an inducible promoter.

Donor bacterium

The donor bacterium may be any suitable bacterium. The donor bacterium may be a bacterium that exists naturally in the GI tract. In one embodiment, the donor bacterium is a commensal bacterium. In another embodiment, the donor bacterium is a probiotic bacterium. The donor bacteria may be a mixture of bacteria, from, e.g., a metagenomic source. For example, the donor bacterium may be from a natural gut microbiota of the gut of a healthy mammal, or from the gut of an unhealthy mammal (e.g., with a gastrointestinal disorder). The unhealthy mammal may have a condition influenced by the GI tract microbiota. For example, the unhealthy mammal may have inflamed GI tract, such as an inflammatory bowel disease (IBD), including ulcerative colitis or Crohn's disease. The unhealthy mammal may have collagenous colitis, lymphocytic colitis, diversion colitis, Behcet's disease, indeterminate colitis, irritable bowel syndrome (IBS, or spastic colon), mucous colitis, microscopic colitis, antibiotic-associated

colitis, constipation, diverticulosis, polyposis coli or colonic polyps.

The donor bacterium may be Gram-positive or Gram-negative.

The donor bacterium may be from any of the following phyla: Bacteroidetes, Firmicutes, Actinobacteria, Proteobacteria, etc.

The donor bacterium may be from any of the following genera: *Bacteroides, Clostridium, Bifidobacterium, Lactobacillales (lactic acid bacteria or LAB), Lactobacillus, Lactococcus, Enterococcus, Streptococcus, Klebsiella, Escherichia, Enterobacter, Peptostreptococcus, Peptococcus, Bacillus, Propionibacteria, Ruminococcus, Gemmiger, Desulfomonas, Salmonella, etc.*

The donor bacterium may be *Bacteroides thetaiotaomicron*, *Clostridium butyricum, Bacteroides fragilis, Bacteroides melaninogenicus, Bacteroides oralis, Enterococcus faecalis, Escherichia coli, Bifidobacterium bifidum, Staphylococcus aureus, Clostridium perfringens, Proteus mirabilis, Clostridium tetani, Clostridium septicum, Pseudomonas aeruginosa, Salmonella enteritidis, Bifidobacterium longum, Bifidobacterium lactis, Bifidobacterium animalis, Bifidobacterium breve, Bifidobacterium infantis, Lactobacillus acidophilus, Lactobacillus casei, Lactobacillus salivarius, Lactococcus lactis, Lactobacillus reuteri, Lactobacillus rhamnosus, Lactobacillus paracasei, Lactobacillus johnsonii, Lactobacillus plantarum, Lactobacillus salivarius,* and *Enterococcus faecium*. U.S. Patent No. 8,591,880.

Non limiting examples of the donor bacterium also include: *Bacillus coagulans, B. lentus, Bacillus licheniformis, B. mesentericus, B. pumilus, Bacillus subtilis, B. natto, Bacteroides amylophilus, Bac. capillosus, Bac. ruminocola, Bac. suis, Bifidobacterium adolescentis, B. animalis, B. breve, B. pseudolongum, B. thermophilum, Enterococcus cremoris, E. diacetylactis, E. intermedius, E. lactis, E. muntdi, E. thermophilus, Kluyveromyces fragilis, L. alimentarius, L. amylovorus, L. crispatus, L. brevis, L. curvatus, L. cellobiosus, Lactobacillus delbrueckii, L. farciminis, L. fermentum, L. gasseri, L. helveticus, L. sakei, L. salivarius, Leuconostoc mesenteroides, Pediococcus damnosus, Pediococcus acidilactici, P. pentosaceus, Propionibacterium freudenreichii, Prop. shermanii, Staphylococcus carnosus, Staph. xylosus, Streptococcus infantarius, Bacteroides uniformis, Streptococcus sanguinis, Streptococcus mutans, Salmonella enterica, Dorea formicigenerans, Strep. salivarius, Streptococcus thermophiles, Strep. Lactis, and E. mundtii.*

DNA library

The DNA library may be a genomic or metagenomic library. A genomic library is a collection of the genomic DNA from a single organism. A metagenomic library is a collection of the genomic DNAs of a mixture of organisms, such as a mixture of microbes (e.g., a mammalian GI tract microbiota).

Procedures for DNA library preparation can be found in, e.g., Molecular Cloning: A Laboratory Manual, 4th Ed., Cold Spring Harbor Laboratory Press, 2012, and Current Protocols in Molecular Biology, John Wiley & Sons, Inc. Those skilled in the art would be able to prepare the above libraries based on these documents.

DNA may be isolated from bacteria by any method well known in the art. For example, DNA extraction may include two or more of the following steps: cell lysis, addition of a detergent or surfactant, addition of protease, addition of RNase, alcohol precipitation (e.g., ethanol precipitation, or isopropanol precipitation), salt precipitation, organic extraction (e.g., phenol-chloroform extraction), solid phase extraction, silica gel membrane extraction, CsCl gradient purification. Various commercial kits (e.g., kits of Qiagen, Valencia, CA) can be used to extract DNA.

Genomic or metagenomic DNA is fragmented prior to library construction. DNA may be fragmented by methods including, but not limited to, sonication, needle shearing, nebulization, acoustic shearing, point-sink shearing and passage through a pressure cell. DNA may also be fragmented by digestion with one or more restriction enzymes.

Following isolation, the DNA fragments may or may not be separated by gel electrophoresis prior to insertion into vectors. The length of the DNA fragments to be used for library construction may range from about 50 base pairs to about 10 kb, from about 50 base pairs to about 2.5 kb, from about 200 base pairs to about 1 kb, from about 0.5 kb to about 10 kb, from about 10 kb to about 50 kb, from about 30 kb to about 40 kb, from about 50 kb to about 100 kb, from about 100 kb to about 200 kb, from about 1 kb to about 8 kb, from about 2 kb to about 6 kb, from about 2 kb to about 5 kb, from about 2 kb to about 4 kb, from about 2 kb to about 3 kb, from about 3 kb to about 5 kb, or longer.

The DNA fragments are then inserted into vectors using, e.g., DNA ligase. Each vector may contain a different insert of DNA. In some embodiments, fragmented DNA is end-repaired

before being ligated to a vector. Fragmented DNAs may be ligated to adapters before being inserted into vectors.

The choice of cloning vector and strategy largely reflects the desired library structure (i.e., insert size and number of clones) and target activities sought. Non-limiting examples of vectors include plasmids, phage lambda, cosmids, fosmids, bacteriophage P1, P1 artificial chromosomes (PACs), and bacterial artificial chromosomes (BACs). Exemplary vectors include GMV1c, pCC1FOS, pWE15, pFOS1, pIndigoBAC536, pWEB, pSMART, pUC18, pBR322 and its derivatives, Lambda ZAP, pHOS2, pUC and its derivatives, pBluescript and its derivatives, pTOPO-XL, and pCF430.

The DNA fragment may be under the control of a constitutive, inducible, and/or tissue-specific promoter, or promoters useful under the appropriate conditions to direct expression of the DNA fragment or a gene.

As used herein, the terms "under the control", "under transcriptional control", "operatively positioned", and "operatively linked" mean that a promoter is in a correct functional location and/or orientation in relation to a nucleic acid sequence, a DNA fragment, or a gene, to control transcriptional initiation and/or expression of that sequence, DNA fragment or gene.

A constitutive promoter is an unregulated promoter that allows for continual transcription of the gene under the promoter's control. Non-limiting examples of constitutive promoters include constitutive E. coli σ^{70} promoters, constitutive E. coli σ^{s} promoters, constitutive E. coli σ^{32} promoters, constitutive E. coli σ^{54} promoters, constitutive E. coli σ^{70} promoters, constitutive B. subtilis σ^{A} promoters, constitutive B. subtilis σ^{B} promoters, T7 promoters, and SP6 promoters. A list of constitutive bacterial promoters may be found in the database of Registry of Standard Biological Parts. They are active in all circumstances in the cell. In one embodiment, the constitutive promoter is pL.

The DNA fragment or a gene may be under the control of an inducible promoter. The transcriptional activity of these promoters is induced by either chemical or physical factors. Chemically-regulated inducible promoters may include promoters whose transcriptional activity is regulated by the presence or absence of oxygen, a metabolite, alcohol, tetracycline, steroids, metal and other compounds. Physically-regulated inducible promoters, including promoters whose transcriptional activity is regulated by the presence or absence of heat, low or high

temperatures, acid, base, or light. In one embodiment, the inducible promoter is pH-sensitive (pH inducible).

The inducer for the inducible promoter may be located in the biological tissue or environmental medium to which the composition is administered or targeted, or is to be administered or targeted. For example, the inducer for the inducible promoter may be located in the mammalian GI tract.

The pH level of a particular biological tissue can affect the inducibility of the pH inducible promoter. See, for example, Boron, et al., Medical Physiology: A Cellular and Molecular Approach. Elsevier/Saunders. (2004), ISBN 1-4160-2328-3, which is incorporated herein by reference.

Examples of acid inducible promoters include, but are not limited to, P170, P1, P3, baiA1, baiA3, lipF promoter, F_1F_0-ATPase promoter, gadC, gad D, glutamate decarboxylase promoter, etc. See, for example, Cotter and Hill, Microbiol. and Mol. Biol. Rev. vol. 67, no. 3, pp. 429-453 (2003); Hagenbeek, et al., Plant Phys., vol. 123, pp. 1553-1560 (2000); Madsen, et al., Abstract, Mol. Microbiol. vol. 56, no. 3, pp. 735-746 (2005); U.S. Pat. No. 6,242,194; Richter, et al., Abstract, Gene, vol. 395, no. 1-2, pp. 22-28 (2007), Mallonee, et al., J. Bacteriol., vol. 172, no. 12, pp. 7011-7019 (1990); each of which is incorporated herein by reference. U.S. Patent No. 8,852,916.

Some non-limiting examples of promoters induced by a change in temperature include P2, P7, and PhS. See, for example, Taylor, et al, Cell, Abstract, vol. 38, no. 2, pp. 371-381 (1984); U.S. Pat. No. 6,852,511, Wang, et al., Biochem. and Biophys. Res. Commun. Abstract, vol. 358, no. 4, pp. 1148-1153 (2007), U.S. Pat. No. 7,462,708, each of which is incorporated herein by reference.

In certain embodiments, the acid inducible promoter is inducible at a pH of about 0.0, about 0.5, about 1.0, about 1.5, about 2.0, about 2.5, about 3.0, about 3.5, about 4.0, about 4.5, about 5.0, about 5.5, about 6.0, about 6.5, about 6.6, about 6.7, about 6.8, about 6.9, or any value therebetween or less.

In some embodiments, the base inducible promoter is inducible at a pH of about 7.1, about 7.5, about 8.0, about 8.5, about 9.0, about 9.5, about 10.0, about 10.5, about 11.0, about 11.5, about 12.0, about 12.5, about 13.0, about 13.5, about 14.0, or any value therebetween or greater.

Examples of inducers that can induce the activity of the inducible promoters also include, but are not limited to, radiation, temperature change, alcohol, antibiotic, steroid, metal, salicylic acid, ethylene, benzothiadiazole, or other compound. In an embodiment, the at least one inducer includes at least one of arabinose, lactose, maltose, sucrose, glucose, xylose, galactose, rhamnose, fructose, melibiose, starch, inunlin, lipopolysaccharide, arsenic, cadmium, chromium, temperature, light, antibiotic, oxygen level, xylan, nisin, L-arabinose, allolactose, D-glucose, D-xylose, D-galactose, ampicillin, tetracycline, penicillin, pristinamycin, retinoic acid, or interferon. Other examples of inducers include, but are not limited to, at least a portion of one of an organic or inorganic small molecule, clathrate or caged compound, protocell, coacervate, microsphere, Janus particle, proteinoid, laminate, helical rod, liposome, macroscopic tube, niosome, sphingosome, vesicular tube, vesicle, unilamellar vesicle, multilamellar vesicle, multivesicular vesicle, lipid layer, lipid bilayer, micelle, organelle, nucleic acid, peptide, polypeptide, protein, glycopeptide, glycolipid, lipoprotein, lipopolysaccharide, sphingolipid, glycosphingolipid, glycoprotein, peptidoglycan, lipid, carbohydrate, metalloprotein, proteoglycan, chromosome, nucleus, acid, buffer, protic solvent, aprotic solvent, nitric oxide, vitamin, mineral, nitrous oxide, nitric oxide synthase, amino acid, micelle, polymer, copolymer, monomer, prepolymer, cell receptor, adhesion molecule, cytokine, chemokine, immunoglobulin, antibody, antigen, extracellular matrix, cell ligand, zwitterionic material, cationic material, oligonucleotide, nanotube, piloxymer, transfersome, gas, element, contaminant, radioactive particle, radiation, hormone, virus, quantum dot, temperature change, thermal energy, or contrast agent. See, for example, Theys, et al., Abstract, Curr. Gene Ther. vol. 3, no. 3 pp. 207-221 (2003), which is incorporated herein by reference.

Recipient bacterium

The recombinant DNAs (e.g., vectors with DNA inserts) are then introduced into recipient bacteria. The recipient bacteria are transformed by any known method, including, but not limited to, electroporation, heat shock, biolistic transformation, and sonic transformation. When the vector is a viral vector, the DNA is introduced into recipient bacteria through transduction.

The recipient bacterium may be any suitable bacterium. In one embodiment, the recipient bacterium has low *in vivo* fitness or may be non-adapted *in vivo*. In another

embodiment, the recipient bacterium has lower *in vivo* fitness than the donor bacterium, allowing for strong selection signals for clones harboring functional donor genes. In yet another embodiment, the recipient bacterium is a commensal bacterium. In still another embodiment, the recipient bacterium is a probiotic bacterium.

The recipient bacterium may be Gram-positive or Gram-negative.

The recipient bacterium may be from any of the following phyla: Bacteroidetes, Firmicutes, Actinobacteria, Proteobacteria, etc.

The recipient bacterium may be from any of the following genera: *Bacteroides, Clostridium, Bifidobacterium, Lactobacillales (lactic acid bacteria or LAB), Lactobacillus, Lactococcus, Enterococcus, Streptococcus, Klebsiella, Escherichia, Enterobacter, Peptostreptococcus, Peptococcus, Bacillus, Propionibacteria, Ruminococcus, Gemmiger, Desulfomonas, Salmonella, etc.*

The recipient bacterium may be *Bacteroides thetaiotaomicron*, *Clostridium butyricum, Bacteroides fragilis, Bacteroides melaninogenicus, Bacteroides oralis, Enterococcus faecalis, Escherichia coli, Bifidobacterium bifidum, Staphylococcus aureus, Clostridium perfringens, Proteus mirabilis, Clostridium tetani, Clostridium septicum, Pseudomonas aeruginosa, Salmonella enteritidis, Bifidobacterium longum, Bifidobacterium lactis, Bifidobacterium animalis, Bifidobacterium breve, Bifidobacterium infantis, Lactobacillus acidophilus, Lactobacillus casei, Lactobacillus salivarius, Lactococcus lactis, Lactobacillus reuteri, Lactobacillus rhamnosus, Lactobacillus paracasei, Lactobacillus johnsonii, Lactobacillus plantarum, Lactobacillus salivarius,* and *Enterococcus faecium*. U.S. Patent No. 8,591,880.

In certain embodiments, the recipient bacteria are the strains *E. coli* K-12, *E. coli* MG1655, *E. coli* HS, or *E. coli* Nissle 1917.

Non limiting examples of the recipient bacterium also include: *Bacillus coagulans, B. lentus, Bacillus licheniformis, B. mesentericus, B. pumilus, Bacillus subtilis, B. natto, Bacteroides amylophilus, Bac. capillosus, Bac. ruminocola, Bac. suis, Bifidobacterium adolescentis, B. animalis, B. breve, B. pseudolongum, B. thermophilum, Enterococcus cremoris, E. diacetylactis, E. intermedius, E. lactis, E. muntdi, E. thermophilus, Kluyveromyces fragilis, L. alimentarius, L. amylovorus, L. crispatus, L. brevis, L. curvatus, L. cellobiosus, Lactobacillus delbrueckii, L. farciminis, L. fermentum, L. gasseri, L. helveticus, L. sakei, L. salivarius, Leuconostoc mesenteroides, Pediococcus damnosus, Pediococcus acidilactici, P. pentosaceus,*

Propionibacterium freudenreichii, Prop. shermanii, Staphylococcus carnosus, Staph. xylosus, Streptococcus infantarius, Bacteroides uniformis, Streptococcus sanguinis, Streptococcus mutans, Salmonella enterica, Dorea formicigenerans, Strep. salivarius, Streptococcus thermophiles, Strep. Lactis, and E. mundtii.

Temporal analysis

The recipient bacteria carrying the genomic library are then introduced to the GI tract of one or more mammals. The mammal may be gnotobiotic. The mammal may be germ-free. The mammal may have an already established microbiota.

The recipient bacteria carrying the genomic or metagenomic library may be enterally administered to a mammal. For example, the recipient bacteria can be introduced by gavage to a mouse.

Stool samples of the mammal are taken at different time points T_1 through T_n (n is an integer and indicates the number of time points for stool sampling). For example, stool samples can be taken at time points T_1 through T_n ranging from about 0 day to about 5 years, from about 0 day to about 3 years, from about 0 day to about 1 year, from about 0 day to about 6 months, from about 0 day to about 3 months, from about 0 day to about 30 days, from about 0 day to about 20 days, from about 0 day to about 15 days, from about 0 day to about 10 days, from about 0 day to about 5 days, or from about 0 day to about 3 days, after the recipient bacteria are introduced to the GI tract.

Stool samples can be taken at various numbers (i.e., the value of n) of time points, including, but not limited to, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more.

The time-series approach can allow discovery of the shifts in population dynamics of clones harboring different gene fragments.

The DNA from the stool samples is extracted and then studied by sequencing.

Sequencing

DNA may be amplified via polymerase chain reaction (PCR) before being sequenced.

The DNA may be sequenced using vector-based primers; or a specific gene is sought by using specific primers. PCR and sequencing techniques are well known in the art; reagents and

equipment are readily available commercially.

Non-limiting examples of sequencing methods include Sanger sequencing or chain termination sequencing, Maxam-Gilbert sequencing, capillary array DNA sequencing, thermal cycle sequencing (Sears et al., Biotechniques, 13:626-633 (1992)), solid-phase sequencing (Zimmerman et al., Methods Mol. Cell Biol., 3:39-42 (1992)), sequencing with mass spectrometry such as matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF/MS; Fu et al., Nat. Biotechnol., 16:381-384 (1998)), and sequencing by hybridization (Chee et al., Science, 274:610-614 (1996); Drmanac et al., Science, 260:1649-1652 (1993); Drmanac et al., Nat. Biotechnol., 16:54-58 (1998)), NGS (next-generation sequencing) (Chen et al., Genome Res. 18:1143-1149 (2008); Srivatsan et al. PloS Genet. 4:e1000139 (2008)), Polony sequencing (Porreca et al., Curr. Protoc. Mol. Biol. Chp. 7; 7.8 (2006), ion semiconductor sequencing (Elliott et al., J. Biomol Tech. 1:24-30 (2010), DNA nanoball sequencing (Kaji et al., Chem Soc Rev 39:948-56 (2010), single-molecule real-time sequencing (Flusberg et al., Nat. Methods 6:461-5 (2010), sequencing by synthesis (e.g., Illumina/Solexa sequencing), sequencing by ligation, sequencing by hybridization, nanopore DNA sequencing (Wanunu, Phys Life Rev 9:125-58 (2012), massively Parallel Signature Sequencing (MPSS); pyro sequencing, SOLiD sequencing (McKeman et al. 2009 Genome Res 19:1527-1541; Shearer et al. 2010 Proc Natl Acad Sci USA 107:21104-21109); shortgun sequencing; Heliscope single molecule sequencing; single molecule real time (SMRT) sequencing. U.S. Patent Publication No. 20140329705.

High-throughput sequencing technologies include, but are not limited to, Illumina/Solex sequencing technology (Bentley et al. 2008 Nature 456:53-59), Roche/454 (Margulies et al. 2005 Nature 437:376-380), Pacbio (Flusberg et al. 2010 Nature methods 7:461-465; Korlach et al. 2010 Methods in enzymology 472:431-455; Schadt et al. 2010 Nature reviews. Genetics 11:647-657; Schadt et al. 2010 Human molecular genetics 19:R227-240; Eid et al. 2009 Science 323:133-138; Imelfort and Edwards, 2009 Briefings in bioinformatics 10:609-618), Ion Torrent (Rothberg et al. 2011 Nature 475:348-352)) and more. For example, Polony technology utilizes a single step to generate billions of "distinct clones" for sequencing. As another example, ion-sensitive field-effect transistor (ISFET) sequencing technology provides a non-optically based sequencing technique. U.S. Patent Publication No. 20140329712.

Gene Analysis

DNA from stool samples collected at different time points during *in vivo* selection is extracted and sequenced to identify genes that are enriched during the *in vivo* selection.

Sequencing of the DNA after *in vivo* selection will result in a collection of individual sequences corresponding to the selected DNA fragments. As used herein, the term "read" refers to the sequence of a DNA fragment obtained after sequencing. In certain embodiments, the reads are paired-end reads, where the DNA fragment is sequenced from both ends of the molecule.

Sequencing reads may be first subjected to quality control to identify overrepresented sequences and low-quality ends. The start and/or end of a read may or may not be trimmed. Sequences mapping to the recipient bacterium may be removed and excluded from further analysis.

Sequencing reads (e.g., the quality-corrected reads) are mapped onto the reference genome of the donor bacterium (bacteria) using any alignment algorithms known in the art. Non-limiting examples of such mapping algorithms include Bowtie; Bowtie2 (Langmead et al. 2009; Langmead et al., Fast gapped-read alignment with Bowtie 2. Nature methods 9(4), 357-9 (2012); Burrows-Wheeler Aligner (BWA, see, Li et al: Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, 26(5), 589-95 (2010)); SOAP2 (Li et al., SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, 25(15), 1966-7 (2009)); GATK; SMRA; PINDEL; SNAP (Zaharia et al., Faster and More Accurate Sequence Alignment with SNAP, arXiv:1111.5572 (2011)]; TMAP1-4; SMALT; and Masai (Siragusa et al., Fast and sensitive read mapping with approximate seeds and multiple backtracking. CoRR abs/1208.4238 (2012)). A recent overview of the alignment algorithms can be found in Li et al., A survey of sequence alignment algorithms for next-generation sequencing. Briefings in Bioinformatics 2010, 11(5), 473-483. U.S. Patent Publication Nos. 20140214334, 20140108323 and 20140315726.

Mathematical algorithms that can be used for alignment also include, the algorithm of Myers and Miller (1988) CABIOS 4:11-17; the local alignment algorithm of Smith et al. (1981) Adv. Appl. Math. 2:482; the global alignment algorithm of Needleman and Wunsch (1970) J. Mol. Biol. 48:443-453; the search-for-local alignment method of Pearson and Lipman (1988) Proc. Natl. Acad. Sci. 85:2444-2448; the algorithm of Karlin and Altschul (1990) Proc. Natl. Acad. Sci. USA 872264, modified as in Karlin and Altschul (1993) Proc. Natl. Acad. Sci. USA

90:5873-5877. Computer implementations of these mathematical algorithms can be utilized for comparison of sequences to determine optimum alignment. Such implementations include, but are not limited to: CLUSTAL in the PC/Gene program (available from Intelligenetics, Mountain View, Calif.); the ALIGN program (Version 2.0) and GAP, BESTFIT, BLAST, FASTA, and TFASTA in the GCG Wisconsin Genetics Software Package, Version 10 (available from Accelrys Inc., 9685 Scranton Road, San Diego, Calif., USA). Alignments using these programs can be performed using the default parameters. The CLUSTAL program is described by Higgins et al. (1988) Gene 73:237-244 (1988); Higgins et al. (1989) CABIOS 5:151-153; Corpet et al. (1988) Nucleic Acids Res. 16:10881-90; Huang et al. (1992) CABIOS 8:155-65; and Pearson et al. (1994) Meth. Mol. Biol. 24:307-331. The ALIGN program is based on the algorithm of Myers and Miller (1988) supra. A PAM120 weight residue table, a gap length penalty of 12, and a gap penalty of 4 can be used with the ALIGN program when comparing amino acid sequences. The BLAST programs of Altschul et al. (1990) J. Mol. Biol. 215:403 are based on the algorithm of Karlin and Altschul (1990) supra. To obtain gapped alignments for comparison purposes, Gapped BLAST (in BLAST 2.0) can be utilized as described in Altschul et al. (1997) Nucleic Acids Res. 25:3389. Alternatively, PSI-BLAST (in BLAST 2.0) can be used to perform an iterated search that detects distant relationships between molecules. See Altschul et al. (1997) supra. In another embodiment, GSNAP (Thomas D. Wu, Serban Nacu "Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010 Apr. 1; 26(7):873-81. 2010) can also be used.

Algorithms and parameters for alignment can be adjusted depending on the type of bacteria selected, the type of target sequence being characterized, and the method of transformation used to introduce the DNA into recipient bacteria.

Mapped reads may be post-processed by removing PCR duplicates (multiple, identical reads), etc.

The sequencing data after in vivo selection can be analyzed statistically.

In vivo selection of beneficial genes that improve gut colonization leads to a measureable enrichment (increase) of their relative abundance across the microbial population. As used herein, the term “enrich” refers to an increase in abundance (or percentage or concentration) of a particular group of genomic DNA fragments. For example, after *in vivo* selection, the DNA

library will contain a higher proportion of DNA fragments or genes than their proportion prior to the *in vivo* selection (enriching) process. A gene enriched at a time point may enhance bacterial fitness in the mammalian GI tract.

When the abundance of a gene during the *in vivo* selection (e.g., at a time point, or multiple time points as discussed herein, of step (d)) is greater than its abundance before the *in vivo* selection (e.g., its abundance in the original library), the gene may be identified to be able to enhance bacterial fitness in the mammalian GI tract. For example, the abundance of a gene during the *in vivo* selection (e.g., at a time point of step (d), such as on day 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30; after 1, 2, 3 or 4 weeks; after 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 or 12 months; or after 1, 2, 3, 4, 5 years or longer, counting from the initiation of the *in vivo* selection) may be at least about 1.2 fold, at least about 1.4 fold, at least about 1.5 fold, at least about 1.8 fold, at least about 2 fold, at least about 3 fold, at least about 4 fold, at least about 5 fold, at least about 8 fold, at least about 10 fold, at least about 15 fold, at least about 20 fold, at least about 25 fold, at least about 30 fold, at least about 35 fold, at least about 40 fold, at least about 50 fold, at least about 60 fold, at least about 70 fold, at least about 80 fold, at least about 90 fold, at least about 100 fold, at least about 200 fold, at least about 250 fold, at least about 300 fold, at least about 400 fold, at least about 500 fold, at least about 600 fold, at least about 700 fold, at least about 800 fold, at least about 900 fold, at least about 1,000 fold, at least about 1,100 fold, at least about 1,200 fold, at least about 1,300 fold, at least about 1,400 fold, at least about 1,500 fold, at least about 1,600 fold, at least about 1,700 fold, at least about 1,800 fold, at least about 1,900 fold, or at least about 2,000 fold, of its abundance before the *in vivo* selection (e.g., its abundance in the library).

Functional relevance of enriched genes may be assessed by metabolic pathway analysis using the KEGG (Kyoto Encyclopedia of Genes and Genomes) and COG databases.

The composition and abundance of the established microbiota after *in vivo* selection can also be studied by sequencing the 16S ribosomal RNA (or 16S rRNA) gene. 16S rRNA is a component of the 30S small subunit of prokaryotic ribosomes.

In one embodiment, changes in a mammalian gut bacterial populations are assessed by fluorescent in situ hybridization (FISH) with 16S rRNA probes. These 16S rRNA probes, specific for predominant classes of the gut microflora (bacteroides, bifidobacteria, clostridia, and lactobacilli/enterococci), are tagged with fluorescent markers. For example, the probes can

include Bif164 (Langendijk et al., Appl. Environ. Microbiol., 61: 3069-3075 (1995)), Bac303 (Manz, Microbiology, 142: 1097-1106 (1996)), His150 (Franks, Appl. Environment. Microbiol., 64: 3336-3345 (1998)), and Lab158 (Harmsen et al., Microbial Ecology Health Disease, 11: 3-12 (1996)). The nucleic acid stain 4'6-diamidino-2-phenylindole (DAPI) may be used for total bacterial counts. Fermentation samples are diluted and fixed in paraformaldehyde. These cells are then washed and re-suspended. The cell suspension is then added to the hybridization mixture and filtered. Hybridization is carried out at appropriate temperatures for the probes. Subsequently, the hybridization mix is vacuum filtered and the filter mounted on a microscope slide and examined using fluorescence microscopy, such that the bacterial groups could be enumerated (Ryecroft et al., J. Appl. Microbiol., 91: 878 (2001)). U.S. Patent No. 8,313,789.

Recombinant Bacteria engineered with gene that enhances fitness

The present invention also involves recombinant bacteria engineered with genes that can enhance bacterial fitness in the mammalian GI tract ("fitness genes"), for example, any gene identified by the present methods. The gene can be exogenous or endogenous. When the gene is endogenous, it can be overexpressed (i.e., having a generally higher expression than the gene in its natural form) and/or constitutively expressed. For example, the overexpressed gene can express its encoded protein at a level at least about 1.2 fold, at least about 1.4 fold, at least about 1.5 fold, at least about 1.8 fold, at least about 2 fold, at least about 3 fold, at least about 4 fold, at least about 5 fold, at least about 6 fold, at least about 7 fold, at least about 8 fold, at least about 10 fold, at least about 15 fold, at least about 20 fold, at least about 25 fold, at least about 30 fold, at least about 35 fold, at least about 40 fold, at least about 50 fold, at least about 60 fold, at least about 70 fold, at least about 80 fold, at least about 90 fold, at least about 100 fold, at least about 200 fold, of the protein expression level of the gene in its natural form.

The gene may be integrated into the bacterial chromosome or is episomal.

Expression of the gene requires that appropriate signals be provided in the vectors, and which include various regulatory elements, such as promoters that drive expression of the genes of interest in host cells.

The gene may be under the control of a constitutive, inducible, and/or tissue-specific promoter, or promoters useful under the appropriate conditions to direct expression of the introduced DNA segment or a gene.

The promoter may be exogenous (heterologous) or endogenous. A promoter may be one naturally-associated with a gene or sequence, or may be obtained by isolating the 5' non-coding sequences located upstream of the coding segment and/or exon. Such a promoter can be referred to as an endogenous promoter. Alternatively, certain advantages may be gained by positioning the coding nucleic acid segment under the control of a recombinant or heterologous promoter, which refers to a promoter that is not normally associated with a nucleic acid sequence in its natural environment. Such promoters may include promoters of other genes, and promoters isolated from any other prokaryotic, viral, or eukaryotic cell, and promoters or enhancers not naturally-occurring, i.e., containing different elements of different transcriptional regulatory regions, and/or mutations that alter expression.

Non-limiting examples of constitutive promoters include constitutive E. coli σ^{70} promoters, constitutive E. coli σ^{s} promoters, constitutive E. coli σ^{32} promoters, constitutive E. coli σ^{54} promoters, constitutive E. coli σ^{70} promoters, constitutive B. subtilis σ^{A} promoters, constitutive B. subtilis σ^{B} promoters, T7 promoters, and SP6 promoters. A list of constitutive bacterial promoters may be found in the database of Registry of Standard Biological Parts. They are active in all circumstances in the cell. In one embodiment, the constitutive promoter is pL.

The gene may be under the control of an inducible promoter. The transcriptional activity of these promoters is induced by either chemical or physical factors. Chemically-regulated inducible promoters may include promoters whose transcriptional activity is regulated by the presence or absence of oxygen, a metabolite, alcohol, tetracycline, steroids, metal and other compounds. Physically-regulated inducible promoters, including promoters whose transcriptional activity is regulated by the presence or absence of heat, low or high temperatures, acid, base, or light. In one embodiment, the inducible promoter is pH-sensitive (pH inducible).

The inducer for the inducible promoter may be located in the biological tissue or environmental medium to which the composition is administered or targeted, or is to be administered or targeted. For example, the inducer for the inducible promoter may be located in the mammalian GI tract.

The pH level of a particular biological tissue can affect the inducibility of the pH inducible promoter. See, for example, Boron, et al., Medical Physiology: A Cellular and Molecular Approach. Elsevier/Saunders. (2004), ISBN 1-4160-2328-3, which is incorporated herein by reference.

Examples of acid inducible promoters include, but are not limited to, P170, P1, P3, baiA1, baiA3, lipF promoter, F_1F_0-ATPase promoter, gadC, gad D, glutamate decarboxylase promoter, etc. See, for example, Cotter and Hill, Microbiol. and Mol. Biol. Rev. vol. 67, no. 3, pp. 429-453 (2003); Hagenbeek, et al., Plant Phys., vol. 123, pp. 1553-1560 (2000); Madsen, et al., Abstract, Mol. Microbiol. vol. 56, no. 3, pp. 735-746 (2005); U.S. Pat. No. 6,242,194; Richter, et al., Abstract, Gene, vol. 395, no. 1-2, pp. 22-28 (2007), Mallonee, et al., J. Bacteriol., vol. 172, no. 12, pp. 7011-7019 (1990); each of which is incorporated herein by reference. U.S. Patent No. 8,852,916.

Some non-limiting examples of promoters induced by a change in temperature include P2, P7, and PhS. See, for example, Taylor, et al, Cell, Abstract, vol. 38, no. 2, pp. 371-381 (1984); U.S. Pat. No. 6,852,511, Wang, et al., Biochem. and Biophys. Res. Commun. Abstract, vol. 358, no. 4, pp. 1148-1153 (2007), U.S. Pat. No. 7,462,708, each of which is incorporated herein by reference.

In an embodiment, the acid inducible promoter is inducible at a pH of about 0.0, about 0.5, about 1.0, about 1.5, about 2.0, about 2.5, about 3.0, about 3.5, about 4.0, about 4.5, about 5.0, about 5.5, about 6.0, about 6.5, about 6.6, about 6.7, about 6.8, about 6.9, or any value therebetween or less.

In an embodiment, the base inducible promoter is inducible at a pH of about 7.1, about 7.5, about 8.0, about 8.5, about 9.0, about 9.5, about 10.0, about 10.5, about 11.0, about 11.5, about 12.0, about 12.5, about 13.0, about 13.5, about 14.0, or any value therebetween or greater.

Examples of inducers that can induce the activity of the inducible promoters also include, but are not limited to, radiation, temperature change, alcohol, antibiotic, steroid, metal, salicylic acid, ethylene, benzothiadiazole, or other compound. In an embodiment, the at least one inducer includes at least one of arabinose, lactose, maltose, sucrose, glucose, xylose, galactose, rhamnose, fructose, melibiose, starch, inunlin, lipopolysaccharide, arsenic, cadmium, chromium, temperature, light, antibiotic, oxygen level, xylan, nisin, L-arabinose, allolactose, D-glucose, D-xylose, D-galactose, ampicillin, tetracycline, penicillin, pristinamycin, retinoic acid, or interferon. Other examples of inducers include, but are not limited to, at least a portion of one of an organic or inorganic small molecule, clathrate or caged compound, protocell, coacervate, microsphere, Janus particle, proteinoid, laminate, helical rod, liposome, macroscopic tube, niosome, sphingosome, vesicular tube, vesicle, unilamellar vesicle, multilamellar vesicle, multivesicular

vesicle, lipid layer, lipid bilayer, micelle, organelle, nucleic acid, peptide, polypeptide, protein, glycopeptide, glycolipid, lipoprotein, lipopolysaccharide, sphingolipid, glycosphingolipid, glycoprotein, peptidoglycan, lipid, carbohydrate, metalloprotein, proteoglycan, chromosome, nucleus, acid, buffer, protic solvent, aprotic solvent, nitric oxide, vitamin, mineral, nitrous oxide, nitric oxide synthase, amino acid, micelle, polymer, copolymer, monomer, prepolymer, cell receptor, adhesion molecule, cytokine, chemokine, immunoglobulin, antibody, antigen, extracellular matrix, cell ligand, zwitterionic material, cationic material, oligonucleotide, nanotube, piloxymer, transfersome, gas, element, contaminant, radioactive particle, radiation, hormone, virus, quantum dot, temperature change, thermal energy, or contrast agent. See, for example, Theys, et al., Abstract, Curr. Gene Ther. vol. 3, no. 3 pp. 207-221 (2003), which is incorporated herein by reference.

A nucleic acid sequence or a gene can be endogenous, or "exogenous" or "heterologous" which means that it is foreign to the cell into which the vector is being introduced or that the sequence or gene is homologous to a sequence in the cell but in a position within the host cell nucleic acid in which the sequence is ordinarily not found.

The genes that can enhance bacterial fitness in the mammalian GI tract may include the genes encoding a glycoside hydrolase, a galactokinase, a glucose/galactose transporter, or any other proteins involved in carbohydrate (e.g., sucrose, galactose, etc.) metabolism/transport. The gene may be from any prokaryote or eukaryote, including the donor bacteria disclosed herein.

In one embodiment, the recombinant bacteria are capable of metabolizing both sucrose and galactose. In another embodiment, the recombinant bacteria are capable of metabolizing sucrose. In yet another embodiment, the recombinant bacteria are capable of metabolizing galactose.

Pre-colonization with sucrose-utilizing probiotic strains to occupy the sucrose niche could also be an effective strategy to resist pathogen colonization.

The fitness gene engineered into the recombinant bacteria may be wild-type or be mutated. When the gene is a mutant, it may be truncated at the 5' end by from about 1 base pair (bp) to about 100 bp, from about 2 bp to about 50 bp, from about 3 bp to about 20 bp, or from about 4 bp to about 10 bp from the start codon. In one embodiment, the protein is *Bacteroides*

thetaiotaomicron glycoside hydrolase and its gene is truncated at the 5' end by about 4 bp from the start codon.

Glycoside hydrolase

A glycoside hydrolase, also called glycosidase or glycosyl hydrolase, is an enzyme that catalyzes the hydrolysis of the glycosidic bond between two or more carbohydrates (e.g., in complex sugars), or between a carbohydrate and a non-carbohydrate moiety. Glycoside hydrolases are typically classified into EC 3.2.1 as enzymes catalyzing the hydrolysis of O- or S-glycosides. Glycoside hydrolases can also be classified according to the stereo-chemical outcome of the hydrolysis reaction: thus they can be classified as either retaining glycoside hydrolases or inverting glycoside hydrolases. Glycoside hydrolases can also be classified as exo- or endo-acting, dependent upon whether they act at the end or in the middle, respectively, of a polysaccharide chain. Glycoside hydrolases may also be classified by sequence or structure based methods. Exemplary glycoside hydrolases include beta-galactosidase (also called beta-gal or β-gal), glucosidase, xylannase, lactase, amylase, chitinase, sucrase, maltase, neuraminidase, invertase, hyaluronidase and lysozyme. Samuel, PNAS, 2006, 103(26) 10011-10016.

The present invention encompasses both wild-type and mutant glycoside hydrolases. The mutant glycoside hydrolase may have conservative amino acid substitutions or functional fragments that do not substantially alter its activity. In certain embodiments, the gene of the glycoside hydrolase is truncated at the 5' end by from about 1 bp to about 50 bp from the start codon. For example, the glycoside hydrolase may be *Bacteroides thetaiotaomicron* glycoside hydrolase and its gene is truncated at the 5' end by about 4 bp from the start codon.

The present invention encompasses glycoside hydrolases from any of Glycoside Hydrolase Families 1 – 128. Henrissat et al., (1995). "Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases". *Proc. Natl. Acad. Sci. U.S.A.* 92(15): 7090–7094. Henrissat et al., (1995). "Structures and mechanisms of glycosyl hydrolases". *Structure* 3 (9): 853–859. Henrissat et al., (June 1996). "Updating the sequence-based classification of glycosyl hydrolases". *Biochem. J.* 316 (2): 695–6. Cantarel et al., (January 2009). "The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics". *Nucleic Acids Res.* 37 (Database issue): D233–8.

Galactokinase

Galactokinase is an enzyme that facilitates the phosphorylation of α-D-galactose to galactose 1-phosphate at the expense of one molecule of ATP. Galactokinase may also phosphorylate 2-deoxy-D-galactose, 2-amino-deoxy-D-galactose, 3-deoxy-D-galactose and D-fucose.

Probiotic Compositions

The present invention also provides for a probiotic composition comprising the present recombinant bacteria. The recombinant bacteria may constitutively express one or more of proteins involved in carbohydrate metabolism and/or transport. The proteins can include a glycoside hydrolase, a galactokinase and a glucose/galactose transporter. The gene of the protein may be under the control of a constitutive or inducible promoter.

Probiotics are microorganisms, or processed compositions of microorganisms which beneficially affect a host. Salminen et al., Probiotics: how should they be defined, Trends Food Sci. Technol. 1999: 10 107-10. U.S. Patent No. 8,216,563.

The present probiotic composition is to be administered enterally, such as oral, sublingual and rectal administrations.

The present probiotic composition can be a food composition, a beverage composition, a pharmaceutical composition, or a feedstuff composition.

The present probiotic composition may comprise a liquid culture. The probiotic composition may be lyophilized, pulverized and powdered. As a powder it can be provided in a palatable form for reconstitution for drinking or for reconstitution as a food additive. The composition can be provided as a powder for sale in combination with a food or drink. The food or drink may be a dairy-based product or a soy-based product. The invention therefore also includes a food or food supplement containing the present composition. Typical food products that may be prepared in the framework of the present invention may be milk-powder based products; instant drinks; ready-to-drink formulations; nutritional powders; milk-based products, such as yogurt or ice cream; cereal products; beverages such as water, coffee, malt drinks; culinary products and soups.

The composition may further contain at least one prebiotic. "Prebiotic" means food substances intended to promote the growth of probiotic bacteria in the intestines. The prebiotic

may be selected from the group consisting of oligosaccharides and optionally contains fructose, galactose, mannose, soy and/or inulin; and/or dietary fibers. The composition of the present invention may further contain prebiotics. Prebiotics may be dietary fibers. Dietary fibers may be selected from the group consisting of fructo-oligosaccharides, galacto-oligosaccharides, xylo-oligosaccharides, isomalto-saccharides, soya oligosaccharides, pyrodextrins, transgalactosylated oligosaccharides, lactulose, beta-glucan, insulin, raffinose, stachyose. Dietary fibers also have the advantage of being resistant to a number of conditions including heating and long storage times. They furthermore may contribute to a treatment in the framework of the present invention by improving gastrointestinal health and by increasing satiety.

The composition can be combined with other adjuvants such as antacids to dampen bacterial inactivation in the stomach. Acid secretion in the stomach could also be pharmacologically suppressed using H2-antagonists or proton pump inhibitors. Typically, the H2-antagonist is ranitidine. Typically the proton pump inhibitor is omeprazole.

The composition of the present invention may further contain protective hydrocolloids (such as gums, proteins, modified starches), binders, film forming agents, encapsulating agents/materials, wall/shell materials, matrix compounds, coatings, emulsifiers, surface active agents, solubilizing agents (oils, fats, waxes, lecithins etc.), adsorbents, carriers, fillers, co-compounds, dispersing agents, wetting agents, processing aids (solvents), flowing agents, taste masking agents, weighting agents, jellifying agents, gel forming agents, antioxidants and antimicrobials.

The composition according to the invention may comprise a source of protein. Any suitable dietary protein may be used, for example animal proteins (such as milk proteins, meat proteins and egg proteins); vegetable proteins (such as soy protein, wheat protein, rice protein, and pea protein); mixtures of free amino acids; or combinations thereof. The proteins may be intact, hydrolyzed, partially hydrolyzed or a mixture thereof.

The composition may also contain a source of carbohydrates and a source of fat. A source of carbohydrate may be added to the composition. Any suitable carbohydrate may be used, for example sucrose, lactose, glucose, fructose, corn syrup solids, maltodextrins, and mixtures thereof. Dietary fiber may also be added if desired.

The pharmaceutical compositions of the present invention can be, e.g., in a solid, semi-solid, or liquid formulation. Compositions can also take the form of tablets, pills, capsules, semisolids, powders, sustained release formulations, emulsions, suspensions, or any other appropriate compositions.

The present composition may be in the form of: an enema composition which can be reconstituted with an appropriate diluent; enteric-coated capsules or microcapsules; powder for reconstitution with an appropriate diluent for naso-enteric infusion, naso-duodenal infusion or colonoscopic infusion; powder for reconstitution with appropriate diluent, flavoring and gastric acid suppression agent for oral ingestion; or powder for reconstitution with food or drink. U.S. Patent Publication No. 20140234260.

The composition may also contain conventional pharmaceutical additives and adjuvants, excipients and diluents, including, but not limited to, water, gelatine of any origin, vegetable gums, ligninsulfonate, talc, sugars, starch, gum arabic, vegetable oils, polyalkylene glycols, flavouring agents, preservatives, stabilizers, emulsifying agents, buffers, lubricants, colorants, wetting agents, fillers, and the like. In all cases, such further components will be selected having regard to their suitability for the intended recipient. U.S. Patent Publication No. 8,741,622.

A sufficient dose of the recombinant bacteria is usually consumed per day in order to achieve successful colonization. The daily dose of probiotics in the composition will depend on the particular person or animal to be treated. Important factors to be considered include age, body weight, sex and health condition. Daily doses generally range from about 10^2 to about 10^{14} cfu (colony forming units), from about 10^2 to about 10^{12} cfu, from about 10^4 to about 10^{12} cfu, from about 10^6 to about 10^{10} cfu, from about 10^6 to about 10^{14} cfu, about 10^7 to about 10^{13} cfu, about 10^{10} to about 10^{14} cfu, about 10^{11} to about 10^{13} cfu, about 1-4x10^{12} cfu, or from about 10^7 to about 10^9 cfu per day. U.S. Patent No. 8,021,656.

The dosage of the present recombinant bacteria in the gut can be adjusted by those skilled in the art to the designated purpose. Any dose showing an effect may be suitable.

Appropriate frequency of administration can be determined by one of skill in the art and can be administered once or several times per day (e.g., twice, three, four or five times daily). The compositions of the invention may also be administered once each day or once every other

day. The compositions may also be given twice weekly, weekly, monthly, or semi-annually. U.S. Patent No. 8,501,686.

Conditions

The present compositions and methods may be used for the treatment and/or prophylaxis of a disorder associated with the presence in the gastrointestinal tract of a mammalian host of abnormal (or an abnormal distribution of) microbiota. The method comprises administering an effective amount of the present composition. For the present methods of identifying genes that enhance bacterial fitness, the donor bacterium may be from a natural gut microbiota of a healthy mammal or a mammal with the following disorders.

Such disorders include but are not limited to those conditions in the following categories:

gastro-intestinal disorders including irritable bowel syndrome (IBS, or spastic colon), and intestinal inflammation, functional bowel disease (FBD), including constipation predominant FBD, pain predominant FBD, upper abdominal FBD, non-ulcer dyspepsia (NUD), gastro-oesophageal reflux, inflammatory bowel disease including Crohn's disease, ulcerative colitis, indeterminate colitis, collagenous colitis, microscopic colitis, chronic Clostridium difficile infection, pseudomembranous colitis, mucous colitis, antibiotic associated colitis, idiopathic or simple constipation, diverticular disease, AIDS enteropathy, small bowel bacterial overgrowth, coeliac disease, polyposis coli, colonic polyps, chronic idiopathic pseudo obstructive syndrome;

chronic gut infections with specific pathogens including bacteria, viruses, fungi and protozoa (e.g., *Clostridum difficile* infection (CDI));

viral gastrointestinal disorders, including viral gastroenteritis, Norwalk viral gastroenteritis, rotavirus gastroenteritis, AIDS related gastroenteritis;

liver disorders such as primary biliary cirrhosis, primary sclerosing cholangitis, fatty liver or cryptogenic cirrhosis;

rheumatic disorders such as rheumatoid arthritis, non-rheumatoid arthritidies, non rheumatoid factor positive arthritis, ankylosing spondylitis, Lyme disease, and Reiter's syndrome;

immune mediated disorders such as glomerulonephritis, haemolytic uraemic syndrome, type 1 or type 2 diabetes mellitus, mixed cryoglobulinaemia, polyarteritis, familial Mediterranean fever, amyloidosis, scleroderma, systemic lupus erythematosus, and Behcets

syndrome;

autoimmune disorders including systemic lupus, idiopathic thrombocytopenic purpura, Sjogren's syndrome, haemolytic uremic syndrome or scleroderma;

neurological syndromes such as chronic fatigue syndrome, migraine, multiple sclerosis, amyotrophic lateral sclerosis, myasthenia gravis, Gillain-Barre syndrome, Parkinson's disease, Alzheimer's disease, Chronic Inflammatory Demyelinating Polyneuropathy, and other degenerative disorders;

psychiatric disorders including chronic depression, schizophrenia, psychotic disorders, manic depressive illness;

regressive disorders including Asperger's syndrome, Rett syndrome, autism, attention deficit hyperactivity disorder (ADHD), and attention deficit disorder (ADD);

sudden infant death syndrome (SIDS), anorexia nervosa; and

dermatological conditions such as, chronic urticaria, acne, dermatitis herpetiformis and vasculitic disorders. U.S. Patent Publication No. 20140234260.

Other metabolic disorders that can be treated or prevented by the above use include obesity, insulin resistance, hyperglycemia, hepatic steatosis, and small intestinal bacterial overgrowth (SIBO), U.S. Patent No. 8,110,177.

The present recombinant bacteria may additionally confer benefits to a subject. These additional benefits are generally known to those skilled in the art and may include managing lactose intolerance, prevention of colon cancer, lowering cholesterol, lowering blood pressure, improving immune function and preventing infections, reducing inflammation and/or improving mineral absorption.

Subjects, which may be treated according to the present invention include all animals which may benefit from the present invention. Such subjects include mammals, preferably humans (infants, children, adolescents and/or adults), but can also be an animal such as dogs and cats, farm animals such as cows, pigs, sheep, horses, goats and the like, and laboratory animals (e.g., rats, mice, guinea pigs, and the like).

The following are examples of the present invention and are not to be construed as limiting.

EXAMPLES

Example 1 Improving microbial fitness in the mammalian gut using *in vivo* temporal functional metagenomics

Elucidating functions of commensal microbial genes in the mammalian gut is challenging because many commensals are recalcitrant to laboratory cultivation and genetic manipulation. We have used TFUMseq (Temporal FUnctional Metagenomics sequencing), a platform to functionally mine bacterial genomes for genes that contribute to fitness of commensal bacteria *in vivo*. Our approach uses metagenomic DNA to construct large-scale heterologous expression libraries that are tracked over time *in vivo* by deep sequencing and computational methods. We have built a plasmid library using the gut commensal *Bacteroides thetaiotaomicron* (Bt) and introduced *Escherichia coli* carrying this library into germfree mice. Population dynamics of library clones revealed that Bt genes conferred significant fitness advantages in *E. coli* over time, including carbohydrate utilization genes, with a Bt galactokinase central to early colonization, and subsequent dominance by a Bt glycoside hydrolase enabling sucrose metabolism coupled with co-evolution of the plasmid library and *E. coli* genome driving increased galactose utilization. Our findings highlight the utility of functional metagenomics for engineering commensal bacteria with improved properties, including expanded colonization capabilities *in vivo*.

To characterize the functions of microbial genes, we have built large-scale shotgun expression libraries, which uses physical shearing or restriction digestion of donor DNA to generate fragments that are cloned into an expression vector and transformed into a recipient bacterial strain for functional analysis. This approach has the advantage that the donor organism need not be culturable or readily genetically manipulable in the laboratory; moreover, it allows investigation of essential genes or those conferring a fitness advantage synergistic with the recipient organism. The use of shotgun libraries for functional metagenomics of mammalian-associated microbiota has been demonstrated *ex vivo*, such as by growing the library in media with different substrates to characterize carbohydrate active enzymes (Tasse *et al*, 2010), prebiotic metabolism (Cecchini *et al*, 2013), glucuronidase activity (Gloux *et al*, 2011), salt tolerance (Culligan *et al*, 2012), and antibiotic resistance genes (Sommer *et al*, 2009), or by using filtered lysates of the library to screen for signal modulation in mammalian cell cultures (Lakhdari *et al*, 2010). This metagenomic shotgun library approach has yet to be demonstrated on a large-scale *in vivo*.

We used high-coverage genetic fragments from the genome of the fully sequenced human gut commensal *Bacteroides thetaiotaomicron* (Bt) (Xu *et al*, 2003) and cloned the fragments into a plasmid library in an *Escherichia coli* K-12 strain. We chose Bt because of its abundance in the mammalian gut, its persistent colonization, and its well-characterized repertoire of catabolic activities, such as sensing polysaccharides and redirecting metabolism to forage on host versus dietary glycans (Sonnenburg *et al*, 2005; Bjursell *et al*, 2006; Martens *et al*, 2008). We subjected the library to *in vitro* and *in vivo* selective pressures, collected output samples at different time points for high-throughput sequencing, and used computational methods to reconstruct the population dynamics of clones harboring donor genes (Figure 1). Our work is an advance over previous studies where our study employs shotgun expression libraries for functional metagenomics *in vivo*. Important features of the mammalian gut are difficult to recapitulate *in vitro*, such as the host immune response. Thus, *in vivo* experiments are essential for investigating the function of commensal microbiota genes in the host. Our study leverages high-throughput sequencing and computational methods to generate detailed dynamics of the entire population subject to selection over time. This kinetic information is crucial for understanding succession events during the inherently dynamic and complex process of host colonization.

Our approach marries two massively parallel strategies to identify genes that are enriched during functional selection. Our strategy involves expression and selection of a functional genomic library followed by characterization of library composition and abundance over time through deep-sequencing. The method is culture-independent so "unculturable" microbes and metagenomic sources can be functionally interrogated without the need for cultivation in the lab.

The present approach tracks an important temporal component of gut colonization by measuring allele frequencies over time to monitor the dynamic changes during colonization.

Results

Library construction and characterization

A 2.2 kb *E. coli* expression vector, GMV1c, was constructed to include the strong constitutive promoter pL and a ribosomal binding site upstream of the cloning site for input DNA fragments (Figure 1). We cloned in 2-5 kb fragments of donor genomic DNA from Bt, and generated a library of ~100,000 members, corresponding to >50X coverage of the donor genome. We sequenced the library on the Illumina HiSeq instrument to confirm sufficient coverage of the

Bt genome (Figure 2A and E1). The distribution of member insert sizes in the input library was verified to be centered around 2-3 kb (Figure 2B), a size range allowing for the full-length representation of almost all Bt genes.

In vitro stability and selection by media condition

To determine vector stability *in vitro*, we performed serial batch passaging of cells carrying GMV1c every one to two days over two weeks in two media conditions: aerobic Luria broth (LB) and anaerobic mouse-chow filtrate (MC). We expected the MC medium and anaerobic conditions to better reflect aspects of the nutritional content and oxygenation status in the mouse gut than rich LB medium in aerobic conditions. In both conditions, the vector was maintained in over 80% of library members without antibiotic selection throughout two weeks of *in vitro* passaging (~300 generations) (Figure 2C), suggesting general stability of the medium copy vector (~40 copies per cell). Clones harboring the empty vector (i.e., plasmid with no Bt insert) were the most fit library member: in both LB and MC conditions, these clones initially constituted 70% of the library and increased to 90% by the end of two weeks, albeit at a slower rate in anaerobic MC.

To identify Bt genes with differential *in vitro* selection in LB and MC conditions relative to the input library, we isolated DNA from Day 0 and Day 6/7 cultures, amplified the inserts by PCR for deep sequencing on the Illumina MiSeq platform and used computational methods to determine donor genes that were differentially enriched or depleted. In each condition, we found a number of significantly enriched or depleted Bt genes (Table 1). At Day 7 in aerobic LB, enriched genes included metabolic enzymes, such as chitobiase (BT_0865), which degrades chitin, and stress response proteins, such as glycine betaine/L-proline transport system permease (BT_1750), which is involved in the import of osmoprotectants glycine betaine or proline that mitigate effects of high osmolarity (Haardt *et al*, 1995). At Day 6 in anaerobic MC, a different set of genes was significantly enriched, particularly the locus consisting of endo-1,4-beta-xylanase (BT_0369), galactokinase (BT_0370), glucose/galactose transporter (BT_0371), and aldose 1-epimerase (BT_0372). These results highlight that our functional metagenomics approach is able to enrich for likely bioactive donor genes that improve fitness of the recipient cells in *in vitro* passaging conditions. Enolase (BT_4572), the only common hit among annotated genes in both media conditions, was found to be depleted relative to the input library. This

enzyme catalyzes the penultimate step of glycolysis, and its overexpression may be toxic in *E. coli* (Usui *et al*, 2012).

Table 1. Bt genes significantly enriched (abundance increased) or depleted (abundance decreased) at Day 7 *in vitro*.

Statistically significant genes ($q < 0.05$) enriched (positive values for the log2 abundance fold changes) or depleted (negative values for the log2 abundance fold changes) at Day 7 relative to Day 0 in the *in vitro* passaging experiment are listed for the anaerobic mouse chow and aerobic Luria broth conditions.

Gene	Gene product	log2(fold change)	q value
Enrichment at Day 7 in anaerobic MC passaging			
BT_0370	galactokinase	3.51	8.30E-05
BT_0372	aldose 1-epimerase	3.21	8.30E-05
BT_0371	glucose/galactose transporter	3.59	1.53E-04
BT_0478	hypothetical protein	3.59	3.70E-04
BT_0369	endo-1,4-beta-xylanase D	2.51	2.63E-03
Enrichment at Day 7 in aerobic LB passaging			
BT_1750	glycine betaine/L-proline transport system permease	9.18	0.00E+00
BT_2055	biopolymer transport protein	3.64	3.03E-05
BT_4358	hypothetical protein	2.84	6.31E-03
BT_1922	N-acetylmuramoyl-L-alanine amidase	2.62	7.24E-03
BT_0659	hypothetical protein	2.59	7.24E-03
BT_4333	hypothetical protein	2.57	7.84E-03
BT_2054	hypothetical protein	2.90	8.97E-03
BT_0757	beta-galactosidase	3.00	1.44E-02
BT_0660	hypothetical protein	2.36	1.48E-02
BT_2732	hypothetical protein	2.48	1.48E-02
BT_2843	integrase	3.34	1.51E-02
BT_3927	hypothetical protein	2.43	1.56E-02
BT_3612	FKBP-type peptidylprolyl isomerase	2.23	2.36E-02

BT_3821	5,10-methylenetetrahydrofolate reductase	2.31	2.36E-02
BT_0865	chitobiase	2.39	2.36E-02
BT_0973	hypothetical protein	2.21	2.36E-02
BT_0676	N-acetylglucosamine-6-phosphate deacetylase	2.17	2.59E-02
BT_2408	LuxR family transcriptional regulator	2.22	2.59E-02
BT_1038	hypothetical protein	2.24	2.59E-02
BT_3985	hypothetical protein	2.16	2.59E-02
BT_2917	hypothetical protein	2.16	2.80E-02
BT_0972	oxidoreductase	2.13	3.17E-02
BT_1923	O-acetylhomoserine (thiol)-lyase	2.43	3.35E-02
BT_1006	nitroreductase	2.10	3.64E-02
BT_1004	hypothetical protein	2.11	3.94E-02
BT_4544	transposase	1.99	4.64E-02
BT_2379	hypothetical protein	5.81	4.64E-02
BT_0974	hypothetical protein	2.08	4.64E-02
BT_0011	hypothetical protein	2.00	4.64E-02
BT_0510	heme biosynthesis protein	2.22	4.64E-02
Depletion at Day 7 in anaerobic MC passaging			
BT_1771	cell surface protein	-3.25	3.70E-04
BT_4572	phosphopyruvate hydratase (enolase)	-3.48	1.32E-03
BT_2959	hypothetical protein	-2.97	5.18E-03
BT_3089	hypothetical protein	-2.73	7.92E-03
BT_3528	hypothetical protein	-2.95	1.69E-02
BT_2051	hypothetical protein	-4.21	1.69E-02
BT_3577	hypothetical protein	-2.16	2.01E-02
Depletion at Day 7 in aerobic LB passaging			
BT_4572	phosphopyruvate hydratase (enolase)	-3.17	2.78E-03
BT_1538	hemagglutinin	-3.46	7.24E-03
BT_3395	acetylglutamate kinase	-2.42	1.48E-02
BT_1817	RNA polymerase ECF-type sigma factor	-2.19	2.36E-02
BT_1818	hypothetical protein	-2.54	2.36E-02

BT_2961	hypothetical protein	-2.32	2.59E-02
BT_4571	RNA polymerase ECF-type sigma factor	-3.91	3.31E-02
BT_2959	hypothetical protein	-2.18	3.45E-02

In vivo library selection in germ-free mice

To investigate *in vivo* gene selection in our library, we inoculated two cohorts of C57BL/6 germ-free mice (n=5 per group) and maintained the mice for 28 days under gnotobiotic conditions. One cohort was colonized with our library; the other cohort with a control GMV1c vector carrying the 5.9 kb luciferase operon (luxCDABE from *Photorhabdus luminescens,* Winson *et al*, 1998). Fecal pellets were collected on days 0.5, 0.75, 1.5, 1.75, 2.5, 3, 4, 7, 10, 14, 21, 25, and 28 after inoculation.

To determine *in vivo* vector stability, we plated fecal pellets on LB, on which *E. coli* either with or without vectors would grow, and on LB+carbenicillin, selective for *E. coli* harboring vectors. Strains carrying the luciferase vector dropped by ~100,000-fold by Day 28 compared to the earliest plated time-point (18 hours), presumably due to negative selective pressures from the energy consumption of the vector-borne luciferase in *E. coli* (Figure 3A). In contrast, our library was well-maintained *in vivo* throughout the 28 days of the experiment, suggesting at least minimal fitness cost to maintain the Bt insert library. Furthermore, unlike in the *in vitro* experiment, where clones containing the empty vector were enriched over time, these clones were virtually absent by the end of the *in vivo* experiments, suggesting positive selection had taken place.

Characterization of *in vivo* library population dynamics

To characterize the entire *in vivo* selected library over time, we extracted DNA from all collected stool samples, PCR amplified the donor inserts, prepared sequencing libraries of the amplicons, and sequenced libraries on the Illumina HiSeq 2500 instrument. Each sample resulted in ~7 million 100 nt paired-end reads. We also Sanger-sequenced vectors from clones directly isolated from stool samples to confirm deep-sequencing results and obtain insights into the structure of full-length inserts.

To obtain a genome-wide view of library selection over time and across the different mice, we calculated an information theoretic measure, termed effective positional diversity,

similar to that commonly used to quantify population diversity in macroscopic and microscopic ecology studies (Jost, 2006; Schloss *et al*, 2009)(Figure 3B). This measure, equal to the exponentiated Shannon entropy over all positions in the Bt genome, reflects how many positions in the donor genome are evenly represented in the population. Effective positional diversity values of the initial library were ~6 Mb, indicating essentially even coverage of the entire Bt genome. From Day 1.75 to Day 7 and continuing until the end of the experiment at Day 28, there was a rapid decline in effective positional diversity, which signifies expansion in the population of clones harboring inserts at a limited number of Bt genomic loci.

To explore the kinetics of gene selection *in vivo*, we plotted the percentage of sequencing reads mapped to genes in the *Bt* genome over time, and examined genes constituting >0.2% of total reads. Prior to inoculation, the read coverage was even over the entire Bt genome and corresponded to <0.2% per gene. Figure 3C provides a representative visualization of gene selection for Mouse 2 (plots for other mice are shown in Figure 6). By 36 hours post-inoculation, five genes, alpha-L-arabinofuranosidase, endo-1,4-beta-xylanase, galactokinase, glucose/galactose transporter, and aldose 1-epimerase (BT_0368 to BT_0372), comprised over half of the reads mapped. At Day 2.5, glucose/galactose transporter (BT_1758) and glycoside hydrolase (BT_1759) became noticeable and continued to increase until they saturated all reads at Day 14. Then, fructokinase (BT_1757) emerged and stabilized at around 6% of the reads throughout the remaining two weeks of the experiment. Abundance fold changes of the BT genes at different time points during in vivo selection are shown in Table 2. The numbers in Table 2 show abundance fold changes at various time points during *in vivo* selection. These observations are generally consistent across all five mice, though the selection kinetics varied slightly (Figure 6). For example, the transition from galactokinase and glucose/galactose transporter (BT_0370 and BT_0371) to glycoside hydrolase (BT_1759) occurred four days earlier in Mouse 5 than in Mouse 2, and the emergence of fructokinase (BT_1757) was detectable only in Mice 2, 4, and 5.

In terms of functional groups rather than individual genes, of the 51.4% Bt genes with COG annotations, those related to carbohydrate transport and metabolism comprised 10% of the input library. Averaged across the five mice, these carbohydrate transport and metabolism genes increased to 25% of reads on Day 0.5, 72% on Day 1.5, and essentially 100% by Day 7 (Figure 7), suggesting the importance of carbohydrate transport and metabolism in *in vivo* fitness.

Table 2. *In vivo* selection of Bt genes

Abundance Fold change		Days (in vivo selection)										
GENES	**Gene Product**	**0.5**	**1.5**	**1.75**	**2.5**	**3**	**4**	**7**	**10**	**14**	**21**	**28**
BT_0297	outer membrane lipoprotein SilC	1.6	2.2	2.4	1.7	0.8	0.2	0.1	0.0	0.0	0.0	0.0
BT_0368	aldose 1-epimerase	1.0	4.3	4.2	3.0	2.0	1.1	8.0	2.9	0.1	0.0	0.1
BT_0369	endo-1,4-beta-xylanase D	2.1	330.7	369.7	458.3	488.7	337.2	356.7	129.5	6.9	1.5	1.1
BT_0370	galactokinase	3.1	972.3	1082.6	1396.4	1334.7	957.9	804.6	423.9	20.7	2.9	1.6
BT_0371	glucose/galactose transporter	2.3	794.7	897.5	1091.9	914.0	646.8	515.2	253.3	14.9	2.1	0.8
BT_0372	aldose 1-epimerase	1.3	359.5	397.3	471.6	308.7	256.3	219.3	167.0	9.9	1.3	0.6
BT_0477	D-glycero-alpha-D-manno-heptose-1,7-bisphosphate 7-phosphatase (*gmhB*)	0.9	1.0	1.1	1.0	0.4	0.1	0.0	0.0	0.0	0.0	0.0
BT_0478	hypothetical protein	0.9	0.7	0.8	0.6	0.3	0.1	0.0	0.0	0.0	0.0	0.0
BT_1236		0.8	1.7	1.9	2.9	2.1	0.7	0.5	0.1	0.0	0.0	0.0
BT_1510	hypothetical protein	0.9	1.2	1.1	0.9	0.4	0.1	0.0	0.0	0.0	0.0	0.0
BT_1511	outer membrane protein OmpA	0.9	1.9	2.3	2.0	0.6	0.2	0.0	0.0	0.0	0.0	0.0
BT_1729		1.3	0.8	0.9	0.4	0.2	0.1	0.2	0.1	0.0	0.0	0.0
BT_1730	dTDP-4-	1.1	0.9	0.9	0.4	0.2	0.1	0.2	0.1	0.0	0.0	0.0

	dehydrorhamnose reductase (*rfbD*; *rmlD*)												
BT_1731	hypothetical protein	0.8	0.6	0.6	0.3	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0
BT_1757	fructokinase	1.1	0.5	0.5	0.5	0.7	0.4	0.1	12.2	145.6	224.7	239.3	
BT_1758	glucose/galactose transporter	45.1	2.9	5.5	44.2	318.6	564.1	500.0	420.0	581.8	635.6	642.8	
BT_1759	glycoside hydrolase	263.6	4.1	7.8	65.7	508.3	944.6	1151.2	1677.5	2000.9	1948.0	1976.1	
BT_1771	cell surface protein	0.9	1.0	0.8	0.7	0.3	0.1	0.0	0.0	0.0	0.0	0.0	
BT_3542		1.1	0.5	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	
BT_4265	GMP synthase (*guaA*)	0.9	0.6	0.5	0.3	0.2	0.1	0.0	0.0	0.0	0.0	0.0	

To rigorously determine the Bt genes that are differentially represented in the population over time and to localize putatively selected regions to specific genes, we applied information theoretic and statistical techniques for longitudinal analysis (Bar-Joseph *et al*, 2003). In our analyses, transient dominance of clones *in vivo* is of particular interest as different genes may confer fitness advantages at distinct stages of host colonization. Further, our experiments capture competition among ~100,000 strains harboring distinct genetic fragments, rather than traditional binary competition experiments. Thus, we are interested in not only clones harboring Bt fragments that show an increase over time in relative abundance, but also those clones that show a significantly slower rate of depletion than other clones. To methodically detect these effects, for every Bt gene, we computed two measures: (1) time-averaged relative abundance (TA-RA), and (2) time-averaged normalized effective coverage (TA-NEC). The TA-RA value is conceptually similar to a time-integrated pharmacological dose value (Byers & Sarver, 2009); in our analysis, it represents the average "dose" of a particular donor gene, relative to all other donor genes present *in vivo* over a period of time. The TA-NEC value quantifies the fraction of the gene that is effectively covered by reads over a period of time. These measures are important to evaluate in tandem, since bystander genetic loci may be differentially abundant in clones (i.e., high TA-RA values) simply because they are contiguous with genes under selection; however, these loci are likely to be detectable as spurious (i.e., low TA-NEC values) because they will often include only fragments of genes.

Genes showing transient selection during early gut colonization

We found 13 Bt genes during the early stage of gut colonization (up to Day 4) with significantly larger than expected TA-RA and TA-NEC values (q-values < 0.05; Table 3). These genes include those coding for enzymes involved in synthesis of extracellular capsular polysaccharides and lipopolysaccharides (LPS), specifically D-glycero-alpha-D-manno-heptose-1,7-bisphosphate 7-phosphatase (*gmhB*) (BT_0477) and dTDP-4-dehydrorhamnose reductase (*rfbD*; *rmlD*) (BT_1730). There are two biosynthesis pathways of nucleotide-activated *glycero-manno*-heptose that result in either L-β-D-heptose or D-α-D-heptose, which serve as precursors or subunits in LPS, S-layer glycoproteins, and capsular polysaccharides (Valvano *et al*, 2002). The *E. coli* GmhB is critical for complete synthesis of the LPS core (Kneidinger *et al*, 2002). The selection for Bt *gmhB* could allow *E. coli* to expand its extracellular glycoprotein display, since

E. coli GmhB is highly selective for β-anomers while Bt GmhB prefers α-anomers during hydrolysis of D-*glycero*-D-*manno*-heptose 1*β*,7-bisphosphate (Wang *et al*, 2010). BT_1730 (*rfbD*; *rmlD*) is involved in dTDP-rhamnose biosynthesis involved in production of O-antigen, a repetitive glycan polymer in LPS, and potentially other cell-membrane components. Deletion of *rmlD* in *Vibrio cholera* results in a severe defect in colonization of an infant mouse model (Chiang & Mekalanos, 1999), and uropathogenic *E. coli* lacking functional RmlD lose serum resistance (Burns & Hull, 1998). Thus, expressing Bt *rmlD* could allow the recipient *E. coli* to alter its antigenicity or resistance to host factors that would impede its initial colonization of the gnotobiotic gut.

Table 3. Statistical testing of *in vivo* selection of Bt genes.

Genes showing significant selection up to Day 4 are in white. Genes showing significant selection up to both Day 4 and Day 28 are bolded. q-values for Day 28 are listed in parentheses.

Gene	Annotation	TA-RA q-value	TA-NEC q-value
BT_0297	outer membrane lipoprotein SilC	3.06E-02	4.08E-04
BT_0370	**galactokinase**	**1.14E-03 (3.25E-03)**	**5.94E-06 (6.95E-09)**
BT_0371	**glucose/galactose transporter**	**1.14E-03 (3.14E-03)**	**3.50E-02 (4.21E-05)**
BT_0477	D-glycero-alpha-D-manno-heptose-1,7-bisphosphate 7-phosphatase (*gmhB*)	1.67E-02	1.32E-02
BT_0478	hypothetical protein	1.77E-03	2.47E-02
BT_1510	hypothetical protein	3.86E-02	1.10E-03
BT_1511	outer membrane protein OmpA	4.38E-02	7.33E-04
BT_1730	dTDP-4-dehydrorhamnose reductase (*rfbD*; *rmlD*)	3.86E-02	3.45E-04
BT_1731	hypothetical protein	4.38E-02	8.80E-03
BT_1757	fructokinase	1.58E-02	2.50E-03

BT_1759	**glycoside hydrolase**	**1.19E-02 (2.48E-07)**	**2.58E-04 (1.21E-09)**
BT_1771	cell surface protein	4.38E-02	4.00E-03
BT_4265	GMP synthase (*guaA*)	3.86E-02	3.33E-02

Several other genes with membrane-associated functions also showed increased selection at Day 4, including outer membrane lipoprotein SilC (BT_0297), cell surface protein (BT_1771), and outer membrane protein OmpA (BT_1511). These genes could confer increased capabilities for *E. coli* to attach to the mucosal surface of the mammalian GI tract, or increased adaptations to the gut chemical environment. For instance, *Bacteroides fragilis* lacking OmpA are more sensitive to SDS, high salt, and oxygen exposure (Wexler *et al*, 2009). In *Bacteroides vulgatus*, OmpA plays a role in intestinal adherence (Sato *et al*, 2010), and in *Klebsiella pneumoniae*, activates macrophages (Soulas *et al*, 2000).

Since nucleotide pools are tightly controlled in *E. coli* (Mehra & Drabble, 1981), the selection for Bt GMP synthase *guaA* (BT_4265) may substantially affect guanine concentration, translation regulation, and cell signaling. Inhibiting GMP synthase induces stationary phase genes in *Bacillus subtilis* (Ratnayake-Lecamwasam *et al*, 2001), and nucleotide concentrations drop when *E. coli* transition from growth to stationary phase (Buckstein *et al*, 2008). These observations suggest that a copy of heterologous *guaA* could enable escape of native tight regulation of the guanine pool to prolong the cell's exponential growth phase. Moreover, extra GMP synthase may further protect *E. coli* from incorporating mutagenic deaminated nucleobases that would interfere with RNA function and gene expression (Pang *et al*, 2012).

Genes showing long-term selection

We found three Bt genes over the entire period of colonization (up to Day 28) with significantly larger than expected TA-RA and TA-NEC values (q-values < 0.05; Table 3); these genes also showed significant selection during early colonization (up to Day 4). All three genes are involved in sugar metabolism and transport, suggesting they may act to unlock more nutrient resources for *E. coli* in the gut. We performed *in vitro* experiments, described below, to further

characterize the functions of these strongly selected loci, centered around a Bt glycoside hydrolase (BT_1759) and galactokinase (BT_0370).

Glycoside hydrolase (BT_1759)

From Day 1.5 to Day 3 in the high-throughput sequencing data, we observed sharply positive selection of glycoside hydrolase (BT_1759), which stabilized and continued to be strongly selected for from Day 4 to Day 28 across all mice (Figure 4A). We confirmed these results with Sanger sequencing, which additionally allowed us to identify exact junctions and directionality of isolated inserts. In clones from Days 7, 14, and 28, we observed the primary selected insert to be 2.5 kb in length, beginning four nucleotides after the annotated glycoside hydrolase (BT_1759) start codon, and ending about one-third of the way into the downstream gene (glucose/galactose transporter). Notably, we also detected other inserts containing different 5' truncated versions of the glycoside hydrolase in the late time points, both in our high-throughput and Sanger sequencing data (Figure 4B).

Sonnenburg et al. previously demonstrated that periplasmic BT_1759 in Bt hydrolyzes smaller fructooligosaccharides and sucrose (Sonnenburg *et al*, 2010). To functionally characterize BT_1759 and surrounding genes when heterologously expressed in *E. coli*, we cloned the CDS (coding DNA sequence) of each into the backbone vector and transformed it into the starting *E. coli* strain. None of the full-length genes conferred growth in M9 minimal media with sucrose as the sole carbon source (Figure 4C). However, clones isolated from mice on Day 28 were able to metabolize sucrose. Furthermore, retransformation of the DNA vectors from these clones into the starting *E. coli* strain also conferred growth on sucrose, indicating that the phenotype was plasmid-borne. Interestingly, sucrose utilization was enabled when we reconstituted the 4 nt truncation found in many of the Day 7 and Day 28 Sanger-sequenced clones into the starting *E. coli* strain. These results suggest that the truncation allows for appropriate processing of the signal sequence to express and localize the Bt enzyme in the periplasmic space of *E. coli*, where sucrose is capable of entering by diffusion.

Galactokinase (BT_0370), glucose/galactose transporter (BT_0371), and native galactokinase reversion

In contrast to the selection profile of glycoside hydrolase (BT_1759) from high-throughput sequencing data, the galactokinase (BT_0370) and glucose/galactose transporter (BT_0371) exhibited an earlier increase in abundance that peaked at Day 2.5 and gradually declined over the remainder of the experiment (Figure 5A). We observed a similar trend in Day 7 clones by Sanger sequencing, with no clones containing BT_0370 or BT_0371 present at Day 28 (Figure 5B).

We confirmed that individually cloned BT_0370 and BT_0371 genes confer galactose utilization in *E. coli* when grown using M9 minimal media supplemented with 0.5% galactose as the sole carbon source (Figure 5C). To our surprise, *E. coli* isolated from mouse stool at later time points were able to grow on galactose even though they carried plasmids with glycoside hydrolase (BT_1759), and not the Bt galactose utilization genes (BT_0370 and BT_0371). However, strains retransformed with BT_1759 were unable to grow on galactose, suggesting that the stool-isolated strains gained the capability to use galactose through mutations independent of the expression plasmid, namely in the recipient *E. coli* genome. After confirmation that our starting *E. coli* strain was *galK*- due to the presence of an insertion sequence (IS2), we hypothesized that stool isolates reverted to *galK*+ via loss of IS2. In stool-isolated clones from Days 7, 14, and 28, we found that the *galK* reversion occurred after Day 7 and was found in >75% of clones in four of five mice at Day 14 (Figure 5D). Interestingly, *E. coli* harboring the insert library exhibited accelerated *galK* reversion in the mouse gut; in the luciferase control mice, there was an overall reversion rate of only 50% by Day 28, as opposed to 100% in the mice that had been inoculated with the Bt library. The genomic *galK* reversion by ~Day 14 suggests that there is early selection for Bt galactokinase (BT_0370), but this foreign gene is subsequently lost as the recipient *E. coli* regain native galactokinase activity, which seems to have a fitness advantage over the heterologously expressed Bt galactokinase gene.

In vivo genomic stability of *E. coli* recipient strain

Given the observed genomic galK reversion, we investigated whether other changes occurred in the *E. coli* genome over the course of our *in vivo* experiments. Genomic stability of non-adapted bacterial cells in the gastrointestinal tract *in vivo* has not been characterized in great detail and microbial mutation rates *in vivo* are not known. We performed whole genome sequencing of 13 *E. coli* isolates from stool of mice inoculated with the library and two *E. coli*

isolates from stool of mice inoculated with the control luciferase construct. Of the isolates from mice that were inoculated with the library, seven were from Day 7 samples containing either BT_1759 or BT_0370 inserts and six were from Day 28 samples containing BT_1759 inserts. In addition to searching for variants in the *E. coli* genome, we also looked for variants on the library plasmid with the known insert locus, and the F plasmid, which was present in the starting *E. coli* strain. Overall, we found single nucleotide variants (SNVs) in only three of the 15 isolates (Table 4).

Table 4. Genetic variants in mouse-isolated clones identified by whole genome sequencing.

Sample	Insert locus & size (kb)	Genomic galK+/-	Variant position on *E. coli* genome	Variant impact & coverage
NEB Turbo control	-	galK-		
Day 7 Mouse 1 clone 3	BT_0370 (4.0)	galK-	SNV 2976657 G>T	galR (R20L) [34/34 reads]
Day 7 Mouse 1 clone 1	BT_1759 (2.5)	galK-		
Day 7 Mouse 2 clone 5	BT_0370 (4.3)	galK-		
Day 7 Mouse 3 clone 1	BT_1759 (3.1)	galK-		
Day 7 Mouse 4 clone 4	BT_0370 (4.1)	galK-		
Day 7 Mouse 5 clone 2	BT_1759 (2.5)	galK+		
Day 7 Mouse 5 clone 4	BT_1759 (3.1)	galK-		
Day 28 Mouse 1 clone 1	BT_1759 (2.5)	galK+	SNV 363100 A>G	lacY (F27S) [173/173 reads]
Day 28 Mouse 2 clone 1	BT_1759 (2.5)	galK+		
Day 28 Mouse 3 clone 1	BT_1759 (2.5)	galK+		
Day 28 Mouse 4 clone 1	BT_1759 (2.5)	galK+		
Day 28 Mouse 5 clone 1	BT_1759 (2.5)	galK+		

Day 28 Mouse 5 clone 4	BT_1759 (4.2)	galK+		
Day 28 Mouse 7 clone 1 lux control	-	galK+		
Day 28 Mouse 10 clone 2 lux control	-	galK-	1. SNV 3991675 G>A 2. SNV 780994 G>A 3. F plasmid SNV 67772 T>C	1. cyaA (G175S) [122/122 reads] 2. intergenic, between lysW and valZ [108/108 reads] 3. traY promoter (-35) [229/229 reads]

One of the isolates from a luciferase control mouse harbored three mutations. One SNV was in the coding sequence of adenylate cyclase *cyaA*, while the other two SNVs were in intergenic regions, between tRNAs *lysW* and *valZ*, and between *traJ* and *traY* on the F plasmid. The functional effects of these SNVs, if any, are unclear. The operon structure of the tRNA region may be *lysT-valT-lysW-valZ-lysYZQ* (Blattner *et al*, 1997), or *valZ-lysY* could be a separate operon as predicted in EcoCyc (Keseler *et al*, 2011), in which case the SNV could affect transcription of the downstream tRNAs. As for the *traY* promoter variant, the -35 hexamer has been documented to be TTTACC (Gaudin & Silverman, 1993). The SNV T>C changes it to CTTACC, which could weaken the promoter to decrease expression of TraY, a DNA-binding protein involved in initiation of DNA transfer during conjugation.

One *E. coli* isolate from a library-inoculated mouse had a genomic change that conferred increased growth on galactose. Isolate 1 from Mouse 1 on Day 28 had a mutation in the lactose / melibiose:H+ symporter, *lacY* (F27S), which is a missense mutation in the first transmembrane region (Guan *et al*, 2002). We did not observe phenotypic differences on MacConkey-lactose plates, since the *E. coli* recipient strain has a deletion in lacZ, and thus all of our isolates were Lac-. However, this *lacY* mutant reached a higher density in M9 galactose compared to other Day 28 isolates, which carried the same plasmid-borne Bt glycoside hydrolase (Figure 8A). This clone also grew to a greater density than *E. coli* recipient strains in which we had cloned the Bt galactokinase operon (BT_0370-BT_0372) (Figure 8B). In fact, the *lacY* transporter can

transport galactose in addition to lactose, and *lacY* mutants have been shown previously to confer faster growth of *E. coli* MG1655 on galactose (Soupene *et al*, 2003).

Remarkably, we found an interaction between the *E. coli* genome and a heterologously expressed Bt gene. Isolate 3 from Mouse 1 on Day 7 had an SNV in the galactose repressor, *galR* (R20L), in its DNA binding domain (Weickert & Adhyat, 1992). *E. coli* GalR binds operator sequences upstream of the *galETK* operon (Weickert & Adhya, 1993), and the amino acid substitution of arginine for leucine could be disruptive to binding. Using MacConkey-galactose plates, we found that this isolate was Gal+, whereas a similar Day 7 clone, which also had a genomic galK- genotype and a Bt galactokinase (BT_0370) insert but no *galR* SNV, exhibited a Gal- phenotype. Since the BT_0370 inserts in the Day 7 clones were not identical (Figure 5B), we re-transformed the plasmids into the starting *E. coli* strain to confirm the phenotype and rule out effects from an underlying chromosomal *galR* mutation. In M9 galactose medium, the *galR_R20L* mutant grew to a higher cell density than a wild-type *galR* strain with the same Bt galactokinase plasmid (Figure 8C). These findings indicate that the *E. coli* genome had co-evolved with the *in vivo* selection of plasmids carrying Bt genes for galactose utilization.

We found no mutations in the library plasmids or Bt genes, and, aside from loss of the IS2 element in the *galK* gene, all other IS elements were intact on the *E. coli* genome. In aggregate, these small number of variants (~0.045 mutations/day) in the *E. coli* recipient strain suggest that outside of genetic loci with selective pressures exerted upon them, the organism remained genetically stable in the mammalian gut over the course of our experiment.

Discussion

We have demonstrated the use of high-throughput *in vivo* screening of genetic fragments from an entire donor genome from a commensal microbe to increase the fitness of a phylogenetically distant bacterial species in the mammalian gut. This is a demonstration of temporal functional metagenomics using shotgun libraries applied to the *in vivo* mammalian gut environment. Our findings attest to the value of a time-series approach, as the shifts in population dynamics of clones harboring different gene fragments would not have been discovered if we had only obtained endpoint data. Further, we introduced computational methods using information theoretic measures and statistical longitudinal analysis techniques that allowed us to identify and localize selection of donor genes over time.

In this study using an *E. coli* plasmid library of Bt genes, we uncovered sequential selection of clones with different carbohydrate utilization genes – first for galactose and then for sucrose metabolism. Galactose plays a substantial role in selection in our experiment, as all three of the observed *E. coli* genomic mutations (in *galK*, *lacY*, and *galR*) affected galactose utilization, and we observed selection for Bt galactokinase (BT_0370) and glucose/galactose transporter (BT_0371) *in vivo*. Galactose is a component of the hemi-cellulose that makes up part of the 15.2% neutral detergent fiber in mouse chow, although galactose composition was not explicitly provided by the manufacturer. Galactose is also a component of mammalian mucin in the GI tract (Juge, 2012). However, our observation that *in vitro* selection occurs for the BT_0370 and BT_0371 galactose utilization locus in MC medium indicates that the mouse chow diet itself is providing sufficient galactose to exert selective pressure at least in part. Once *E. coli* restored native galactokinase (*galK*) activity in its genome through loss of IS2, Bt genes that catabolized a second carbon source, sucrose, became dominant. Sucrose is a dominant simple carbohydrate in mouse chow, present at 0.71% (w/w) in comparison to 0.22% for glucose and fructose. Per Freter's nutrient-niche hypothesis, which described substrate-level competition and substrate-limited population levels (Freter *et al*, 1983), our results suggest that galactose is preferred over sucrose, and that a clone capable of utilizing both carbon sources will outcompete clones capable of only using one of the sources. Nutrient-based niches have been documented in the mammalian GI tract, including the varying sugar preferences among commensal and pathogenic *E. coli* strains (Maltby *et al*, 2013), and polysaccharide utilization loci (PULs) in *Bacteroides* species that promote long-term colonization (Lee *et al*, 2013). In fact, the enterohemorrhaghic *E. coli* strain EDL933 can use sucrose, while commensal *E. coli* strains K-12 MG1655, HS, and Nissle 1917 cannot (Maltby *et al*, 2013). Incorporating sucrose utilization, such as through the truncated Bt glycoside hydrolase (BT_1759) identified in this study, could enhance retention of probiotic *E. coli* strains. Pre-colonization with sucrose-utilizing probiotic strains to occupy the sucrose niche could also be an effective strategy to resist pathogen colonization.

Bt has been investigated previously using transposon mutagenesis systems coupled to mouse gut colonization experiments (Goodman *et al*, 2009), facilitating comparison of our results to the prior study. Goodman et al. found no difference in abundances of galactokinase (BT_0370) mutants *in vitro* but BT_0370 mutants were underrepresented *in vivo*. In contrast, in our study, the Bt galactokinase was selected for not only *in vivo*, but also *in vitro*. Furthermore,

Goodman et al. found dTDP-4-dehydrorhamnose reductase (BT_1730) and GMP synthase (BT_4265) mutants were underrepresented both *in vitro* and *in vivo*. However, in our study, BT_1730 and BT_4265 seemed to confer fitness only *in vivo*. The *in vitro* discrepancies may be a result of slightly different culturing and media conditions. The *in vivo* results are in agreement for BT_0370, BT_1730, and BT_4265, though the other genes we identified in our experiments were not significantly altered in representation in the transposon mutagenesis experiments, highlighting the different capabilities of the two approaches.

Overall, our goal is to build better commensal microbes, as it provides a general approach to functionally identifying genes from metagenomic DNA that enhance microbial fitness *in vivo*. Going forward, there are two primary considerations for designing future experiments: the choice of the bacterial strain to receive the donor plasmid library, and the mammalian host environment. In this study, we used a cloning strain of *E. coli* as the recipient bacteria which offer ease of use and a robust, high-quality library. This strain has inactivated restriction systems, preventing underrepresentation of DNA inserts in the library that may contain otherwise recognized methylated sites from the donor source. Further, the putative unfitness of the strain *in vivo*, compared to a wild-type commensal strain, allows for stronger selection signals from clones harboring functional donor genes. As we saw, the recipient strain also plays a role in the co-evolution of the insert library and the strain's own genome. We observed a genomic change, specifically the *galK* reversion, driving the shift in library selection from Bt galactokinase (BT_0370) to Bt glycoside hydrolase (BT_1759). Furthermore, we found single nucleotide variations in *E. coli galR* and *lacY* loci that boosted galactose utilization in individual clones harboring functional Bt genes. Given that co-evolution drives genomic changes in the recipient strain, using a well-characterized recipient strain may facilitate mechanistic interpretation of these changes.

The state of the mammalian host is also an important variable in our approach. In this work, germfree mice were mono-associated with the library. We expect selection results will differ in the setting of mice pre-colonized with a microbiota due to changes in nutrient availability and other ecological interactions, including competition or syntrophy. For instance, co-colonization experiments have demonstrated that probiotic strains and commensal bacteria adapt their substrate utilization. Bt shifts its metabolism from mucosal glycans to dietary plant polysaccharides when in the presence of *Bifidobacterium animalis*, *Bifidobacterium longum*, or

Lactobacillus casei (Sonnenburg *et al*, 2006). *Bacteroides* species are also known to engage in public-goods based syntrophy by releasing outer membrane vesicles (OMVs) that contain surface glycoside hydrolases or polysaccharide lyases (Rakoff-Nahoum *et al*, 2014). These enzymes catabolize large polysaccharides into smaller units, which can then be utilized by other species in the community. Given the complexities of multispecies bacterial communities, the ability of our strategy to track large numbers of clones over time will be important for detecting relevant genes that confer a fitness advantage within dynamically changing communities.

Our results suggest several future studies. Potential investigations could use total metagenomic DNA from stool samples, rather than DNA from culturable organisms. Another area of interest would be probing community composition and dynamics of selection in different regions of the gut. These studies would provide insights into biogeographical niches coupled with temporal data provided by our method. The genes identified in our study can be used to directly build a better probiotic strain. One could incorporate a metagenomic plasmid library into a probiotic strain and introduce the strain into a complex host-bacterial community to isolate genes that increase the strain's fitness *in vivo*. We have already identified sucrose utilization as an important and feasible trait to incorporate into an enhanced probiotic strain. Ultimately, our studies could enable the rational design of probiotic or commensal strains for various clinical applications, such as resisting pathogen colonization, compensating for a high-fat/high-sucrose Western diet, or tempering host autoimmunity.

Materials and Methods

Bacterial strains and growth conditions

Bacteroides thetaiotaomicron VPI-5482 (ATCC # 29148) was grown anaerobically in a rich medium based on Brain Heart Infusion with other supplements added. The genomic library was maintained in an *Escherichia coli* K-12 strain, NEB Turbo (New England Biolabs, Ipswich, MA). *E. coli* strains were grown in Luria broth (LB) and supplemented with carbenicillin (final concentration 100 μg/mL) as needed. For anaerobic growth, an anaerobic jar (GasPak System, Becton Dickinson, Franklin Lakes, NJ) was used.

Mouse chow (MC) filtrate was prepared by adding 150 mL deionized water to 8 g of crushed mouse chow (Mouse Breeding Diet 5021, LabDiet, St. Louis, MO). The mixture was heated at 95°C for 30 minutes with mixing, passed through a 0.22 μm filter, and autoclaved. The

sterility of the MC filtrate was confirmed by incubating at 37°C in aerobic and anaerobic conditions and observing no growth after several days.

Library generation

Bacteroides thetaiotaomicron genomic DNA was isolated (DNeasy Blood & Tissue Kit, Qiagen, Venlo, Netherlands), fragmented by sonication to 3-5 kb (Covaris E210, Covaris, Woburn, MA), and size-selected and extracted by gel electrophoresis (Pippin Prep, Sage Sciences). The fragments were end-repaired (End-It DNA End-Repair Kit, Epicenter, Madison, WI) and cloned into a PCR-amplified GMV1c backbone vector via blunt-end ligation. The reaction was transformed into NEB Turbo electrocompetent *E. coli* cells (New England Biolabs). The library size was quantified by counting colonies formed on selective media (LB carbenicillin) after plating a fraction of the transformed cells. To assess the size of inserts successfully cloned into the library, we picked colonies for PCR amplification using primers ver2_f/r (Table 5) that flanked the insert site. We further confirmed the presence of inserts by submitting amplified inserts for Sanger sequencing (Genewiz, South Plainfield, NJ) and aligning sequences with the donor *B. thetaiotaomicron* genome.

Table 5. Primers used in the study.

Name	**Sequence (5' -> 3')**
A_L	AGGACGCACTGACCGAATT
A_R	TTTATTTGATGCCTCTAGCACGC
ver2_f	TTTACTTTGCAGGGCTTCCC
ver2_r	ACTGAGCCTTTCGTTTTATTTGATG
galK16_chk_f	CCTGCCACTCACACCATTCAG
galK16_chk_r	TGGGCGCATCGAGGGA
GMV_amp_f	AACAAGCTTGATATCGAATTCCTGC
GMV_amp_r	GACGGTACCTTTCTCCTCTTTAATGA

Plasmid retention

Individual stool pellets from Days 0.75, 1.5, 1.75, 2.5, 4, 10, 14, 21, 25, and 28 were homogenized in 10% PBS and plated on LB agar with or without carbenicillin (carb). To obtain accurate counts, platings were performed in triplicate and repeated at 100X dilutions if the plates were overgrown. Plasmid retention was calculated as the number of colonies grown on LB-carb plates divided by the number of colonies grown on LB only plates.

In vitro selection

After inoculating the library in LB or MC broth, the cultures were passaged by diluting at 20X into fresh media. LB cultures were grown in aerobic conditions with shaking and passaged every day for two weeks. MC cultures were grown in anaerobic conditions without shaking and passaged every two days for two weeks, since the cultures took more time to reach saturation compared to the LB condition.

In vivo selection

C57BL/6 gnotobiotic mice were housed in the Center for Clinical and Translational Metagenomics gnotobiotic core facility. Germ-free mice were orally gavaged with ~2 x 10^8 CFU of bacteria in a volume of 200 μL on Day 0. Mice inoculated with the library were separately housed. Fecal pellets were collected at 0.5, 0.75, 1.5, 1.75, 2.5, 3, 4, 7, 10, 14, 21, 25, and 28 days post-inoculation and stored at -80°C in 10% PBS buffer.

Colony PCR and Sanger sequencing

Individual colonies were isolated from stool samples streaked onto LB agar with carbenicillin (100 μg/mL). Colonies were grown overnight at 37°C in a 96-well plate with 200 μL of LB+carbenicillin. 0.8 μL of the culture was added to a total PCR reaction volume of 20 μL. The PCR mix (KAPA HiFi HotStart ReadyMix PCR Kit, Kapa Biosystems, Wilmington, MA) contained primers ver2_f/r (Table 5) that flanked the insert site. PCR amplicons were submitted for sequencing (Genewiz) and the insert sequence was mapped back to the *B. thetaiotaomicron* genome using BLASTn.

Primers for genotyping the *galK* locus on the *E. coli* genome are listed in Table 5. The presence or absence of IS2 in *galK* was confirmed using primers galK16_chk_f/r that flanked the expected insertion site in *galK*.

DNA extraction and PCR amplification of inserts for Illumina sequencing

DNA was extracted from collected samples in the *in vitro* experiment using the DNeasy Blood & Tissue Kit (Qiagen). Inserts were PCR amplified using primers ver2_f/r (Table 5) in KAPA HiFi HotStart Mix (Kapa Biosystems) and purified with Agencourt AMPure XP beads (Beckman Coulter, Indianapolis, IN) at a beads:sample volumetric ratio of 0.5:1. The amplicons were prepared for sequencing using the Nextera kit (Illumina, San Diego, CA).

For all fecal samples from the *in vivo* experiment, the QIAamp DNA Stool Mini Kit (Qiagen) was used. Isolated DNA was digested with PspXI and AvrII enzymes (New England Biolabs) prior to purification with QIAquick PCR Purification Kit (Qiagen) and subsequent PCR amplification with primers A_L and A_R (Table 5) in KAPA HiFi HotStart Mix. The PCR reaction was purified with AMPure beads at a beads:sample volumetric ratio of 0:5:1.

Initially, in our sequencing of the *in vitro* samples, we observed a high fraction (30-45%) of reads mapping to the backbone vector and fewer reads (20%) mapping to the *B. thetaiotaomicron* genome. Given the large (>3 kb) insert sizes of these libraries, traditional amplification methods evidently over-amplify the smaller vector backbone (2 kb), thereby overwhelming vectors containing actual genomic inserts. We therefore optimized the sample preparation protocol by incorporating a double digestion strategy prior to PCR amplification of the inserts. The two restriction sequences selected were the least common (of all available sites on the plasmid) in the *B. thetaiotaomicron* genome. We used restriction sites PspXI and AvrII that flanked the insert site on the backbone vector prior to PCR-amplification of the insert with primers A_L and A_R. These two enzymes had a minimal number of restriction sites (29 for PspXI and 62 for AvrII) in the *B. thetaiotaomicron* genome. Double digestion appears to eliminate the dominating band corresponding to the backbone vector. With this new protocol, in our subsequent *in vivo* sequencing, we observed <4% of reads mapping to the backbone vector and >90% of reads mapping to the *B. thetaiotaomicron* genome.

High-throughput sequencing and analysis of *in vitro* library selection data

Samples were sequenced on the MiSeq (Illumina) instrument at the Molecular Biology Core Facilities of the Dana-Farber Cancer Institute. Due to the PCR amplification protocol prior to optimization (see previous section), we observed large amounts of *E. coli* plasmid DNA in our

sequencing reads. To maximize the reads aligned to the *B. thetaiotaomicron* genome, we aggressively trimmed low quality bases and removed sequences mapping to the *E. coli* genome or with length shorter than 20 nt. The reference genome of *B. thetaiotaomicron* (NC_004663 and NC_004703) was downloaded from the NCBI nucleotide database. Due to the aggressive preprocessing of reads described, the length of trimmed sequences was shorter than 50 nt. Therefore, Bowtie (Langmead *et al*, 2009) was applied instead of Bowtie2 for higher sensitivity. Default parameters were used for building a Bowtie index with the *B. thetaiotaomicron* chromosome and plasmid sequences. Paired-end reads were aligned to the reference genome with parameter –X 300 using Bowtie. SAM files from the Bowtie alignment were converted to indexed and sorted BAM files using SAMtools (Li, et al. 2009). Cuffdiff (Trapnell *et al*, 2013) was applied to test differential representation of genes (i.e., the library grown in rich medium at time 0 versus the library grown in rich medium at day 7, and the library grown in MC medium at time 0 versus the library grown in MC medium at day 7).

High-throughput sequencing and processing of *in vivo* library selection data

B. thetaiotaomicron genomic DNA inserts were amplified from isolated *E. coli* plasmids using our improved PCR protocol (see above). After Nextera sequencing library preparation, paired-end reads of 101 bp length were generated on the HiSeq 2500 (Illumina) instrument at the Baylor College of Medicine Alkek Center for Metagenomics and Microbiome Research. All reads passed quality control (base quality >30) using FastQC (Babraham Bioinformatics). To eliminate plasmid DNA sequences in reads, the reads were trimmed using custom Perl scripts that removed all flanking regions matching 15bp of the plasmid DNA on the 5' and 3' ends of *B. thetaiotaomicron* insert fragment. Reads less than 20bp after trimming were discarded, and the others were matched as pairs with the forward read and reverse reads.

Sequencing reads were mapped onto the reference genome of *B. thetaiotaomicron* using Bowtie2 (Langmead *et al*, 2009). Default parameters were used for building the Bowtie2 index using the *B. thetaiotaomicron* chromosome and plasmid sequences, and for aligning reads to the reference sequence. SAM files generated from Bowtie2 alignment were converted to indexed and sorted BAM files using SAMtools (Li *et al*, 2009). In SAMtools, 'mpileup' with parameter '-B' was used to obtain the depth of coverage of the reference genome. Across all samples, the

mean of the mapped bases to the *B. thetaiotaomicron* genome was 1.17 x 10^9, with a minimum of 4.31 x 10^8 and maximum of 2.49 x 10^9 bases per sample.

Statistical analyses of *in vivo* selection data

Analyses were performed using custom scripts written in Matlab (MathWorks, Natick, MA).

The effective positional diversity (EPD), a genome-wide measure of library selection, was calculated using the formula:

$$\mathrm{EPD(t)} = e^{-\sum_{i=1}^{P} r_{ti} \ln r_{ti}}$$

Here, r_{ti} represents the fraction of reads at time t mapping to nucleotide i in a reference sequence totaling P nucelotides (e.g., the Bt genome).

The time-averaged relative abundance (TA-RA), a gene-level measure of library selection, was calculated using the formula:

$$\mathrm{TA\text{-}RA(g,t_1,t_2)} = \int_{t_1}^{t_2} \frac{f_g(\tau)d\tau}{t_2 - t_1}$$

Here, t_1 and t_2 denote the bounds of the time-interval of interest, and f_g represents a continuous-time function for gene g. The function f_g was estimated as follows. We fit a cubic smoothing spline, using the Matlab function csaps, applied to the log fold change in Fragments Per Kilobase per Million mapped reads (FPKM) for gene g at each time-point t (i.e., the FPKM value at time-point t divided by the FPKM value for the gene in the starting library). The smoothing spline was used to account for non-uniform temporal sampling and noise in the data.

The time-averaged normalized effective coverage (TA-NEC), a gene-level measure of coverage, was calculated using the formula:

$$\mathrm{TA\text{-}NEC(g,t_1,t_2)} = \int_{t_1}^{t_2} \frac{h_g(\tau)d\tau}{(t_2 - t_1)l_g}$$

Here, l_g denotes the length of gene g, and h_g represents a continuous-time function for gene g. The function h_g was estimated as follows. We fit a cubic smoothing spline, using the Matlab function csaps, applied to the effective coverage, EC(g,t) for the gene at each time-point:

$$\mathrm{EC(g,t)} = e^{-\sum_{i=s_g}^{s_g+l_g} r_{ti} \ln r_{ti}}$$

Here, s_g denotes the start of the gene.

To detect genes with significantly higher than expected selection, we performed a one-sided *t*-test on Box-Cox transformed TA-RA and TA-NEC values, and corrected for multiple hypothesis testing using the Matlab function mafdr. To estimate the relevant null hypotheses for the *t*-tests, while taking into account possible biases due to differential representation of genes in the input library, we used a robust regression algorithm (Matlab function robustfit) in which the input library value served as the independent variable, and the TA-RA or TA-NEC value served as the dependent variable.

Whole genome sequencing of isolated clones from the *in vivo* library selection

Whole genome sequencing of *E. coli* recipient isolates from seven Day 7 clones, six Day 28 clones, and two Day 28 luciferase control clones was performed on the MiSeq (Illumina) instrument after Nextera (Illumina) sequencing library preparation at the Molecular Biology Core Facilities of the Dana-Farber Cancer Institute. The raw data was processed with Millstone, which combines BWA alignment, GATK for BAM realignment and cleanup, and SnpEff for variant effect prediction. Reads were aligned to *E. coli* K-12 DH10B as well as MG1655 to identify any variants not in common with the starting library strain NEB Turbo. The average genome coverage of each sequenced strain ranged from 20 to 140X. Alignments were also performed against the F plasmid (which is present in the starting recipient strain) and a library plasmid with the expected insert (which we had characterized in Sanger-sequencing of individual clones).

Growth assays

Cells were pre-conditioned by growth in minimal media (M9) supplemented with 0.2% glucose prior to inoculation in growth assays. Then 1 μL of the culture was inoculated into a final volume of 200 μL of M9 supplemented with 0.2%, unless otherwise noted, of a sole carbon source, such as glucose, fructose, starch, lactose, galactose, or sucrose. When needed, MacConkey base agar with a final concentration of 1% lactose or galactose was also used to characterize lactose or galactose utilization.

Example 2 Identifying genes that improve microbial fitness in the healthy and inflamed gut using *in vivo* temporal functional metagenomics

The temporal functional metagenomics of the present invention will be used to systematically dissect the genetic determinants underling microbial colonization of the healthy and inflamed gut and to understand the long-term adaptation of microbes to the mammalian gut. This can help understand how these genetic determinants may play a role in the maintenance of healthy and development of diseased states.

The temporal functional metagenomics marries two massively parallel strategies to identify genes that are enriched among an expression library of metagenomic DNA during functional selection. The approach involves a) construction of a DNA fragment library from a donor microbe or a metagenomic source, b) transformation and heterologous expression of the library in a recipient microbe, c) application of selective pressure on the population over time, and d) characterization of the changes in library composition and abundance by deep-sequencing. When applying this method to gut colonization, we can quantitatively determine if and what genes from donor genomic or metagenomic sources can enhance the colonization of a less adapted recipient microbe.

Our results in Example 1 show a statistically significant enrichment for certain metabolic and transporter genes cloned into *Escherichia coli* from *Bacteroides thetaiotaomicron* that confer improved fitness during *in vitro* batch growth. We introduced genomic fragments of the well adapted commensal *Bacteroides thetaiotaomicron (Bt)* into a poorly-adapted non-pathogenic *Escherichia coli* and selected the population in a cohort of germ-free mice. Throughout 30 days of colonization and *in vivo* selection, we identified a number of *Bt* genes that were significantly enriched in the *E. coli* population. Deep sequencing of the population revealed that several *Bt* metabolic and transporter genes enhanced the ability of *E. coli* to better colonize the mouse gut,

including those for galactose transport, levan utilization, and membrane stability (Figures 3 – 7). These results highlight the potential for the temporal functional metagenomics to identify gut colonization factors (GCFs) and use them to improve retention of poorly adapted microbes.

The following research will be conducted: discover genetic determinants that promote microbial colonization of the healthy gut using genome-wide approaches; identify genetic factors in inflammation-associated microbiota that lead to retention in the chronically diseased gut and develop counteracting strategies; and characterize and understand the long-term adaptation and coevolution between microbial colonizers and the mammalian gut throughout neonatal development.

Several techniques in functional genomics, synthetic biology, deep-sequencing, and gnotobiotic mice models will be integrated. This research will accelerate the development of new and better therapies (e.g. engineered probiotics) against diseases of GI tract.

Aim 1. To systematically identify and characterize genes from commensal microbes that promote improved colonization in the healthy gut of gnotobiotic and conventional mice.

A temporal functional metagenomics approach will be used to gain insight into what types of genes from the microbiome gene pool can specifically improve microbial fitness of a gut colonizer. Specifically, I ask what genes when horizontally acquired can significantly increase the fitness of non-pathogenic *Escherichia coli* in the gut. We will use *E. coli,* a facultative anaerobe, as a suitable Gram-negative (GN) model because it is normally found in the gut at low levels, reflecting its lower fitness compared to other dominant commensals. Given the prevalence of lateral gene transfer in sharing genes between gut microbiota, it is also important to understand the ways in which pathogenic *Escherichia* strains (e.g. *E. coli* O157H7) can enhance their ability to remain in the gut through horizontal gene transfer.

We will first build and transform a genomic library composed of 3-5 kb uniformly fragmented genomic DNA from *C. butyricum (Cb)* into the probiotic recipient *L. plantarum (Lp)* or *Bacillus subtilis (Bs)*. The *Cb-Lp* library (or *Cb-Bs* library) will then be introduced by a single gavage to colonize the gut of healthy germ-free mice. *In vivo* selection of beneficial *C. butyricum* genes that improve gut colonization of *L. plantarum* (or *Bacillus subtilis*) will lead to a measureable enrichment in gene abundance in the microbiome population *in vivo*. We will extract the *in vivo* selected *Cb-Lp* library (or *Cb-Bs* library) from fecal pellets collected daily and use the Illumina deep-sequencing platform to identify and quantify genes that are enriched

during *in vivo* selection. Functional relevance of enriched GCF genes will be assigned through KEGG and COG metabolic pathway analysis.

This aim includes construction and *in vitro* characterization of *Cb-Bs* genomic library, and i*n vivo* mice selection of *Cb-Bs* library and deep-sequencing of fecal output.

GN/GP systems: Since *E. coli* will serve as a Gram-negative (GN) recipient for Gram-negative DNA sources, we will also construct a Gram-positive (GP) expression system in parallel using *Bacillus subtilis*, which is not a native colonizer of the gut. Alternatively, the Gram(+) bacteria *Lactobacillus plantarum* (ATCC 14917) may be used as a recipient. This system can be developed for other probiotics including *L. reuteri, L. rhamnosus,* and *B. longum.*

Clostridum butyricum will be used as a representative Gram-positive donor as it is a natural gut colonizer and has been shown to interfere with the growth of gut pathogen *C. difficile*. Woo et al. Inhibition of the cytotoxic effect of Clostridium difficile in vitro by Clostridium butyricum MIYAIRI 588 strain. *J Med Microbiol* 11: 1617-25 (2011).

Metagenomic donor DNA sources: Based on validation of the *Cb-Lp* library (or *Cb-Bs* library) generation protocol, we will further use metagenomic DNA from natural gut microflora as a source of donor DNA to build expression libraries in *Lp* or *Bs* to identify GCFs that can improve its gut retention.

Diet perturbations: In addition to the standard low-fat, plant polysaccharide-rich diet fed to the mice, other modified diets will be used to assess selection for different genes by different dietary regimens. These additional diets includes: high-fat, high-sugar (Western diet) and combinations of defined macro-nutrients (casein for protein, corn oil for fat, cornstarch for poly-saccharide, and sucrose for simple sugar) described previously. Faith et al., Predicting a human gut microbiota's response to diet in gnotobiotic mice. *Science* 333:101-104 (2011).

Alternative *in vivo* environments: We will further apply this *in vivo* temporal functional metagenomics strategy on conventionally grown mice with an already established microbiota to test if any genes from the donor source can enhance the establishment and infiltration of a poor-colonizer into a native microbial gut community. We will also assess the native microbiota using deep 16S sequencing to assess any changes in native microbiome composition.

Additional genes and gene families that are enriched during *in vivo* selection from different genomic and metagenomic sources under various environmental perturbations will be

discovered. We further anticipate that different gut colonization factors (GCFs) will be selected depending on different dietary and nutritional inputs given to the mammalian animal. Differences in library enrichment between germ-free and conventional mice will highlight the different selective forces of a naïve un-colonized gut and a richly established gut. Further individual characterizations will be done on clones that are significantly enriched to determine the molecular basis of its positive effects on gut colonization.

Aim 2. To identify specific genetic factors in inflammation-associated microbiota that lead to colonization and retention of the chronically diseased gut and to develop strategies to reverse the diseased state by enhancing better re-colonization of the dysbiotic gut.

Recent studies have found that depletion in *Bacteroidetes* and certain *Clostridia* groups are linked to inflammatory bowel diseases (IBD) and ulcerative colitis. Frank et al., Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. USA* 104: 13780-5 (2007). Conversely, *Bacillus* and *Enterobacteriales* are enriched in these diseased GI tracts. What are the genetic determinants that allow inflammation-associated microbiota to colonize and remain in the gut? Are there ways to reverse this negative outcome? To tackle this problem, murine models of IBD will be used to conduct a genome-wide analysis of genetic determinants of colonization in the inflamed gut.

This aim will include construction and characterization of H/I-D/R (healthy/inflamed - donor/recipient) genomic libraries, *in vivo* mice selection of H/I-D/R library and sequencing of fecal output, and i*n vivo* selection using *C. diff* model to test pathogen exclusion.

Healthy/Inflamed Donor/Recipient (HIDR) study: Our donor library will be built using fecal samples from healthy human volunteers and from IBD patients. These libraries will be put into both our Gram-positive (*B. subtilis*) and Gram-negative (*E. coli*) expression systems. We will transplant these libraries into two C57BL/6 mice groups, a healthy gut group and an inflamed gut group. Prior to microbiota inoculation, well-established chemical routes will be used to induce chronic colitis by addition of dextran sodium sulfate (DSS) to the drinking water. Our temporal functional metagenomics will then be applied to determine 1) whether there are differences in gene enrichment between metagenomic DNA originating from healthy versus IBD patients in the healthy mouse gut and 2) how those differences change in the chronically inflamed mice. This cross-donor cross-environment comparison will highlight the differences in

gene composition of health and IBD microbiome and the ability of health and inflamed gut to specifically select for those alleles.

Therapeutics to reverse *C. diff* infections: *Clostridum difficile* infection (CDI) is a growing clinical challenge with very high rates of reoccurrence and is responsible for >14,000 deaths per year in the US. In order to reverse colitis such as those caused by persistent CDI, we aim to develop specific strategies to improve the fitness of probiotic or engineered commensal bacteria that can more effectively competing with the pathogen. In fact, a similar strategy through fecal transplantation to recolonize the diseased gut with microbiota from healthy donors has shown clinical success in reversing CDI. Brandt et al., Fecal microbiota transplantation for recurrent *clostridium difficile* infection. *J Clin Gastroenterol.* 45:159-67 (2011). Our goal is to enhance probiotic infiltration and establishment amongst the diseased microflora in the gut and to outcompete the pathogen. We will first isolate high-fitness (HF) strains enriched in the cohort that contained IBD donor DNA and selected in the colitic gut *in vivo* from the HIDR study. These HF strains represent clones that can effectively grow in the inflamed gut. We will then reinoculate HF strains into mice infected with persistent *C. difficile* and measure the ability of the HF strains to compete with *C. diff.* In parallel, we will inoculate a GN/GP library from a healthy donor into the *C. diff*-associated murine model and apply the temporal functional metagenomics to identify commensal genes that are enriched in the presence of a persistent infection. We will measure *C. difficile* titers to assess selection against the pathogen and enrichment of our desired higher fitness strains. In addition to assessing the microbiota throughout colonization, other murine immunological assays can also be done to measure serum IgG, IgA, and T-cell activation levels.

The genetic differences that enable certain microbes to better propagate in the inflamed gut than the healthy gut will be dissected. These alleles will be used to develop strategies to counteract reoccurring *C. difficile* infections using enhanced probiotics to more effectively compete with pathogens *in vivo*.

Aim 3: To characterize the long-term adaptation of microbial gut colonizers in gnotobiotic mice as a model for understanding neonatal colonization and host-microbe coevolution throughout development and maturation.

Recent studies have highlighted the intimate link between microbial colonization and intestinal maturation during neonatal development. Koenig et al., Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci.* 108:4578-4585 (2011). The gut microbiome plays a key role in sensitizing the immune system to properly distinguish harmless bacteria from pathogens. While microbial composition in the gut fluctuates significantly within the first year of birth, the microbiota stabilizes soon after and appears to become specifically associated with the individual over time, forming a "personalized microbiome." Little is however know about the long-term adaptation and evolution of these microbes in a living organism. A laboratory example of microbial long-term evolution (LTE) spanned two decades, which constitute only a fraction of the time that microbes inhabit the body during an average lifespan. Blount et al., Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature* 489:513-518 (2012). There are some fundamental questions about LTE *in vivo*. What is the rate of microbial evolution in a new and naïve gut environment? What are the mutations that arise over time to improve microbial fitness? How reproducible is evolution *in vivo*? If and how does the population diversify spatially and metabolically to better utilize different resource? To address these questions, long-term evolution of *B. thetaiotaomicron* in the gut of live mice will be studied with new technologies in next-generation sequencing and analysis.

This Aim will include conducting long-term evolution experiments *in vivo*, isolate resulting clones and perform deep-sequencing and analysis of evolved strains, and phenotypic assays for clone fitness versus wild-type.

Long-Term Evolution *in vivo*: The use of gnotobiotic murine model will allow us to start with a defined and mono-associated microbial community, which we otherwise do not have in conventionally raised mice. *B. theta* is chosen because it is a dominant gut microbe in the Bacteroides phylum with a large array of metabolic capabilities. Xu et al., A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. *Science*. 299:2074-6 (2003). In addition, the microbe appears to provide protective effects against colonization of certain pathogens. The *B. theta* genome is sequenced, well-annotated, and amenable to manipulation through variety of genetic technique for downstream analysis. Goodman et al. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe*. 6:279–289 (2009). We will inoculate germ-free mice with *B. theta* and track the evolution of this microbe in the gut,

both phenotypically and genotypically. Isolates will be obtained from fecal matter at regular intervals, with a subset selected for more detailed analysis. We will use whole-genome sequencing (WGS) of microbial isolates to identify the molecular changes that have taken place in the genome and temporally trace the origins of any adaptive radiation. We will further analyze the evolved isolates using Biolog Phenotype MicroArrays to assess phenotypic and metabolic changes. Evolved isolates will be competed with the ancestral strain to determine changes in fitness both *in vitro* and *in vivo*. We can further carry out transcriptomics to determine changes in transcriptional profiles.

Reproducibility and vertical transmission: We will inoculate three mice with the identical ancestral strain to test for reproducibility of evolution. The LTE will also be conducted throughout the lifespan of the first-generation of mice and re-inoculated into a second generation upon natural termination of the first. Several important evolutionary phases are likely at play including initial colonization, short-term adaptation, long-term competition, and re-adaptation to a new host by vertical transmission.

Diet & antibiotic resistance: Host diet and nutritional availability is a key factor during LTE. We will maintain the same diet throughout this LTE experiment using a defined mouse chow (DMC) composed of defined macronutrients – casein for protein, corn oil for fat, cornstarch for polysaccharide, and sucrose for simple sugar, amino acids and vitamins. A defined rich diet is important for limiting batch-to-batch variation in mouse chow to control fluctuations in selective pressure and allow us to better measure nutritional utilization. Presently, an alarming level of antibiotics including fluoroquinolones (Ciprofloxacin) is being used during food processing leading to subinhibitory concentrations in our foods. Exposure to these antibiotics can lead to adaptive mutations by gut microbes to develop resistance, which reduces clinical efficacy of antibiotics. We will determine whether sub-inhibitory concentrations of Cipro in drinking water of our *B. theta*-associated mice can lead to increased resistance phenotypes. Isolates will undergo WGS analysis to determine resistance alleles and MIC levels will be measured *in vitro*.

In vitro LTE: Concurrent to the *in vivo* LTE experiment, we will conduct a parallel *in vitro* LTE experiment via daily passaging in an anaerobic chamber. The same ancestor and defined mouse chow will be used. The goal is to identify differences between host-microbe interactions vs. simply batch culture growth.

We expect to gain significant insights into the mutations that commensal bacteria access to better adapt to the host and provide a clearer picture of evolution in action. This work will form the basis of subsequent studies to investigate the evolution of pathogens and acquisition of multi-drug resistance.

Example 3 Introduction of recombinant high-fitness bacteria to the mammalian gut

The aim of this study is to determine the mechanism by which gut colonization factors (GCF) improve gut retention and affect microbe-microbe dynamics, and demonstrate that probiotics engineered with GCFs can more effectively infiltrate the established microbiota of a mammalian gut.

Based on the GCFs identified in Example 1 and Example 2, we will apply detailed characterization to quantify the level of fitness benefit that different GCFs impart on the recipient probiotic bacteria. These characterizations will initially be performed *in vitro*. Our results (Example 1) show that certain metabolic factors including those associated with polysaccharide utilization can improve gut-retention and provide colonization resistance against other bacteria (potentially pathogens). We will apply phenotypic assays including growth profiling of cells containing GCFs on various defined nutritional media (e.g. polysaccharides and glycans found in the diet). Co-culturing competition with fluorescently labeled wild-type *L. plantarum* will quantitatively determine the degree to which GCFs improve probiotic fitness.

We hypothesize that GCFs will not only improve colonization of probiotic bacteria, but can potentially impact the compositional structure of the gut microflora. The goal is to enhance probiotic infiltration and establishment amongst the natural microflora to improve function and stability of the gut microbiome. Toward this effort, we will introduce different GCF-enriched "high-fitness" (HF) probiotic bacteria to a cohort of conventionally associated mice that contain a complex microflora. This synthetic biology approach will take advantage of the fact that the HF probiotics will be engineered to more competitively infiltrate the native microflora, which is a current challenge in probiotic therapies. The HF strains will be introduced into mice specifically associated with a pre-defined microflora (e.g. Altered Schaedler Flora) or a conventional microflora. We will profile the impact of the HF probiotic on the composition and abundance of the established microbiota using deep 16S sequencing based on well-established protocols. Dietary conditions may be changed as an additional variable to assess its impact on

altering infiltration of foreign microbes into the native microbiome. 16S profiling of the native microbiome will inform population-level changes post-inoculation of a new microbe to determining whether a probiotic strain can significantly alter the native microbiota.

This work aims to assess how individual genes (or gene groups) can directly impact the rest of the gut microbiota to functionally dissecting these complex microbe-microbe interactions.

This approach is culture-independent, meaning that "unculturable" microbes and metagenomic sources can be used as donor DNA for our system without the need for cultivation in the lab. Second, our approach identifies gain-of-function phenotypes, in contrast to other loss-of-function methods using genome-wide transposon mutagenesis. Third, the proposed approach applies functional metagenomics to an *in vivo* system. Finally, this approach tracks an important temporal component of gut colonization by measuring allele frequencies over time to monitor the dynamic changes during probiotic colonization.

References

Bäckhed F, Ley RE, Sonnenburg JL, Peterson DA & Gordon JI (2005) Host-Bacterial Mutualism in the Human Intestine. *Science (80-.).* **307:** 1915–1920

Bar-Joseph Z, Gerber G, Jaakkola T, Gifford D & Simon I (2003) Continuous representations of time series gene expression data. *J. Comput. Biol.* **3-4:** 341–356

Bjursell MK, Martens EC & Gordon JI (2006) Functional genomic and metabolic studies of the adaptations of a prominent adult human gut symbiont, Bacteroides thetaiotaomicron, to the suckling period. *J. Biol. Chem.* **281:** 36269–79

Blattner FR, Plunkett G, Bloch C a, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick H a, Goeden M a, Rose DJ, Mau B & Shao Y (1997) The complete genome sequence of Escherichia coli K-12. *Science* **277:** 1453–62

Buckstein MH, He J & Rubin H (2008) Characterization of nucleotide pools as a function of physiological state in Escherichia coli. *J. Bacteriol.* **190:** 718–26

Burns S & Hull S (1998) Comparison of Loss of Serum Resistance by Defined Lipopolysaccharide Mutants and an Acapsular Mutant of UropathogenicEscherichia coli O75: K5. *Infect. Immun.* **66:** 4244–4253

Byers JP & Sarver JG (2009) Pharmacokinetic Modeling. In *Pharmacology: Principles and Practice*, Hacker M Messer W & Bachmann K (eds) pp 201–277. Elsevier

Cecchini DA, Laville E, Laguerre S, Robe P, Leclerc M, Doré J, Henrissat B, Remaud-Siméon M, Monsan P & Potocki-Véronèse G (2013) Functional metagenomics reveals novel pathways of prebiotic breakdown by human gut bacteria. *PLoS One* **8:** e72766

Chiang S & Mekalanos J (1999) rfb mutations in Vibrio cholerae do not affect surface production of toxin-coregulated pili but still inhibit intestinal colonization. *Infect. Immun.* **67:** 976–980

Cho I, Blaser M. The human microbiome: at the interface of health and disease, *Nat Rev Genet* 13:260-70 (2012).

Culligan EP, Sleator RD, Marchesi JR & Hill C (2012) Functional metagenomics reveals novel salt tolerance loci from the human gut microbiome. *ISME J.* **6:** 1916–25

Dethlefsen L, McFall-Ngai M & Relman DA (2007) An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature* **449:** 811–8

Esvelt K, Wang HH. Genome-scale engineering for systems and synthetic biology. *Mol Sys Biol.* 9:641, (2013).

Freter R, Brickner H, Botney M, Cleven D & Aranki A (1983) Mechanisms that control bacterial populations in continuous-flow culture models of mouse large intestinal flora. *Infect. Immun.* **39:** 676

Gaudin HM & Silverman PM (1993) Contributions of promoter context and structure to regulated expression of the F plasmid fraV promoter in Escherichia coii K-12. *Mol. Microbiol.* **8:** 335–342

Gill S, Pop M, DeBoy R & Eckburg P (2006) Metagenomic analysis of the human distal gut microbiome. *Science* **312:** 1355–1359

Gloux K, Berteau O, El Oumami H, Béguet F, Leclerc M & Doré J (2011) A metagenomic β-glucuronidase uncovers a core adaptive function of the human intestinal microbiome. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl :** 4539–46

Goodman AL, McNulty NP, Zhao Y, Leip D, Mitra RD, Lozupone C a, Knight R & Gordon JI (2009) Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe* **6:** 279–89

Guan L, Murphy FD & Kaback HR (2002) Surface-exposed positions in the transmembrane helices of the lactose permease of Escherichia coli determined by intermolecular thiol cross-linking. *Proc. Natl. Acad. Sci. U. S. A.* **99:** 3475–80

Haardt M, Kempf B, Faatz E & Bremer E (1995) The osmoprotectant proline betaine is a major substrate for the binding-protein-dependent transport system ProU of Escherichia coli K-12. *Mol. Gen. Genet.* **246:** 783–6

Hooper L V (2004) Bacterial contributions to mammalian gut development. *Trends Microbiol.* **12:** 129–34

Huttenhower C, Gevers D, Knight R & HMP (2012) Structure, function and diversity of the healthy human microbiome. *Nature* **486:** 207–14

Isaacs FJ, Carr PA, Wang HH, et al. Precise manipulation of chromosomes in vivo enables genome-wide codon replacement. *Science* 333: 348-53 (2011).

Jost L (2006) Entropy and diversity. *Oikos* **113:** 363–375

Juge N (2012) Microbial adhesins to gastrointestinal mucus. *Trends Microbiol.* **20:** 30–9

Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, Muñiz-Rascado L, Bonavides-Martinez C, Paley S, Krummenacker M, Altman T, Kaipa P, Spaulding A, Pacheco J, Latendresse M, Fulcher C, Sarker M, Shearer AG, Mackie A, Paulsen I, Gunsalus RP, et al (2011) EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Res.* **39:** D583–90

Kneidinger B, Marolda C, Graninger M, Zamyatina A, McArthur F, Kosma P, Valvano MA & Messner P (2002) Biosynthesis pathway of ADP-L-glycero-β-D-manno-heptose in Escherichia coli. *J. Bacteriol.* **184:** 363–369

Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, Knight R. Experimental and analytical tools for studying the human microbiome. *Nat Rev Genet.* 13:47-58 (2012).

Lakhdari O, Cultrone A, Tap J, Gloux K, Bernard F, Ehrlich SD, Lefèvre F, Doré J & Blottière HM (2010) Functional metagenomics: a high throughput screening method to decipher microbiota-driven NF-κB modulation in the human gut. *PLoS One* **5:** 1–10

Langmead B, Trapnell C, Pop M & Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10:** R25

Lee SM, Donaldson GP, Mikulski Z, Boyajian S, Ley K & Mazmanian SK (2013) Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature* **501:** 426–9

Ley RE, Peterson DA & Gordon JI (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124:** 837–48

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G & Durbin R (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–9

Maltby R, Leatham-Jensen MP, Gibson T, Cohen PS & Conway T (2013) Nutritional basis for colonization resistance by human commensal Escherichia coli strains HS and Nissle 1917 against E. coli O157:H7 in the mouse intestine. *PLoS One* **8:** e53957

Martens EC, Chiang HC & Gordon JI (2008) Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host Microbe* **4:** 447–57

Mee M, Wang HH. Engineering ecosystems and synthetic ecologies. *Mol. BioSyst.* **8:**2470-83 (2012).

Mehra R & Drabble W (1981) Dual Control of the gua Operon of Escherichia coli K12 by Adenine and Guanine Nucleotides. *J. Gen. Microbiol.* **123:** 27–37

Van Opijnen T, Bodi KL & Camilli A (2009) Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat. Methods* **6:** 767–72

Pang B, McFaline JL, Burgis NE, Dong M, Taghizadeh K, Sullivan MR, Elmquist CE, Cunningham RP & Dedon PC (2012) Defects in purine nucleotide metabolism lead to substantial incorporation of xanthine and hypoxanthine into DNA and RNA. *Proc. Natl. Acad. Sci. U. S. A.* **109:** 2319–24

Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss J a, Bonazzi V, McEwen JE, Wetterstrand K a, Deal C, Baker CC, Di Francesco V, Howcroft TK, Karp RW, Lunsford RD, Wellington CR, Belachew T, Wright M, Giblin C, David H, et al (2009) The NIH human microbiome project. *Genome Res.* **19:** 2317–2323

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T & others (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464:** 59–65

Rakoff-Nahoum S, Coyne MJ & Comstock LE (2014) An ecological network of polysaccharide utilization among human intestinal symbionts. *Curr. Biol.* **24:** 40–9

Ratnayake-Lecamwasam M, Serror P, Wong K-W & Sonenshein A (2001) Bacillus subtilis CodY represses early-stationary-phase genes by sensing GTP levels. *Genes ...* **15:** 1093–1103

Sato K, Kumita W, Ode T, Ichinose S, Ando A, Fujiyama Y, Chida T & Okamura N (2010) OmpA variants affecting the adherence of ulcerative colitis-derived Bacteroides vulgatus. *J. Med. Dent. Sci.* **57:** 55–64

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ & Weber CF (2009) Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75:** 7537–7541

Sommer MOA, Dantas G & Church GM (2009) Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* **325:** 1128–31

Sonnenburg ED, Zheng H, Joglekar P, Higginbottom SK, Firbank SJ, Bolam DN & Sonnenburg JL (2010) Specificity of polysaccharide use in intestinal bacteroides species determines diet-induced microbiota alterations. *Cell* **141:** 1241–52

Sonnenburg JL, Chen CTL & Gordon JI (2006) Genomic and metabolic studies of the impact of probiotics on a model gut symbiont and host. *PLoS Biol.* **4:** e413

Sonnenburg JL, Xu J, Leip DD, Chen C-H, Westover BP, Weatherford J, Buhler JD & Gordon JI (2005) Glycan foraging in vivo by an intestine-adapted bacterial symbiont. *Science* **307:** 1955–9

Soulas C, Baussant T, Aubry J-P, Delneste Y, Barillat N, Caron G, Renno T, Bonnefoy J-Y & Jeannin P (2000) Cutting Edge: Outer Membrane Protein A (OmpA) Binds to and Activates Human Macrophages. *J. Immunol.* **165:** 2335–2340

Soupene E, Heeswijk WC Van, Plumbridge J, Stewart V, Bertenthal D, Lee H, Prasad G, Paliy O, Charernnoppakul P & Kustu S (2003) Physiological Studies of Escherichia coli Strain MG1655 : Growth Defects and Apparent Cross-Regulation of Gene Expression. *J. Bacteriol.* **185:** 5611–5626

Stappenbeck TS, Hooper L V & Gordon JI (2002) Developmental regulation of intestinal angiogenesis by indigenous microbes via Paneth cells. *Proc. Natl. Acad. Sci. U. S. A.* **99:** 15451–5

Tasse L, Bercovici J, Pizzut-Serin S, Robe P, Tap J, Klopp C, Cantarel BL, Coutinho PM, Henrissat B, Leclerc M, Doré J, Monsan P, Remaud-Simeon M & Potocki-Veronese G (2010) Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. *Genome Res.* **20:** 1605–12

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL & Pachter L (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31:** 46–53

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R & Gordon JI (2007) The human microbiome project. *Nature* **449:** 804–10

Usui Y, Hirasawa T, Furusawa C, Shirai T, Yamamoto N, Mori H & Shimizu H (2012) Investigating the effects of perturbations to pgi and eno gene expression on central carbon metabolism in Escherichia coli using (13)C metabolic flux analysis. *Microb. Cell Fact.* **11:** 87

Valvano M, Messner P & Kosma P (2002) Novel pathways for biosynthesis of nucleotide-activated glycero-manno-heptose precursors of bacterial glycoproteins and cell surface polysaccharides. *Microbiology* **148:** 1979–1989

Walker AW, Duncan SH, Louis P & Flint HJ (2014) Phylogeny, culturing, and metagenomics of the human gut microbiota. *Trends Microbiol.* **22:** 267–74

Wang HH, Isaacs FJ, Carr PA, Sun ZZ, Xu G, Forest CR, Church GM. Programming cells by multiplex genome engineering and accelerated evolution. *Nature* 460:894-8, (2009).

Wang HH, Kim HB, Cong L, Jeong JH, Bang D, Church GM. Genome-scale Promoter Engineering by Co-Selection MAGE. *Nat Methods* 9: 591-3 (2012).

Wang L, Huang H, Nguyen HH, Allen KN, Mariano PS & Dunaway-Mariano D (2010) Divergence of biochemical function in the HAD superfamily: D-glycero-D-manno-heptose-1,7-bisphosphate phosphatase (GmhB). *Biochemistry* **49:** 1072–81

Weickert MJ & Adhya S (1993) Control of transcription of gal repressor and isorepressor genes in Escherichia coli. *J. Bacteriol.* **175:** 251–8

Weickert MJ & Adhyat S (1992) Isorepressor of the gal Regulon in Escherichia coli. *J. Mol. Biol.* **226:** 69–83

Wexler HM, Tenorio E & Pumbwe L (2009) Characteristics of Bacteroides fragilis lacking the major outer membrane protein, OmpA. *Microbiology* **155:** 2694–706

Winson MK, Swift S, Hill PJ, Sims CM, Griesmayr G, Bycroft BW, Williams P & Stewart GS (1998) Engineering the luxCDABE genes from Photorhabdus luminescens to provide a bioluminescent reporter for constitutive and promoter probe plasmids and mini-Tn5 constructs. *FEMS Microbiol. Lett.* **163:** 193–202

Xu J, Bjursell MK, Himrod J, Deng S, Carmichael LK, Chiang HC, Hooper L V & Gordon JI (2003) A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. *Science* **299:** 2074–6

The scope of the present invention is not limited by what has been specifically shown and described hereinabove. Those skilled in the art will recognize that there are suitable alternatives to the depicted examples of materials, configurations, constructions and dimensions. Numerous references, including patents and various publications, are cited and discussed in the description of this invention. The citation and discussion of such references is provided merely to clarify the description of the present invention and is not an admission that any reference is prior art to the invention described herein. All references cited and discussed in this specification are incorporated herein by reference in their entirety. Variations, modifications and other implementations of what is described herein will occur to those of ordinary skill in the art without departing from the spirit and scope of the invention. While certain embodiments of the present invention have been shown and described, it will be obvious to those skilled in the art that changes and modifications may be made without departing from the spirit and scope of the invention. The matter set forth in the foregoing description and accompanying drawings is offered by way of illustration only and not as a limitation.

What is claimed is:

1. A method of identifying genes that enhance bacterial fitness in the gastrointestinal (GI) tract, the method comprising the steps of:
(a) constructing a genomic or metagenomic library comprising fragments of the genome of at least one donor bacterium;
(b) introducing the library to recipient bacteria;
(c) introducing the recipient bacteria carrying the library into the GI tract of a mammal;
(d) taking stool samples of the mammal at different time points T_1 through T_n where n is an integer;
(e) isolating DNA from the stool samples of step (d); and
(f) sequencing the DNA of step (e).

2. The method of claim 1, wherein after step (f) at least one gene is identified to enhance bacterial fitness in the mammalian GI tract, where the gene has an abundance at a time point of step (d) that is at least 10 fold of its abundance in the library.

3. The method of claim 1, wherein the donor bacterium belongs to genera *Bacteroides* or *Clostridium*.

4. The method of claim 3, wherein the donor bacterium is *Bacteroides thetaiotaomicron* or *Clostridium butyricum*.

5. The method of claim 1, wherein the at least one donor bacterium is from a natural gut microbiota.

6. The method of claim 5, wherein the natural gut microbiota is from the gut of a healthy mammal, or from the gut of a mammal with inflamed GI tract.

7. The method of claim 6, wherein the mammal with inflamed GI tract has an inflammatory bowel disease (IBD) or irritable bowel syndrome (IBS).

8. The method of claim 7, wherein the IBD is ulcerative colitis or Crohn's disease.

9. The method of claim 6, wherein the healthy mammal, or the mammal with inflamed GI tract, is a human subject.

10. The method of claim 1, wherein the recipient bacteria belong to phyla Bacteroidetes, Firmicutes, Proteobacteria, or Actinobacteria.

11. The method of claim 1, wherein the recipient bacteria belong to genera *Bacteroides*, *Clostridium, Escherichia, Bacillus, Lactobacillus*, or *Bifidobacterium*.

12. The method of claim 1, wherein the recipient bacteria are *Escherichia coli (E. coli)*, *Bacillus subtilis (B. subtilis)*, *Lactobacillus plantarum (L. plantarum), Lactobacillus reuteri (L. reuteri), Lactobacillus rhamnosus (L. rhamnosus), Bifidobacterium longum (B. longum), Bacteroides thetaiotaomicron*, *Clostridium butyricum, Bacteroides fragilis, Bacteroides melaninogenicus, Bacteroides oralis, Bacteroides amylophilus, Clostridium butyricum, Clostridium perfringens, Clostridium tetani,* or *Clostridium septicum.*

13. The method of claim 1, wherein the mammal is a mouse.

14. The method of claim 13, wherein the mouse is germ-free before introduction of the recipient bacteria.

15. The method of claim 13, wherein the mouse is healthy.

16. The method of claim 13, wherein the mouse has inflamed GI tract.

17. The method of claim 1, wherein the DNA is fragmented by sonication or digestion by at least one restriction enzyme.

18. The method of claim 1, wherein the fragments of the donor bacterial genome are under the control of a constitutive promoter in the recipient bacteria.

19. The method of claim 1, wherein the fragments of the donor bacterial genome are under the control of an inducible promoter in the recipient bacteria.

20. The method of claim 1, wherein the time points T_1 through T_n range from about 0 day to about 30 days after step (c).

21. The method of claim 1, wherein the DNA is sequenced by deep sequencing, or Sanger sequencing.

22. A probiotic composition comprising recombinant bacteria comprising a gene encoding a protein selected from the group consisting of a glycoside hydrolase, a galactokinase and a glucose/galactose transporter, wherein the gene is heterologous, or the gene is endogenous and overexpresses the protein.

23. The probiotic composition of claim 22, wherein the gene is under the control of a constitutive promoter.

24. The probiotic composition of claim 22, wherein the gene is under the control of an inducible promoter.

25. The probiotic composition of claim 22, wherein the protein is wild-type or a mutant.

26. The probiotic composition of claim 22, wherein the gene is truncated at the 5' end by from about 1 base pair (bp) to about 50 bp from the start codon.

27. The probiotic composition of claim 22, wherein the protein is *Bacteroides thetaiotaomicron* glycoside hydrolase and its gene is truncated at the 5' end by about 4 bp from the start codon.

28. The probiotic composition of claim 22, wherein the gene is integrated into the bacterial chromosome or is episomal.

29. The probiotic composition of claim 22, wherein the recombinant bacteria are capable of metabolizing both sucrose and galactose.

30. The probiotic composition of claim 22, wherein the composition is selected from the group consisting of a food composition, a beverage composition, a pharmaceutical composition, and a feedstuff composition.

31. The probiotic composition of claim 22, wherein the composition is a dairy product.

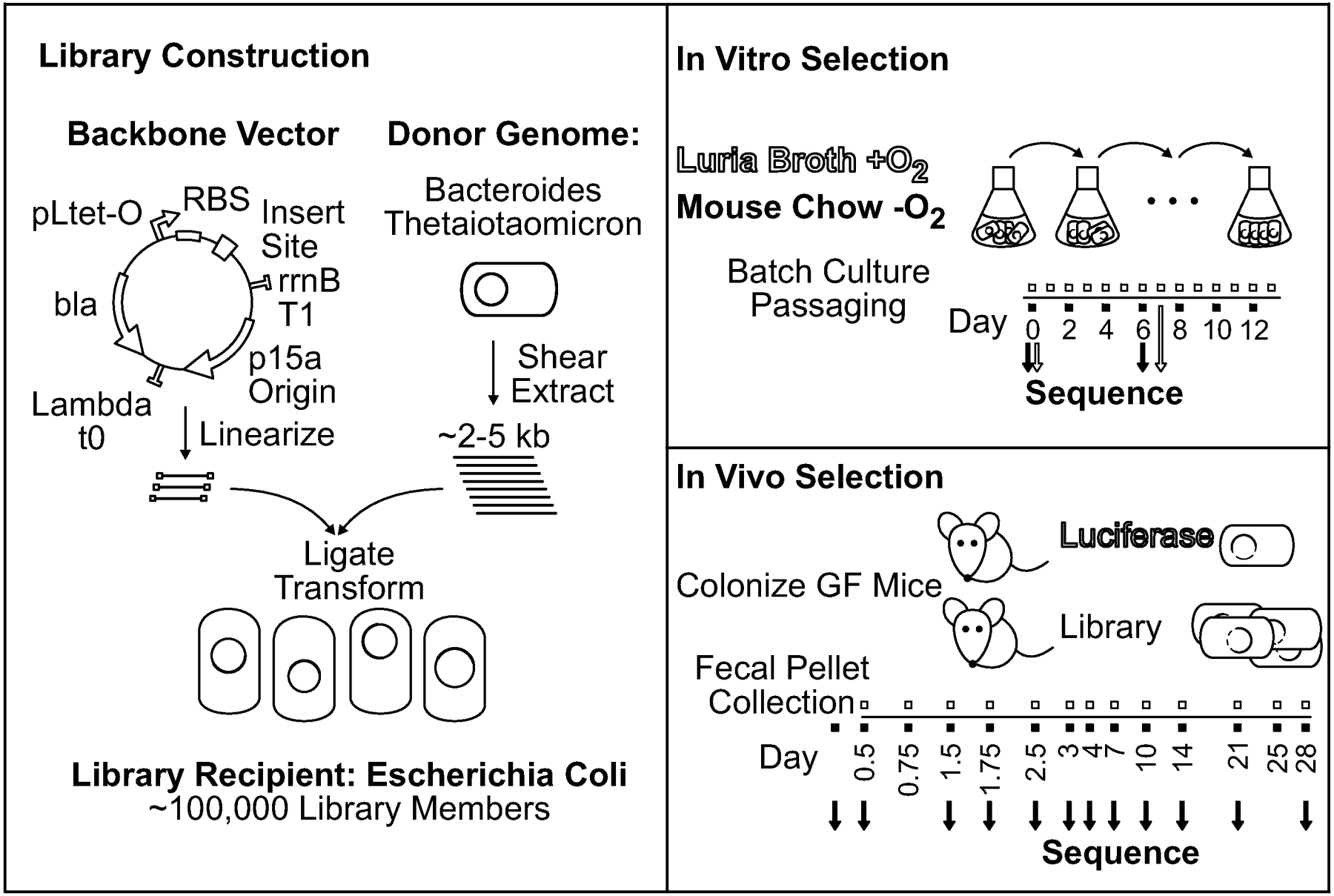
Library Construction
Backbone Vector
Donor Genome:
pLtet-O
RBS
Insert Site
rrnB T1
bla
p15a Origin
Lambda t0
Linearize
Bacteroides Thetaiotaomicron
Shear Extract
~2-5 kb
Ligate Transform
Library Recipient: Escherichia Coli
~100,000 Library Members
In Vitro Selection
Luria Broth +O2
Mouse Chow -O2
Batch Culture Passaging
Day
0 2 4 6 8 10 12
Sequence
In Vivo Selection
Luciferase
Colonize GF Mice
Library
Fecal Pellet Collection
Day
0.5 0.75 1.5 1.75 2.5 3 4 7 10 14 21 25 28
Sequence

FIG. 1

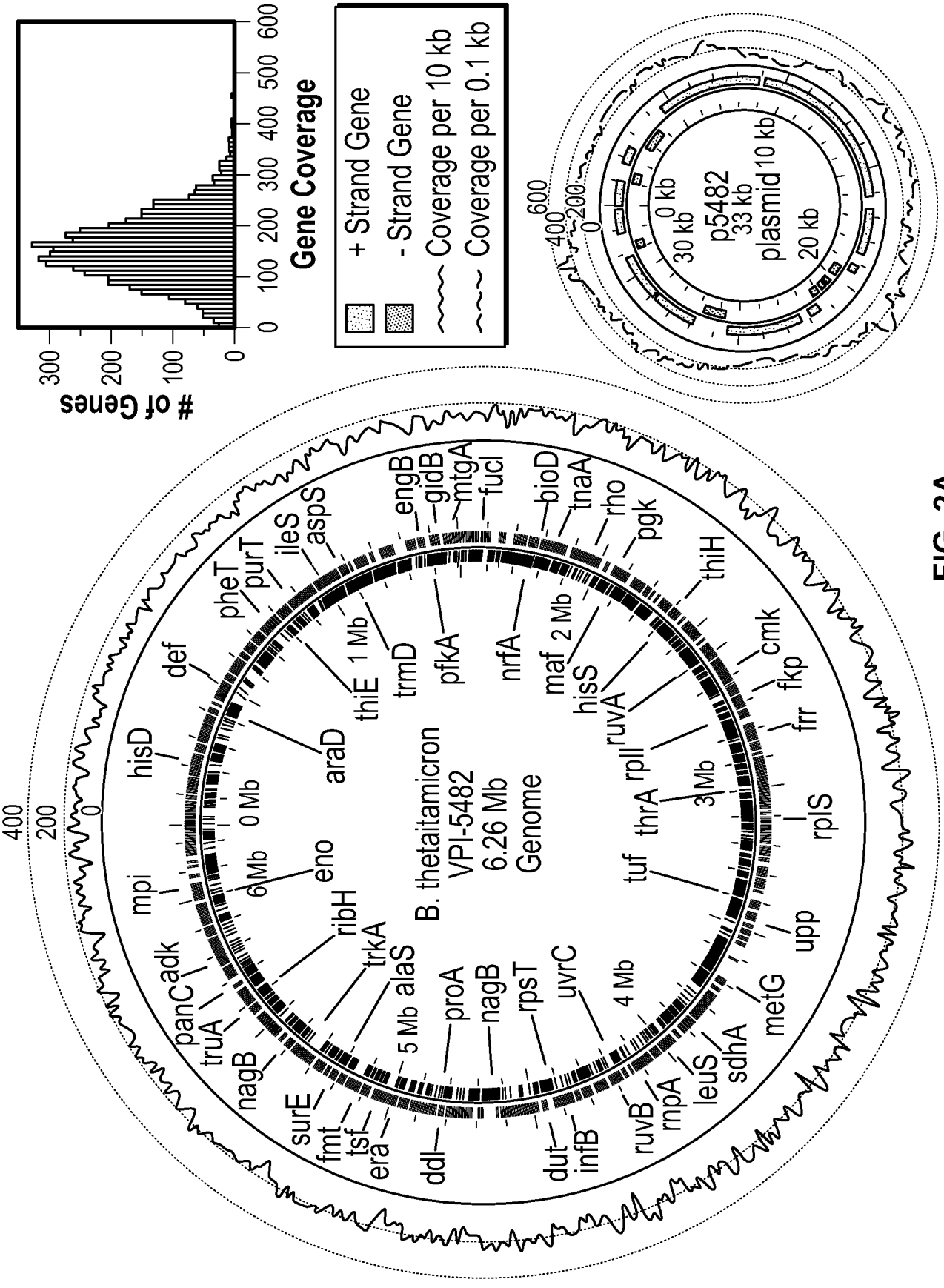
Gene Coverage
# of Genes
+ Strand Gene
- Strand Gene
Coverage per 10 kb
Coverage per 0.1 kb
p5482
33 kb
plasmid
B. thetaitamicron
VPI-5482
6.26 Mb
Genome
0 Mb
1 Mb
2 Mb
3Mb
4 Mb
5 Mb
6Mb
hisD
def
pheT
purT
ileS
aspS
engB
gidB
mtgA
fucI
bioD
tnaA
rho
pgk
thiH
cmk
fkp
frr
rplS
upp
metG
sdhA
leuS
rnpA
ruvB
infB
dut
uvrC
rpsT
nagB
proA
pfkA
trmD
thiE
araD
eno
ribH
trkA
alaS
surE
fmt
tsf
era
ddl
nagB
truA
panC
adk
mpi
nrfA
maf
hisS
ruvA
rplI
thrA
tuf

FIG. 2A

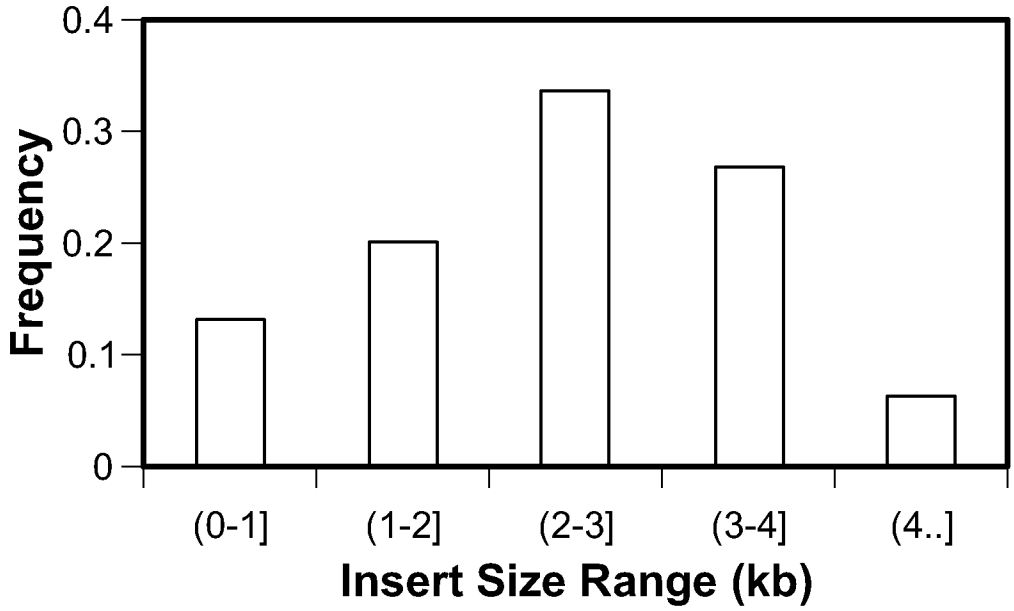
0.4
0.3
0.2
0.1
0
Frequency
(0-1]
(1-2]
(2-3]
(3-4]
(4..]
Insert Size Range (kb)

FIG. 2B

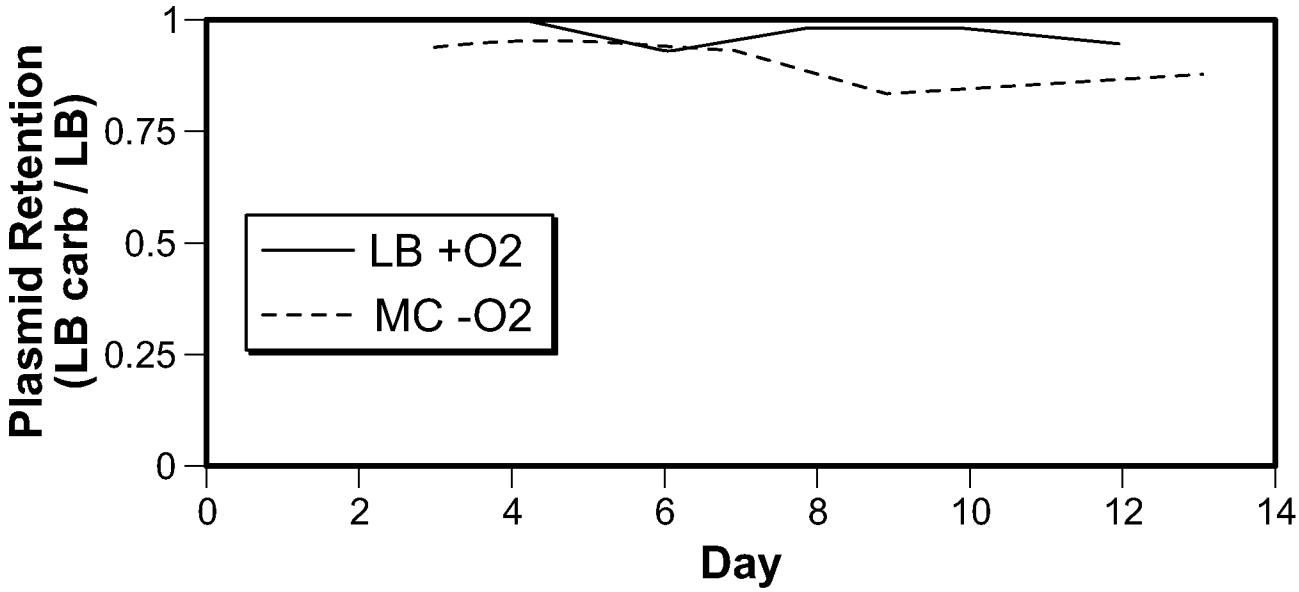
1
0.75
0.5
0.25
0
Plasmid Retention
(LB carb / LB)
LB +O2
MC -O2
0
2
4
6
8
10
12
14
Day

FIG. 2C

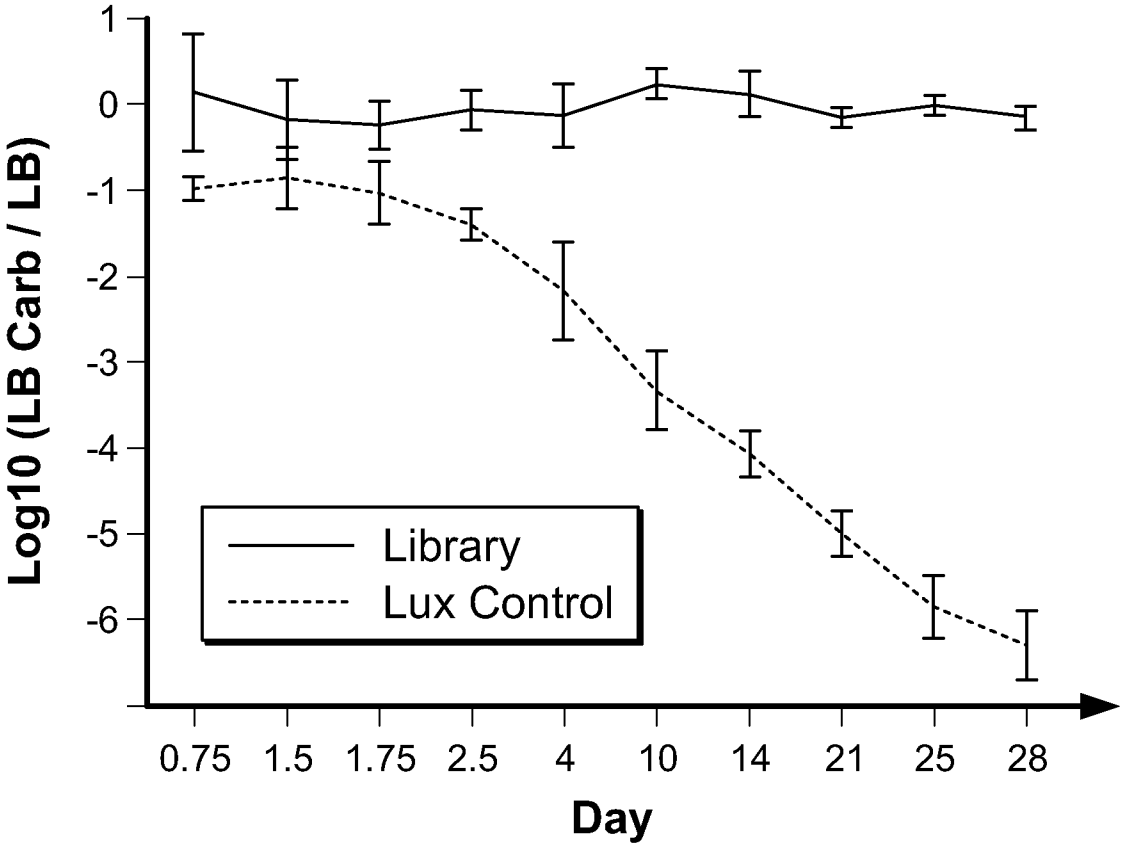
Log10 (LB Carb / LB)
1
0
-1
-2
-3
-4
-5
-6
Library
Lux Control
0.75
1.5
1.75
2.5
4
10
14
21
25
28
Day

FIG. 3A

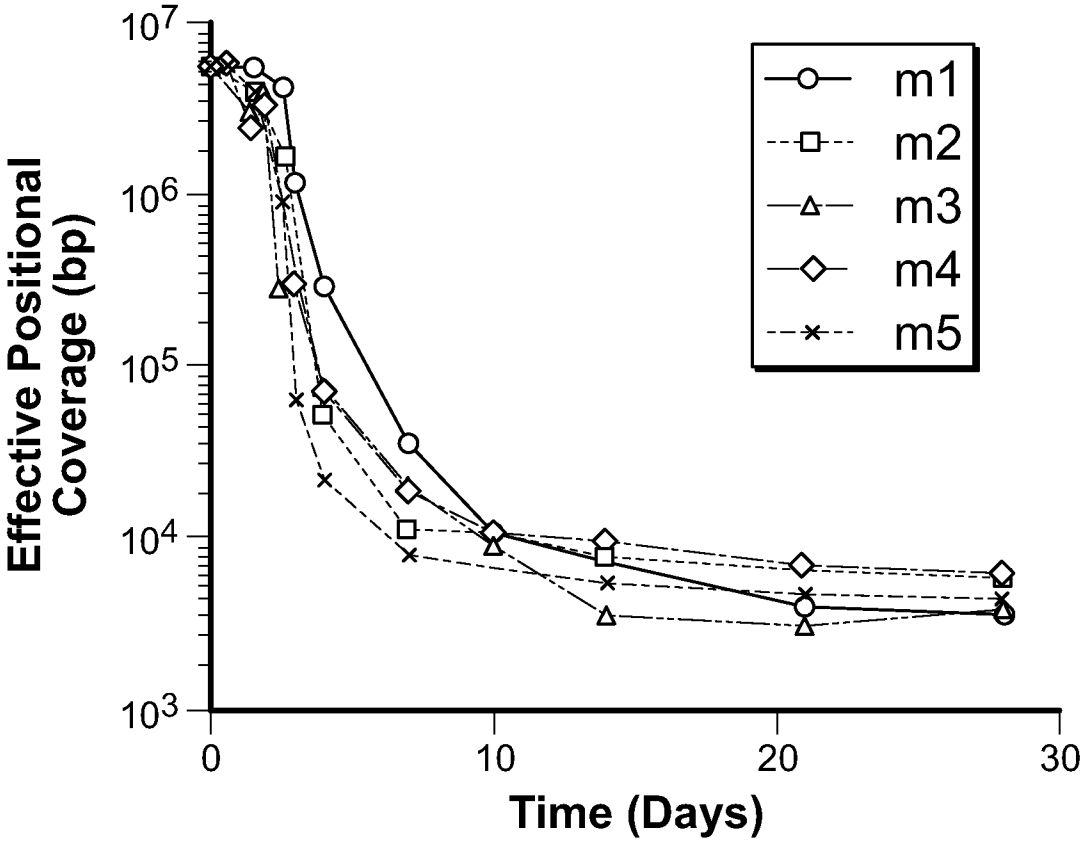
Effective Positional Coverage (bp)
10^7
10^6
10^5
10^4
10^3
m1
m2
m3
m4
m5
0
10
20
30
Time (Days)

FIG. 3B

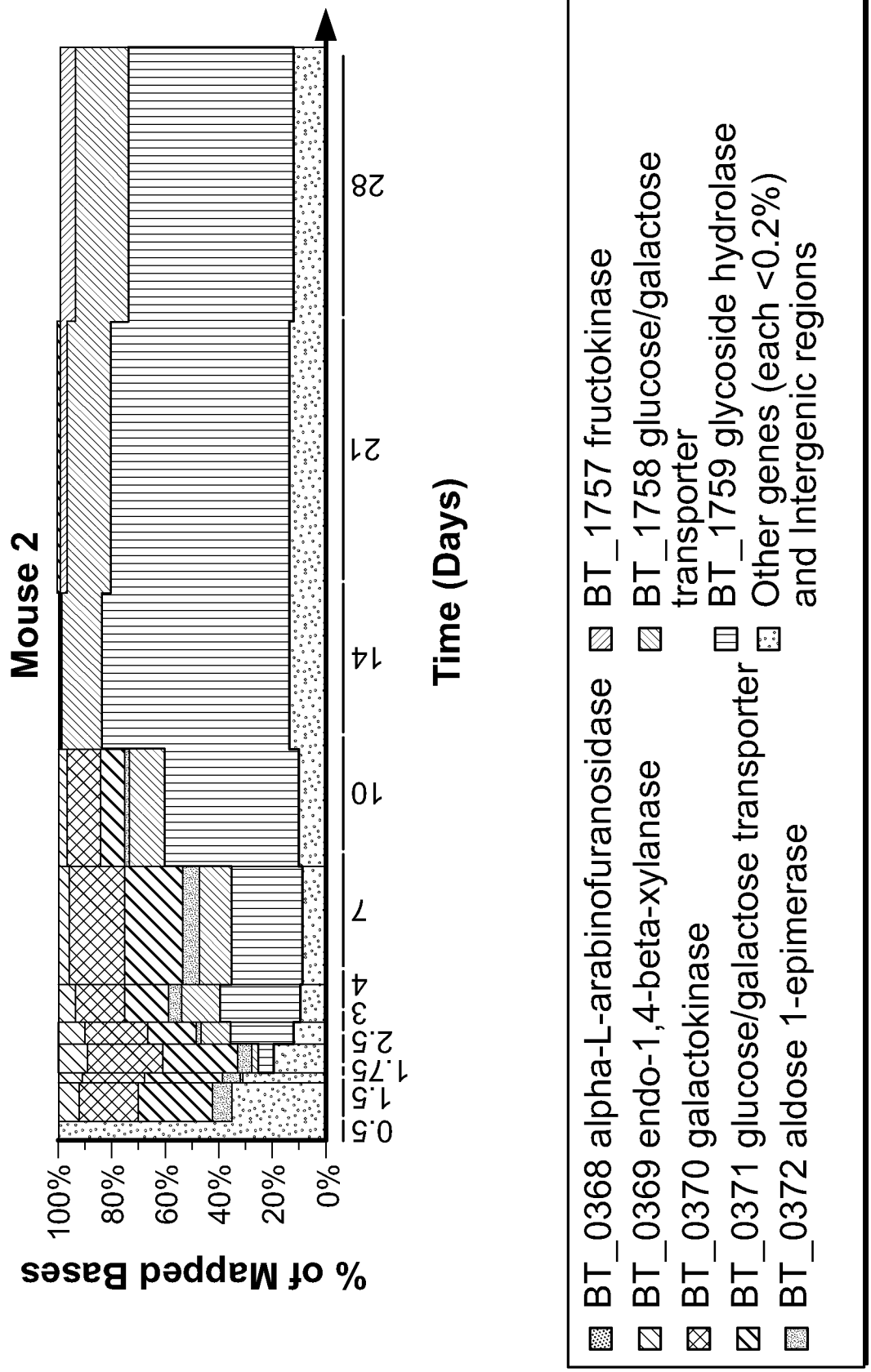
Mouse 2
% of Mapped Bases
100%
80%
60%
40%
20%
0%
0.5
1.5
1.75
2.5
3
4
7
10
14
21
28
Time (Days)
BT_0368 alpha-L-arabinofuranosidase
BT_0369 endo-1,4-beta-xylanase
BT_0370 galactokinase
BT_0371 glucose/galactose transporter
BT_0372 aldose 1-epimerase
BT_1757 fructokinase
BT_1758 glucose/galactose transporter
BT_1759 glycoside hydrolase
Other genes (each <0.2%) and Intergenic regions

FIG. 3C

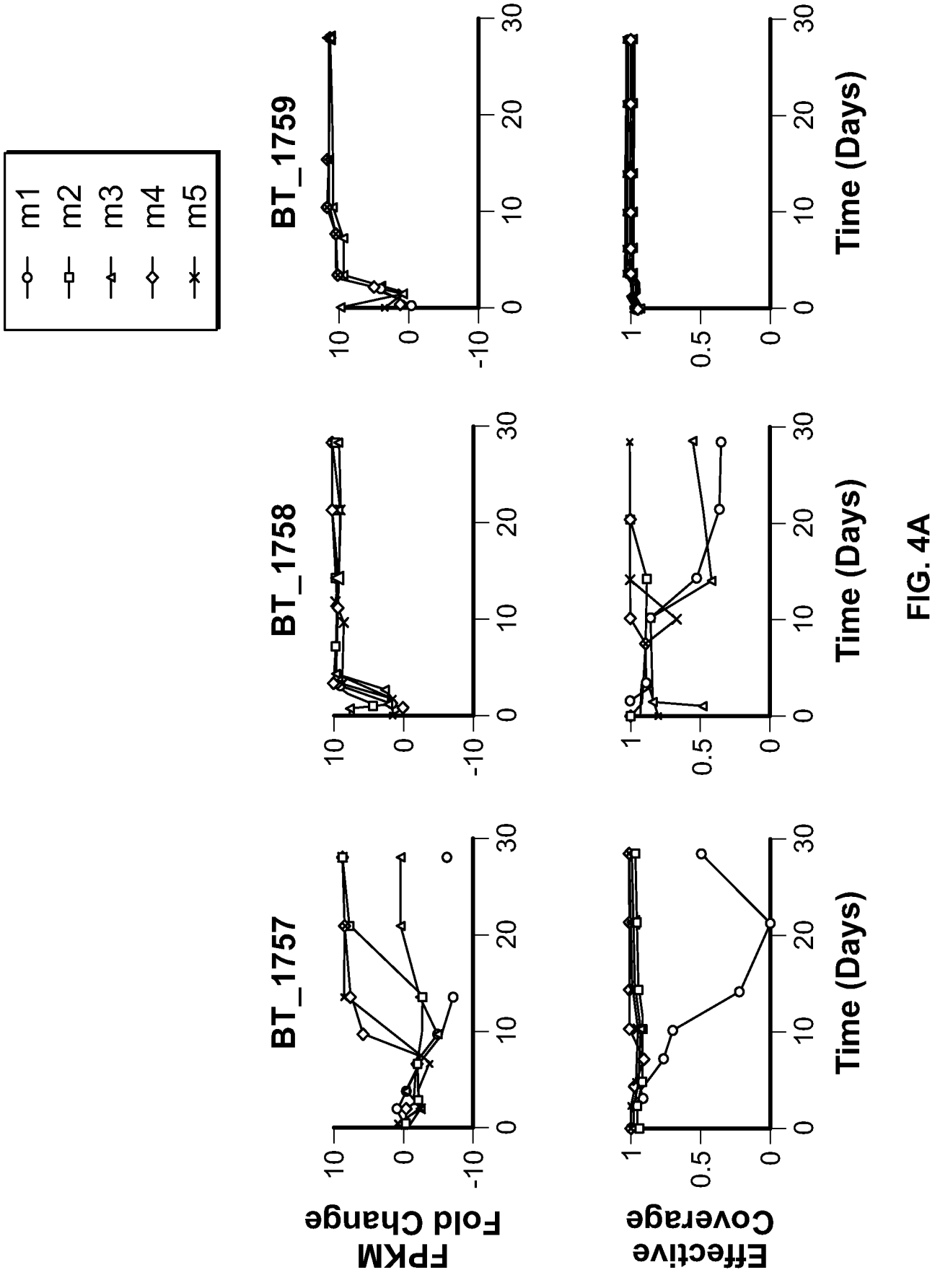
m1
m2
m3
m4
m5
BT_1757
BT_1758
BT_1759
FPKM Fold Change
Effective Coverage
Time (Days)
10
0
-10
1
0.5
0
0
10
20
30

FIG. 4A

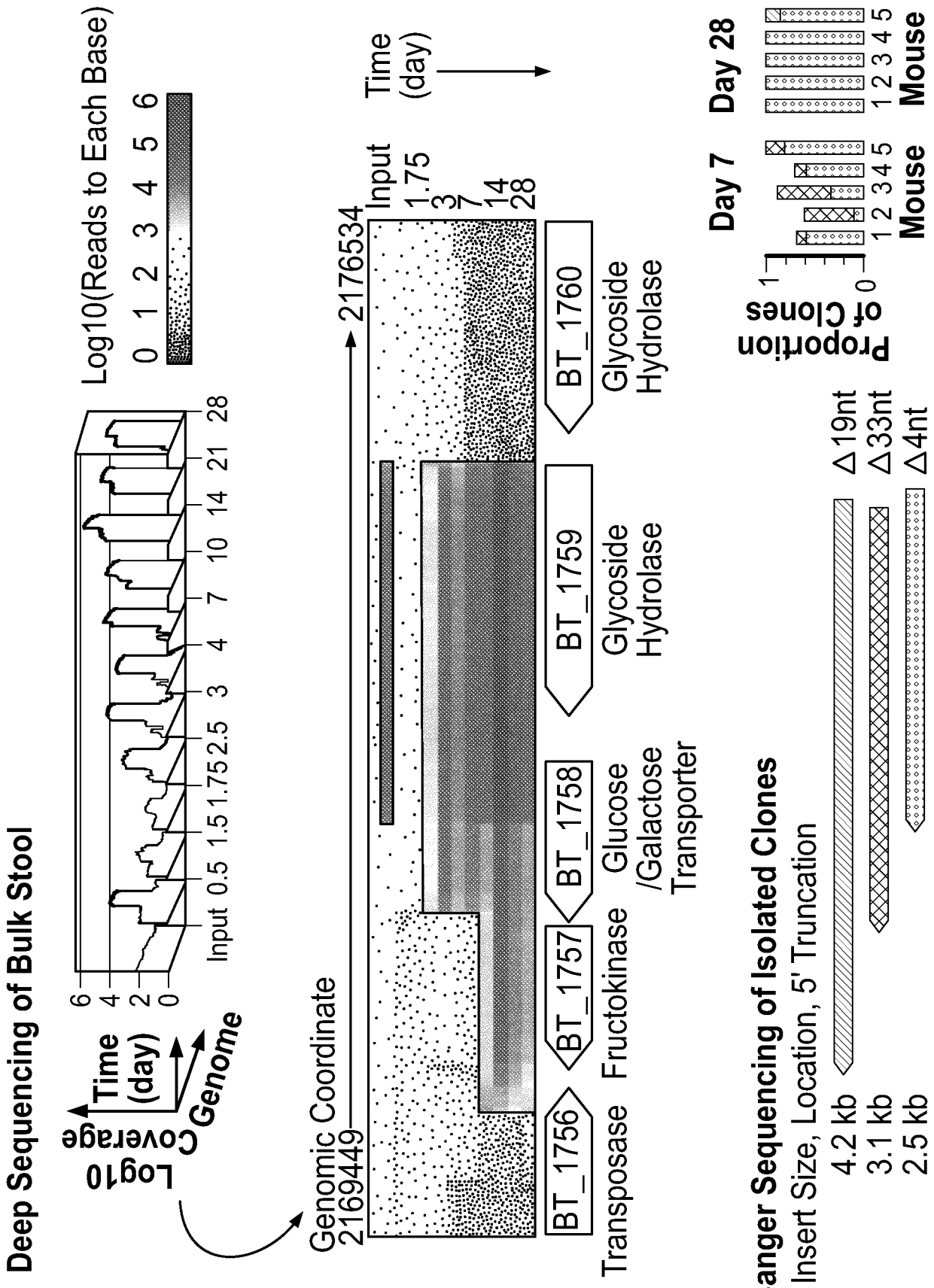
Deep Sequencing of Bulk Stool
Log10 Coverage
Time (day)
Genome
Input 0.5 1.5 1.75 2.5 3 4 7 10 14 21 28
Log10(Reads to Each Base)
0 1 2 3 4 5 6
Genomic Coordinate
2169449
2176534
Input
1.75
3
7
14
28
Time (day)
BT_1756
BT_1757
BT_1758
BT_1759
BT_1760
Transposase
Fructokinase
Glucose /Galactose Transporter
Glycoside Hydrolase
Glycoside Hydrolase
Sanger Sequencing of Isolated Clones
Insert Size, Location, 5' Truncation
4.2 kb
3.1 kb
2.5 kb
Δ19nt
Δ33nt
Δ4nt
Day 7
Day 28
Proportion of Clones
1
0
1 2 3 4 5
Mouse
1 2 3 4 5
Mouse

FIG. 4B

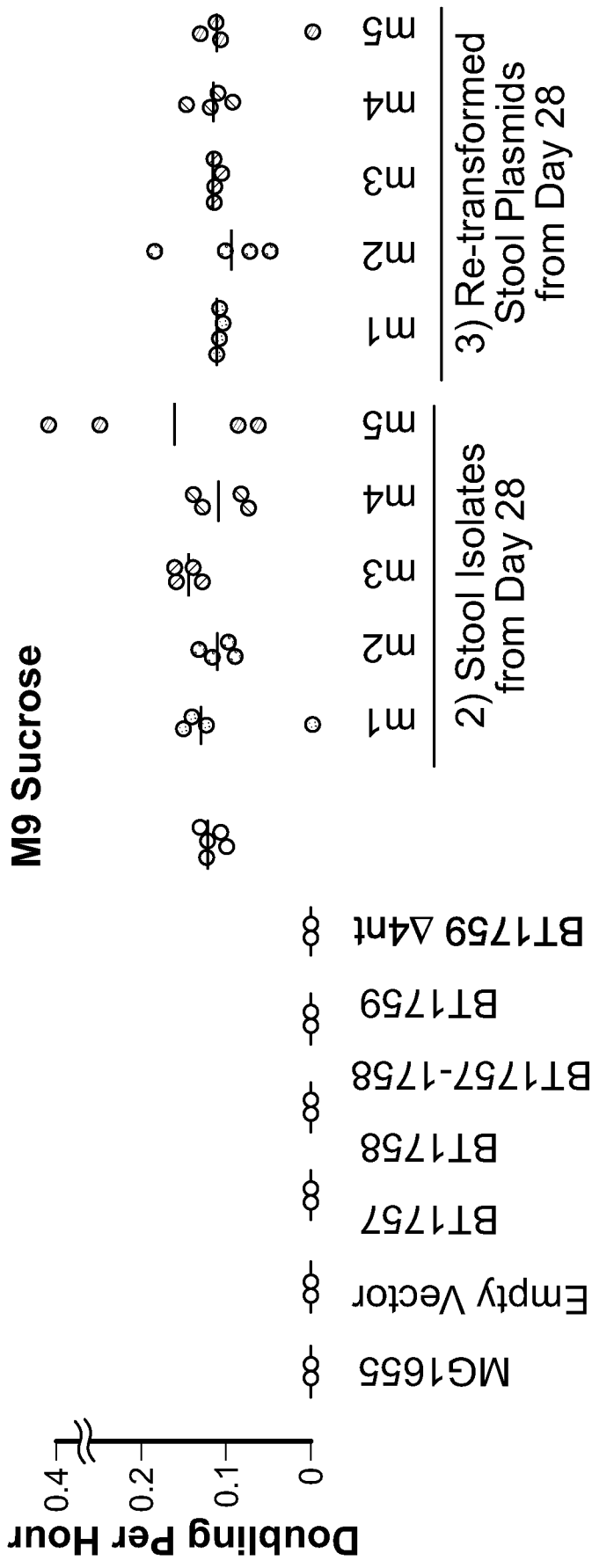
M9 Sucrose
Doubling Per Hour
0.4
0.2
0.1
0
MG1655
Empty Vector
BT1757
BT1758
BT1757-1758
BT1759
BT1759 Δ4nt
1) Cloned CDS
m1
m2
m3
m4
m5
2) Stool Isolates from Day 28
m1
m2
m3
m4
m5
3) Re-transformed Stool Plasmids from Day 28

FIG. 4C

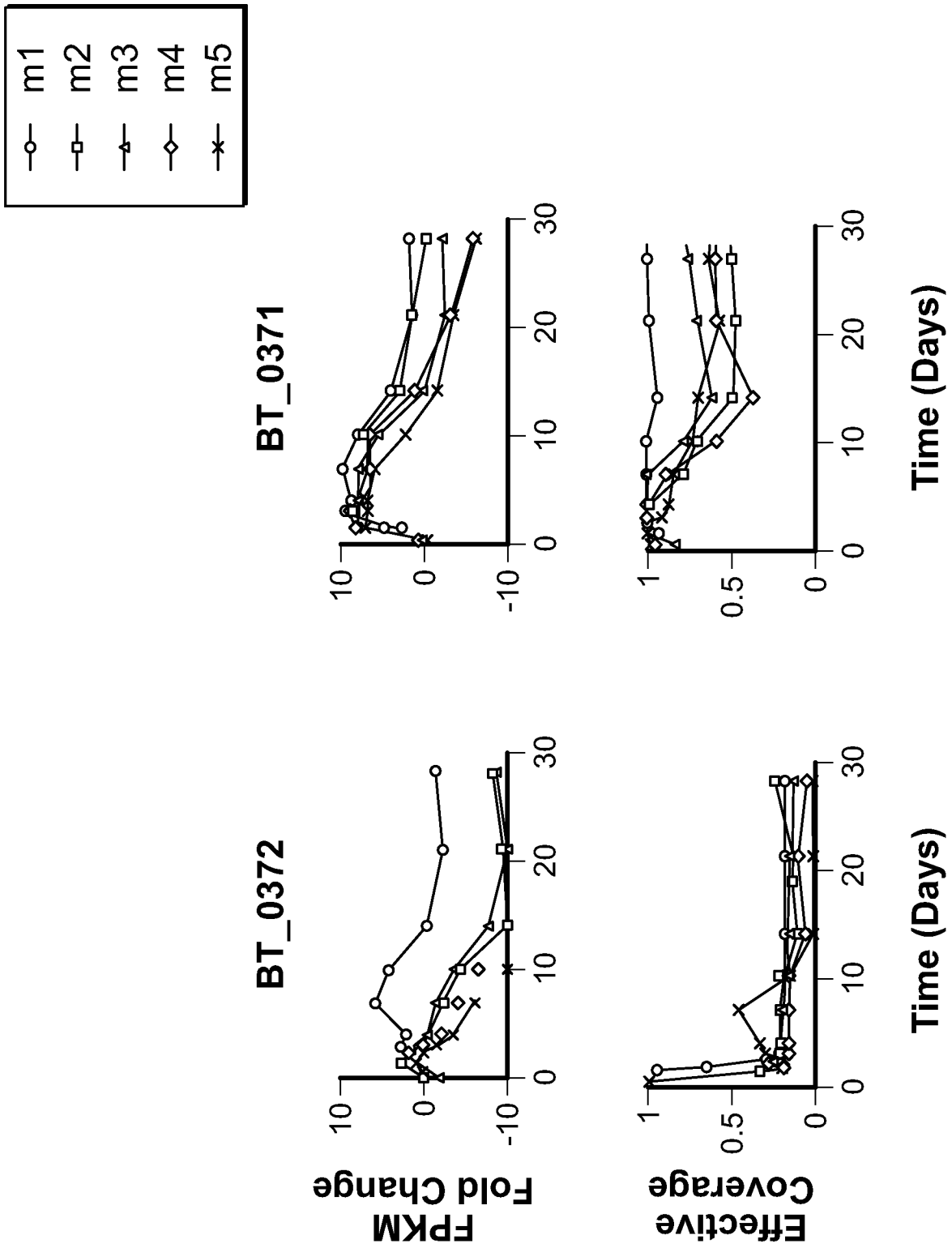
m1
m2
m3
m4
m5
BT_0371
BT_0372
FPKM Fold Change
Effective Coverage
Time (Days)
Time (Days)
10
0
-10
1
0.5
0
0
10
20
30

FIG. 5A

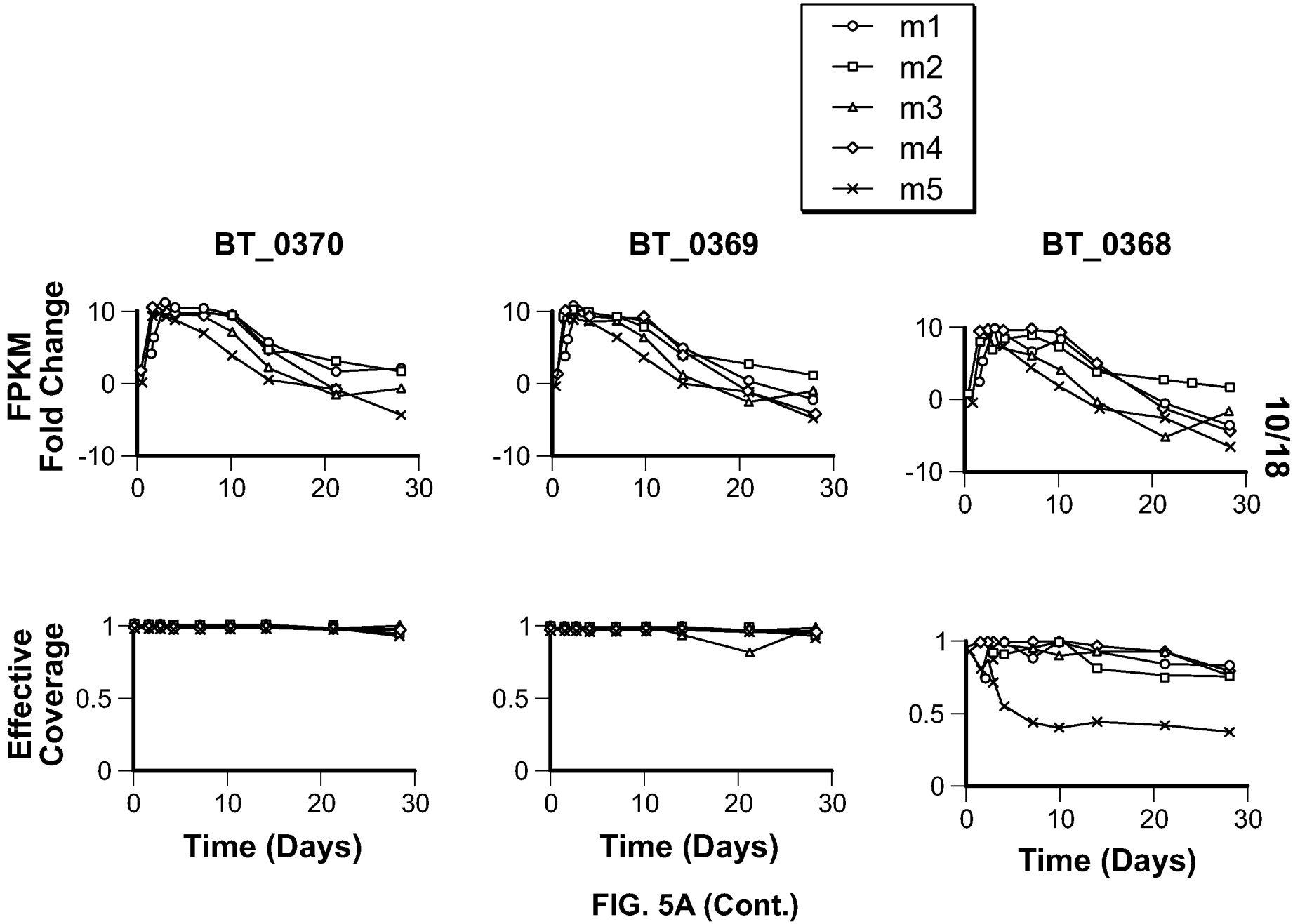
m1
m2
m3
m4
m5
BT_0370
BT_0369
BT_0368
FPKM Fold Change
10
0
-10
0
10
20
30
Effective Coverage
1
0.5
0
Time (Days)

FIG. 5A (Cont.)

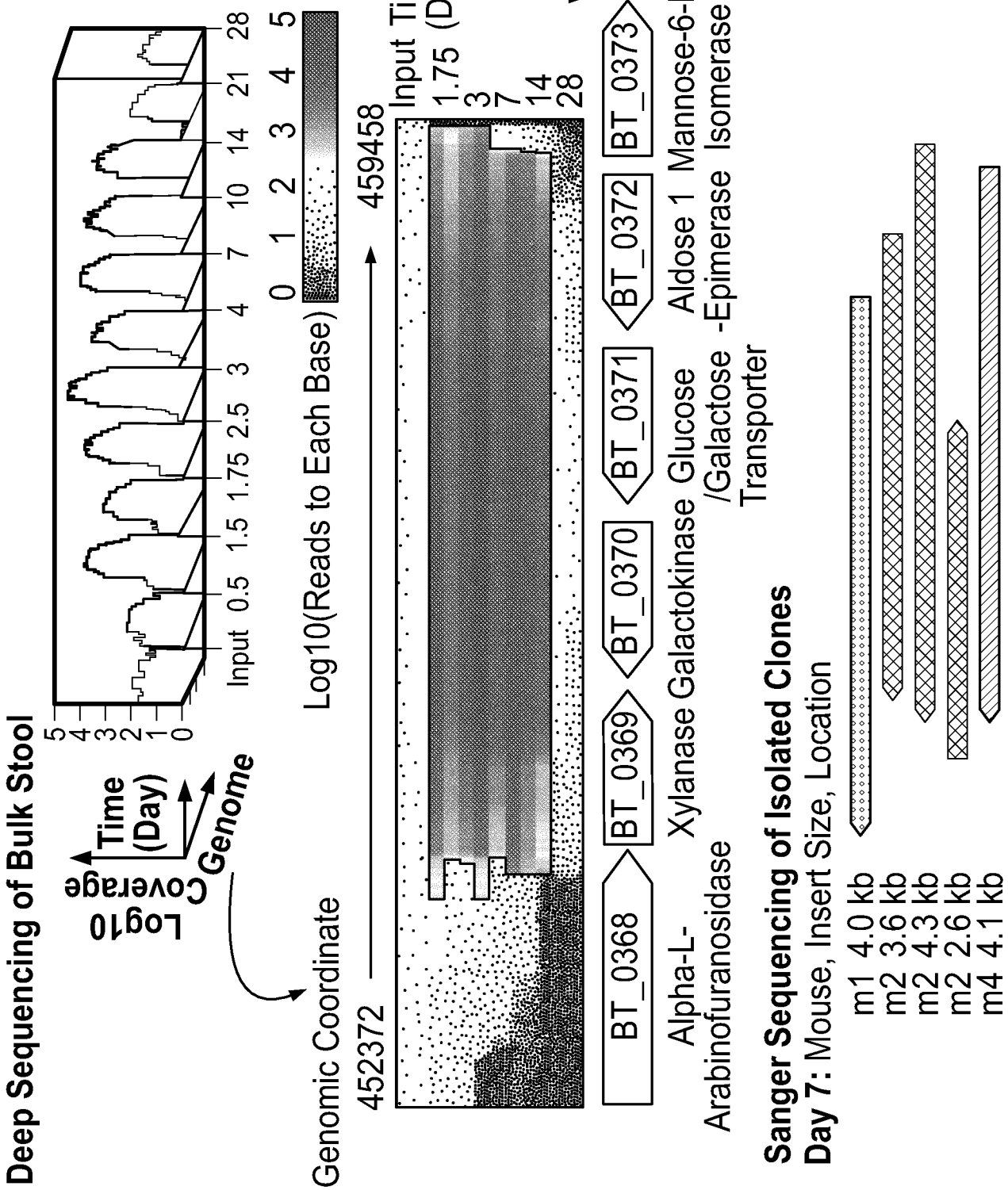
Deep Sequencing of Bulk Stool
Log10 Coverage
Time (Day)
Genome
Input 0.5 1.5 1.75 2.5 3 4 7 10 14 21 28
0 1 2 3 4 5
Log10(Reads to Each Base)
Genomic Coordinate
452372
459458
Input Time
1.75 (Day)
3
7
14
28
BT_0368
BT_0369
BT_0370
BT_0371
BT_0372
BT_0373
Alpha-L-Arabinofuranosidase
Xylanase
Galactokinase
Glucose /Galactose Transporter
Aldose 1 -Epimerase
Mannose-6-P Isomerase
Sanger Sequencing of Isolated Clones
Day 7: Mouse, Insert Size, Location
m1 4.0 kb
m2 3.6 kb
m2 4.3 kb
m2 2.6 kb
m4 4.1 kb

FIG. 5B

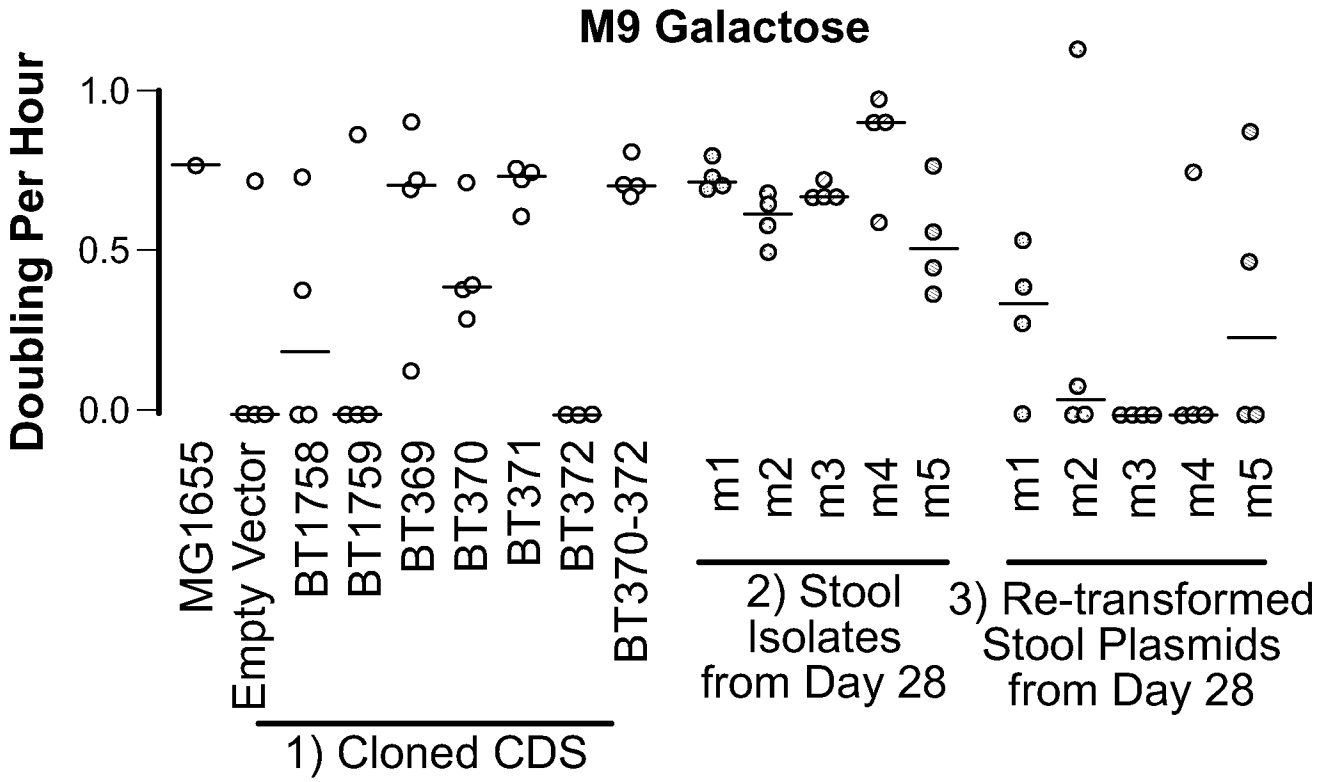
M9 Galactose
Doubling Per Hour
1.0
0.5
0.0
MG1655
Empty Vector
BT1758
BT1759
BT369
BT370
BT371
BT372
BT370-372
m1
m2
m3
m4
m5
m1
m2
m3
m4
m5
1) Cloned CDS
2) Stool Isolates from Day 28
3) Re-transformed Stool Plasmids from Day 28

FIG. 5C

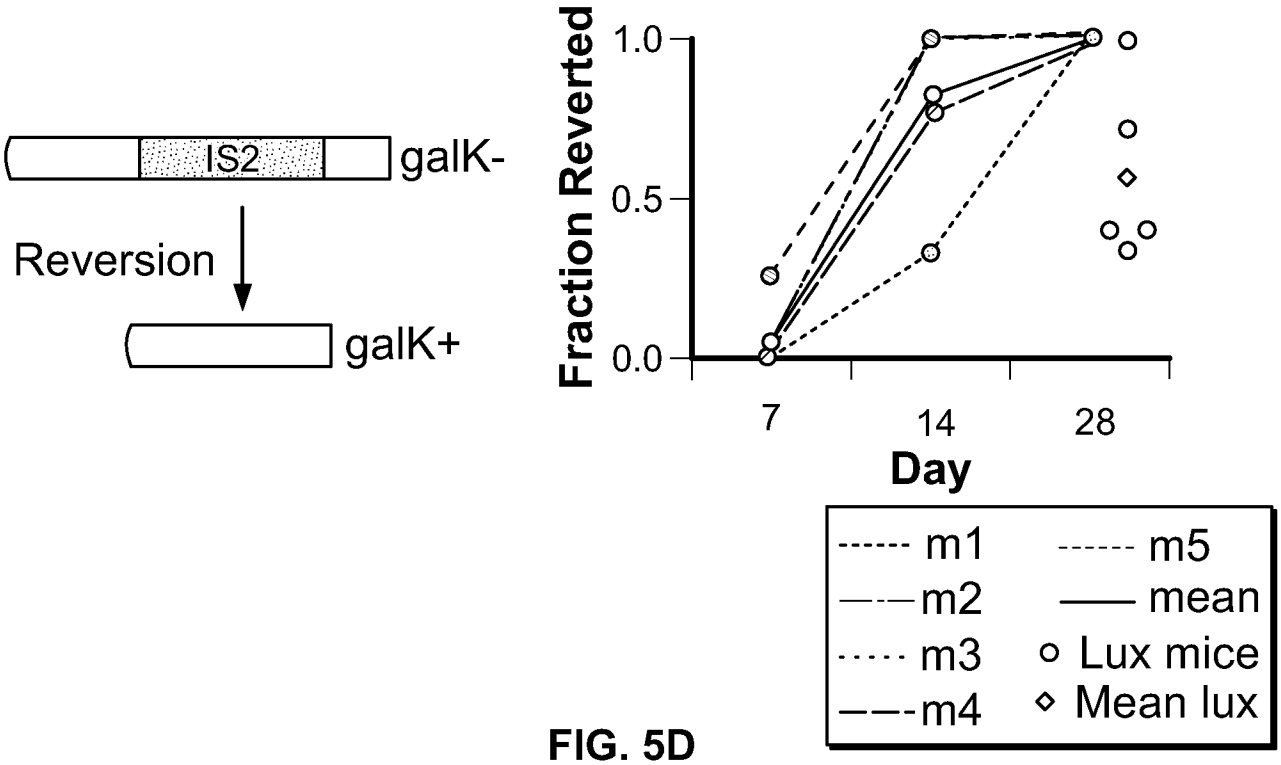
IS2
galK-
Reversion
galK+
Fraction Reverted
1.0
0.5
0.0
7
14
28
Day
m1
m2
m3
m4
m5
mean
Lux mice
Mean lux

FIG. 5D

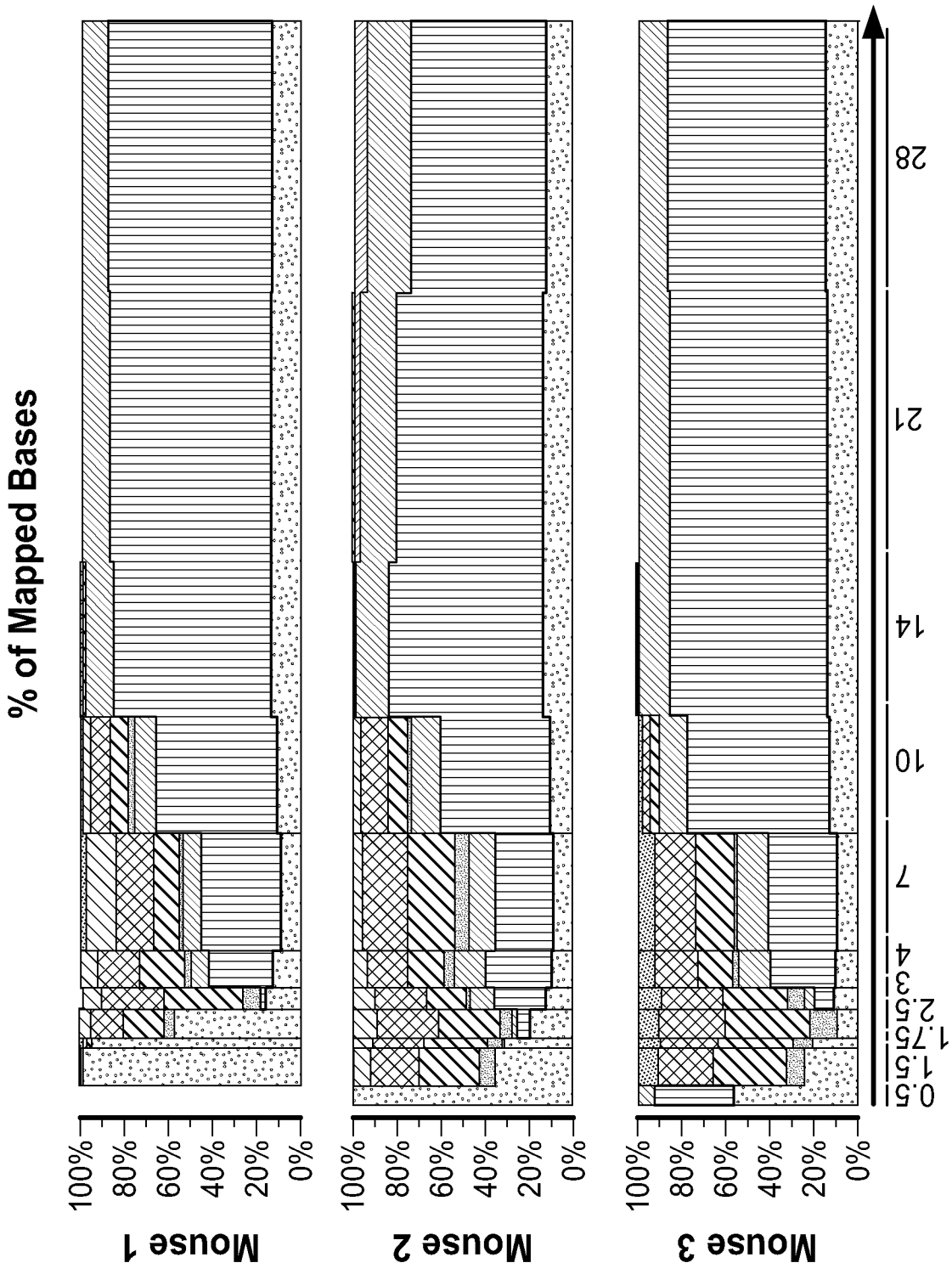
% of Mapped Bases
100%
80%
60%
40%
20%
0%
Mouse 1
Mouse 2
Mouse 3
0.5
1.5
1.75
2.5
3
4
7
10
14
21
28
Time (Days)

FIG. 6

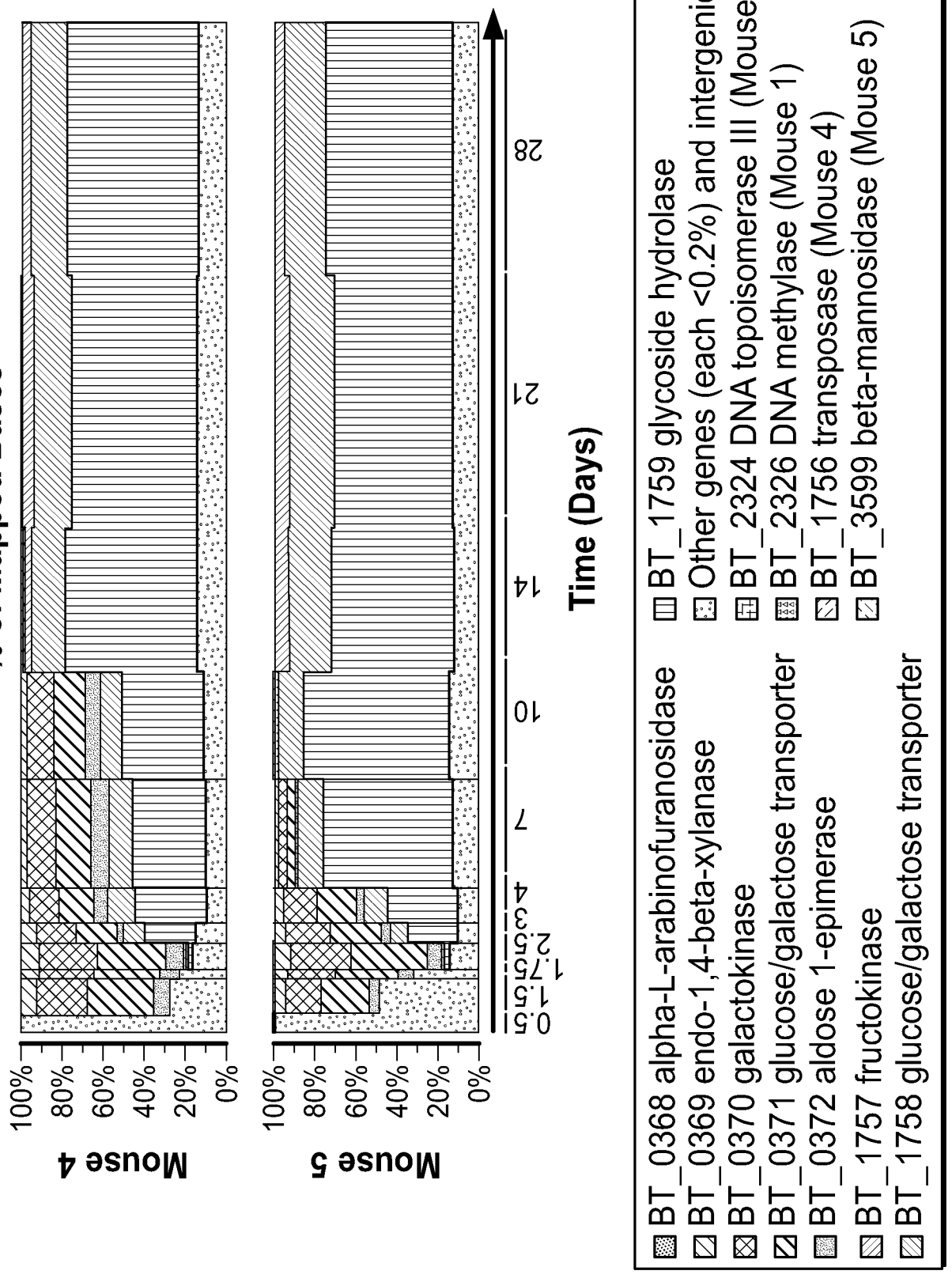
% of Mapped Bases
Mouse 4
Mouse 5
100%
80%
60%
40%
20%
0%
0
1.75
2.33
4
7
10
14
21
28
Time (Days)
BT_0368 alpha-L-arabinofuranosidase
BT_0369 endo-1,4-beta-xylanase
BT_0370 galactokinase
BT_0371 glucose/galactose transporter
BT_0372 aldose 1-epimerase
BT_1757 fructokinase
BT_1758 glucose/galactose transporter
BT_1759 glycoside hydrolase
Other genes (each <0.2%) and intergenic regions
BT_2324 DNA topoisomerase III (Mouse 1)
BT_2326 DNA methylase (Mouse 1)
BT_1756 transposase (Mouse 4)
BT_3599 beta-mannosidase (Mouse 5)

FIG. 6 (Cont.)

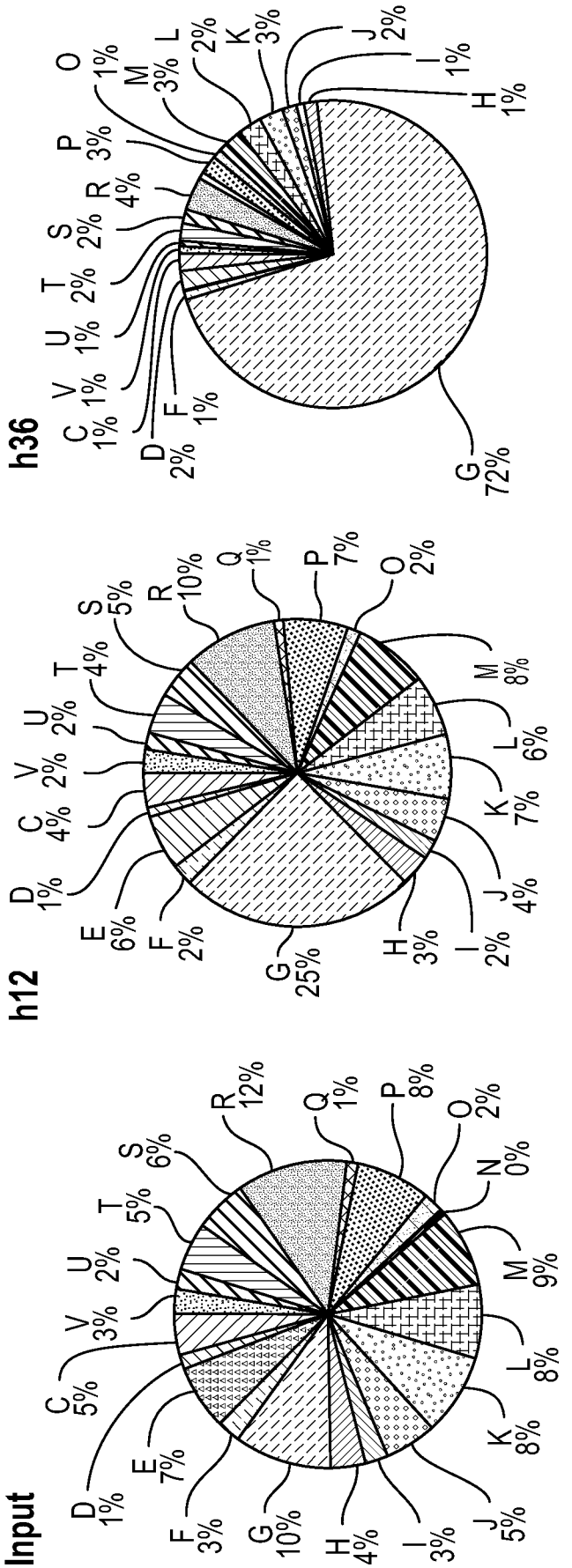
h36
G 72%
D 2%
C 1%
V 1%
U 1%
T 2%
S 2%
R 4%
P 3%
O 1%
M 3%
L 2%
K 3%
J 2%
I 1%
H 1%
F 1%
h12
G 25%
E 6%
D 1%
C 4%
V 2%
U 2%
T 4%
S 5%
R 10%
Q 1%
P 7%
O 2%
M 8%
L 6%
K 7%
J 4%
I 2%
H 3%
F 2%
Input
D 1%
C 5%
V 3%
U 2%
T 5%
S 6%
R 12%
Q 1%
P 8%
O 2%
N 0%
M 9%
L 8%
K 8%
J 5%
I 3%
H 4%
G 10%
F 3%
E 7%

FIG. 7

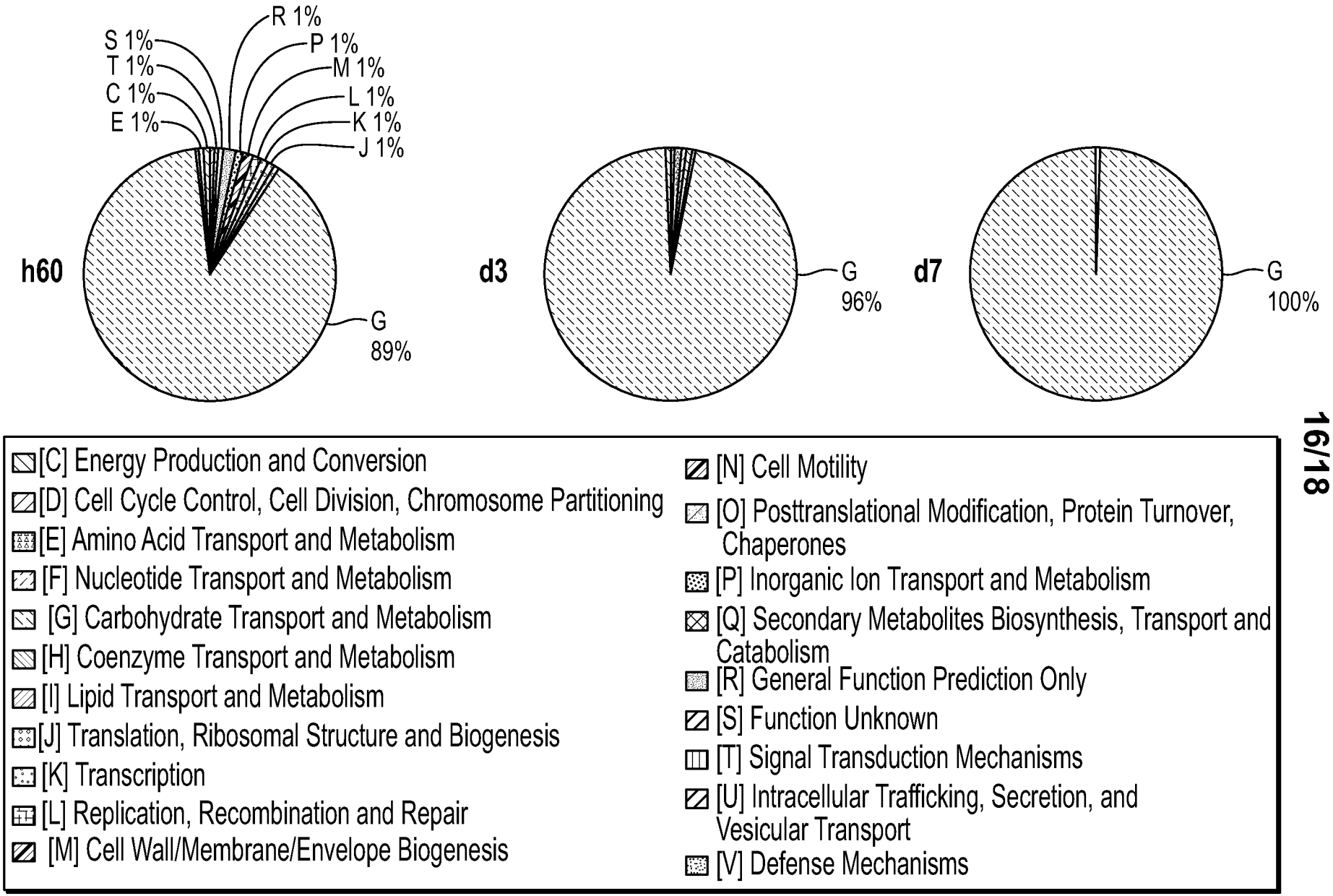
h60
S 1%
T 1%
C 1%
E 1%
R 1%
P 1%
M 1%
L 1%
K 1%
J 1%
G
89%
d3
G
96%
d7
G
100%
[C] Energy Production and Conversion
[D] Cell Cycle Control, Cell Division, Chromosome Partitioning
[E] Amino Acid Transport and Metabolism
[F] Nucleotide Transport and Metabolism
[G] Carbohydrate Transport and Metabolism
[H] Coenzyme Transport and Metabolism
[I] Lipid Transport and Metabolism
[J] Translation, Ribosomal Structure and Biogenesis
[K] Transcription
[L] Replication, Recombination and Repair
[M] Cell Wall/Membrane/Envelope Biogenesis
[N] Cell Motility
[O] Posttranslational Modification, Protein Turnover, Chaperones
[P] Inorganic Ion Transport and Metabolism
[Q] Secondary Metabolites Biosynthesis, Transport and Catabolism
[R] General Function Prediction Only
[S] Function Unknown
[T] Signal Transduction Mechanisms
[U] Intracellular Trafficking, Secretion, and Vesicular Transport
[V] Defense Mechanisms

FIG. 7 (Cont.)

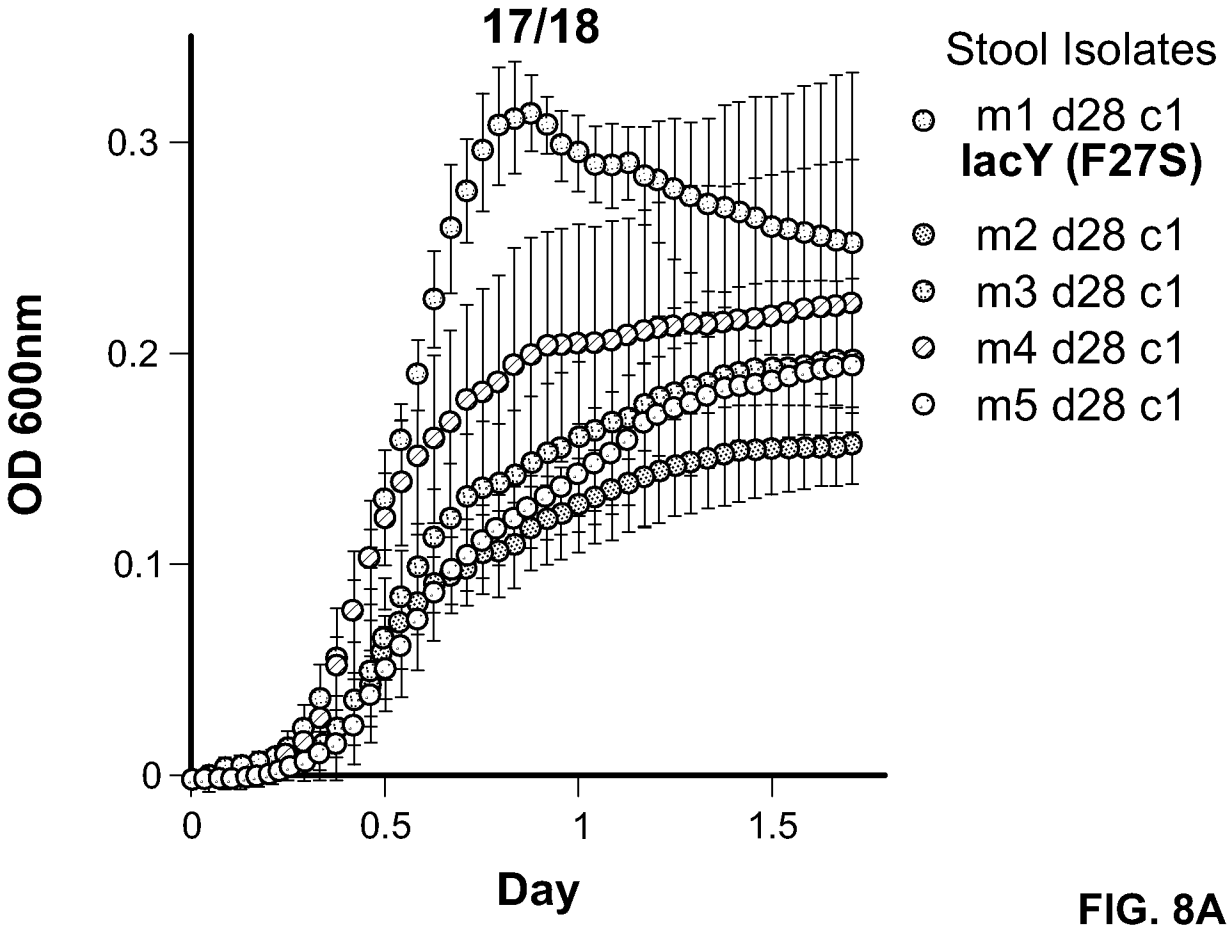
17/18
OD 600nm
0.3
0.2
0.1
0
0
0.5
1
1.5
Day
Stool Isolates
m1 d28 c1
lacY (F27S)
m2 d28 c1
m3 d28 c1
m4 d28 c1
m5 d28 c1

FIG. 8A

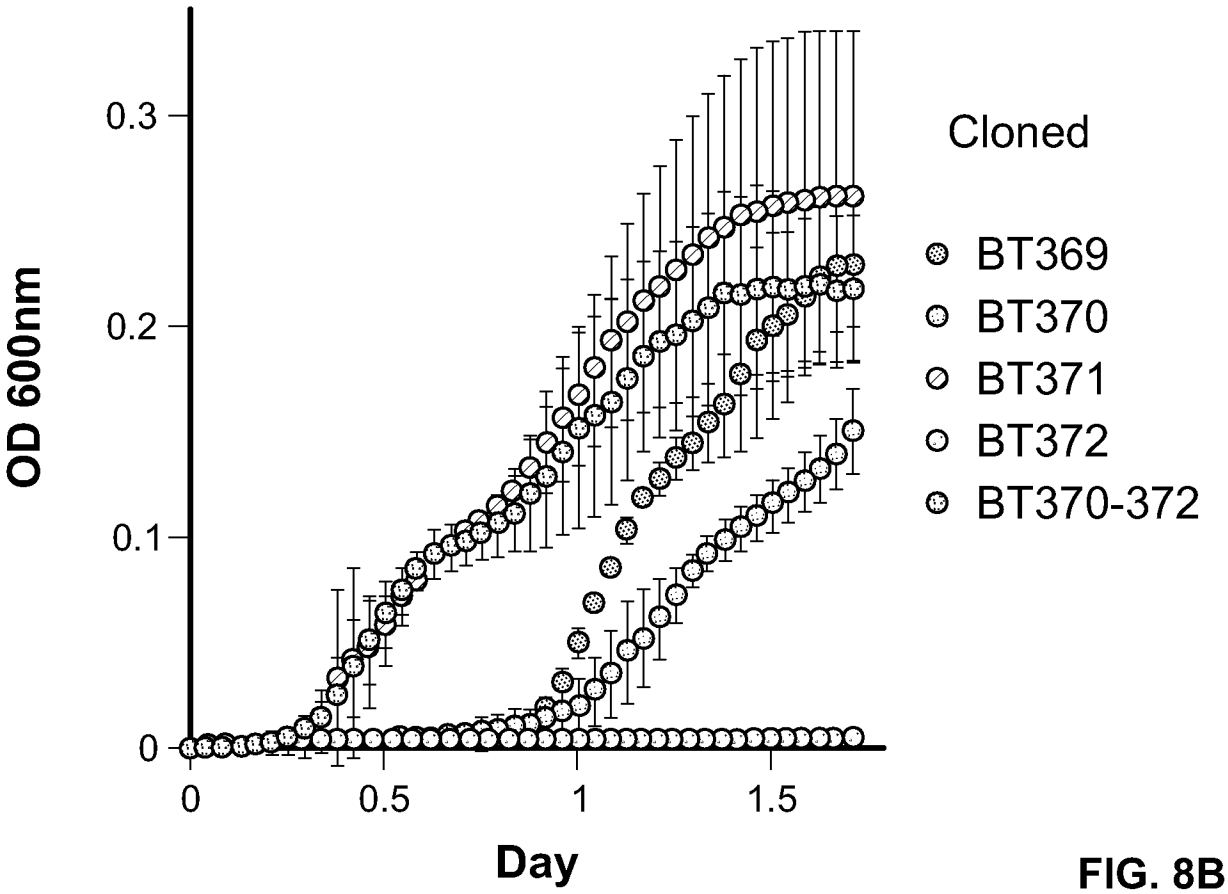
OD 600nm
0.3
0.2
0.1
0
0
0.5
1
1.5
Day
Cloned
BT369
BT370
BT371
BT372
BT370-372

FIG. 8B

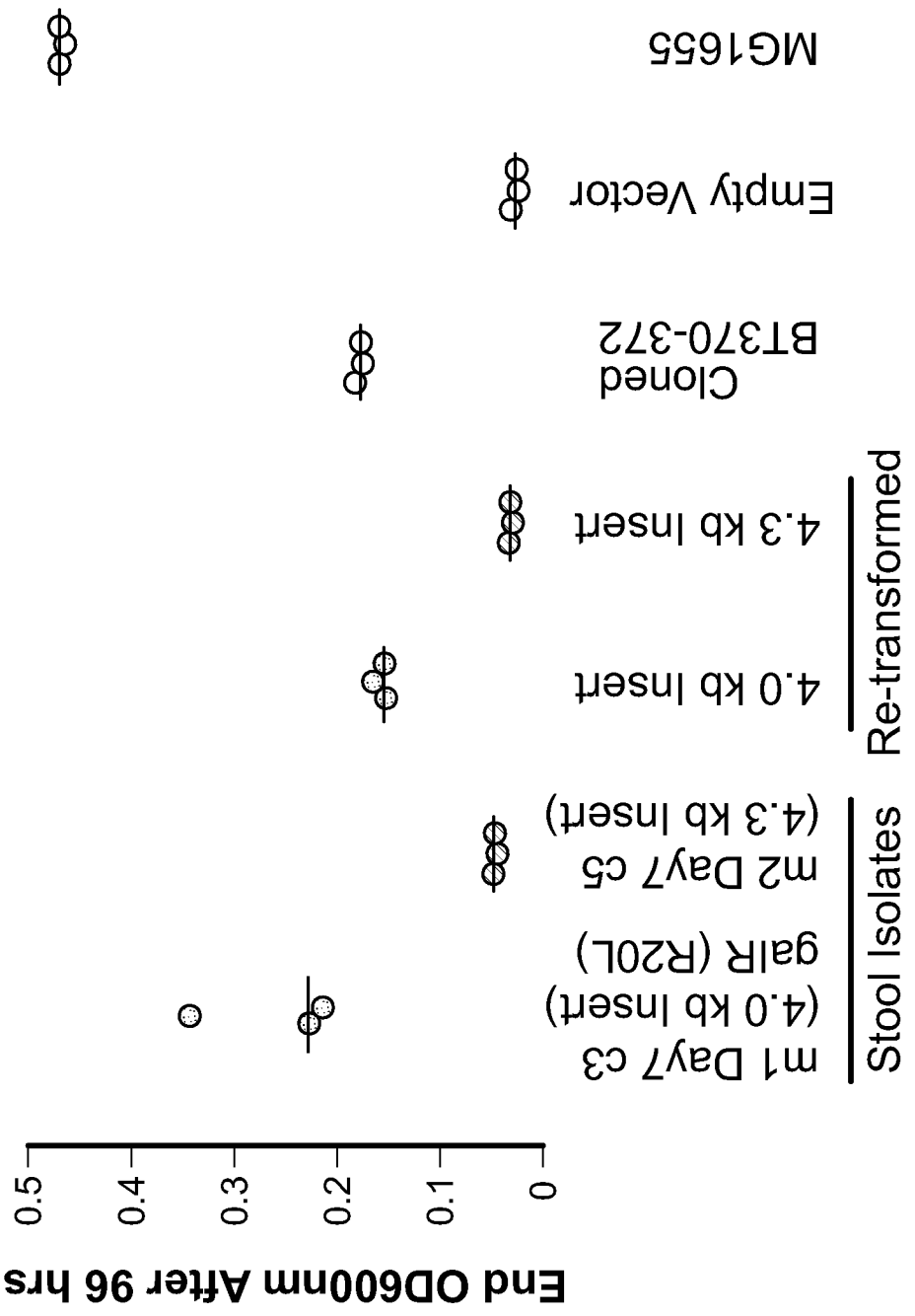
End OD600nm After 96 hrs
0.5
0.4
0.3
0.2
0.1
0
Stool Isolates
m1 Day7 c3 (4.0 kb Insert)
galR (R20L)
m2 Day7 c5 (4.3 kb Insert)
Re-transformed
4.0 kb Insert
4.3 kb Insert
Cloned BT370-372
Empty Vector
MG1655

FIG. 8C

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2014/066173

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - C12Q 1/68 (2015.01)
CPC - A61K 35/741 (2015.01)
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC(8) - A23K 1/00, 35/66; A61P 1/00; C07K 14/33; C12N 1/20, 15/09, 15/10; C12Q 1/68; C12R 1/01, 1/145 (2015.01)
USPC - 424/93.2; 426/2, 61, 648; 435/252.3, 252.4; 506/10, 14; 514/5.7

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
CPC - A61K 35/74, 35/741, 35/745, 35/747; C07K 14/33, 14/335; C12N 1/20; C12Q 1/6806 (2015.01) (keyword delimited)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

PatBase, PubMed, Google Patents, Google
Search terms used: GALK GK1 galactose kinase galactokinase probiotic microbiat* metagenom* librar* glycoside hydrolases glycosidases glycosyl hydrolases glucose galactose transporter

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	RIESENFELD et al. 'METAGENOMICS: Genomic Analysis of Microbial Communities,' Annual Review of Genetics, 14 July 2004 (14.07.2004), Vol. 38, Pgs. 525-552. entire document	1-31
A	WO 2012/122522 A2 (WASHINGTON UNIVERSITY) 13 September 2012 (13.09.2012) entire document	1-31
A	US 2010/0183559 A1 (VAN SINDEREN et al) 22 July 2010 (22.07.2010) entire document	1-31
A	US 2012/0164275 A1 (JANZEN et al) 28 June 2012 (28.06.2012) entire document	1-31
A	US 2011/0053273 A1 (BENDERS et al) 03 March 2011 (03.03.2011) entire document	1-31
P, Y	FAITH et al. 'Identifying Gut Microbe-Host Phenotype Relationships Using Combinatorial Communities in Gnotobiotic Mice,' Science Translational Medicine, 22 January 2014 (22.01.2014), Vol. 6, Pgs. 1-24. entire document	1-31

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
24 January 2015	12 FEB 2015
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Blaine R. Copenheaver PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774