2010

Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors

Andrew K. Benson
University of Nebraska-Lincoln, abenson1@unl.edu

Scott A. Kelly
University of North Carolina

Ryan Legge
University of Nebraska-Lincoln

Fangrui Ma
University of Nebraska-Lincoln, fangrui.ma@gmail.com

Soo Jen Low
University of Nebraska, Lincoln

See next page for additional authors

Benson, Andrew K.; Kelly, Scott A.; Legge, Ryan; Ma, Fangrui; Low, Soo Jen; Kim, Jaehyoung; Zhang, Min; Oh, Phaik Lyn; Nehrenberg, Derrick; Huab, Kunjie; Kachman, Stephen D.; Moriyama, Etsuko N.; Walter, Jens; Peterson, Daniel A.; and Pomp, Daniel, "Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors" (2010). *Faculty Publications from the Center for Plant Science Innovation*. 49.
https://digitalcommons.unl.edu/plantscifacpub/49

Authors
Andrew K. Benson, Scott A. Kelly, Ryan Legge, Fangrui Ma, Soo Jen Low, Jaehyoung Kim, Min Zhang, Phaik Lyn Oh, Derrick Nehrenberg, Kunjie Huab, Stephen D. Kachman, Etsuko N. Moriyama, Jens Walter, Daniel A. Peterson, and Daniel Pomp

Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors

Andrew K. Benson[a,1], Scott A. Kelly[b], Ryan Legge[a], Fangrui Ma[a], Soo Jen Low[a], Jaehyoung Kim[a], Min Zhang[a], Phaik Lyn Oh[a], Derrick Nehrenberg[b], Kunjie Hua[b], Stephen D. Kachman[c], Etsuko N. Moriyama[d], Jens Walter[a], Daniel A. Peterson[a], and Daniel Pomp[b,e]

[a]Department of Food Science and Technology and Core for Applied Genomics and Ecology, University of Nebraska, Lincoln, NE 68583-0919; [b]Department of Genetics, Carolina Center for Genome Science, University of North Carolina, Chapel Hill, NC 27599-7545; [c]Department of Statistics, University of Nebraska, Lincoln, NE 68583-0963; [d]School of Biological Sciences and Center for Plant Science Innovation, University of Nebraska, Lincoln, NE 68588-0118; and [e]Department of Nutrition, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC 27599-7461

Edited by Trudy F. C. Mackay, North Carolina State University, Raleigh, NC, and approved September 8, 2010 (received for review June 10, 2010)

In vertebrates, including humans, individuals harbor gut microbial communities whose species composition and relative proportions of dominant microbial groups are tremendously varied. Although external and stochastic factors clearly contribute to the individuality of the microbiota, the fundamental principles dictating how environmental factors and host genetic factors combine to shape this complex ecosystem are largely unknown and require systematic study. Here we examined factors that affect microbiota composition in a large (*n* = 645) mouse advanced intercross line originating from a cross between C57BL/6J and an ICR-derived outbred line (HR). Quantitative pyrosequencing of the microbiota defined a core measurable microbiota (CMM) of 64 conserved taxonomic groups that varied quantitatively across most animals in the population. Although some of this variation can be explained by litter and cohort effects, individual host genotype had a measurable contribution. Testing of the CMM abundances for cosegregation with 530 fully informative SNP markers identified 18 host quantitative trait loci (QTL) that show significant or suggestive genome-wide linkage with relative abundances of specific microbial taxa. These QTL affect microbiota composition in three ways; some loci control individual microbial species, some control groups of related taxa, and some have putative pleiotropic effects on groups of distantly related organisms. These data provide clear evidence for the importance of host genetic control in shaping individual microbiome diversity in mammals, a key step toward understanding the factors that govern the assemblages of gut microbiota associated with complex diseases.

16S rDNA | pyrosequencing | quantitative trait loci mapping | microbiome phenotyping | population

Humans are born with a sterile gastrointestinal (GI) tract that is rapidly colonized by successive waves of microorganisms until a dense microbial population stabilizes at about the time of weaning (1). This population is dominated by thousands of bacterial species that belong to a small number of phyla (2–4). Despite conservation at the highest taxonomic ranks, the composition of the adult gut microbiota varies dramatically from individual to individual, including differences in the relative ratios of dominant phyla and variation in genera and species found in an individual host (4). Once established, these compositional features are highly resilient to perturbation (5). Although the mechanism of this homeostasis is unknown, it suggests a "top down" model for assembly of the symbiotic microbial community that is largely determined by the host.

A mechanistic insight into the assembly of the gut microbiota is immediately relevant to our understanding of complex human diseases (6). Obesity (7), coronary heart disease (8), diabetes (9), and inflammatory bowel disease (10) have all been associated with composition of gut microbiota. These diseases are well understood to be multifactorial, with both environmental and genetic components (11–13), and the contribution of the gut microbiota is currently viewed as an environmental factor (14). Although a number of studies have suggested that composition of the gut microbiota may be subject to host genetic forces, existing evidence is conflicting and confounded by the genetic diversity of vertebrate (especially human) populations and strong environmental effects (15–19).

To study the combination of environmental and host genetic factors that shape composition of the gut microbiota, we investigated a large murine intercross model in which genetic background can be systematically evaluated while environmental factors are carefully controlled. In this model, we quantified variation in taxonomic composition of gut microbiota and estimated the effects of maternal environment and host genotype. We used quantitative trait loci (QTL) analysis to test whether specific taxa cosegregate as quantitative traits with linked genomic markers. Using sophisticated methods for quantitative microbiota analysis and a suitably large number of genomic polymorphic markers, we have identified significant QTL that control variability in the abundances of different taxa in the mouse gut microbiome. We found that gut microbiota composition as a whole can be understood as a complex, polygenic trait influenced by combinations of host genomic loci and environmental factors.

Results

Core Measurable Microbiota in the G_4 Intercross Population. The availability of a large murine advanced intercross line (AIL) mapping population developed and maintained in a controlled environment (20) gave us a unique opportunity to examine the distribution of gut microbial taxa in a population of known pedigree. The random and sequential intercrossing over multiple generations in the AIL population increases the chance of recombination; as a result, AILs offer greater mapping resolution and narrower confidence intervals compared with a typical F_2 mapping population (21). The breeding protocol that created the AIL used in our study effectively expanded the mapping space 3-fold from that of a standard murine map (20).

The microbiota were phenotyped by pyrosequencing of 16S rDNA, generating a detailed and quantitative estimate of the

Author contributions: A.K.B., S.D.K., E.N.M., and D.P. designed research; S.A.K., R.L., F.M., S.J.L., J.K., M.Z., P.L.O., D.N., K.H., and D.P. performed research; F.M., J.K., and E.N.M. contributed new reagents/analytic tools; A.K.B., S.A.K., R.L., F.M., S.J.L., P.L.O., S.D.K., E.N.M., J.W., D.A.P., and D.P. analyzed data; and A.K.B., S.A.K., S.D.K., E.N.M., J.W., D.A.P., and D.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: abenson1@unl.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1007028107/-/DCSupplemental.

taxonomic composition of gut microbiota across the entire population of AILs. To accommodate this massive amount of data and to estimate covariation of phylogenetically related taxa up and down taxonomic ranks, we used the CLASSIFIER algorithm to predict relative abundances of organisms (22). The CLASSIFIER, which assigns taxonomic rank to sequence reads by matching distributions of nucleotide substrings to a model defined from sequences of known microorganisms, detected 420 genera, 143 families, 53 orders, 24 classes, and 16 phyla in the 645 samples sequenced. The relative abundances of the major phyla (Firmicutes, 30–70%; Bacteriodetes, 10–40%; Proteobacteria, 1–15%; Actinobacteria, Tenericutes, TM7, and Verrucomicrobia, 0.1–0.5%) were very similar to those reported for cecal sampling from murine models (7). CLASSIFIER assignments were validated by SEQMATCH (Table S1). Many genera were found in only a few animals; only a small number of genera were distributed quantitatively across most or all animals (Fig. 1*A*). These taxa—ones that are largely conserved and that vary quantitatively, and whose abundance can be accurately estimated from pyrosequencing data—were the focus of our analysis. Data from multiple technical repeats of five different samples (Fig. 1*B*) identified a minimum of 30 sequence reads for a given taxon as the threshold for quantitative repeatability. This threshold was subsequently applied as an average of 30 reads per bin across the entire mapping population. We define the resulting 19 genera and a total of 64 different taxonomic groups as a *core measurable microbiota* (CMM) (Table S2). Although the CMM genera represent only a small portion of the 420 total genera that we detected, they account for >90% of the sequence reads that were assigned to a genus by the CLASSIFIER, and thus define taxa that constitute a significant portion of the identifiable and quantifiable portion of the total microbiota. The CMM are log-normally distributed across the mapping population (Fig. 1*C*), with most genera distributed in a relatively narrow range of relative abundances and a small number of taxa, such as *Turicibacter*, showing a broader range (Table S2).

Litter and Cohort Have Significant Effects on Gut Microbiota Composition. If the relative abundances of the CMM are considered as complex traits, then the variation represented in their log-normal distributions would be a result of both environmental factors and host genetics. Given the well-defined nature of this large, segregating AIL population, our pyrosequencing data gave us the opportunity to evaluate systematically the relative contribution of separate apparent forces, such as the maternal environment and host genetics, a task that has not yet been accomplished in such a population.

As expected, environmental effects were readily observed by a mixed-model analysis (Table S3), which included fixed effects for parent of origin and sex along with random effects for cohort and family (nested with parent of origin) and litter (nested with cohort). On average, cohort accounted for 26% of the variation in taxa of the CMM (Table S4). Family and litter each accounted for about 5% of the variation in taxa of the CMM, with over half of the taxa showing litter effects that were significantly different from 0 ($P < 0.05$) (Table S3). Whereas variation between families and variation within litter include both a genetic component and an environmental component, variation between litters within a family includes only an environmental component, thereby leaving host genetics to explain significant proportions of the variation.

Composition of the Gut Microbiota Behaves as a Polygenic Trait. We used QTL analysis to assess the degree to which host genotype contributes to the variation in CMM across the AIL mapping population. The proportion (Prop) of each CMM taxon at each taxonomic rank was treated as an individual trait and tested for cosegregation with 530 fully informative SNP markers. Although AILs enhance mapping resolution, the complex breeding history of our study population falsified the assumption of independence

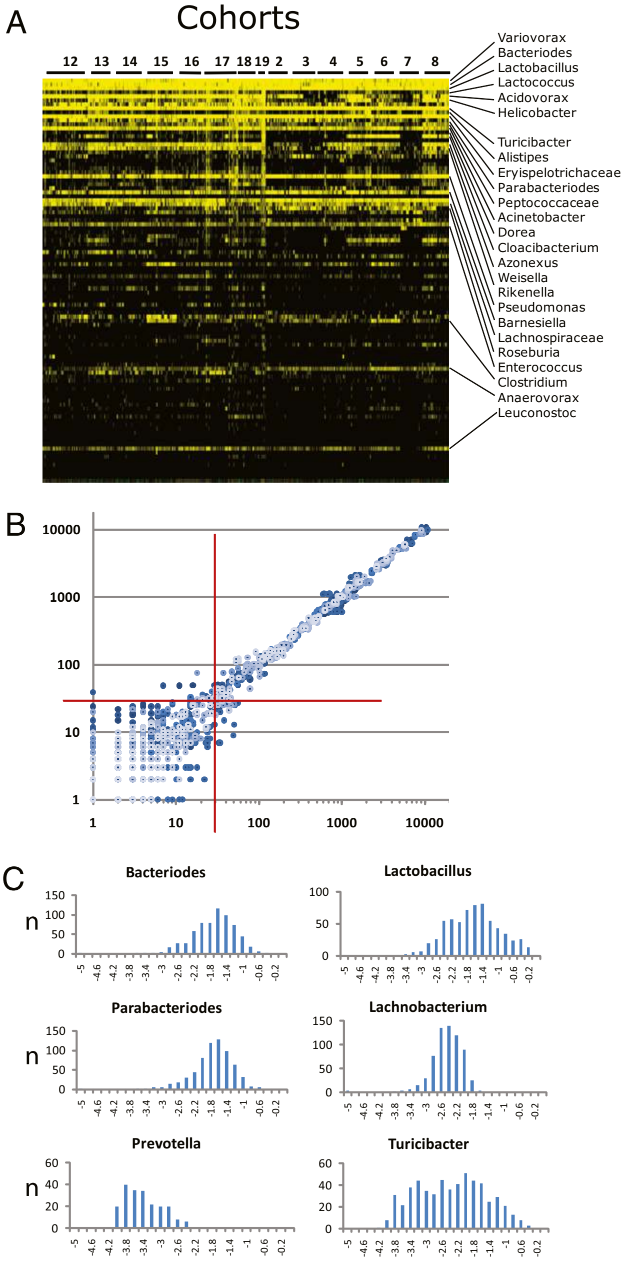

Fig. 1. Characterization of the gut microbiota across the AIL population. (*A*) A heat map of the relative abundance of the top 100 genera identified in the G_4 AIL population. Vertical columns represent individual animals; horizontal rows depict genera. Genera of interest are indicated. Black indicates absent taxa. (*B*) A scatterplot generated from pairwise combinations of data from technical repeats from five different samples. 16S rDNA from each sample was amplified with three different sets of bar-coded primers. Processed and filtered sequences from each barcode–sample combination were then assigned taxonomy by CLASSIFIER. Sequence counts for each taxonomic bin were log-transformed and plotted for all pairwise combinations of the three repeats for each sample. Axes are the log10-transformed values for total sequence reads of each taxon. The red crosshairs indicate the 30-read threshold. Above this number, correlation reaches >0.998: below this number, correlation dissipates rapidly. (*C*) Histograms of the frequency distribution of selected CMM taxa across the 645 animals. The histograms were plotted from log10-transformed values of the proportion (Prop) of sequence reads for each taxon (i.e., number of reads for that taxon/total sequence reads for a given animal). Thus, each histogram depicts the number of animals (*y* axis) with log10-transformed Prop values (*x* axis) for the given taxon.

among individuals and made conventional mapping strategies inappropriate. To overcome this problem, we used the genome reshuffling for advanced intercross permutation (GRAIP) procedure, which estimates parental (F_3 in our case) genotypes and uses a permutation scheme to simulate sets of F_3 progenitors (23). From these progenitor sets, recombination and inheritance are simulated, creating randomized G_4 populations (n = 50,000) that respect the original family structure while removing any association between genotype and phenotype. QTL analyses are then performed on the original and GRAIP-permuted populations. Locus-specific and genome-wide empirical P values are estimated using the distribution of P values from the permuted maps.

With the GRAIP procedure, 26 out of 64 taxonomic groups of organisms from the CMM showed association with 13 significant QTL (LOD $\geq$ 3.9; $P < 0.05$) and 5 additional suggestive QTL (LOD $\geq$ 3.5; $P < 0.1$). Results for significant and suggestive QTL and associated data are shown in Tables S5 and S6. QTL positions relative to the genomic markers and the phylogenetic relationships of the corresponding taxa are illustrated in Fig. 2. Each QTL individually accounted for 1.6–9.0% of the total phenotypic variation; average additive effects were frequently significant, and dominance effects were especially large for the Proteobacteria. Genetic control is exerted across the entire phylogenetic space of the gut microbiota, with at least one taxon from each of the four major phyla mapping to a significant QTL. The QTL were dispersed over eight chromosomes, with multiple QTL mapping to MMU1, MMU7, and MMU10 (Fig. 2). This pattern of cosegregation in our intercross population now provides direct evidence that composition of the gut microbiota as a whole is heritable as a complex, polygenic trait.

Host genetic control appears to focus largely on the tips of the phylogenetic tree. This phenomenon was particularly apparent in diverse groups of organisms (e.g., Bacteriodetes, Clostridia, Bacilli) in which QTL were observed only at the genus and family levels. Phylum- or class-level QTL were apparent only in the Actinobacteria, Erysipilotrichi, and Epsilon classes of the Proteobacteria, which were each dominated numerically by a few taxa (e.g., Coriobacteriaceae within the Actinobacteria, Turicibacter within the Erysipilotrichi, *Helicobacter* within the Epsilon) that accounted for the QTL signal.

QTL for Host-Adapted Species of Lactobacilli. Among the CMM organisms, only the genera *Helicobacter* and *Lactobacillus* are known to form close physical associations with host tissues, a characteristic that would be expected to be modulated by host factors. Significant QTL were detected for *Helicobacter*, but no QTL were identified for *Lactobacillus* (Table S1). Lactobacilli form dense cell layers on the murine forestomach epithelium, and its isolates' adherence phenotypes have been shown to be host-specific (24, 25); *L. reuteri* even comprises host-adapted subpopulations (26). This degree of host adaptation at the species level and below, and the fact that no QTL were detected at the genus level, led us to speculate that it may be precisely at the lower taxonomic ranks that host genetic control over *Lactobacilli* is exerted. To test for cosegregation at the species level, we mapped as individual traits the relative abundance of three groups with 97% identity: *L. reuteri*, *L. johnsonii*/*L. gasseri*, and *L. animalis*/*L. murinus* (Fig. S1). Indeed,

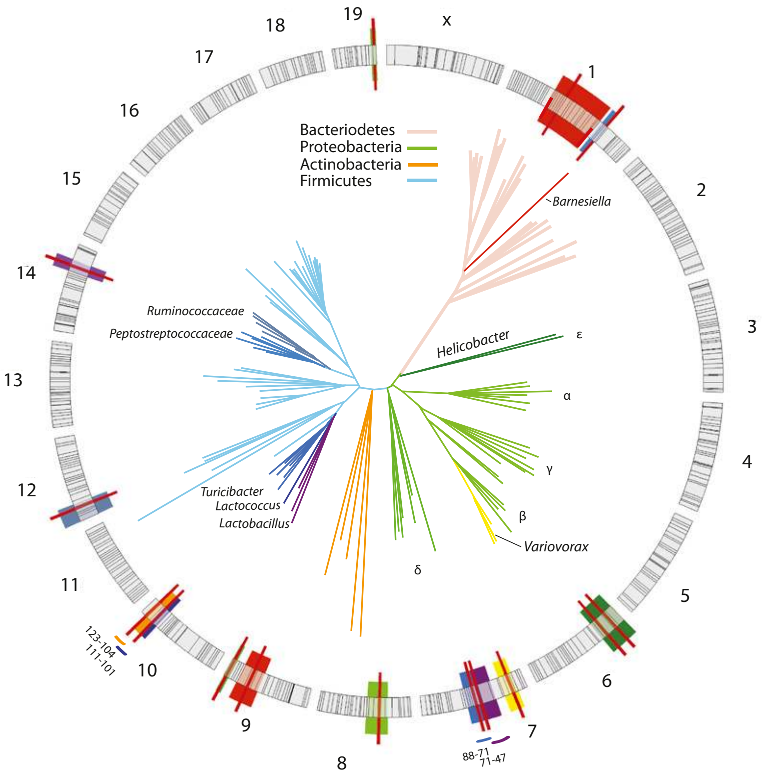

Fig. 2. QTL mapping of the murine gut microbiota. The circular diagram depicts the 19 murine autosomes and X chromosome drawn to scale. Black lines mark the positions of the SNPs used for QTL mapping. QTL confidence intervals are shaded in colors that correspond to the branches of the organism(s) in the phylogenetic tree. QTL peaks are marked by solid red lines. Color-coded bars outside the circle indicate confidence intervals of adjacent QTL. Coordinates of the confidence intervals (in Mb) are also indicated. The representative phylogenetic tree was derived from 100,000 sequences randomly drawn from the total data set of 5.2. The sampled sequences were clustered with CD-Hit; representative sequences of the most abundant 200 clusters were used for phylogenetic analysis by the neighbor-joining method. Major phyla are color-coded.

the *L. johnsonii*/*gasseri* group segregated with two significant QTL on MMU14 and MMU7 (Table S5), implying that intimate associations between the host and its microbiota are subject to heritable genetic factors.

Some QTL Have Pleiotropic Effects on the Gut Microbial Taxa. Several QTL appear to have pleiotropic effects on multiple taxa and these effects can be divided into three groups. The first group includes QTL that affect relatively closely related organisms, such as the QTL for *L. johnsonii*/*gasseri* on MMU7 (peak at 66 Mb), which is adjacent to the QTL for *Turicibacter* (peak at 73 Mb), with overlapping confidence intervals. Colocalization of these QTL implies that MMU7 may encode a gene that influences both taxa, or that this region contains linked genes that, individually or in combination, affect gut microbiota composition.

The QTL for the phylum Proteobacteria exemplifies the second type of pleiotropy. Here the peak and confidence interval for a QTL on MMU6 at 28 Mb are nearly identical to those of a *Helicobacter* QTL. Thus, this single phylum-level QTL may have significant effects on the ability of *Helicobacter* to colonize the murine GI tract along with a broader effect on the entire Proteobacteria population. This finding underscores the importance of testing for cosegregation at different levels of taxonomic hierarchy. A second QTL on MMU8 was also associated with the phylum Proteobacteria, distinct from all other QTL for lower taxonomic ranks of Proteobacteria, implying that the relative abundance of an entire Phylum can be controlled by a single genomic locus.

Finally, a third type of pleiotropy can be found for the genus *Lactococcus* (phylum Firmicutes) and the family Coriobacteriaceae (phylum Actinobacteria). These QTL colocalize in the 104–123 Mb region of MMU10, with peaks at 107 Mb and 119 Mb, respectively. These organisms, unlike those in the first two groups of pleiotropic QTL, have a very distant phylogenetic relationship. Nonetheless, they show a positive correlation in the data set and have either shared gene action or overlapping QTL, with significant dominance effects of the C57BL/6J allele (Table S5). Thus, the effect of these colocalizing QTL was to cause positive correlation between the relative abundances of Coriobacteriaceae and *Lactococcus*, illustrating the significance of host genetic influence on the population structure in the gut.

Discussion

From an essentially sterile state at birth, the gut ecosystem develops rapidly as microbes successively colonize vacant niches. In humans, this period of succession persists until 18–24 mo of age, when the gut microbiota attains its "adult-like" composition and begins to behave as a highly individualized climax community (1, 27, 28). Despite tremendous diversity of the gut microbial species, many of which are sparsely distributed between individual hosts, recent work has revealed that a core of >50 taxa are found in nearly half of human subjects sampled (29, 30). This finding is consistent with the observations in our large murine population under controlled conditions (Fig. 1*A*). Our discovery that the CMM taxa, which are some of the most abundant organisms in the GI tract, are subject to host genetic control now supports the concept of a core microbiome as a universal feature among vertebrate hosts, with the relative abundances of CMM taxa collectively behaving as a complex polygenic trait. This glimpse of the host genetic architecture underpinning gut microbiota composition was attained under the highly controlled environmental conditions of our murine intercross population, and shows that these genetic effects are broadly distributed across the dominant CMM phyla (Fig. 2) and can influence very specific groups of organisms or have pleiotropic effects on diverse taxonomic groups.

Establishment of this murine model and demonstration of heritability are important steps toward experimental paradigms that can define the mechanisms which drive the assembly of the microbiota in individuals. As an example, we again turn to the colocalized QTL for the Coriobacteriaceae and *Lactococcus* that span a 15-Mb region on MMU10 (Fig. 2). As shown in Fig. 3*A*, these QTL are closely positioned and control Gram-positive organisms, which is consistent with several genes in this region, namely *Irak3*, which modulates MyoD88-dependent peptidoglycan (PGN)-stimulated responses of the TLR2 pathway (31), and the two primary murine lysozyme genes, *Lyz1* and *Lyz2* (32). The same interval also contains genes encoding IFN-γ (*Ifng*) and IL-

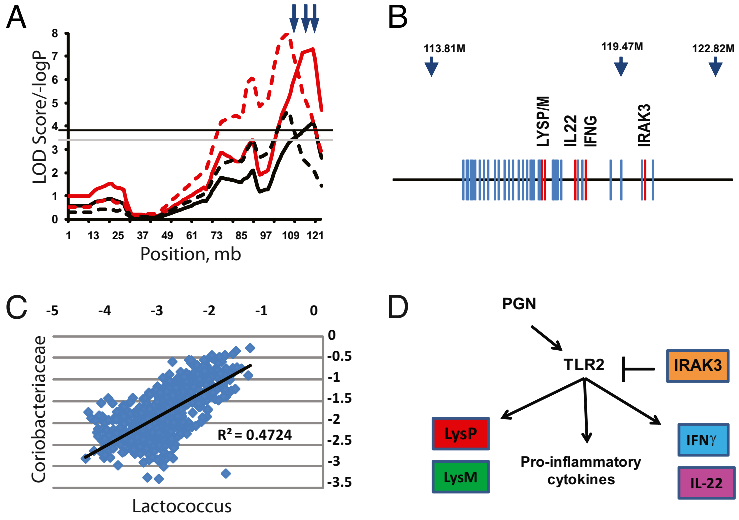

Fig. 3. Fine structure of the genomic region of the significant QTL on chromosome 10. (*A*) The simple mapping output (red lines) and GRAIP permutation output (black lines) for QTL analysis of Coriobacteriaceae (solid lines) and *Lactococcus* (dashed lines). Genome-wide GRAIP-adjusted significance thresholds were generated from 50,000 permutations. Thus for the GRAIP output, a minimum possible *P* value with 50,000 permutations is 0.00002 (1/50,000), so the maximum −log *P* is 4.7. The black and gray horizontal lines represent the permuted 95% and 90% LOD thresholds, respectively. Arrows at the top show the relative positions of the three SNP markers nearest the QTL. (*B*) A scaled gene map of the QTL region. Arrows indicate SNP markers and their positions (in Mb). Genes are marked by blue; genes of interest are in red. (*C*) A scatterplot of log-transformed Prop values from the Coriobacteriaceae and the *Lactococcus* taxon bins of the 645 animals used in the study. (*D*) The combined functional pathways of the genes of interest in the QTL across multiple cell types. The bar from IRAK3 to TLR2 represents direct action. Arrows represent the relative influence of each gene and not necessarily direct gene action.

22 (*Il22*), which play substantial roles in mucosal immunity, where they shape T cell development and elicit antibacterial responses in intestinal epithelial cells (33, 34). Lactococci have only recently been observed in the GI tract through pyrosequencing data, but members of the Coriobacteriaceae (e.g., *Eggerthella*, *Enterorhabdus*) are associated with mouse models of inflammatory disease (35, 36). The significance of this QTL is underscored by the strong correlation of these two taxa (Fig. 3*C*) due, at least in part, to the QTL effect.

The *Il22* gene is duplicated in the C57BL/6J genome, making it tempting to speculate that this duplication at least partially accounts for the MMU10 QTL effect. Indeed, in G_4 progeny homozygous for the C57BL/6J allele of the JAX0030095 marker (at 119 Mb, adjacent to *Il22*), the Coriobacteriaceae and *Lactococcus* are both significantly less abundant (Fig. S3). Although this result would be anticipated, it is not clear whether the duplicated gene, which is truncated, is actually functional (37). Given the collective antimicrobial functions of genes within this cluster, an alternative explanation is that cumulative allelic variation in several candidate genes in this region accounts for the overall QTL effect, as has been previously observed for several QTL that were dissected into subregions through congenic analysis (38, 39). The mapping power of our approach will increase as we continue into later generations of the AIL (now at G_{10}). Moreover, new genetic resource populations that will soon be available, such as the Collaborative Cross (40, 41), will increase the genomic search space, ultimately allowing the discovery of new QTL for gut microbiota and the refinement of QTL signals to fewer candidate genes.

Fundamentally, the pattern of host genetic control that we observed is consistent with the broader effects of evolutionary divergence of the gut microbiota composition across many host species (2–4). Specifically, the effects of host genetics, like those of host speciation, involve all dominant phyla and favor selection at the tips of the phylogenetic tree. Such patterns could be predicted to emerge from host speciation events that involve concerted divergence of complex sets of loci (e.g., different QTL) and corresponding stepwise changes in the microbial populations they control. This could explain the evolution of highly specialized mammalian organs (e.g., foregut, hindgut, ceca) that harness microbes for fermentation of fibrous plant materials (42). By exerting top-down selection pressure, host genetic control would subdue microbial competition within the gut ecosystem to promote microbes that benefit the host at the cost of their own competitive fitness. This view is consistent with the suggestion that the adaptive immune system has specifically evolved in vertebrates to regulate and maintain beneficial microbial communities (43). Important insights into this question will clearly emerge from QTL analyses across multiple host species.

Beyond the fundamental significance for host–microbe interactions, demonstrating that heritable traits affect the gut microbiota also may shed new light on our understanding of complex diseases. In many ways, the gut microbiota does behave as an environmental factor implicated in fat storage (14) or immune system development (44–46). However, our work shows that the gut microbiota can now be viewed as an environmental factor that itself is controlled in part by host genetic factors and potentially by interactions between host and microbial genomes. This view implies that genetic predisposition to complex diseases may be manifested in part by a predisposition to aberrant patterns of microbial colonization, which in turn contribute to disease processes. This concept is reinforced by recent studies in monogenic models showing that both aberrations in gut microbiome composition and characteristics of complex diseases can be caused by a single null mutation (9, 36, 47, 48). Moreover, it is interesting to point out that *Turicibacter*, *Barnesiella*, and members of the Coriobacteriaceae—taxa that we have now shown to be controlled by QTL—are associated with complex disease characteristics in murine models (36, 49); in each instance, the confidence intervals of our QTL overlap known QTL for complex diseases. For example, the QTL for *Turicibacter* of MMU7 overlaps the HCS1 QTL for susceptibility to murine hepatocellular carcinomas (50), whereas the QTL for Coriobacteriaceae on MMU10 overlaps the Scc9 locus associated with murine susceptibility to colon tumors (51). The QTL on MMU1 for *Barnesiella* also overlaps the conserved gene ATG16L, and this region is syntenic with the ATG16L region of the human chromosome 2 (234Mb region) recently shown to be associated with Crohn's disease (52). Although these discoveries were made in different genetic backgrounds, and the confidence intervals of each QTL contain many genes, it will be interesting to see if any of these loci have pleiotropic effects on both microbiota abundance and disease. Conversely, for complex diseases whose genetic architecture is already well defined, such as the >200 QTL mapped for traits related to obesity (53), our discovery now begs the question of whether some of these QTL could manifest their phenotypes through their effects on gut microbiome composition and, if so, which organisms they affect.

Similarly, the CMM concept can now be translated to genome-wide association studies in humans, in which dense panels of well-defined genomic markers can be tested for association with CMM characteristics. We believe that, with highly refined data from murine models, mapping heritable genetic factors controlling gut microbiome composition will ultimately be an important tool for studying disease. This strategy is also applicable to agriculturally relevant food animals, where host genetic control is likely to be implicated in colonization by zoonotic pathogens as well as organisms important for ruminal fermentations and feed intake phenotypes.

Methods

Animal Population. A moderately (G_4) advanced intercross line (AIL) was bred from reciprocal crosses between the inbred strain C57BL/6J and the ICR-derived HR line (54). In brief, F_3 breeding pairs produced multiple litters to expand ($n = 815$) the G_4 population, with staggered mating to reduce intergroup age variation. To accommodate phenotyping constraints, G_4 individuals were divided into 19 consecutive cohorts of ~45 mice each, with approximately even numbers of both sexes. After weaning, G_4 animals were group-caged by sex and provided ad libitum access to a repeatable synthetic diet (Research Diet D10001) and water. At ~8 wk of age, mice were caged individually; the following day, fecal samples were collected and stored at −30 °C.

Deep Pyrosequencing of the Gut Microbiota. DNA extraction from fecal pellets and pyrosequencing have been described previously (55). The V1-V2 region of the 16S rRNA gene was amplified using bar-coded fusion primers with the Roche-454 A or B Titanium sequencing adapters (see *SI Methods*). Of the 709 G_4 animals' samples, robust PCR products were obtained from 645 samples. Pooled and gel-purified amplicon products were sequenced using Roche-454 GS FLX Titanium chemistry.

Pyrosequencing Data Processing Pipelines. Raw reads were filtered according to length and quality criteria (see *SI Methods*). Filter-pass reads were parsed into sample-barcoded bins and uploaded to a publicly accessible MySQL database (http://cage.unl.edu). More than 5.2 million quality-filtered reads were obtained from 645 samples, an average of 8,000 reads per animal. Reads were assigned taxonomic status with a parallelized version of the multi-CLASSIFIER algorithm (22), and reads in each taxonomic bin were normalized as the absolute proportion (Prop) of the total number of reads for each sample (see *SI Methods*). These Prop values for each taxon were used as "traits" for QTL analysis.

To confirm taxonomic assignments, we randomly sampled 40,000 sequences from genus-level bins and checked best-hits from the RDP database using SeqMatch (Table S1). In addition, we validated the quantitative nature of the pyrosequencing data by qPCR using *Lactobacillus*-specific primers (56), which yielded highly significant correlation ($r > 0.64$; Fig. S2).

QTL Analysis. Prop values of microbial taxa were log10-transformed, and for animals for which no counts were obtained for a given taxon, a value of 0.5/total reads was log10-transformed and used. Each individual microbial "trait" was then evaluated for location and magnitude of QTL. Complete descriptions of the marker genotyping and the final set of SNPs ($n = 530$, with an average spacing of 4.7 Mb) used in the QTL analyses are provided elsewhere (20). To account for the G_4 family structure (nonindependence of individuals), we used the GRAIP procedure (23), as described previously (20). Details of the QTL analysis are presented in *SI Methods*.

ACKNOWLEDGMENTS. We thank the members of the Ribosomal Database Project at Michigan State University for kindly providing the parallelized CLASSIFIER and training data sets. We gratefully acknowledge Theodore Garland, Jr., University of California-Riverside, for providing the original HR mice that contributed to creation of the G$_4$ population. We benefited significantly from discussions with David Threadgill (North Carolina State University). We thank Nina Murray (University of Nebraska-Lincoln) for editing the manuscript and creating the artwork. This project was supported by National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Grants RC1DK087346 (to A.K.B. and D. P.) and DK076050 (to D.P.). Some phenotypes were collected using the Animal Metabolism Phenotyping core facility at the University of North Carolina's Nutrition Obesity Research Center (funded by NIDDK Grant DK056350).

1. Tannock GW (2007) What immunologists should know about bacterial communities of the human bowel. *Semin Immunol* 19:94–105.
2. Dethlefsen L, McFall-Ngai M, Relman DA (2007) An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature* 449:811–818.
3. Ley RE, et al. (2008) Evolution of mammals and their gut microbes. *Science* 320: 1647–1651.
4. Ley RE, Peterson DA, Gordon JI (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* 124:837–848.
5. Antonopoulos DA, et al. (2009) Reproducible community dynamics of the gastrointestinal microbiota following antibiotic perturbation. *Infect Immun* 77:2367–2375.
6. Carroll IM, Threadgill DW, Threadgill DS (2009) The gastrointestinal microbiome: A malleable, third genome of mammals. *Mamm Genome* 20:395–403.
7. Ley RE, et al. (2005) Obesity alters gut microbial ecology. *Proc Natl Acad Sci USA* 102: 11070–11075.
8. Fava F, Lovegrove JA, Gitau R, Jackson KG, Tuohy KM (2006) The gut microbiota and lipid metabolism: Implications for human health and coronary heart disease. *Curr Med Chem* 13:3005–3021.
9. Wen L, et al. (2008) Innate immunity and intestinal microbiota in the development of type 1 diabetes. *Nature* 455:1109–1113.
10. Frank DN, et al. (2007) Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci USA* 104:13780–13785.
11. Lindgren CM, et al.; Wellcome Trust Case Control Consortium Procardis Consortia Giant Consortium (2009) Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLoS Genet* 5:e1000508.
12. Schmidt C, et al. (2008) A meta-analysis of QTL for diabetes-related traits in rodents. *Physiol Genomics* 34:42–53.
13. van Heel DA, et al.; Genome Scan Meta-Analysis Group of the IBD International Genetics Consortium (2004) Inflammatory bowel disease susceptibility loci defined by genome scan meta-analysis of 1952 affected relative pairs. *Hum Mol Genet* 13: 763–770.
14. Bäckhed F, et al. (2004) The gut microbiota as an environmental factor that regulates fat storage. *Proc Natl Acad Sci USA* 101:15718–15723.
15. Zoetendal EG, et al. (2001) The host genotype affects the bacterial community in the human gastronintestinal tract. *Microb Ecol Health Dis* 13:129–134.
16. Van de Merwe JP, Stegeman JH, Hazenberg MP (1983) The resident faecal flora is determined by genetic characteristics of the host: Implications for Crohn's disease? *Antonie van Leeuwenhoek* 49:119–124.
17. Turnbaugh PJ, et al. (2009) A core gut microbiome in obese and lean twins. *Nature* 457:480–484.
18. Khachatryan ZA, et al. (2008) Predominant role of host genetics in controlling the composition of gut microbiota. *PLoS One* 3:e3064.
19. Deloris Alexander A, et al. (2006) Quantitative PCR assays for mouse enteric flora reveal strain-dependent differences in composition that are influenced by the microenvironment. *Mamm Genome* 17:1093–1104.
20. Kelly SA, et al. (2010) Genetic architecture of voluntary exercise in an advanced intercross line of mice. *Physiol Genomics* 42:120–200.
21. Darvasi A, Soller M (1995) Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics* 141:1199–1207.
22. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261–5267.
23. Peirce JL, et al. (2008) Genome Reshuffling for Advanced Intercross Permutation (GRAIP): Simulation and permutation for advanced intercross population analysis. *PLoS One* 3:e1977.
24. Fuller R, Brooker BE (1974) Lactobacilli which attach to the crop epithelium of the fowl. *Am J Clin Nutr* 27:1305–1312.
25. Savage DC, Dubos R, Schaedler RW (1968) The gastrointestinal epithelium and its autochthonous bacterial flora. *J Exp Med* 127:67–76.
26. Oh PL, et al. (2010) Diversification of the gut symbiont *Lactobacillus reuteri* as a result of host-driven evolution. *ISME J* 4:377–387.
27. Palmer C, Bik EM, DiGiulio DB, Relman DA, Brown PO (2007) Development of the human infant intestinal microbiota. *PLoS Biol* 5:e177.
28. Mackie RI, Sghir A, Gaskins HR (1999) Developmental microbial ecology of the neonatal gastrointestinal tract. *Am J Clin Nutr* 69:1035S–1045S.
29. Qin J, et al.; MetaHIT Consortium (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464:59–65.
30. Tap J, et al. (2009) Towards the human intestinal microbiota phylogenetic core. *Environ Microbiol* 11:2574–2584.
31. Nakayama K, et al. (2004) Involvement of IRAK-M in peptidoglycan-induced tolerance in macrophages. *J Biol Chem* 279:6629–6634.
32. Markart P, et al. (2004) Comparison of the microbicidal and muramidase activities of mouse lysozyme M and P. *Biochem J* 380:385–392.
33. De Kimpe SJ, Kengatharan M, Thiemermann C, Vane JR (1995) The cell wall components peptidoglycan and lipoteichoic acid from *Staphylococcus aureus* act in synergy to cause shock and multiple organ failure. *Proc Natl Acad Sci USA* 92: 10359–10363.
34. Zheng Y, et al. (2008) Interleukin-22 mediates early host defense against attaching and effacing bacterial pathogens. *Nat Med* 14:282–289.
35. Clavel T, et al. (2009) Isolation of bacteria from the ileal mucosa of TNFdeltaARE mice and description of *Enterorhabdus mucosicola* gen. nov., sp. nov. *J Syst Evol Microbiol* 59:1805–1812.
36. Ye J, et al. (2008) Bacteria and bacterial rRNA genes associated with the development of colitis in IL-10(-/-) mice. *Inflamm Bowel Dis* 14:1041–1050.
37. Dumoutier L, Van Roost E, Ameye G, Michaux L, Renauld JC (2000) IL-TIF/IL-22: Genomic organization and mapping of the human and mouse genes. *Genes Immun* 1: 488–494.
38. Farber CR, Medrano JF (2007) Dissection of a genetically complex cluster of growth and obesity QTLs on mouse chromosome 2 using subcongenic intercrosses. *Mamm Genome* 18:635–645.
39. Jerez-Timaure NC, Eisen EJ, Pomp D (2005) Fine mapping of a QTL region with large effects on growth and fatness on mouse chromosome 2. *Physiol Genomics* 21: 411–422.
40. Churchill GA, et al.; Complex Trait Consortium (2004) The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat Genet* 36: 1133–1137.
41. Chesler EJ, et al. (2008) The Collaborative Cross at Oak Ridge National Laboratory: Developing a powerful resource for systems genetics. *Mamm Genome* 19:382–389.
42. Stevens CE, Hume ID (1998) Contributions of microbes in vertebrate gastrointestinal tract to production and conservation of nutrients. *Physiol Rev* 78:393–427.
43. McFall-Ngai M (2007) Adaptive immunity: Care for the community. *Nature* 445:153.
44. Wang Q, et al. (2006) A bacterial carbohydrate links innate and adaptive responses through Toll-like receptor 2. *J Exp Med* 203:2853–2863.
45. Mazmanian SKL, Liu CH, Tzianabos AO, Kasper DL (2005) An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell* 122:107–118.
46. Mazmanian SK, Round JL, Kasper DL (2008) A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature* 453:620–625.
47. Vijay-Kumar M, et al. (2010) Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. *Science* 328:228–231.
48. Turnbaugh PJ, et al. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444:1027–1031.
49. Presley LL, Wei B, Braun J, Borneman J (2010) Bacteria associated with immunoregulatory cells in mice. *Appl Environ Microbiol* 76:936–941.
50. Gariboldi M, et al. (1993) Chromosome mapping of murine susceptibility loci to liver carcinogenesis. *Cancer Res* 53:209–211.
51. van Wezel T, Ruivenkamp CA, Stassen AP, Moen CJ, Demant P (1999) Four new colon cancer susceptibility loci, Scc6 to Scc9 in the mouse. *Cancer Res* 59:4216–4218.
52. Parkes M, et al.; Wellcome Trust Case Control Consortium (2007) Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat Genet* 39:830–832.
53. Pomp D, Allan MF, Wesolowski SR (2004) Quantitative genomics: Exploring the genetic architecture of complex trait predisposition. *J Anim Sci* 82(E-Suppl):E300–312.
54. Kelly SA, et al. (2010) Parent-of-origin effects on voluntary exercise levels and body composition in mice. *Physiol Genomics* 40:111–120.
55. Martínez I, et al. (2009) Diet-induced metabolic improvements in a hamster model of hypercholesterolemia are strongly linked to alterations of the gut microbiota. *Appl Environ Microbiol* 75:4175–4184.
56. Walter J, et al. (2001) Detection of *Lactobacillus*, *Pediococcus*, *Leuconostoc*, and *Weissella* species in human feces by using group-specific PCR primers and denaturing gradient gel electrophoresis. *Appl Environ Microbiol* 67:2578–2585.

Supporting Information

Benson et al. 10.1073/pnas.1007028107

SI Methods

Pyrosequencing. The V1-V2 region of the 16S rRNA gene was amplified using bar-coded fusion primers with the Roche-454 A or B titanium sequencing adapters (in italics), followed by a unique 8-base barcode sequence (B) and finally the 5′ ends of primer A-8FM (5′-*CCATCTCATCCCTGCGTGTCTCCGACTCAG*BBBBBBBBAGAGTTTGATCMTGGCTCAG) and of primer B-357R (5′-*CCTATCCCCTGTGTGCCTT-GGCAGTCTCAG*BBBBBBBBCTGCTGCCTYCCGTA-3′). All PCR reactions were quality- controlled for amplicon saturation by gel electrophoresis; band intensity was quantified against standards using GeneTools software (Syngene). For each region of a two-region picotiter plate, amplicons from 48 reactions were pooled in equal amounts and gel-purified. The resulting products were quantified using PicoGreen (Invitrogen) and a Qubit fluorometer (Invitrogen) before sequencing using Roche-454 GS FLX titanium chemistry.

Data Processing Pipeline. The raw data from the 454 pyrosequencing machine were first processed through a quality filter that removed unqualified sequence reads that did not meet the following criteria:

1. A complete forward primer and barcode
2. ≤ 2 "N" in a sequence read, where N is equivalent to an interrupted and resumed signals from sequential flows
3. 200 nt $\leq$ sequence length $\leq$ 500 nt
4. Average quality score ≥ 20.

After filtering, each read was trimmed to remove 3′ adapter and primer sequences and was parsed by barcode. The corresponding .QUAL file also was updated to remove quality scores from reads not passing quality filters. The files are associated with sample information in a hierarchical manner in MySQL tables. The processed data and the MySQL database tables are stored on a database server and available to the public at http://cage.unl.edu.

Given the massive size of the pyrosequencing data set and the need to normalize the taxonomy across the entire data set in a hierarchical fashion, a limited number of current algorithms could be modified and implemented. The CLASSIFIER algorithm assigns taxonomic status to each sequence read based on a covariance model developed from a training set (1). This algorithm is capable of processing very large data sets and was recently shown to provide adequate taxonomic assignments to pyrosequencing data (2). We implemented a parallelized version of the CLASSIFIER (kindly provided by the Center for Microbial Ecology, Ribosomal Database Project at Michigan State University), using the standard threshold of 0.8, with reads classified down to the lowest level until the score <0.8, at which point reads are classified as "unclassified" at the next-higher taxonomic rank.

The hierarchical output data from the from CLASSIFIER were further processed by computing the absolute proportion of each sequence, calculated as

$$absolute\ proportion = \frac{\#reads\ of\ a\ taxon}{total\ number\ of\ reads\ in\ a\ sample}$$

The absolute proportion is referred to as the Prop value. The multi-CLASSIFIER algorithm, proportion calculation, and assembly of the Prop table for the entire data set were performed sequentially on a Linux cluster of computer nodes, with the jobs controlled by the PBS portable batch system. The data were partitioned into a number of smaller groups, and the calculations were computed independently in a cluster node for each group, with the final results compiled when all were complete. At a threshold of 0.8, the data from all 645 animals in our data set included 420 different genera, 143 families, 53 orders, 24 classes, and 16 phyla that contained at least one assigned sequence. Of the 420 observed genera, 47 genera accounted for >99% of the sequences, and 19 accounted for >90% of the sequences.

To test the robustness of the CLASSIFIER algorithm, we compared the CLASSIFIER-based taxonomic assignments to the RDP database using SEQMATCH. Samples of 40,000 sequences assigned to one of several representative taxa were chosen and compared with the RDP database using the SEQMATCH program. Results for the top hits were compiled and are reported in Table S1.

Details of the QTL Analysis. QTL analyses generated *P* values for the original population and the GRAIP-permuted populations (n = 50,000); these were performed on log-transformed traits using the multiple-imputation method (3) within R/qtl (4). Statistical models included parent-of-origin type [i.e., whether a G_4 individual was descended from a progenitor (F_0) cross HR♀ X B6♂ or B6♀ X HR♂, coded as 1 or 0, respectively] and parity (i.e., order of litters from individual F_3 dams). The X chromosome was treated as an autosome, because R/qtl assumes a F_2 population and requires the identity of the cross direction. The output from R/qtl was then used to calculate locus-specific *P* values as described previously (5). Locus-specific *P* values were calculated for each marker of the original data set, using the value of that specific marker in each of the permuted maps at each locus as a null distribution. The null distribution for each marker was compared with the value for the original G_4 mapping data set to generate locus-specific *P* values at marker positions. These *P* values were interpolated onto the genome based on known physical positions of markers and placed on a scaffold at regular physical intervals. Finally, genome-wide, adjusted *P* values were computed by creating an ordered list of the minimum possible *P* values (or highest −log *P*, LOD) from each GRAIP-permuted map. Because we used 50,000 permutations, the minimum possible *P* value was 0.00002 (1/50,000) and the maximum –log *P* was 4.7. The 95th percentile (P = 0.05; LOD $\geq$ 3.9) and 90th percentile (P = 0.1; LOD $\geq$ 3.5) defined significant and suggestive loci, respectively. Confidence intervals were approximated by 1 LOD drop support intervals (relative to the GRAIP-permuted LOD score). Standard linear regression was used to estimate the percent variation by fitting the imputed QTL marker genotypes; the additive and dominance QTL effects were calculated using R/qtl.

1. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261–5267.
2. Liu Z, DeSantis TZ, Andersen GL, Knight R (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res* 36:e120.
3. Sen S, Churchill GA (2001) A statistical framework for quantitative trait mapping. *Genetics* 159:371–387.
4. Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19:889–890.
5. Peirce JL, et al. (2008) Genome Reshuffling for Advanced Intercross Permutation (GRAIP): Simulation and permutation for advanced intercross population analysis. *PLoS One* 3:e1977.
6. Walter J, et al. (2001) Detection of *Lactobacillus*, *Pediococcus*, *Leuconostoc*, and *Weissella* species in human feces by using group-specific PCR primers and denaturing gradient gel electrophoresis. *Appl Environ Microbiol* 67:2578–2585.

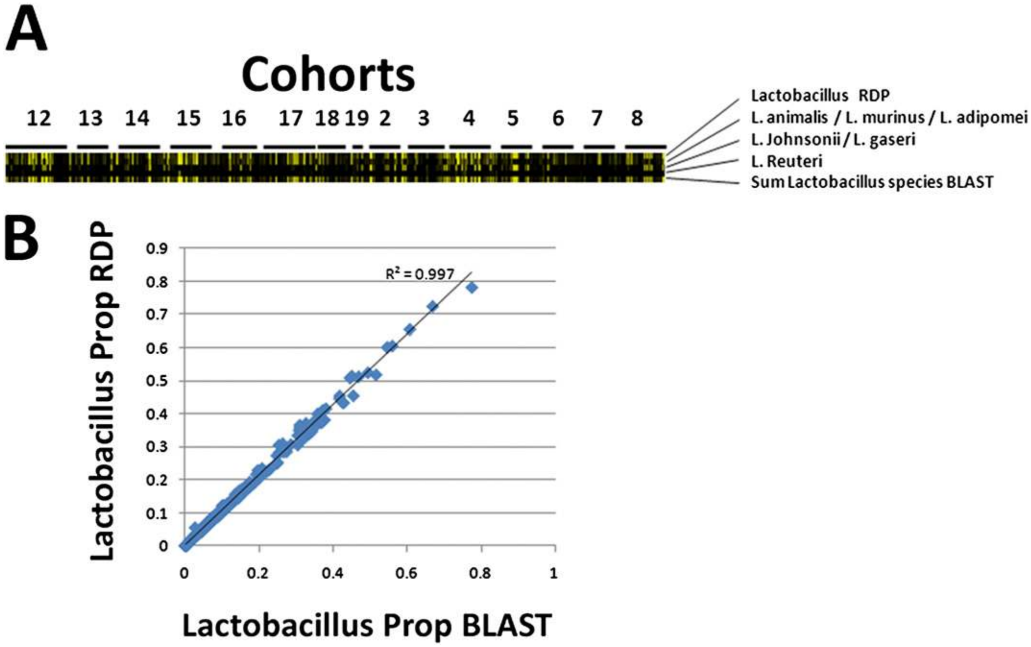

Fig. S1. BLAST Analysis of *Lactobacillus* species. To analyze the *Lactobacillus* at the species level, Bioedit v7.0.9 was used to perform a local nucleotide BLAST (blastn) search using the murine *Lactobacillus* type strain sequences: *L. animalis (*AB326350.1), *L. apodemi* (AJ871178.1), *L. murinus* (AB326349.1), *L. reuteri* (CP000705.1), *L. gasseri* (CP000413.1), and *L. johnsonii* (ACGR01000047.1). These sequences were trimmed to ∼340 nucleotides to match the length of the V1-V2 amplicons and used as queries against entire sets of read sequences from each sample with a 97% identity threshold for species assignment. The number of each *Lactobacillus* species hits for each sample was then divided by the total number of reads and used as the Prop value for the sample. (*A*) A heat map depicting the relative abundance of BLAST hit distribution for the species groups of *L. animalis/murinus/, L. johnsonii/gasseri*, and *L. reuteri*. The top row depicts the relative abundance of the genus *Lactobacillus* from the CLASSIFIER algorithm, and the bottom row shows the pooled cumulative Prop for BLAST hits of all three species groups. (*B*) A scatterplot of relative Prop of *Lactobacillus* from the RDP CLASSIFIER versus the cumulative Prop of the *Lactobacillus* species groups from the BLAST analysis.

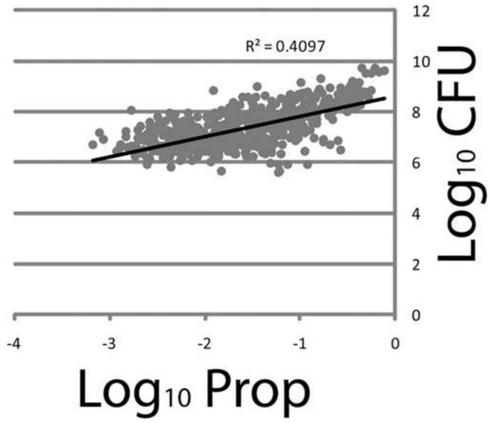

Fig. S2. Correlation between pyrosequencing and qPCR estimates for *Lactobacillus* in the G_4 AIL population. To quantify organisms in the Lactobacilli group, real-time qPCR was performed using a Mastercycler ep *realplex* (Eppendorf) and the group-specific primers Lac1 and Lac2 described previously (6). The primers target the 16S rDNA of *Lactobacillus* spp., *Pediococcus* spp., *Leuconostoc* spp., and *Weissella* spp., and result in a product length of 340 bp. The reaction mixture (25 μL) consisted of 1× QuantiFast SYBR Green PCR Master Mix (Qiagen), 25 pmol of each primer, template DNA, and RNase-free water. The amplification program was an initial denaturation at 94 °C for 2 min, followed by 35 cycles of denaturation at 94 °C for 30 s, annealing at 61 °C for 1 min, and extension at 68 °C for 1 min. A melting curve analysis was performed after each run. Standard curves were generated from 10-fold serial dilutions of DNA extracted from pure cultures of *L. reuteri* (DSM 20016^T) and *L. gasseri* (ATCC 33323^T). A plot of the threshold cycles (C_t) vs. bacterial counts (CFU/mL) resulted in a linear relationship with a correlation coefficient (*r*) of −0.989 (R^2 = 0.98). The total number of bacteria (CFU/g) for each stool sample was determined by interpolation of the standard curve. Both standards and samples were run in duplicate, and the counts were averaged. To measure the linear relationship between pyrosequencing and qPCR, a correlation analysis was performed on the amount of bacteria quantified by each method. Specifically, the bacterial counts (in $\log_{10}$ CFU/mL) obtained by qPCR was plotted against the $\log_{10}$ proportion of *Lactobacillus, Leuconostoc, Pediococcus*, and *Weissella* reads over the total reads for each sample. A significant correlation ($P < 0.0001$) was obtained, with $r = 0.625$.

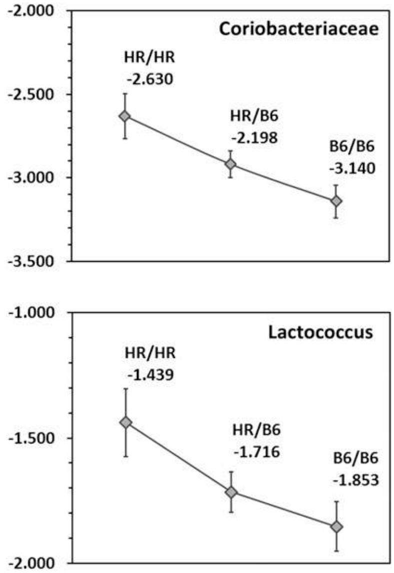

Fig. S3. Association of alleles at the JAX0030095 marker on MMU10 with the relative abundance of Coriobacteriaceae and *Lactococcus*. The log10-transformed Prop values for the family Coriobacteriaceae and the genus *Lactococcus* were averaged for each combination of JAX0030095 alleles. Alleles and the average log10-transformed Prop values are indicated above the relevant data points. Error bars indicate 2 SDs.

Table S1. SEQMATCH best hits of selected taxonomic bins from CLASSIFIER output

Taxonomic rank	Taxa*	Top organisms†	Taxa represented‡	Counts§	Prop total¶	Prop top hits**	Average S_ab score††
Genus	*Variovorax*	Variovorax paradoxus; Iso1; AY127900	Variovorax	15,303	0.382575	0.4481244	0.9253395
	40k‡‡	Variovorax sp. TUT1027; AB098595	Variovorax	10,450	0.26125	0.3060119	0.9306991
		Uncultured eubacterium WD2115; AJ292627	Variovorax	4,424	0.1106	0.1295499	0.9046336
		Variovorax paradoxus S110; CP001635	Variovorax	3,972	0.0993	0.1163138	0.9478978
				34,149	**0.853725**		
Genus	*Helicobacter*	*Helicobacter* ganmani; ES-5; AY561831	*Helicobacter*	38,664	0.840522	0.8623043	0.8908472
	46k‡‡	*Helicobacter* hepaticus; AJ007931	*Helicobacter*	4,944	0.107478	0.1102636	0.8031028
		Uncultured bacterium; L-123; EU622666	*Helicobacter*	646	0.014043	0.0144074	0.8069954
		Uncultured bacterium; MD2_aap36e09; EU508632	*Helicobacter*	584	0.012696	0.0130247	0.928738
				44,838	**0.974739**		
Genus	*Bacteroides*	*Bacteroides* acidifaciens; AB021157	*Bacteroides*	6,672	0.128308	0.3118048	0.8855073
	52k‡‡	Uncultured bacterium; SWPT13_aaa01g04; EF096855	*Bacteroides*	6,455	0.124135	0.3016637	0.8953075
		Uncultured bacterium; HY1_h06_1; EU458381	Odoribacter	4,220	0.081154	0.1972147	0.7443513
		Uncultured bacterium; K80N2_04b08; EU454172	*Bacteroides*	4,051	0.077904	0.1893168	0.906758
				21,398	**0.4115**		
Genus	*Parabacteroides*	Uncultured bacterium; lean2_aaa01f09; EF096000	Parabacteroides	17,825	0.445625	0.6227509	0.8851896
	40k‡‡	Uncultured bacterium; SJTU_A2_04_88; EF403654	Parabacteroides	4,034	0.10085	0.1409356	0.9071552
		Uncultured bacterium; RL246_aai73h07; DQ793582	Parabacteroides	3,733	0.093325	0.1304196	0.9290656
		Uncultured bacterium; WF16S_154; EU939416	Parabacteroides	3,031	0.075775	0.1058939	0.9160894
				28,623	**0.715575**		
Genus	*Marinilabilia*	Uncultured bacterium; HD5++50; EU791010	Barnesiella	10,475	0.261875	0.5164423	0.8619934
	40k‡‡	Uncultured bacterium; nbt15e03; FJ893065	Barnesiella	6,548	0.1637	0.3228319	0.8312596
		Uncultured bacterium; mcbc135; AM932661	Odoribacter	1,842	0.04605	0.090815	0.7321471
		Uncultured bacterium; C20_j04; AY991881	Odoribacter	1,418	0.03545	0.0699108	0.8218131
				20,283	**0.507075**		
Genus	*Alistipes*	Uncultured bacterium; WD3_aak03b12; EU510226	Alistipes	8,234	0.20585	0.2924006	0.8541191
	40k‡‡	Uncultured bacterium; cc_74; GQ175415	Alistipes	7,759	0.193975	0.2755327	0.8011231
		Uncultured bacterium; WD4_aal37e01; EU510373	Alistipes	7,640	0.191	0.2713068	0.8777465
		Uncultured bacterium; 16saw34-1g01.w2k; EF603689	Alistipes	4,527	0.113175	0.1607599	0.8833928
				28,160	**0.704**		
Genus	*Rikenella*	Uncultured bacterium; WD3_aak01e03; EU510108	Unclassified Bacteroidales	30,563	0.72769	0.8172799	0.8550692
	42k‡‡	Uncultured bacterium; C21_e10; AY993107	Unclassified Bacteroidales	2,858	0.068048	0.0764253	0.8142789
		Uncultured bacterium; cc_96; GQ175429	Rikenella	2,277	0.054214	0.0608889	0.6285806
		Uncultured bacterium; 2.16F; EU655924	Unclassified Bacteroidales	1,698	0.040429	0.0454059	0.7522668
				37,396	**0.890381**		
Family	Peptostreptococcaceae	Uncultured bacterium; R-9612; FJ880565	Sporacetigenium	14,429	0.327932	0.5388982	0.8899369
	44k‡‡	Uncultured bacterium; MD23_2aaa04g05; EU507538	Sporacetigenium	5,811	0.132068	0.2170308	0.9109019
		Uncultured bacterium; MD18_aaa01c10; EU506158	Sporacetigenium	4,063	0.092341	0.151746	0.9240386
		Uncultured bacterium; MD19_aaa01c03; EU506401	Sporacetigenium	2,472	0.056182	0.0923249	0.897591
				26,775	**0.608523**		
Genus	*Lactococcus*	Lactococcus lactis subsp. cremoris; YIT 2007 (ATCC 19257); AB008214	Lactococcus	10,278	0.25695	0.3364982	0.8874481
	40k‡‡	Lactococcus lactis subsp. cremoris SK11; CP000425	Lactococcus	8,971	0.224275	0.2937074	0.8944487
		Uncultured bacterium; 1–5D; EU289440	Lactococcus	6,533	0.163325	0.2138882	0.896064

Table S1. Cont.

Taxonomic rank	Taxa*	Top organisms[†]	Taxa represented[‡]	Counts[§]	Prop total[¶]	Prop top hits**	Average S_ab score[††]
		Lactococcus lactis subsp. lactis; RO6; AF515224	Lactococcus	4,762	0.11905	0.1559062	0.9152703
				30,544	**0.7636**		
Genus	*Roseburia*	Uncultured bacterium; RL184_aao65g01; DQ809900	Roseburia	9,117	0.227925	0.4072635	0.8149241
	40k[‡‡]	Uncultured bacterium; CRWD2_aaa03d03; EU503700	Roseburia	8,557	0.213925	0.3822478	0.8707943
		Uncultured bacterium; CRWD5_aaa04f02; EU504227	Unclassified Lachnospiraceae	2,445	0.061125	0.10922	0.9149681
		Uncultured bacterium; K74N1_19e08; EU455153	Unclassified Clostridiales	2,267	0.056675	0.1012687	0.9076903
				22,386	**0.55965**		
Genus	Turicibacter	Uncultured bacterium; control_7 d-F2; EF406422	Turicibacter	18,682	0.44481	0.4727705	0.8986114
	42k[‡‡]	Uncultured bacterium; infected_7 d-E1; EF406660	Turicibacter	17,816	0.42419	0.4508553	0.8995282
		Uncultured bacterium; R-6524; FJ880085	Turicibacter	1,621	0.038595	0.0410214	0.9036231
		Uncultured bacterium; R-9107; FJ881096	Turicibacter	1,397	0.033262	0.0353528	0.8997015
				39,516	**0.940857**		
Order	Coriobacteriales	Uncultured bacterium; C18_f09_1; EF614565	Unclassified Coriobacteriaceae	2,951	0.210786	0.4407767	0.8278333
	14k[‡‡]	Uncultured bacterium; MD2_aap35a10; EU508535	Asaccharobacter	1,520	0.108571	0.2270351	0.9184368
		Coriobacteriaceae bacterium B7; DQ789120	Unclassified Coriobacteriaceae	1,201	0.085786	0.1793876	0.9162223
		Uncultured bacterium; SWPT20_aaa03a06; EF097741	Unclassified Coriobacteriaceae	1,023	0.073071	0.1528006	0.9142815
				6,695	**0.478214**		
Family	Coriobacteriaceae	Uncultured bacterium; C18_f09_1; EF614565	Unclassified Coriobacteriaceae	2,951	0.210786	0.4407767	0.8278333
	14k[‡‡]	Uncultured bacterium; MD2_aap35a10; EU508535	Asaccharobacter	1,520	0.108571	0.2270351	0.9184368
		Coriobacteriaceae bacterium B7; DQ789120	Unclassified Coriobacteriaceae	1,201	0.085786	0.1793876	0.9162223
		Uncultured bacterium; SWPT20_aaa03a06; EF097741	Unclassified Coriobacteriaceae	1,023	0.073071	0.1528006	0.9142815
				6,695	**0.478214**		

*All sequences from respective taxa assigned by CLASSIFIER were extracted. At least 40,000 random sequences were selected from each taxon and analyzed by RDP SEQMATCH. Taxa with fewer than 40,000 sequences were analyzed to completion.

[†] Top four bacteria with the most sequence matches to the RDP SeqMatch database for the given taxa.

[‡] The lowest taxonomic rank assigned by RDP SeqMatch for the given top organism.

[§] The number of matches to the database for the given top organism.

[¶] The proportion of sequences matching the given top organism divided by the total number of sequences pooled for analysis of the given taxa.

**The proportion of the sequences matching the given top organism divided by the compiled amount of sequences making up all four top organisms for the given taxa.

[††] The average of RDP SeqMatch score (S_ab). The S_ab score is the number of (unique) 7-base oligomers shared between the query sequence and a given RDP sequence divided by the lowest number of unique oligos in either of the two sequences.

[‡‡] The total number of sequences pooled for RDP SeqMatch analysis for the given taxa.

Table S2. **Descriptive statistics for CMM traits measured in the G_4 population**

		Average*	SD	Min	Max
Phylum	Actinobacteria	−2.67739	0.590533	−4.39794	−0.39324
Class	Actinobacteria	−2.67739	0.590533	−4.39794	−0.39324
Subclass	Coriobacteridae	−2.94055	0.584882	−4.39794	−0.54654
Order	Coriobacteriales	−2.94055	0.584882	−4.39794	−0.54654
Suborder	Coriobacterineae	−2.94055	0.584882	−4.39794	−0.54654
Family	Coriobacteriaceae	−2.94055	0.584882	−4.39794	−0.54654
Subclass	Actinobacteridae	−3.28907	0.685807	−4.65758	−0.39482
Phylum	Bacteroidetes	−0.64014	0.322578	−2.21247	−0.07597
Class	Flavobacteria	−3.24884	0.768164	−4.63827	−1.20137
Order	Flavobacteriales	−3.24884	0.768164	−4.63827	−1.20137
Family	Flavobacteriaceae	−3.24962	0.768473	−4.63827	−1.20137
Class	Bacteroidetes	−0.86399	0.340153	−2.55486	−0.30995
Order	Bacteroidales	−0.86399	0.340153	−2.55486	−0.30995
Family	Rikenellaceae	−1.45665	0.361402	−2.86902	−0.78168
Genus	*Odoribacter*	−2.69635	0.658767	−4.55284	−1.61057
Genus	*Alistipes*	−1.82236	0.403583	−3.16052	−0.96128
Genus	*Rikenella*	−3.0305	0.75621	−4.55284	−1.60478
Family	Bacteroidaceae	−1.81256	0.524582	−4.25181	−0.56101
Genus	*Bacteroides*	−1.8127	0.524608	−4.25181	−0.56101
Family	Porphyromonadaceae	−1.83477	0.483651	−4.25181	−0.69437
Genus	*Parabacteroides*	−1.83713	0.483785	−4.25181	−0.69497
Phylum	Proteobacteria	−1.29749	0.441019	−2.74642	−0.19835
Class	Epsilonproteobacteria	−2.12995	0.931753	−4.65758	−0.50796
Order	Campylobacterales	−2.12999	0.931747	−4.65758	−0.50796
Family	Helicobacteraceae	−2.14262	0.94777	−4.65758	−0.50813
Genus	*Helicobacter*	−2.15126	0.950455	−4.65758	−0.51233
Class	Deltaproteobacteria	−2.18371	0.669213	−4.21467	−0.81547
Class	Alphaproteobacteria	−2.79909	0.689063	−4.60206	−1.0188
Order	Rhizobiales	−3.16046	0.793923	−4.65758	−1.25349
Class	Gammaproteobacteria	−2.38611	0.639983	−4.23657	−0.21968
Order	Pseudomonadales	−2.82208	0.638174	−4.45593	−0.21992
Order	Enterobacteriales	−2.83163	0.598018	−4.38722	−1.44615
Family	Enterobacteriaceae	−2.83163	0.598018	−4.38722	−1.44615
Class	Betaproteobacteria	−2.2471	0.633134	−4.05552	−0.41858
Order	Burkholderiales	−2.38423	0.667937	−4.05552	−0.42322
Family	Comamonadaceae	−2.4195	0.676075	−4.05552	−0.43046
Genus	*Variovorax*	−2.66522	0.751411	−4.25181	−0.43876
Phylum	Firmicutes	−0.27565	0.143062	−1.06802	−0.0228
Class	Bacilli	−1.20876	0.500503	−2.47638	−0.10353
Order	Lactobacillales	−1.23337	0.502172	−2.5085	−0.10555
Family	Lactobacillaceae	−1.73651	0.687982	−4.08619	−0.10924
Genus	*Lactobacillus*	−1.74217	0.687414	−4.08619	−0.11181
Family	Leuconostocaceae	−2.67244	0.558744	−4.45593	−1.14704
Genus	*Weissella*	−2.80507	0.626531	−4.65758	−1.21328
Family	Streptococcaceae	−1.73707	0.5654	−3.28651	−0.28054
Genus	*Lactococcus*	−1.75409	0.572448	−3.28651	−0.28122
Order	Bacillales	−2.96039	0.641711	−4.36653	−0.69596
Class	Erysipelotrichi	−2.41441	0.808503	−4.27572	−0.42666
Order	Erysipelotrichales	−2.41441	0.808503	−4.27572	−0.42666
Family	Erysipelotrichaceae	−2.41441	0.808503	−4.27572	−0.42666
Genus	*Turicibacter*	−2.69515	0.992398	−4.55284	−0.42707
Class	Clostridia	−0.42739	0.192067	−1.52896	−0.07883
Order	Clostridiales	−0.43079	0.192643	−1.53304	−0.08154
Family	Lachnospiraceae	−0.70714	0.232744	−2.04648	−0.26048
Genus	*Lachnobacterium*	−3.35505	0.842755	−4.5376	−0.92087
Genus	*Dorea*	−2.38523	0.446171	−4.18709	−1.11065
Genus	*Lachnospiraceae Incertae Sedis*	−2.55034	0.383489	−4.25964	−1.68721
Genus	*Roseburia*	−2.89953	0.590418	−4.65758	−0.5072
Family	Peptostreptococcaceae	−2.84529	0.996391	−4.58503	−0.27802
Genus	*Peptostreptococcaceae Incertae Sedis*	−2.85821	0.998041	−4.58503	−0.28497
Family	Ruminococcaceae	−1.5107	0.244235	−2.61386	−0.62938
Family	Clostridiaceae	−3.44843	0.821475	−4.72125	−0.70714
Subfamily	Clostridiaceae 1	−3.4492	0.821532	−4.72125	−0.70714
Genus	*Clostridium*	−3.55616	0.770174	−4.72125	−0.86786

*Prop values of 0 were replaced with 0.5/total reads. and all Prop values were log10-transformed for descriptive statistics.

Table S3. Mixed-model analysis of CMM traits with an across-taxa FDR < 0.05

Rank	Taxon	Source of variation*	*P* value†	FDR‡
Phylum	Proteobacteria	Parent of origin	0.0022	0.017925
Class	Deltaproteobacteria	Parent of origin	<0.0001	0.000453
Class	Epsilonproteobacteria	Parent of origin	<0.0001	0.000453
Order	Campylobacterales	Parent of origin	<0.0001	0.000453
Family	Clostridiaceae	Parent of origin	0.0112	0.049683
Family	Helicobacteraceae	Parent of origin	<0.0001	0.000453
Family	Peptostreptococcaceae	Parent of origin	0.0115	0.049683
Family	Ruminococcaceae	Parent of origin	0.0006	0.005208
Subfamily	Clostridiaceae 1	Parent of origin	0.0111	0.049683
Genus	*Dorea*	Parent of origin	0.0025	0.018041
Genus	*Helicobacter*	Parent of origin	<0.0001	0.000453
Genus	*Lachnobacterium*	Parent of origin	0.006	0.035161
Genus	*Lachnospiraceae Incertae sedis*	Parent of origin	0.0005	0.005208
Genus	*Peptostreptococcaceae Incertae sedis*	Parent of origin	0.0116	0.049683
Genus	*Rikenella*	Parent of origin	0.0049	0.031611
Phylum	Actinobacteria	Sex	0.0024	0.016972
Class	Actinobacteria	Sex	0.0024	0.016972
Class	Epsilonproteobacteria	Sex	0.0073	0.033396
Class	Erysipelotrichi	Sex	0.0068	0.033396
Subclass	Coriobacteridae	Sex	0.0006	0.006676
Order	Bacillales	Sex	0.0108	0.04052
Order	Campylobacterales	Sex	0.0073	0.033396
Order	Coriobacteriales	Sex	0.0006	0.006676
Order	Erysipelotrichales	Sex	0.0068	0.033396
Suborder	Coriobacterineae	Sex	0.0006	0.006676
Family	Coriobacteriaceae	Sex	0.0006	0.006676
Family	Erysipelotrichaceae	Sex	0.0068	0.033396
Family	Helicobacteraceae	Sex	0.0092	0.036682
Family	Peptostreptococcaceae	Sex	<0.0001	0.002721
Genus	*Helicobacter*	Sex	0.0082	0.035158
Genus	Peptostreptococcaceae Incertae sedis	Sex	<0.0001	0.002721
Genus	Turicibacter	Sex	0.0015	0.013717
Phylum	Actinobacteria	Cohort	<0.0001	0.000025
Phylum	Bacteroidetes	Cohort	<0.0001	0
Phylum	Firmicutes	Cohort	<0.0001	0.000025
Phylum	Proteobacteria	Cohort	<0.0001	0.000121
Subclass	Actinobacteridae	Cohort	0.0013	0.002488
Subclass	Coriobacteridae	Cohort	<0.0001	0.00002
Order	Enterobacteriales	Cohort	<0.0001	0.000001
Order	Flavobacteriales	Cohort	0.0002	0.000369
Order	Lactobacillales	Cohort	0.0007	0.001321
Order	Pseudomonadales	Cohort	<0.0001	0.000063
Order	Rhizobiales	Cohort	<0.0001	0.000197
Suborder	Coriobacterineae	Cohort	<0.0001	0.00002
Class	Bacteroidetes	Dam	0.0004	0.011485
Order	Bacteroidales	Dam	0.0004	0.011485
Phylum	Actinobacteria	Litter	0.0009	0.004262
Phylum	Proteobacteria	Litter	0.0017	0.006141
Class	Actinobacteria	Litter	0.0009	0.004262
Class	Deltaproteobacteria	Litter	0.0104	0.025511
Class	Epsilonproteobacteria	Litter	<0.0001	0.001275
Class	Erysipelotrichi	Litter	0.0001	0.001275
Class	Flavobacteria	Litter	0.0004	0.002262
Subclass	Actinobacteridae	Litter	0.0017	0.006141
Subclass	Coriobacteridae	Litter	0.0019	0.006141
Order	Bacillales	Litter	0.011	0.026008
Order	Burkholderiales	Litter	0.0064	0.016302
Order	Campylobacterales	Litter	<0.0001	0.001275
Order	Coriobacteriales	Litter	0.0019	0.006141
Order	Erysipelotrichales	Litter	0.0001	0.001275
Order	Flavobacteriales	Litter	0.0004	0.002262
Order	Pseudomonadales	Litter	0.0155	0.03545
Order	Rhizobiales	Litter	0.0007	0.003648

Table S3. Cont.

Rank	Taxon	Source of variation*	*P* value[†]	FDR[‡]
Suborder	Coriobacterineae	Litter	0.0019	0.006141
Family	Clostridiaceae	Litter	0.0056	0.01482
Family	Comamonadaceae	Litter	0.0164	0.036246
Family	Coriobacteriaceae	Litter	0.0019	0.006141
Family	Erysipelotrichaceae	Litter	0.0001	0.001275
Family	Flavobacteriaceae	Litter	0.0004	0.002262
Family	Helicobacteraceae	Litter	0.0001	0.001275
Subfamily	Clostridiaceae 1	Litter	0.0055	0.01482
Genus	*Clostridium*	Litter	0.0031	0.009582
Genus	*Dorea*	Litter	0.0198	0.042126
Genus	*Helicobacter*	Litter	0.0002	0.001275
Genus	*Lachnobacterium*	Litter	0.0055	0.01482
Genus	*Turicibacter*	Litter	0.0002	0.001275
Genus	*Weissella*	Litter	0.0204	0.042126

*Abbreviated notation for sources of variation: cohort for cohort(parent of origin), dam for dam(parent of origin), and litter for litter (parent of origin*cohort*dam).

[†] The *P* value is the probability of obtaining a larger F value in the individual taxon analysis.

[‡] FDR is the across-taxa false discovery rate adjusted *P* value calculated separately for each source of variation.

Table S4. REML estimated variance components of CMM traits

Rank	Taxon	Proportion of total variance*				Variance†			
		Cohort	Family	Litter	Residual	Cohort	Family	Litter	Residual
Phylum	Actinobacteria	0.39788	0	0.08413	0.51798	0.13479	0	0.0285	0.1755
Phylum	Bacteroidetes	0.27611	0.06455	0	0.65935	0.03189	0.00745	0	0.07615
Phylum	Firmicutes	0.32759	0.06585	0	0.60656	0.00601	0.00121	0	0.01114
Phylum	Proteobacteria	0.41389	0.04939	0.08617	0.45055	0.05838	0.00697	0.01215	0.06355
Class	Actinobacteria	0.39788	0	0.08413	0.51798	0.13479	0	0.0285	0.1755
Class	Alphaproteobacteria	0.38916	0.00749	0.04531	0.55804	0.16557	0.00319	0.01928	0.2374
Class	Bacilli	0.24173	0.04747	0.01526	0.69553	0.06647	0.01305	0.0042	0.1913
Class	Bacteroidetes	0.26202	0.07093	0	0.66705	0.03386	0.00917	0	0.08619
Class	Betaproteobacteria	0.25345	0.01973	0.03592	0.6909	0.07182	0.00559	0.01018	0.1958
Class	Clostridia	0.05253	0.04713	0	0.90035	0.00187	0.00167	0	0.03199
Class	Deltaproteobacteria	0.18166	0	0.1135	0.70485	0.0465	0	0.02905	0.1804
Class	Epsilonproteobacteria	0.17242	0.16151	0.16859	0.49747	0.11118	0.10414	0.10871	0.3208
Class	Erysipelotrichi	0.1029	0.04101	0.12422	0.73187	0.07042	0.02807	0.08501	0.5009
Class	Flavobacteria	0.36772	0	0.08908	0.54319	0.17613	0	0.04267	0.2602
Class	Gammaproteobacteria	0.46878	0.01188	0.05323	0.46612	0.19338	0.0049	0.02196	0.1923
Order	Bacillales	0.21766	0.03951	0.08784	0.65499	0.08841	0.01605	0.03568	0.266
Order	Bacteroidales	0.26202	0.07093	0	0.66705	0.03386	0.00917	0	0.08619
Order	Burkholderiales	0.27349	0	0.0543	0.67221	0.08385	0	0.01665	0.2061
Order	Clostridiales	0.05309	0.04697	0	0.89994	0.0019	0.00168	0	0.03214
Order	Coriobacteriales	0.42787	0	0.08143	0.49069	0.15041	0	0.02863	0.1725
Order	Enterobacteriales	0.41728	0.00428	0.05511	0.52333	0.14701	0.00151	0.01941	0.1844
Order	Erysipelotrichales	0.1029	0.04101	0.12422	0.73187	0.07042	0.02807	0.08501	0.5009
Order	Flavobacteriales	0.36772	0	0.08908	0.54319	0.17613	0	0.04267	0.2602
Order	Lactobacillales	0.23713	0.04833	0.01495	0.6996	0.06562	0.01337	0.00414	0.1936
Order	Pseudomonadales	0.35992	0.00417	0.06987	0.56604	0.1392	0.00161	0.02702	0.2189
Order	Rhizobiales	0.42862	0	0.07539	0.496	0.22995	0	0.04044	0.2661
Suborder	Coriobacterineae	0.42787	0	0.08143	0.49069	0.15041	0	0.02863	0.1725
Family	Bacteroidaceae	0.30932	0.00968	0.06251	0.6185	0.08616	0.0027	0.01741	0.1723
Family	Clostridiaceae	0.21729	0.02276	0.08559	0.67436	0.144	0.01509	0.05672	0.4469
Family	Comamonadaceae	0.27078	0	0.04536	0.68385	0.08496	0	0.01423	0.2146
Family	Coriobacteriaceae	0.42787	0	0.08143	0.49069	0.15041	0	0.02863	0.1725
Family	Enterobacteriaceae	0.41728	0.00428	0.05511	0.52333	0.14701	0.00151	0.01941	0.1844
Family	Erysipelotrichaceae	0.1029	0.04101	0.12422	0.73187	0.07042	0.02807	0.08501	0.5009
Family	Flavobacteriaceae	0.36838	0	0.08826	0.54336	0.17655	0	0.0423	0.2604
Family	Helicobacteraceae	0.14769	0.17841	0.16148	0.51241	0.0982	0.11863	0.10737	0.3407
Family	Lachnospiraceae	0.06745	0.03029	0.01847	0.88379	0.00366	0.00164	0.001	0.04789
Family	Lactobacillaceae	0.08002	0.07373	0	0.84625	0.0404	0.03722	0	0.4273
Family	Leuconostocaceae	0.47525	0.00714	0.06711	0.45051	0.15215	0.00229	0.02148	0.1442
Family	Peptostreptococcaceae	0.16267	0.01352	0.05315	0.77067	0.15868	0.01318	0.05184	0.7518
Family	Porphyromonadaceae	0.19482	0.09877	0.01168	0.69472	0.05108	0.0259	0.00306	0.1822
Family	Rikenellaceae	0.26116	0.05894	0	0.67991	0.03695	0.00834	0	0.09619
Family	Ruminococcaceae	0.08045	0.05935	0	0.8602	0.00437	0.00322	0	0.04669
Family	Streptococcaceae	0.46151	0.02891	0.05591	0.45368	0.14802	0.00927	0.01793	0.1455
Subfamily	Clostridiaceae 1	0.21721	0.02305	0.08595	0.67379	0.14392	0.01527	0.05695	0.4465
Genus	*Alistipes*	0.30995	0	0.05017	0.63988	0.05545	0	0.00898	0.1145
Genus	*Bacteroides*	0.30922	0.00943	0.06259	0.61876	0.08614	0.00263	0.01744	0.1724
Genus	*Clostridium*	0.22068	0.01223	0.09866	0.66843	0.13105	0.00726	0.05859	0.3969
Genus	*Dorea*	0.18903	0.07916	0.07951	0.6523	0.03365	0.01409	0.01415	0.1161
Genus	*Helicobacter*	0.15294	0.17651	0.16352	0.50703	0.1031	0.11899	0.11023	0.3418
Genus	*Lachnobacterium*	0.11745	0.11622	0.08846	0.67787	0.0778	0.07698	0.05859	0.449
Genus	*Lachnospiraceae Incertae sedis*	0.08521	0.10581	0	0.80897	0.01086	0.01349	0	0.1031
Genus	*Lactobacillus*	0.07871	0.07461	0	0.84668	0.03968	0.03762	0	0.4269
Genus	*Lactococcus*	0.46108	0.02927	0.0544	0.45525	0.15159	0.00962	0.01788	0.1497
Genus	*Odoribacter*	0.13062	0.02303	0.035	0.81134	0.06023	0.01062	0.01614	0.3741
Genus	*Parabacteroides*	0.19493	0.09813	0.0125	0.69444	0.0511	0.02572	0.00328	0.182
Genus	*Peptostreptococcaceae Incertae sedis*	0.15946	0.01438	0.05114	0.77502	0.1563	0.0141	0.05013	0.7597
Genus	*Rikenella*	0.1678	0.02807	0.05314	0.75099	0.08676	0.01452	0.02748	0.3883
Genus	*Roseburia*	0.0487	0.12158	0.08587	0.74385	0.01454	0.0363	0.02564	0.2221
Genus	*Turicibacter*	0.09938	0.04869	0.12768	0.72425	0.09663	0.04734	0.12415	0.7042
Genus	*Variovorax*	0.33673	0.00279	0.04756	0.61292	0.13608	0.00113	0.01922	0.2477
Genus	*Weissella*	0.52388	0.00223	0.0754	0.3985	0.1991	0.00085	0.02865	0.1514

*Proportion of total variance is the variance divided by the sum of the cohort, family, litter, and residual variances.

†Variances were estimated using REML with a linear mixed model that included fixed effects for parent of origin and sex and random effects for cohort(parent of origin), family(parent of origin), and litter(parent of origin*cohort*family).

Table S5. QTL detected and respective statistics for Prop1 traits

	Trait	Nearest marker	Chromosome	Peak position, Mb	Naive LOD	GRAIP LOD*	95% CI, Mb[†]	% Var[‡]	Additive ± SE[§]	Dominance ± SE[§]
Phylum	Actinobacteria									
Subclass	Coriobacteridae	JAX00300375	10	119	7.2	3.9**	104–123	5.7	0.20 ± 0.03[¶]	–0.03 ± 0.05
Order	Coriobacteriales	JAX00300375	10	119	7.1	4.0**	105–122	5.7	0.20 ± 0.03[¶]	–0.03 ± 0.05
Suborder	Coriobacterineae	JAX00300375	10	119	7.0	3.9**	104–123	5.7	0.20 ± 0.03[¶]	–0.03 ± 0.05
Family	Coriobacteriaceae	JAX00300375	10	119	7.3	4.2**	106–122	5.7	0.20 ± 0.03[¶]	–0.03 ± 0.05
Phylum	Proteobacteria	JAX00139228	6	28	8.6	4.1**	–40	1.5	–0.05 ± 0.02	0.08 ± 0.03
		JAX00666793	8	43	8.6	4.1**	33–63	3.2	–0.08 ± 0.02	0.12 ± 0.03
Class	Epsilonproteobacteria	JAX00603343	6	13	9.2	4.7**	–39	1.7	–0.03 ± 0.05	0.24 ± 0.07
Order	Campylobacterales	JAX00603343	6	13	9.2	4.7**	–39	1.7	–0.03 ± 0.05	0.24 ± 0.07
Family	Helicobacteraceae	JAX00603343	6	13	8.7	4.4**	–39	1.7	–0.03 ± 0.05	0.24 ± 0.07
Genus	*Helicobacter*	JAX00603343	6	13	8.8	4.4**	–39	1.6	–0.02 ± 0.05	0.24 ± 0.08
Class	Deltaproteobacteria	JAX00480903	19	56	5.1	3.9**	54-	2.5	–0.10 ± 0.04	0.17 ± 0.05
Class	Gammaproteobacteria	JAX00707462	9	119	6.2	3.6	117-	4.0	–0.14 ± 0.04	0.19 ± 0.05
Order	Pseudomonadales	JAX00707462	9	119	6.8	3.8	117-	4.4	–0.14 ± 0.04	0.21 ± 0.05
Class	Betaproteobacteria	JAX00633165	7	19	8.6	4.7**	15–29	6.3	–0.22 ± 0.03[¶]	–0.10 ± 0.05
Order	Burkholderiales	JAX00633165	7	19	10.7	4.7**	14–33	7.9	–0.26 ± 0.04[¶]	–0.11 ± 0.05
Family	Comamonadaceae	JAX00633165	7	19	10.8	4.7**	13–34	7.9	–0.26 ± 0.04[¶]	–0.12 ± 0.05
Genus	*Variovorax*	JAX00633165	7	19	9.7	4.7**	14–28	7.2	–0.27 ± 0.04[¶]	–0.16 ± 0.06
Phylum	Firmicutes									
Species	*L.johnsonii/L.gasseri* 97%	JAX00641805	7	66	6.8	4.7**	47–71	4.7	–0.27 ± 0.05[¶]	–0.11 ± 0.07
		JAX00387018	14	93	5.8	4.7**	86–103	3.9	–0.23 ± 0.05[¶]	–0.16 ± 0.07
Family	Streptococcaceae	JAX00022058	10	107	8.0	4.7**	101–111	7.0	0.21 ± 0.03[¶]	–0.05 ± 0.04
Genus	*Lactococcus*	JAX00022058	10	107	8.0	4.7**	100–111	7.0	0.21 ± 0.03[¶]	–0.05 ± 0.05
Class	Erysipelotrichi	JAX00643377	7	73	6.4	4.0**	65–88	5.0	–0.24 ± 0.04[¶]	0.03 ± 0.06
Order	Erysipelotrichales	JAX00643377	7	73	6.5	4.2**	67–87	5.0	–0.24 ± 0.04[¶]	0.03 ± 0.06
Family	Erysipelotrichaceae	JAX00643377	7	73	6.5	4.0**	66–88	5.0	–0.24 ± 0.04[¶]	0.03 ± 0.06
Genus	*Turicibacter*	JAX00643377	7	73	7.1	4.6**	71–88	5.3	–0.30 ± 0.05[¶]	0.09 ± 0.08
Family	Peptostreptococcaceae	JAX00010715	1	148	5.8	3.8	143–150	4.4	–0.25 ± 0.05[¶]	0.16 ± 0.08
Genus	*Peptostreptococcaceae IS*	JAX00010715	1	148	5.7	3.7	143–150	4.3	–0.25 ± 0.05[¶]	0.17 ± 0.08
Family	Ruminococcaceae	JAX00327082	12	17	5.5	4.4**	–26	3.4	0.06 ± 0.01[¶]	0.04 ± 0.02
Phylum	Bacteriodetes									
Genus	*Barnesiella*	JAX00005735	1	80	10.7	4.7**	63–139	9.0	–0.23 ± 0.03[¶]	0.14 ± 0.05
		JAX00173791	9	87	4.6	3.5	72–104	3.4	–0.14 ± 0.04	0.14 ± 0.05

*LOD exceeding the 95% (P = 0.05, LOD ≥ 3.9) permutation threshold are denoted by **; other QTL exceeded the 90% (P = 0.1, LOD ≥ 3.5) threshold.

[†] Confidence intervals for QTL positions were obtained using a 1.0 LOD drop in Mb (relative to the GRAIP-permuted LOD score).

[‡] Percentage of phenotypic variance accounted for by the QTL effect.

[§] For additive and dominance effects: positive values indicate increasing effect of the HR allele or increasing effect of the heterozygote, respectively.

[¶] Indicates that additive and/or dominance effects were statistically significant at $P < 0.0$.

Table S6. Genotype (C57BL/6J = BB; HR = AA) frequencies (% of total calls) at a given SNP location

Trait		SNP	MMU	% of BB	% of BA	% of AA
Subclass	Coriobacteridae	JAX00300375	10	30.3	51.6	18.1
Order	Coriobacteriales	JAX00300375	10	30.3	51.6	18.1
Suborder	Coriobacterineae	JAX00300375	10	30.3	51.6	18.1
Family	Coriobacteriaceae	JAX00300375	10	30.3	51.6	18.1
Genus	*Odoribacter*	JAX00005735	1	23.0	44.7	32.3
		JAX00173791	9	22.1	53.6	24.3
Phylum	Proteobacteria	JAX00139228	6	29.2	45.6	25.2
		JAX00666793	8	27.9	47.0	25.1
Class	Epsilonproteobacteria	JAX00603343	6	32.3	45.4	22.2
Order	Campylobacterales	JAX00603343	6	32.3	45.4	22.2
Family	Helicobacteraceae	JAX00603343	6	32.3	45.4	22.2
Genus	*Helicobacter*	JAX00603343	6	32.3	45.4	22.2
Class	Deltaproteobacteria	JAX00480903	19	23.0	55.3	21.7
Class	Gammaproteobacteria	JAX00707462	9	29.5	47.6	22.8
Order	Pseudomonadales	JAX00707462	9	29.5	47.6	22.8
Class	Betaproteobacteria	JAX00633165	7	21.5	50.6	27.9
Order	Burkholderiales	JAX00633165	7	21.5	50.6	27.9
Family	Comamonadaceae	JAX00633165	7	21.5	50.6	27.9
Genus	*Variovorax*	JAX00633165	7	21.5	50.6	27.9
Family	Streptococcaceae	JAX00022058	10	37.4	43.2	19.4
Genus	*Lactococcus*	JAX00022058	10	37.4	43.2	19.4
Species	*L.johnsonii/L.gasseri* 97%	JAX00641805	7	24.7	48.1	27.3
		JAX00387018	14	23.9	54.0	22.1
Class	Erysipelotrichi	JAX00643377	7	26.9	45.2	27.9
Order	Erysipelotrichales	JAX00643377	7	26.9	45.2	27.9
Family	Erysipelotrichaceae	JAX00643377	7	26.9	45.2	27.9
Genus	*Turicibacter*	JAX00643377	7	26.9	45.2	27.9
Family	Peptostreptococcaceae	JAX00010715	1	29.8	41.3	28.9
Genus	*Peptostreptococcaceae IS*	JAX00010715	1	29.8	41.3	28.9
Family	Ruminococcaceae	JAX00327082	12	22.8	48.2	29.0