CHAPTER 12

Environmental Biotechnology

SUMMARY

Environmental biotechnology has opened up new avenues for the identification and study of previously undiscovered life forms, such as viruses, bacteria, and other, more obscure genetic elements. These life forms and genetic elements may contain genes for novel proteins of metabolisms that could be useful in many areas of science.

Whole genomes from an environmental sample are studied using metagenomics. The genomes from all species present within a microscopic community are examined simultaneously. Many of the techniques used in metagenomics correspond to those already discussed in previous chapters, including PCR, RT-PCR, DNA sequencing, and microarray analysis. Identification of new sources of antibiotics, enzymes that degrade pollutants, and those that can synthesize new products is possible with metagenomics.

The DNA present in environmental samples must be enriched before it can be analyzed. In SIP, samples are mixed with stable isotopes with the goal that some become incorporated into the DNA of metabolically active cells present in the sample. The DNA containing the isotope is then separated from other, non-labeled DNA and cloned into vectors to construct metagenomic libraries. In BrdU enrichment, the compound 5-bromo-2-deoxyuridine (BrdU) is added to environmental samples, which can be incorporated directly into DNA or RNA of actively growing cells. A variation on SIP, called RNA-SIP, uses the small subunit ribosomal RNA (SSU rRNA) labeled with an isotope to isolate RNA present in the environmental sample. This is advantageous over the other methods because it is not specific to metabolically active organisms and can also indicate the presence of dormant microbes. Two different environmental samples can be compared using suppressive subtractive hybridization (SSH) by isolating total DNA from each sample. Each DNA sample is attached to separate linkers, denatured, and then rehybridized. Sequences that are similar between the two samples will reanneal, leaving only the unique sequences to anneal back upon themselves, which would then be further analyzed for novel genes.

Traditionally, organisms were identified from the environment by sequencing their 16S rRNA sequences, which could be done without first culturing the organisms. The downside to this method is that it does not provide any information about the genetics or metabolism of the new organism. Other sequence-dependent techniques, such as the generation of metagenomic libraries, help elucidate the entire genomes of novel organisms present in environmental samples. Entire genomes from many different microbes are fragmented and cloned into vectors and then propagated in *E. coli*. In some cases, total sample mRNA is extracted and converted to cDNA, which is then cloned into the vectors. In PCR, primers complementary to the genes of previously known, pollutant-degrading enzymes help to identify related proteins in a metagenomic library. Examining regions next to known genes can also identify novel proteins. Microarrays provide a global approach to identify different types of organisms in an environmental sample. The overall goal of the preceding techniques is to obtain novel genes, some of which might encode proteins capable of degrading environmental pollutants.

In addition to sequence-based approaches, functional and activity-based evaluation of metagenomic libraries can also yield proteins with novel properties. Expression vectors containing appropriate features for expression allow the properties of the protein products of genes to be further analyzed. Genes in metagenomic libraries could also be linked to reporter genes, such as GFP. It is important to keep in mind the limitations of host cells when expressing foreign proteins. Some overexpressed proteins may be toxic to the surrogate system or may be sequestered within inclusion bodies.

Metagenomics has provided researchers with a means to study organisms that are symbionts with a host. Culturing these microbes outside their host is often difficult and sometimes

impossible. Examining the genomes of the symbionts on a global scale has helped identify new properties of the symbionts, such as the presence of flagella in the bacterial symbionts of deep-sea tube worms.

Natural attenuation of pollutants through the process of bioremediation has made great advances due to metagenomics. Naturally occurring microbes often have very diverse pathways, which contain novel enzymes. These microbes can be found in diverse habitats as well. The uses for pollutant-degrading microbes are immense. Oil spills, nuclear and chemical waste spills, and other types of pollutants could be broken down into harmless by-products. Sometimes one organism completely eliminates the pollutant. In other instances, a couple of microorganisms might need to be present to complete the job. Occasionally, the microbes plus their sources of energy might need to be applied to a polluted area together.

A majority of the world's energy production is from fossil fuels, and this use is not expected to decline. On the contrary, demand for energy is expected to increase. These levels are not sustainable, however, and will result in the depletion of nonrenewable resources. As a result, biofuels and other types of biologically derived energy resources are being studied. They include ethanol from corn or sugarcane and also biodiesel from soybeans and oil palm. Ethanol is used as an additive to gasoline to improve emissions. The production of biodiesel replaces regular diesel fuel and is not an additive. Using algae to create biofuels is less costly and does not use a food source for the reactions. Microalgae are small photosynthetic creatures, and macroalgae are large multi-cellular, plant-like organisms (seaweeds). Many macroalgae have large sugar contents and could be fermented to ethanol. Others have a large lipid content that could be converted to biodiesel.

Algae are also capable of producing biogases, including biohydrogen and biomethane. High-carbohydrate macroalgae can be converted to biomethane during fermentation. Cyanobacteria (photosynthetic bacteria) can produce hydrogen from sewage waste and organic molecules by fermentation.

In microbial fuel cells, bacteria catalyze specific reactions at the anode or cathode to produce an electrical current. Electrode-reducing bacteria oxidize substrates, and electrons are transferred to an anode. Electrode-oxidizing bacteria strip electrons from the cathode to reduce other substances. Oxygen availability influences the capabilities of the organisms to perform these reactions and the types of compounds that must be present at either the cathode or anode, depending on the microorganisms. Many of the substrates can be obtained from wastewater, which aids in bioremediation. Although microbial fuel cells are not yet optimized, there is great potential for this technology.

Case Study A Systematic Analysis of Biosynthetic Gene Clusters in the Human Microbiome Reveals a Common Family of Antibiotics

Mohamed S. Donia et al. (2014). *Cell* 158, 1402–1414.

Hundreds and thousands of strains of bacteria live within and on the human body. The microbiome is different among people and the microorganisms present at specific sites on the same body. Human microbiome projects have characterized "healthy" microbiomes and identified disease variants as well, such as those for Crohn's disease, bacterial vaginosis, obesity, and diabetes. Small molecules that easily diffuse help mediate host-microbe interactions. Symbiosis, mediated by these small signaling molecules, has been observed between bacteria and multiple hosts, including marine invertebrates, nematodes, insects, and plants. Other studies have shown a correlation between bacteria small molecule signaling and human diseases such as colon cancer, as well as immunosuppression of gut mucosal immune response.

The authors of this study investigated the human microbiome with regards to small molecule signaling and biosynthetic capacity. They performed systematic analysis of the microbiome in an attempt to identify biosynthetic gene clusters.

Investigating the genomes of the human microbiota usually yields large amounts of data. Why is this aspect of the human microbiome important? How many biosynthetic gene clusters were revealed in this study? These gene clusters were involved in which cellular processes?

Examining the metagenomics of the human microbiome has the potential to identify natural products that may be of use. In this study, more than 14,000 biosynthetic gene clusters were predicted, ranging from the small molecule classes of polysaccharides, nonribosomal peptides, polyketides, peptides, siderophores, and hybrids. Of these predicted clusters, the authors used shotgun metagenomic sequencing data from the Human Microbiome Project and calculated that 3118 biosynthetic gene clusters are present in healthy microbiomes.

Does the body site location affect the type of biosynthetic gene clusters that were identified? If so, how? Which areas were richest in biosynthetic gene clusters? Which class of biosynthetic gene cluster was abundant in the analyses?

Yes. The biosynthetic gene cluster types changed with regards to the origin of the sample. The oral and gut microbiomes represent the richest biosynthetic gene clusters, probably due to the overabundance of microbes in these locations. Skin, airway, and urogenital tract samples represent fewer biosynthetic gene clusters, likely because of the lower diversity of microbes in these locations. The most abundant biosynthetic gene cluster class in the gut and oral cavity is for saccharide biosynthesis. RiPPs (ribosomally encoded and post-translationally modified peptides) were moderate and represented in every location. Other gene clusters, including NRPS and PKS, were moderately present but more depleted than the other classes of biosynthetic gene clusters.

What microbes are represented in these analyses? Are these taxa different from similar experiments using nonhuman environmental isolates instead? If so, how is this significant in terms of research?

Most biosynthetic gene clusters in the human microbiome are harbored by *Bacteroides*, *Parabacteroides*, *Corynebacterium*, *Rothia*, and *Ruminococcus*. These are not the same taxa from nonhuman environmental isolates. High numbers of biosynthetic gene clusters from nonhuman environmental isolates generally include *Streptomyces*, *Bacillus*, *Pseudomonas*, *Burkholderia*, and *Myxococcus*. Even the typical genera from the human microbiota (*Escherichia*, *Lactobacillus*, *Haemophilus*, and *Enterococcus*) harbor few biosynthetic gene clusters. The biosynthetic gene clusters from the human microbiome are an underserved reservoir for potential small-molecular products.

The authors identified four examples of widely distributed biosynthetic gene clusters that met a specific set of criteria. What were the criteria used to identify these examples? Briefly describe the distribution, location, and class of these four widely distributed biosynthetic gene clusters (BGC).

These four examples were identified based on one of three criteria: predominance in a body site, resemblance to a known biosynthetic gene cluster, or cosmopolitan distribution of the cluster in body site and/or class. The first BGC is widely distributed in gram-positive bacteria within the gut microbiome. Members within this BGC might bind an aliphatic amino acid or other small molecule substrates. Saccharide BGCs are the most abundant family of clusters and are distributed in all five body locations, although they are particularly predominant in the gut with phylum Bacteroidetes harboring the largest numbers of clusters. In the oral cavity, complex polyketides with similarity to proteins from marine isolates having antibacterial and antitumor activity were found in members of Actinobacteria. RiPPs were widely distributed and variable within the human microbiota. Three subclasses exist for RiPPs: lantibiotics in the Firmicutes and Actinobacteria within all sites, thiazole/oxazole-modified microcins in the oral cavity, and thiopeptides, which have potent antibacterial activity against gram-positive bacteria.

The authors state that no thiopeptide biosynthetic gene cluster or small-molecule product has ever been characterized from the human microbiome. Which thiopeptides were characterized in this paper? What is the distribution, and what are some of the properties of these thiopeptides?

The authors discovered a total of 13 thiopeptide gene clusters from their methodology. One of these was found in the human gut samples, and seven were found in the oral cavity sample. Based on flanking sequence analysis, the gut thiopeptide gene cluster likely belongs to a *Eubacterium* species. Two of the thiopeptide gene clusters in the oral cavity belong to Actinomyces. The other five in the oral cavity belong to *Streptococcus*. Transposons and phage integrases were present within

Case Study A Systematic Analysis of Biosynthetic Gene Clusters in the Human Microbiome Reveals a Common Family of Antibiotics—cont'd

the clusters. One cluster was even present on a plasmid. These properties suggest that the thiopeptide gene clusters are mobile and can undergo horizontal gene transfer.

Of the 13 thiopeptide BGCs identified in this study, only 3 of these BGCs were similar to known thiopeptides. Furthermore, the researchers chose only one to characterize. What were the reasons for selecting the thiopeptide *bgc66* for characterization? What are some of the characteristics of *bgc66*?

The thiopeptide *bcg66* is similar to thiocillin of *Bacillus cereus*. Furthermore, *bgc66* is harbored by the vaginal isolate, *Lactobacillus gasseri*, which is easier to genetically manipulate than the organisms harboring the other thiopeptide gene clusters. These characteristics contributed to the reason the researchers chose *bgc66*. The researchers determined that the *bgc66* product was likely post-translationally modified. Additionally, they determined the structure of the *bgc66* product, which they named lactocillin. Interestingly, the structure of lactocillin contains an indolyl-S-cysteine residue that is inserted by an enzyme that is not encoded within the thiopeptide gene cluster. Also, lactocillin has a rare, free carboxylic acid at the C terminus and lacks any oxygen-requiring post-translational modifications, probably because this organism lives in an anaerobic environment.

Genes within the thiopeptide gene clusters do not encode the enzyme that inserts the indolyl-S-cysteine into lactocillin. Where are the genes for this enzyme located?

The *lcl* gene cluster contains the genes that encode the enzyme responsible for adding indolyl-S-cysteine to lactocillin. This gene cluster has limited distribution in the human microbiota, including in the vaginal microbiota where *Lactobacillus gasseri* is present. The authors broadened their search to all available metatranscriptomic data sets and found that lcl gene clusters were present in oral samples, particularly within supragingival plaque samples, and potentially expressed in humans.

The authors identified biosynthetic gene clusters within the human microbiota that may be involved in small-molecule signaling. As a consequence of their search for BGCs, the authors identified a thiopeptide gene cluster that produces a potent antibiotic against common pathogens of the vagina. The results of this study will serve as a reservoir for future research into the host-microbe and microbe–microbe interactions via small molecules.

Cell

A Systematic Analysis of Biosynthetic Gene Clusters in the Human Microbiome Reveals a Common Family of Antibiotics

Mohamed S. Donia,[1] Peter Cimermancic,[1] Christopher J. Schulze,[2] Laura C. Wieland Brown,[3] John Martin,[4] Makedonka Mitreva,[4] Jon Clardy,[5] Roger G. Linington,[2] and Michael A. Fischbach[1,*]

[1]Department of Bioengineering and Therapeutic Sciences and the California Institute for Quantitative Biosciences, University of California, San Francisco, San Francisco, CA 94158, USA

[2]Department of Chemistry and Biochemistry, University of California, Santa Cruz, Santa Cruz, CA 95064, USA

[3]Department of Chemistry, Indiana University, Bloomington, IN 47405, USA

[4]The Genome Institute, Department of Medicine and Department of Genetics, Washington University School of Medicine, St. Louis, MO 63108, USA

[5]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA

*Correspondence: fischbach@fischbachgroup.org

http://dx.doi.org/10.1016/j.cell.2014.08.032

SUMMARY

In complex biological systems, small molecules often mediate microbe-microbe and microbe-host interactions. Using a systematic approach, we identified 3,118 small-molecule biosynthetic gene clusters (BGCs) in genomes of human-associated bacteria and studied their representation in 752 metagenomic samples from the NIH Human Microbiome Project. Remarkably, we discovered that BGCs for a class of antibiotics in clinical trials, thiopeptides, are widely distributed in genomes and metagenomes of the human microbiota. We purified and solved the structure of a thiopeptide antibiotic, lactocillin, from a prominent member of the vaginal microbiota. We demonstrate that lactocillin has potent antibacterial activity against a range of Gram-positive vaginal pathogens, and we show that lactocillin and other thiopeptide BGCs are expressed in vivo by analyzing human metatranscriptomic sequencing data. Our findings illustrate the widespread distribution of small-molecule-encoding BGCs in the human microbiome, and they demonstrate the bacterial production of drug-like molecules in humans.

INTRODUCTION

The human microbiome is composed of hundreds of bacterial species and thousands of strains, and its composition differs from person to person and between different body sites of the same individual (Human Microbiome Project Consortium, 2012b). During the last decade, tremendous efforts have been made to sequence isolates of the human microbiota and metagenomic samples from various body sites (Human Microbiome Project Consortium, 2012a, 2012b; Nelson et al., 2010; Qin et al., 2010). These studies have yielded a basic understanding of the "healthy" human microbiome and have correlated deviations from the healthy state to maladies such as obesity, diabetes, bacterial vaginosis, and Crohn's disease (Gajer et al., 2012; Gevers et al., 2012, 2014; Ravel et al., 2011; Turnbaugh et al., 2009). Several recent studies have begun to examine the human microbiome from a functional point of view, where direct molecular interactions between host and microbe are revealed (An et al., 2014; Hsiao et al., 2013; Mazmanian et al., 2005; Mazmanian et al., 2008; Nougayrède et al., 2006; Wieland Brown et al., 2013; Wyatt et al., 2010).

Diffusible and cell-associated small molecules often mediate host-microbe interactions in complex environments. Examples of small-molecule-mediated interactions have been revealed in symbioses between bacteria and insects (Oh et al., 2009), marine invertebrates (Kwan et al., 2012), nematodes (McInerney et al., 1991), and plants (Long, 2001). In addition, several studies have explored the role of small molecules in interactions between microbiota and the mammalian host. For example, *Staphylococcus aureus* pyrazinones were shown to be inducers of bacterial virulence (Wyatt et al., 2010), the *Escherichia coli* metabolite colibactin was found to contribute to colon cancer (Nougayrède et al., 2006), and polysaccharide A from *Bacteroides fragilis* has been shown to suppress the gut mucosal immune response (Mazmanian et al., 2005, 2008). Recently, we and others showed that *Bacteroides fragilis* produces the canonical CD1d ligand α-galactosylceramide, revealing a specific mechanism by which the gut microbiota are capable of modulating host natural killer T cell function (An et al., 2014; Wieland Brown et al., 2013). Another recent study correlated the prevalence of hepatic cancer in mice to deoxycholic acid, a secondary bile acid produced by certain members of the gut microbiota (Yoshimoto et al., 2013). A recent metatranscriptomic study showed the expression of genes matching the clusters of orthologous groups (COG) category "secondary metabolites biosynthesis, transport, and catabolism," which is consistent with the possibility of small-molecule production but could also indicate

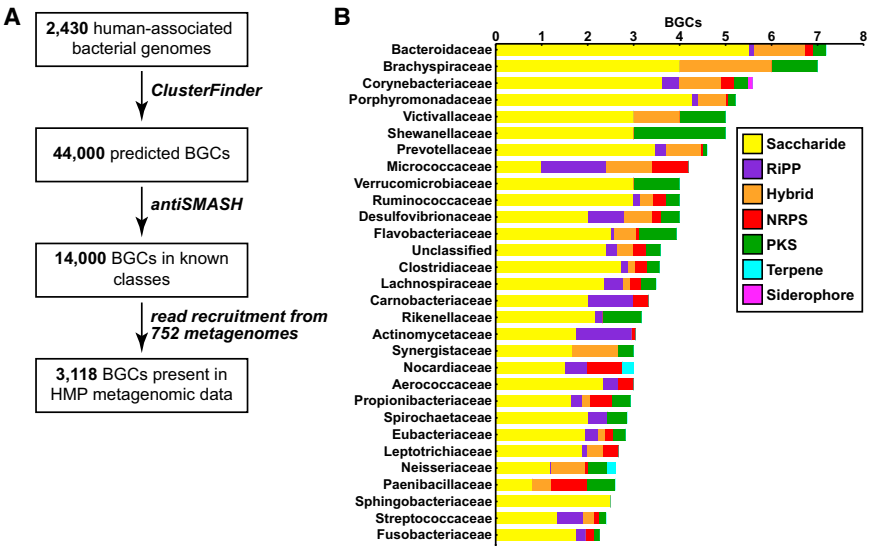


Figure 1. Overview of BGCs in the Human Microbiome

(A) Computational and experimental workflow for the identification of BGCs from human-associated bacteria.

(B) A bar graph showing the top 30 families by average BGC abundance from the human microbiome and the average number and type of BGCs discovered in isolates of each genus using the workflow in (A). See also Figure S1 and Data S1 for the full data set of predicted BGCs.

the expression of catabolic and/or transport genes unrelated to biosynthesis (Leimena et al., 2013). These examples raise the question of whether there exists a much larger set of bacterially produced molecules that mediate microbiota-host interactions. Due to the complexity of the human microbiome and its vast coding potential, a more systematic approach is needed to explore small-molecule-mediated interactions between humans and their microbiota.

In this study, we explored the biosynthetic capacity of the human microbiome by performing the first systematic identification and analysis of its biosynthetic gene clusters. Unlike previous approaches that have focused on one compound or bacterial strain at a time, our approach allows the global analysis of biosynthetic gene clusters (BGCs) that encode small molecules in thousands of isolates of the human microbiota. By measuring the representation of these BGCs in human metagenomic samples, we can assess the small-molecule coding capacity of a community, generating powerful hypotheses about which molecules might mediate microbe-host and microbe-microbe interactions in a particular community and how their prevalence differs among individuals. To illustrate the utility of this approach, we used a combination of chemistry, genetics, metagenomics, and metatranscriptomics to study a family of gene clusters that is widely distributed in the human microbiome, including the characterization of its small-molecule product and the analysis of its prevalence among body sites and individuals.

RESULTS

A Systematic Approach Identifies 3,118 Biosynthetic Gene Clusters that Are Present in Human Metagenomic Samples

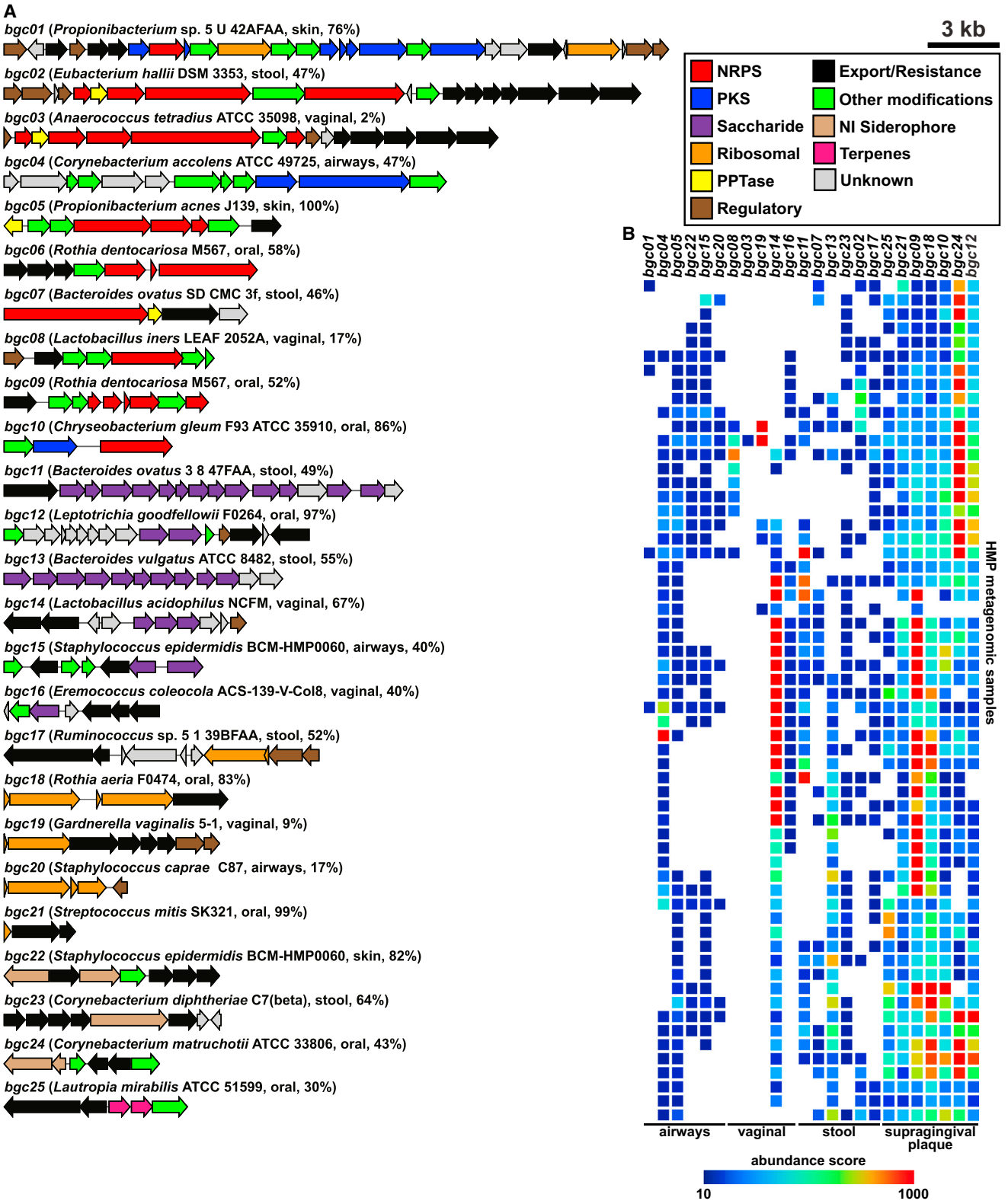
The identification of biosynthetic gene clusters in bacterial genome sequences has become a powerful tool for natural product discovery. We began by using an algorithm we recently developed, ClusterFinder, to systematically analyze 2,430 reference genomes of the human microbiota from a range of body sites (Cimermancic et al., 2014). ClusterFinder detected >14,000 biosynthetic gene clusters (average of 6 gene clusters per genome) for a broad range of small-molecule classes, including saccharides, nonribosomal peptides (NRPs), polyketides (PKs), ribosomally encoded and posttranslationaly modified peptides (RiPPs), NRPs-independent siderophores, and hybrids thereof (Figure 1 and Figure S1 available online).

Reasoning that the small-molecule products of BGCs that are widely distributed in the human population are most likely to mediate conserved microbe-host and microbe-microbe interactions, our next goal was to identify the subset of BGCs that is commonly found in healthy individuals. To achieve this goal, we examined the representation of these gene clusters in shotgun metagenomic sequencing data generated by the Human Microbiome Project (HMP, 752 samples collected from five main body sites of healthy subjects) (Human Microbiome Project Consortium, 2012b). Briefly, we used mblastx (Davis et al., 2013) to detect metagenomic reads that match the >14,000 predicted BGCs (see Experimental Procedures for a more detailed description of how gene clusters were quantified in metagenomic samples). Using this method, we calculated that a subset of 3,118 biosynthetic gene clusters (22%) are present in the microbiomes of healthy individuals; all subsequent analyses focused on this subset (Figures 1 and S1 and Data S1).

An Overview of BGC Types and Distribution in the Human Microbiome

The identification of >3,000 biosynthetic gene clusters in the human microbiome was remarkable, given that almost nothing is known about their small-molecule products or biological activities. We next examined the distribution and abundance of each of these BGCs in the 752 HMP samples.

The gut and oral cavity are by far the richest in BGCs, reflecting their increased microbial abundance relative to the other body sites (see also Figure S2A, see Experimental Procedures) (Human Microbiome Project Consortium, 2012b). A typical gut harbors 599 clusters, whereas a typical oral cavity harbors 1,061 clusters; notably, the SD of each of these values is large (152 and 143, respectively), pointing to a far greater degree of gene cluster diversity in some individuals than others. Typical skin, airway, and urogenital tract samples harbor many fewer BGCs on average (101, 31, and 41, respectively, see also Table S1).



(legend on next page)

The smaller number of BGCs in the skin, airway, and urogenital tract communities could be a result of the lower microbial diversity in these communities in comparison to the oral and gut microbiota (Human Microbiome Project Consortium, 2012b). In addition, fewer samples were sequenced from skin, airways, and urogenital tract communities (27, 65, and 94 samples, respectively) than from oral and gut communities (415 and 147, respectively), possibly affecting the total number of BGCs that were found to be present in these body sites.

The gene cluster classes in the human microbiota differ in important ways from those in non-human-associated bacteria (see also Figure S2B). Even in light of the predominance of saccharides among environmental bacteria (Cimermancic et al., 2014), saccharides are significantly enriched in the microbiota (see also Figure S2B), making them by far the most abundant BGC class in the gut and oral cavity (average of 443 and 662 saccharide clusters per sample, respectively). RiPPs, which are modestly enriched, are broadly distributed in every body site. Although nonribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) gene clusters are significantly depleted in the microbiota, they are still present at moderate levels (average of 57 PKS and 19 NRPS clusters in a gut sample and 129 PKS and 46 NRPS clusters in an oral cavity sample; see also Table S1), and notable exemplars are widely distributed in the healthy human population (see below). Most of the BGCs from nonhuman environmental isolates are harbored by members of the abundant genera *Streptomyces*, *Bacillus*, *Pseudomonas*, *Burkholderia*, and *Myxococcus* (Cimermancic et al., 2014); in contrast, most human-associated BGCs are harbored by members of the abundant human-associated genera *Bacteroides*, *Parabacteroides*, *Corynebacterium*, *Rothia*, and *Ruminococcus* (up to seven BGCs per genome), pointing to undermined taxa that should be rich BGC sources for experimental mining efforts (Figure 1 and see also Figure S1). Many of these genomes harbor large BGCs despite being only 2 to 3 Mb, suggesting an important ecological role for their small-molecule products. Some common genera of the human microbiota, including *Escherichia*, *Lactobacillus*, *Haemophilus*, and *Enterococcus*, harbor fewer than two BGCs per genome. Taken together, these findings indicate that the human microbiome harbors a rich and diverse array of BGCs that is mostly distinct in source and composition from that of nonhuman isolates.

The best-studied gene clusters are not widely distributed among healthy individuals. With the exception of the pyrazinone BGC from *S. aureus* (found in >9% of the airways samples), two of the most thoroughly studied BGCs from human-associated bacteria are found in <5% of the samples from their body site of origin (colibactin from *E. coli* and polysaccharide A from *Bacteroides fragilis* (Mazmanian et al., 2005; Nougayrède et al., 2006; Wyatt et al., 2010). In contrast, nearly all of the BGCs from our data set that are widely distributed in healthy humans (present in >10% of the samples from the body site of origin) have never been studied or even described; a selected subset of these clusters is shown in Figure 2 (*bgc01–bgc25*). These results illustrate how little is known about the small-molecule products of the most common BGCs in human-associated niches.

Intriguingly, a smaller subset of 519 clusters is present in >50% of the gut samples; similar ''common'' BGC subsets are present in the oral cavity (582 clusters), skin (65 clusters), urogenital tract (16 clusters), and airways (11 clusters) (see also Table S1). In addition, several widely distributed families of closely related BGCs were revealed by our analysis. Among these BGC families, four examples stood out because of their predominance in a body site, resemblance to a known BGC, or cosmopolitan distribution: a family of NRPS BGCs from the gut, a pair of PKS BGCs from the oral cavity, Bacteroidetes saccharides, and RiPPs.

A Large Family of Nonribosomal Peptide BGCs Is Widely Distributed in the Gut Community

The first BGC family we identified is a set of NRPS clusters that are found exclusively in gut isolates (hereafter *bgc26–bgc54*). Surprisingly, members of this family are harbored by species of a wide variety of Gram-positive (*Clostridium*, *Ruminococcus*, *Eubacterium*, *Lachnospiraceae*, *Blautia*) and Gram-negative genera (*Bacteroides* and *Desulfovibrio*). In addition to human gut isolates, this NRPS family is found in isolates from the chicken gut (*Bacteroides barnesiae* DSM 18169) and bovine rumen (*Methanobrevibacter ruminantium* M1) (Figure 3). Remarkably, 137/149 (92%) of the HMP stool samples contained at least one member of this family of NRPS clusters (see also Figure S3), whereas none or very few of the airway (0/94), oral (2/406), vaginal (2/65), and skin (1/27) metagenomic samples harbored a member of this family. The wide distribution and broad representation of this NRPS family suggests its importance for gut microbial inhabitants. Members of this BGC family fall into two main groups: the first consists of three NRPS modules, whereas the second is composed of only two (Figure 3). The predicted substrate of the penultimate adenylation domain is an aliphatic amino acid (Leu, Ile, Val, or Ala), whereas the other adenylation domains have predicted substrates that vary widely among the clusters, suggesting that the small-molecule products of these BGCs may be a family of closely related molecules.

A Family of Complex Polyketides from Oral Actinobacteria Are Widely Distributed in the Oral Cavity

The most widely distributed multimodular PKS in the oral cavity is a pair of closely related ~80 kb clusters of the trans-AT type

Figure 2. A Selected Subset of BGCs from the Human Microbiota

(A) 25 selected BGCs from the human microbiota, spanning each of the body sites (gut, vagina, airways and skin, and oral cavity), BGC types (PKS, NRPS, RiPPs, terpenes, NI siderophores, and saccharides), and prevalent bacterial phyla (Actinobacteria, Bacteroidetes, Firmicutes, and Proteobacteria). The label of each gene cluster indicates its source organism, body site of origin, and the percentage of HMP samples harboring this cluster in its body site of origin. All but two of these BGCs are present in more than 10% of the samples from their body site of origin, indicating that they are widely distributed among healthy subjects.

(B) Heat map showing the representation of BGCs from (A) in a subset of 60 selected HMP metagenomic samples from four body sites. The color of the cells in the heat map represents an abundance score ranging from 10 (blue) to 1,000 (red) (key shown to the bottom, see Extended Experimental Procedures for calculation of abundance scores and see also Figure S2 and Table S1).

(hereafter *bgc54–bgc55*). These clusters, which are found in two different species of oral Actinobacteria (*Propionibacterium propionicum* F0230a and *Actinomyces timonensis* DSM 23838), are strikingly similar in their domain architecture to a BGC from the marine isolate *Streptomyces sp.* A7248 that encodes the production of the homodimeric macrolide SIA7248 (Figure 4) (Zou et al., 2013). Despite this similarity, the enzymes encoded by the three clusters share only ~40% identity, and no mobile elements were found in either of the human-associated clusters. SIA7248 is chemically similar to the marinomycins, which are known to have potent antibacterial and antitumor activity (Kwon et al., 2006). Our computational analysis of PKS architecture and AT domain selectivity predicts that the small-molecule products of both PKS clusters will be nearly identical to the marinomycins and SIA7248, except for an alternative loading moiety activated by an acyl-CoA ligase domain that is absent in the SIA7248 cluster. Notably, *bgc54* is widely distributed in healthy subjects (34% of the HMP supragingival plaque samples), indicating that a close relative of a complex marine bacterial metabolite might be common in the oral cavity.

Bacteroidetes Saccharides Are Predominant and Variable in the Gut

Saccharide BGCs are the most abundant family in each of the five body sites and are particularly predominant in the gut (74% of all BGCs in a typical HMP stool sample are saccharides; see also Table S1). The phylum Bacteroidetes harbors the largest number of saccharide BGCs (Cimermancic et al., 2014), two of which have been structurally and functionally characterized (polysaccharide A and B; Baumann et al., 1992; Mazmanian et al., 2005, 2008; Tzianabos et al., 1993). Two observations were notable. First, among the most common saccharide BGCs were capsular polysaccharide loci from *B. vulgatus* and *B. ovatus* (see also Figure S4), which differ considerably from the better-known examples in *B. fragilis*, suggesting that structurally distinct capsular polysaccharides might be remarkably common in the healthy human population. Second, ten HMP stool samples that show similar taxonomic composition by MetaPhlAn profiling (Segata et al., 2012) harbor largely distinct sets of saccharide BGCs, with only a small number of BGCs common among the samples (see also Figure S4). Thus, knowledge about the taxonomic composition of gut communities from 16S rDNA analysis or metagenomic classification tools does not reveal the inherent diversity of BGCs in these communities.

Ribosomally Synthesized, Posttranslationally Modified Peptides Are among the Most Widely Distributed and Variable BGCs Encoded by the Human Microbiota

BGCs encoding modified RiPPs are widely distributed among the human microbiota. Three subclasses of RiPPs were particularly notable: lantibiotics, which were numerous, variable, and broadly distributed throughout the Firmicutes and Actinobacteria from all of the major body sites; thiazole/oxazole-modified microcins (TOMMs), which were prevalent in the oral cavity; and thiopeptides. Although a small set of lantibiotics and TOMMs have been isolated from members of the human microbiota (Chikindas et al., 1995; Gonzalez et al., 2010; Lee et al., 2008), most are produced by pathogens or rare commensals.

Thiopeptides are highly modified RiPPs that have potent antibacterial activity against Gram-positive species. A semisynthetic member of this class, LFF571 (Novartis), is currently in phase II clinical trials for treating *Clostridium difficile* infection (LaMarche et al., 2012). Although a thiopeptide BGC was previously predicted by our group and others in the genome of the skin isolate *Propionibacterium acnes* (Brzuszkiewicz et al., 2011; Wieland Brown et al., 2009), no thiopeptide BGC or small-molecule product from the human microbiome has ever been characterized experimentally. Remarkably, in the analysis reported here, we discovered thiopeptide-like BGCs in isolates from every human body site (*Lactobacillus gasseri*, urogenital; *Propionibacterium acnes*, skin; *Streptococcus downei* and *S. sobrinus*, oral; and *Enterococcus faecalis*, gut). Additionally, we discovered a thiopeptide gene cluster in the porcine gut isolate *Lactobacillus johnsonii* PF01, indicating that thiopeptide BGCs are found in the microbiota of animals other than humans (Figure 5A).

We next turned to the question of whether we could discover additional thiopeptide gene clusters in metagenomic data that were not present in any sequenced bacterial genome. To address this question, we mined HMP metagenomic assemblies for the presence of additional thiopeptide gene clusters; remarkably, we discovered eight additional thiopeptide BGCs (7 complete, 1 partial) using this strategy, increasing the number of thiopeptide BGCs identified here to 13 (*bgc56-bgc68*) (Figure 5A) (see Experimental Procedures). Of these, we found one in human gut metagenomic samples (*bgc61*) and seven in human oral metagenomic data sets (*bgc58*, *bgc59*, *bgc62*, *bgc64*, *bgc65*, *bgc67*, and *bgc68*).

An analysis of the genes flanking these clusters suggests that *bgc61* resides in a prominent member of the human gut microbiome (*Eubacterium rectale*), whereas two of the oral metagenomic thiopeptide gene clusters are found in *Actinomyces* and five others are harbored by *Streptococcus* (see also Table S2). Transposases and phage integrases are unusually prevalent among the thiopeptide BGCs (70%), and in at least one case (*bgc66*), we could show bioinformatically and experimentally that the cluster exists on a plasmid, suggesting a potential for mobility (see below, and note that the thiopeptide BGC from the porcine gut isolate *Lactobacillus johnsonii* PF01 [Lee et al., 2011] also exists on a plasmid). Four of the 13 thiopeptide BGCs are present in >20% of the HMP samples at one of the body sites, and 155/406 HMP oral samples (38%) harbor at least one thiopeptide BGC (Figure 5C and Table S2). For example, *bgc61* is present in 11% of gut samples and *bgc65* is present in 34% of oral samples, indicating that there exist widely distributed biosynthetic gene clusters in the healthy human population that are not found in the reference genome database.

In order to gain more insight into the evolution of these human-associated thiopeptide BGCs, we performed a detailed bioinformatic analysis of their biosynthetic genes and precursor peptides. Genomic and metagenomic thiopeptide BGCs fall into six subfamilies in which members of each subfamily harbor similar precursor peptides (Figures 5A and 5B). Notably, the BGC precursor peptides and enzymes that posttranslationally modify them cluster largely according to their body site of origin rather than the phylogeny of their host (Figures 5A, 5B, and S5). Moreover, the distribution of the oral thiopeptide clusters across

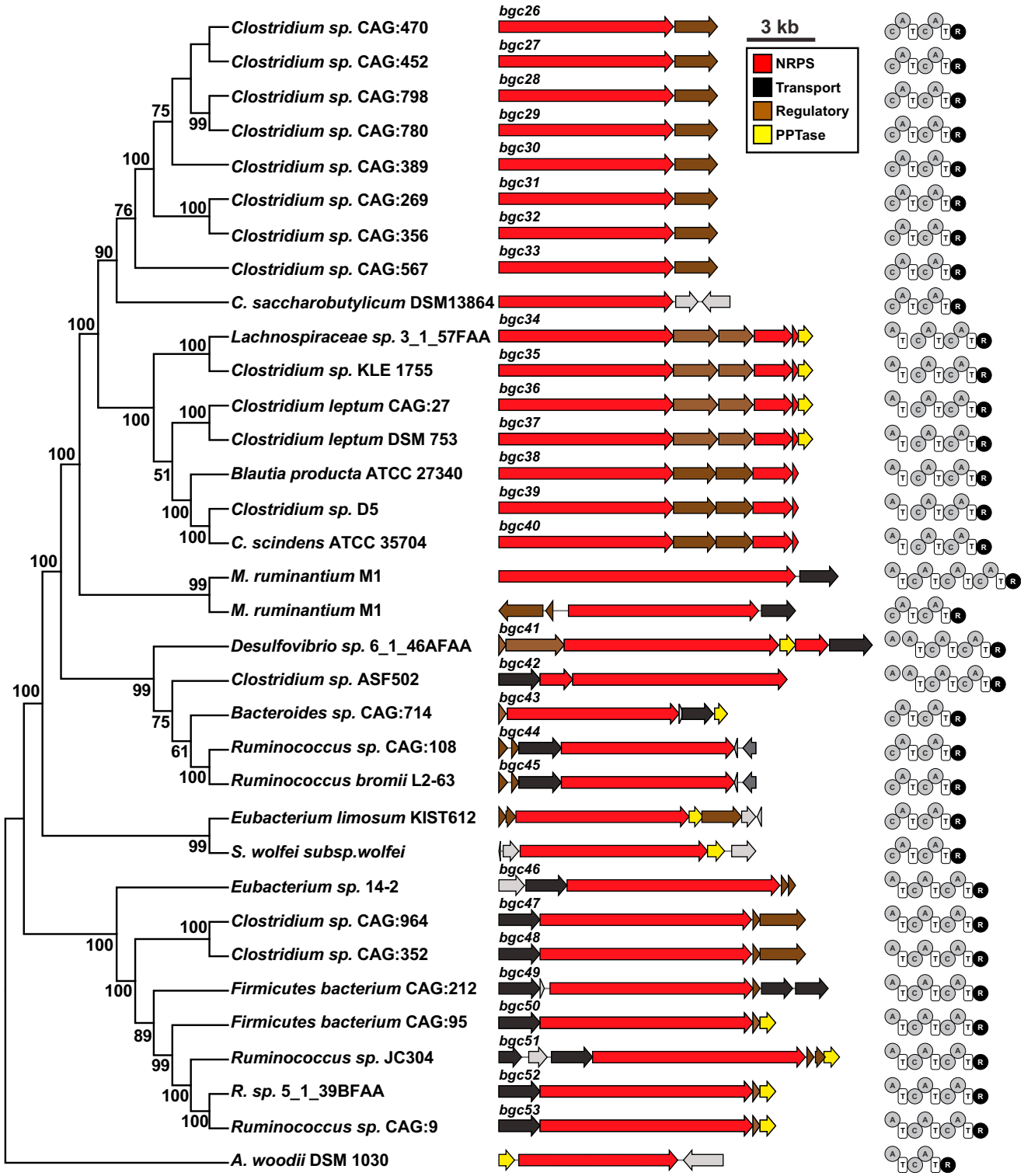


Figure 3. An Abundant Family of NRPS BGCs Is Found Exclusively in Gut Isolates and Stool Metagenomes

Left, a phylogenetic tree (Maximum Parsimony, MEGA5) based on the main NRPS gene of 28 clusters from human bacterial gut isolates, two from bovine rumen archaeal isolates, and four from environmental isolates (see Experimental Procedures). The numbers next to the branches represent the percentage of replicate trees in which this topology was reached in a bootstrap test of 1,000 replicates. Middle, schematic of each BGC (see also Figure S3 for a full heat map showing the prevalence and abundance of each member of this NRPS family in HMP stool samples). Right, domain organization of the NRPS genes of each cluster (A, adenylation domain; C, condensation domain; T, thiolation domain; R, terminal reductase domain).

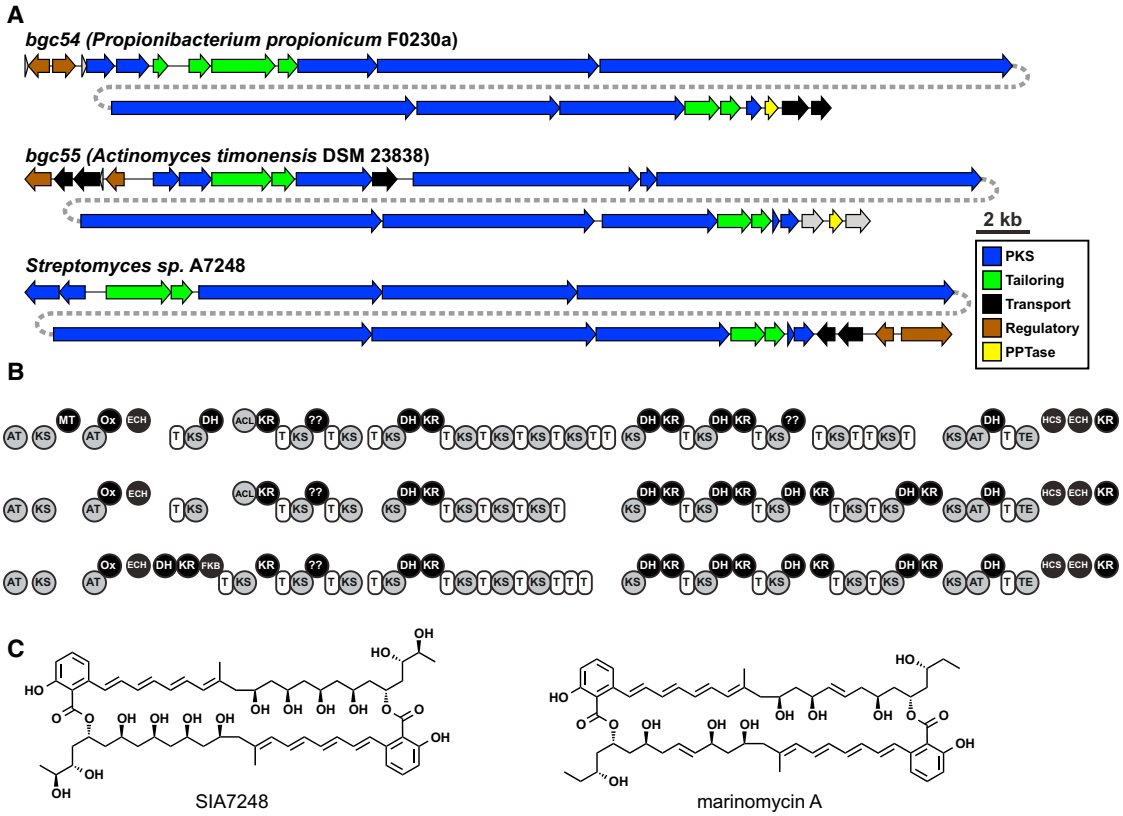


Figure 4. A Family of Complex PKS BGCs Is Prevalent in the Human Oral Cavity

(A) Related PKS BGCs in human oral actinobacteria *P. propionicum* F0230a and *A. timonensis* DSM 23838 and the marine actinobacterium *Streptomyces* sp. A7248. The label of each BGC indicates its source organism.

(B) Domain organization of the three BGCs shown in (A) (AT, acyltransferase domain; KS, ketosynthase domain; MT, methyltransferase domain; Ox, oxidation domain; ECH, enoyl-CoA hydratase domain; DH, dehydratase domain; KR, ketoreductase domain; ACL, acyl-CoA ligase domain; T, thiolation domain, HCS, HMG-CoA synthase domain; FKB, FkbH-like protein. Note that the domain architecture is remarkably conserved among the three pathways, despite low amino acid sequence identity (~40%).

(C) Structures of SIA7248, the product of the *Streptomyces* sp. A7248 BGC shown in (A), and the related molecule marinomycin A.

the sample set indicates that most of the samples harbor only one thiopeptide BGC (Figure 5C). Taken together, these results are consistent with the possibility that thiopeptide gene clusters are mobile and that they undergo horizontal transfer among the microbiota within a body site (Smillie et al., 2011).

Discovery and Characterization of a Thiopeptide from the Human Microbiome

We next set out to purify and solve the structure of a thiopeptide from the human microbiome. Of the 13 thiopeptide BGCs described here, three resembled BGCs for known thiopeptides (*bgc56* is homologous to the berninamycin BGC from *Streptomyces bernensis*, and one of its two putative precursor peptides is identical to that of berninamycin [Malcolmson et al., 2013]). Likewise, the genetic organization and precursor peptide sequences of *bgc68* and *bgc66* are similar to those of the TP-1161 (*Nocardiopsis sp.* TFS65-07) and thiocillin (*Bacillus cereus* ATCC 14579) BGCs, respectively (Engelhardt et al., 2010; Liao et al., 2009; Wieland Brown et al., 2009). Because there is more precedent for the genetic manipulation of *Lactobacillus* than *Propionibacterium* and *bgc68* does not exist in any of the reference strains, we selected *bgc66* for chemical and biological characterization.

The reference genome in which *bgc66* resides is a vaginal isolate of *Lactobacillus gasseri* from a subject in Texas (*L. gasseri* JV-V03, HMP DACC); *L. gasseri* is one of the four *Lactobacillus* species that commonly dominate the vaginal community (Ravel et al., 2011). In order to determine whether the small-molecule product of *bgc66* is produced under conditions of laboratory culture, we constructed an insertional mutant of the gene harboring a YcaO domain (locus tag: HMPREF0514_11730) by single crossover mutagenesis (see Experimental Procedures). By comparing the HPLC and LC-MS profiles of organic extracts of the wild-type and *bgc66::HMPREF0514_11730* mutant strains, we identified a single peak that is present in the wild-type but missing in the insertional mutant (Figure 6A). This peak had an absorption spectrum consistent with a compound that harbors a trithiazoylpyridine core (Figure 6A and see also Figure S6). High-resolution mass spectrometry enabled us to calculate the empirical formula $C_{51}H_{45}N_{13}O_{10}S_7$ (observed $[M+H]^+$ at *m/z*

1224.15184, calculated $[M+H]^+$ 1224.15354, Δppm = 1.3). This mass and empirical formula deviated from the computationally predicted mass and empirical formula for the *bgc66* product (1058.0667 and $C_{42}H_{34}N_{12}O_8S_7$, respectively), suggesting that the *bgc66* product harbored at least one unknown posttranslational modification.

To obtain multimilligram quantities of the *bgc66* product for structural characterization, we performed a 50 L cultivation of *L. gasseri* JV-V03. Our initial attempts to perform NMR experiments on the isolated product failed due to its low solubility in NMR solvents and apparent instability. Hypothesizing that the *bgc66* product harbored a free carboxylic acid that was partly responsible for its insolubility and instability, we treated the product extracted from a second 60 L cultivation with TMS-diazomethane to convert any free carboxylic acids to methyl esters. The resulting product had a mass consistent with the addition of a single methyl group and exhibited greatly improved solubility and stability. After extensive purification by HPLC, the purified product (0.5 mg) was characterized using a series of 1D and 2D NMR experiments, high-resolution tandem mass spectrometry (MS/MS and MS^n), and isotope labeling experiments. Collectively, these data enabled us to determine the structure of the *bgc66* product, to which we assign the name lactocillin (see Experimental Procedures, Figures 6B and S6, and Data S2).

Lactocillin harbors a canonical 26-membered thiopeptide scaffold, with a trithiazolylpyridine core and a macrocycle with four cysteine-derived heterocycles and a single dehydrobutyrine residue. Three structural features set it apart from thiocillin. (1) The presence of an indolyl-*S*-cysteine residue at position 8, a feature that is structurally and regiochemically reminiscent of the 3-methylindolyl and quinaldic acid moieties of nosiheptide and thiostrepton, respectively (Just-Baringo et al., 2014; Kelly et al., 2009; Liao et al., 2009; Yu et al., 2009). Interestingly, the enzymes predicted to install the indolyl-*S*-cysteine in lactocillin (*lclI*, a thiolation domain, *lclJ*, an adenylation domain, and *lclK*, an alpha-beta hydrolase) are unrelated to any genes in the nosiheptide or thiostrepton BGCs, indicating that it may have arisen by convergent evolution. Lactocillin also (2) harbors a free carboxylic acid at the C terminus, a rare feature among thiopeptides (Just-Baringo et al., 2014) and (3) lacks any oxygen-requiring posttranslational modifications, reflecting the anaerobic conditions under which *L. gasseri* thrives in the vaginal community (Figure 6B).

Next, we set out to determine whether lactocillin has a spectrum of activity similar to other thiopeptides, which are potent antibiotics against Gram-positive, but not Gram-negative, bacteria. To answer this question, we tested purified lactocillin against commonly observed pathogens and commensals from the vaginal community. Lactocillin was active against *Staphylococus aureus*, *Enterococcus faecalis*, *Gardnerella vaginalis*, and *Corynebacterium aurimucosum*, but not *Escherichia coli*, an activity spectrum similar to that of other thiopeptides and suggestive of a potential role in protecting the vaginal microbiota against pathogen invasion. Interestingly, lactocillin was inactive against the vaginal commensals *Lactobacillus jensenii* JV-V16, *Lactobacillus crispatus* JV-V01, and *L. gasseri* SV-16A-US (a different strain from the lactocillin producer) at 425 nM, raising the possibility that the vaginal microbiota have evolved resistance against a compound they commonly encounter (Table 1 and see Experimental Procedures).

Metatranscriptomic Data Analysis Reveals that the *lcl* Cluster Is Expressed in Humans

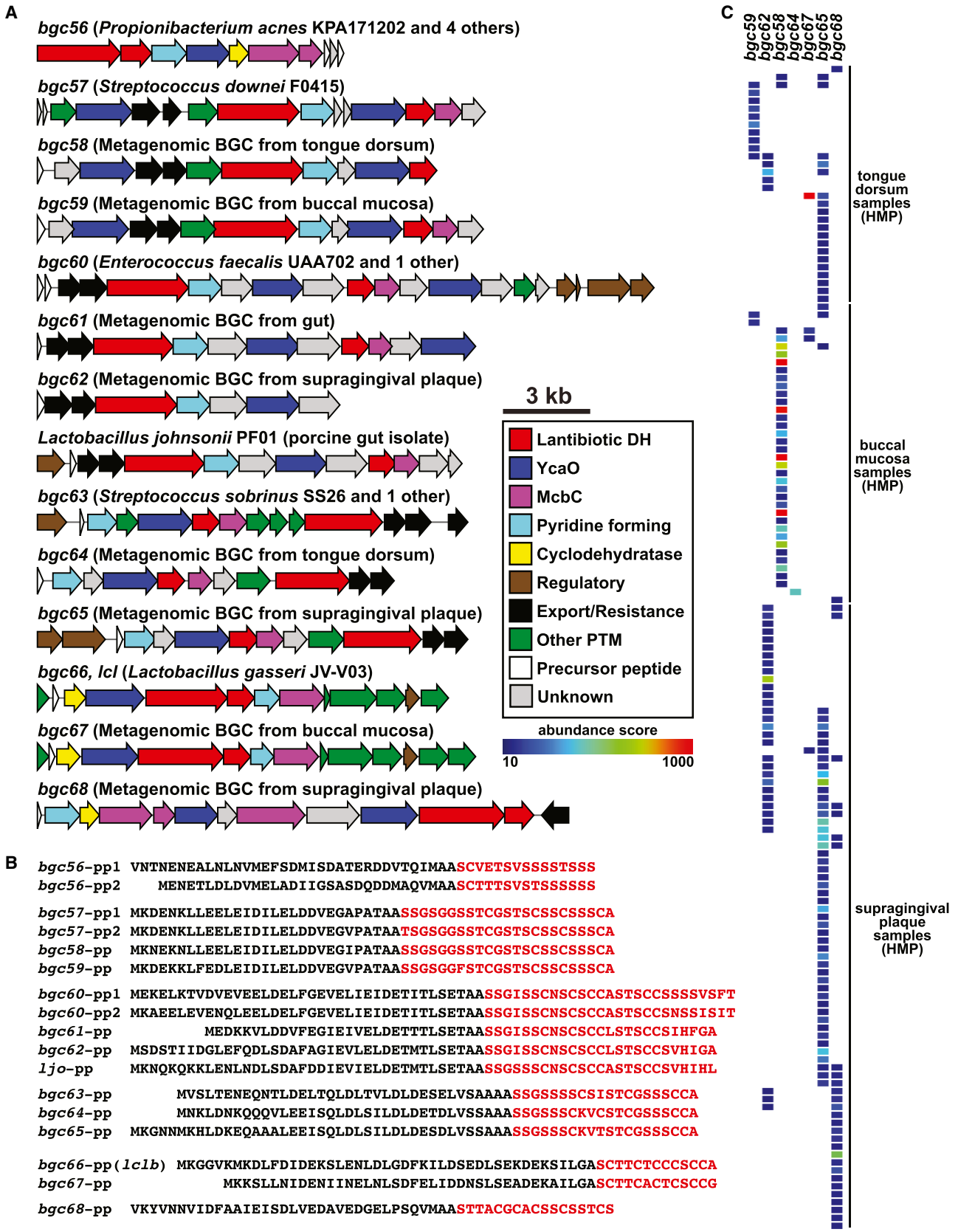
We next examined the distribution pattern of the *lcl* gene cluster in the human microbiota. In contrast to other human-associated thiopeptide BGCs (Figures 5A and 5C), the *lcl* gene cluster appears to have a limited distribution in genomic and metagenomic data sets: *L. gasseri* JV-V03 is the only one of the ten sequenced *L. gasseri* genomes that harbors the cluster (see Extended Experimental Procedures), and none of the six vaginal metagenomic samples in which *L. gasseri* is the dominant strain contain reads that match the *lcl* cluster. However, building on the finding that *lcl* resides on a plasmid (Figure 6C) and that some BGCs are present in more than one body site (see also Figure S2), we broadened our search to publicly available metatranscriptomic data sets from the oral cavity (see Experimental Procedures). To our surprise, we found that two thiopeptide gene clusters—*lcl* and *bgc65*—were covered by oral metatranscriptomic reads in multiple samples (in total, reads matching *lcl* and *bgc65* were found in 3/38 and 11/38 of the supragingival plaque metatranscriptomic samples, respectively) (see Experimental Procedures, Figure 6D and see also Figure S7). In addition, the rest of genes on the plasmid in which the *lcl* cluster resides were also present in oral metatranscriptomic samples, further emphasizing its mobility (see also Figure S7). These results show that thiopeptide gene clusters are actively transcribed in human samples, suggesting a potential role in mediating microbe-microbe interactions in the communities in which they are expressed.

DISCUSSION

Here, we systematically identify BGCs in the genomes of human isolates, examine their representation in metagenomic and metatranscriptomic samples, identify widely distributed BGC families, and characterize the structure and activity of their small-molecule products. Our approach introduces a method for identifying and prioritizing BGCs for experimental characterization and can easily be generalized to other BGC families. The resulting database of BGCs that were detected in human samples (see also Data S1) will serve as a resource for future studies that aim to discover small-molecule-mediated interactions in the human microbiota.

Our approach has three important limitations. First, we rely on sequenced isolates of the human microbiota to predict the initial set of BGCs that we later use to recruit reads from whole-genome shotgun metagenomic samples. Although the ~2,400 bacterial genomes that we analyze here span all of the human-associated bacterial phyla and represent each major body site, they are biased toward gut and oral residents and easily cultivated species, for example.

Second, in examining the representation of BGCs in human metagenomic samples, we relied on the HMP whole-genome shotgun metagenomic data sets, which were sampled from fewer body sites and subsites than the larger 16S data set. Fewer samples were sequenced from skin than other body sites (27 versus at least 60 for all other body sites), and these samples were all obtained from one skin subsite (retroauricular crease). We anticipate that the number of skin-associated BGCs will



(legend on next page)

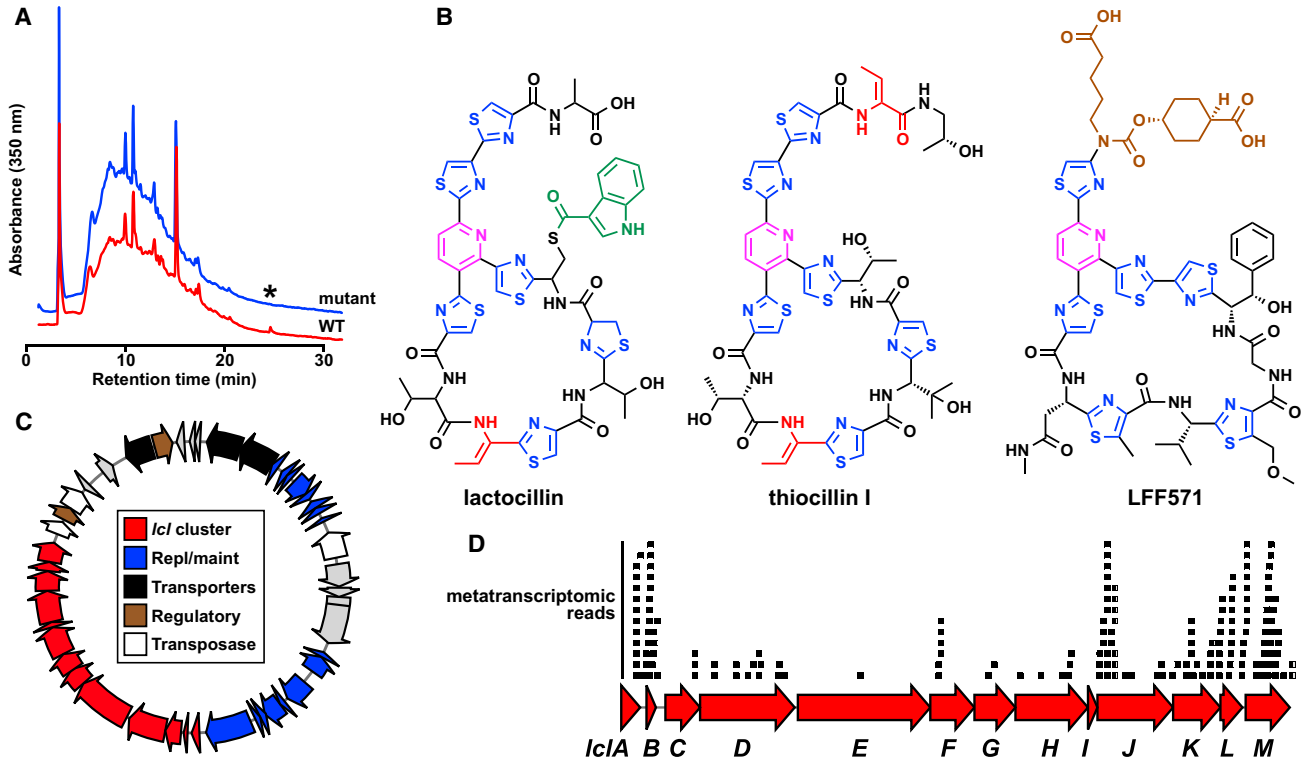


Figure 6. Characterization of an Antibiotic, Lactocillin, from a Human Vaginal Isolate

(A) HPLC analysis of organic extracts of cell pellets from wild-type (red) and *lclD* insertional mutant (blue) strains of *L. gasseri* JV-V03, monitored at 350 nm. An asterisk indicates the HPLC peak corresponding to lactocillin.

(B) Planar structure of lactocillin (see Experimental Procedures and see also Figure S6 and Data S2 for details about its purification and structural elucidation), the *Bacillus cereus* antibiotic thiocillin, and the clinical candidate LFF571 (Phase II, Novartis). Note the structural similarities among the three thiopeptides.

(C) Plasmid harboring the lactocillin BGC (see Experimental Procedures for details of the experimental closure of the circular plasmid). The lactocillin gene cluster (red) occupies ~30% of the plasmid; other elements on the plasmid include plasmid replication and maintenance genes (blue), transporters (black), transcription regulators (brown), and transposases and phage integrases (white).

(D) Raw oral metatranscriptomic reads were recruited to the *lcl* cluster using blastn and aligned using Geneious. Each black bar represents one read. Note that the precursor peptide *lclB* is among the most deeply covered genes in the *lcl* cluster, as anticipated for a RiPP pathway (see Experimental Procedures and see also Figure S7 for metatranscriptomic analysis of the whole *lcl* plasmid and of *bgc65*).

rise as the number and diversity of whole-genome shotgun metagenomic data sets increases.

Third, we show that metatranscriptomic reads matching the *lcl* cluster exist in RNA-seq data from the human oral cavity, suggesting that the *lcl* cluster is expressed in humans. To prove conclusively that lactocillin is produced in humans, it would need to be directly detected in human vaginal or oral samples using sensitive analytical techniques. The activity of lactocillin in its native context will need to be further explored in colonization experiments in which a lactocillin producer is compared to an isogenic strain deficient in lactocillin production.

Landmark studies of gene clusters encoding catabolic pathways from the microbiota have shown the power of connecting genes to microbiome-related functions. Sonnenburg et al. (2010) and Larsbrink et al. (2014) have dissected the function of gene clusters from *Bacteroides* that catabolize fructans and xyloglucans, respectively. Both studies lend important insights into the role of specialized catabolic modules in competition for nutrient niches in the gut community, and they demonstrate that a mechanistic understanding of gene cluster function yields predictive power into how members of the gut community will respond to a change in diet. The current

Figure 5. Thiopeptide BGCs Are Widespread in Isolates and Metagenomes of All Main Human Body Sites

(A) Five thiopeptide BGCs from human isolates, eight thiopeptide BGCs from human metagenomes, and one thiopeptide BGC from a pig isolate are shown. The label of each BGC indicates its body site of origin.

(B) Precursor peptides corresponding to the thiopeptide BGCs shown in (A). Note that the precursor peptides fall into six subgroups of nearly identical sequences. The structural portion of the precursor peptide is shown in red (see also Figure S5 for a phylogenetic analysis of thiopeptide BGCs).

(C) A heat map showing the representation and abundance of six oral thiopeptide BGCs in HMP metagenomic oral samples (see also Table S2 for the quantification of all thiopeptide BGCs in all HMP samples). Note that, although each BGC shown in the heat map is well represented in the oral cavity, most samples harbor only one thiopeptide BGC.

Table 1. Minimum Inhibitory Concentration of Lactocillin against Vaginal and Oral Pathogens and Commensals

Strain	Phylum	Description	Lactocillin MIC (nM)
Staphylococcus aureus	Firmicutes	pathogen	42[a]
Escherichia coli	Proteobacteria	commensal	NA[b]
Enterococcus faecalis	Firmicutes	pathogen	425
Lactobacillus jensenii	Firmicutes	vaginal commensal	NA[b]
Lactobacillus gasseri[c]	Firmicutes	vaginal commensal	NA[b]
Lactobacillus crispatus	Firmicutes	vaginal commensal	NA[b]
Corynebacterium aurimucosum	Actinobacteria	vaginal pathogen	42[a]
Finegoldia magna	Firmicutes	vaginal pathogen	NA[b]
Gardnerella vaginalis	Actinobacteria	vaginal pathogen	212
Streptococcus sanguinis	Firmicutes	oral commensal	212
Streptococcus sobrinus	Firmicutes	oral commensal/ pathogen	85
Streptococcus mutans	Firmicutes	oral commensal/ pathogen	425

See Experimental Procedures for details. *Streptococcus sanguinis* and *Gardnerella vaginalis* were partially inhibited at 80 nM lactocillin but not completely until the next concentration tested (212 nM). *Streptococcus mutans* was partially inhibited at 212 nM but not completely until the next concentration tested (425 nM).

[a]Lowest concentration tested was 42 nM.

[b]Highest concentration tested was 425 nM.

[c]Note that the strain of *L. gasseri* tested was SV-16A-US, which is a different strain from the lactocillin producer (JV-V03).

state of knowledge of how BGCs in the human microbiota function is comparatively less well developed but holds great promise in yielding similar insights into microbe-host and microbe-microbe interactions.

The small-molecule products of BGCs are widely used in the clinic, and they constitute much of the chemical language of interspecies interactions. Our data highlight the fact that there exist hundreds of widely distributed BGCs of unknown function in the human microbiome, and they provide a template for future experimental efforts to discover biologically active small molecules from the microbiota (see also Data S1 for a full data set of human-associated BGCs). These molecules represent a promising starting point for studying microbe-host interactions at the level of molecular mechanism and potentially a rich source of therapeutics.

EXPERIMENTAL PROCEDURES

Computational Analysis of BGCs from the Human Microbiome

Genome sequences of 2,430 bacterial strains isolated from humans were obtained from JGI-IMG, and BGCs were predicted using ClusterFinder (Cimermancic et al., 2014). The initial set of 44,000 putative BGCs was then analyzed by antiSMASH (Medema et al., 2011), which classified a subset of 14,000 BGCs into a known category. A database containing the amino acid sequence of each gene in these 14,000 BGCs was constructed and queried against the processed (post-QC) reads of 752 HMP metagenomic samples using mblastx (Davis et al., 2013). Abundance scores of genes and gene clusters were calculated, and 3,118 BGCs were found to be present in at least one metagenomic sample (see also Extended Experimental Procedures and Data S3 for details about abundance score calculations, thresholds, and additional computational analyses).

Identification of Thiopeptide BGCs from HMP Metagenomic Data

A database containing the amino acid sequence of 24 homologs of TclM (the enzyme responsible for pyridine ring formation in thiocillin) was constructed. This data set was then used as a query for tblastn searches against metagenomic assemblies of the 752 HMP samples. Contigs that generated hits to multiple thiopeptide genes were analyzed using antiSMASH, and results were verified manually. A thiopeptide BGC was called only when genes for all the essential posttranslational modifications were identified in addition to the precursor peptide (see also Extended Experimental Procedures).

Generation of an Insertional Mutant in *L. gasseri* JV-V03 and Verification of Lactocillin Production

L. gasseri JV-V03 was cultivated in an anaerobic chamber (80% N_2, 10% CO_2, and 10% H_2) in MRS broth (BD) at 37°C. A 1,000 bp fragment of *lclD* was PCR amplified and cloned into pKM082, a suicide vector harboring an erythromycin resistance gene. *L. gasseri* JV-V03 was transformed with this vector by electroporation, and transformants were selected on erythromycin and verified using PCR (see also Extended Experimental Procedures and Table S3). Cell pellets from the wild-type strain and insertional mutants were grown in 1 L MRS broth, extracted with methanol, and analyzed using HPLC and LC-MS monitoring at 350 and 220 nm.

Large-Scale Production and Derivatization of Lactocillin

A 60 L culture of *L. gasseri* JV-V03 was grown for 2 days at 37°C. Cell pellets were harvested and extracted with 3 L methanol. Organic extracts were dried using rotary evaporation and fractionated on a C18 Sep Pak column using a step gradient of 20%, 40%, 60%, 80%, and 100% methanol in water. Lactocillin eluted in the 80% and 100% fractions, as determined by LC-MS. Semipurified lactocillin from the 80% fraction was methylated using TMS-diazomethane, and the methylated lactocillin product was purified by preparative HPLC (see also Extended Experimental Procedures).

Structural Elucidation of Lactocillin Methyl Ester

The structure of purified lactocillin methyl ester was solved using a combination of MS experiments (HRMS, HRMS/MS, and $HRMS^n$) and 1D and 2D NMR spectroscopy experiments (1D ^{1}HNMR, gCOSY, HSQC, HMBC, and ROESY). In addition, heavy isotope feeding experiments were performed to confirm the identity of the indolyl acyl group in the structure (see also Extended Experimental Procedures).

Purification and MIC Determination of Lactocillin

Lactocillin was purified by preparative HPLC from the 100% Sep Pak fraction and quantified using HPLC. The minimal inhibitory concentration of purified lactocillin against a suite of commensal and pathogenic vaginal and oral isolates was tested using a range of concentrations from 42 to 425 nM. A vehicle-alone control was used in each experiment (see also Extended Experimental Procedures).

Metatranscriptomic Analysis of Thiopeptide BGCs

A database containing the nucleotide sequences of all human-associated thiopeptide BGCs was generated. Raw Illumina reads from 38 oral metatranscriptomic samples were compared to this database using blastn, and hits with expectation values < $1e^{-10}$ were recruited to the clusters using Geneious.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, seven figures, three data files, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2014.08.032.

AUTHOR CONTRIBUTIONS

M.S.D., C.J.S., R.G.L., and M.A.F. designed the research and analyzed the data, with substantial input from P.C., M.M., and J.C. M.S.D., C.J.S., and L.C.W.B. performed the experimental research. M.S.D., P.C., and J.M. performed the computational research. M.S.D. and M.A.F. wrote the manuscript.

ACKNOWLEDGMENTS

We are indebted to Marnix Medema (Max Planck Institute for Marine Microbiology), Amrita Pati (JGI), and members of the Fischbach Group for helpful advice. We thank Krishna Parsawar and Chad Nelson (University of Utah) and Jeff Johnson (UCSF) for help with mass spectrometry experiments and Mark Kelly (UCSF) for help with NMR experiments. This work was supported by a Howard Hughes Medical Institute Predoctoral Fellowship (P.C.), NIH grant TW006634 (R.G.L.), a Medical Research Program Grant from the W.M. Keck Foundation (M.A.F.), a Fellowship for Science and Engineering from the David and Lucile Packard Foundation (M.A.F.), DARPA award HR0011-12-C-0067 (M.A.F.), Program for Breakthrough Biomedical Research (M.A.F.), and NIH grants OD007290, AI101018, AI101722, and GM081879 (M.A.F.). M.A.F. is on the scientific advisory boards of NGM Biopharmaceuticals and Warp Drive Bio.

Received: April 17, 2014
Revised: May 30, 2014
Accepted: August 20, 2014
Published: September 11, 2014

REFERENCES

An, D., Oh, S.F., Olszak, T., Neves, J.F., Avci, F.Y., Erturk-Hasdemir, D., Lu, X., Zeissig, S., Blumberg, R.S., and Kasper, D.L. (2014). Sphingolipids from a symbiotic microbe regulate homeostasis of host intestinal natural killer T cells. Cell *156*, 123–133.

Baumann, H., Tzianabos, A.O., Brisson, J.R., Kasper, D.L., and Jennings, H.J. (1992). Structural elucidation of two capsular polysaccharides from one strain of Bacteroides fragilis using high-resolution NMR spectroscopy. Biochemistry *31*, 4081–4089.

Brzuszkiewicz, E., Weiner, J., Wollherr, A., Thürmer, A., Hüpeden, J., Lomholt, H.B., Kilian, M., Gottschalk, G., Daniel, R., Mollenkopf, H.J., et al. (2011). Comparative genomics and transcriptomics of Propionibacterium acnes. PLoS ONE *6*, e21581.

Chikindas, M.L., Novák, J., Driessen, A.J., Konings, W.N., Schilling, K.M., and Caufield, P.W. (1995). Mutacin II, a bactericidal antibiotic from Streptococcus mutans. Antimicrob. Agents Chemother. *39*, 2656–2660.

Cimermancic, P., Medema, M.H., Claesen, J., Kurita, K., Wieland Brown, L.C., Mavrommatis, K., Pati, A., Godfrey, P.A., Koehrsen, M., Clardy, J., et al. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell *158*, 412–421.

Davis, C., Kota, K., Baldhandapani, V., Gong, W., Abubucker, S., Becker, E., Martin, J., Wylie, K.M., Khetani, R., Hudson, M.E., et al. (2013). mBLAST: Keeping up with the sequencing explosion for (meta)genome analysis. J. Data Mining Genomics Proteomics *4*, 3.

Engelhardt, K., Degnes, K.F., and Zotchev, S.B. (2010). Isolation and characterization of the gene cluster for biosynthesis of the thiopeptide antibiotic TP-1161. Appl. Environ. Microbiol. *76*, 7093–7101.

Gajer, P., Brotman, R.M., Bai, G., Sakamoto, J., Schutte, U.M., Zhong, X., Koenig, S.S., Fu, L., Ma, Z.S., Zhou, X., et al. (2012). Temporal dynamics of the human vaginal microbiota. Sci. Transl. Med. *4*, 132ra152.

Gevers, D., Knight, R., Petrosino, J.F., Huang, K., McGuire, A.L., Birren, B.W., Nelson, K.E., White, O., Methé, B.A., and Huttenhower, C. (2012). The Human Microbiome Project: a community resource for the healthy human microbiome. PLoS Biol. *10*, e1001377.

Gevers, D., Kugathasan, S., Denson, L.A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., Schwager, E., Knights, D., Song, S.J., Yassour, M., et al. (2014). The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe *15*, 382–392.

Gonzalez, D.J., Lee, S.W., Hensler, M.E., Markley, A.L., Dahesh, S., Mitchell, D.A., Bandeira, N., Nizet, V., Dixon, J.E., and Dorrestein, P.C. (2010). Clostridiolysin S, a post-translationally modified biotoxin from Clostridium botulinum. J. Biol. Chem. *285*, 28220–28228.

Hsiao, E.Y., McBride, S.W., Hsien, S., Sharon, G., Hyde, E.R., McCue, T., Codelli, J.A., Chow, J., Reisman, S.E., Petrosino, J.F., et al. (2013). Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. Cell *155*, 1451–1463.

Human Microbiome Project Consortium (2012a). A framework for human microbiome research. Nature *486*, 215–221.

Human Microbiome Project Consortium (2012b). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214.

Just-Baringo, X., Albericio, F., and Álvarez, M. (2014). Thiopeptide antibiotics: retrospective and recent advances. Mar. Drugs *12*, 317–351.

Kelly, W.L., Pan, L., and Li, C. (2009). Thiostrepton biosynthesis: prototype for a new family of bacteriocins. J. Am. Chem. Soc. *131*, 4327–4334.

Kwan, J.C., Donia, M.S., Han, A.W., Hirose, E., Haygood, M.G., and Schmidt, E.W. (2012). Genome streamlining and chemical defense in a coral reef symbiosis. Proc. Natl. Acad. Sci. USA *109*, 20655–20660.

Kwon, H.C., Kauffman, C.A., Jensen, P.R., and Fenical, W. (2006). Marinomycins A-D, antitumor-antibiotics of a new structure class from a marine actinomycete of the recently discovered genus "marinispora". J. Am. Chem. Soc. *128*, 1622–1632.

LaMarche, M.J., Leeds, J.A., Amaral, A., Brewer, J.T., Bushell, S.M., Deng, G., Dewhurst, J.M., Ding, J., Dzink-Fox, J., Gamber, G., et al. (2012). Discovery of LFF571: an investigational agent for Clostridium difficile infection. J. Med. Chem. *55*, 2376–2387.

Larsbrink, J., Rogers, T.E., Hemsworth, G.R., McKee, L.S., Tauzin, A.S., Spadiut, O., Klinter, S., Pudlo, N.A., Urs, K., Koropatkin, N.M., et al. (2014). A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. Nature *506*, 498–502.

Lee, S.W., Mitchell, D.A., Markley, A.L., Hensler, M.E., Gonzalez, D., Wohlrab, A., Dorrestein, P.C., Nizet, V., and Dixon, J.E. (2008). Discovery of a widely distributed toxin biosynthetic gene cluster. Proc. Natl. Acad. Sci. USA *105*, 5879–5884.

Lee, J.H., Chae, J.P., Lee, J.Y., Lim, J.S., Kim, G.B., Ham, J.S., Chun, J., and Kang, D.K. (2011). Genome sequence of Lactobacillus johnsonii PF01, isolated from piglet feces. J. Bacteriol. *193*, 5030–5031.

Leimena, M.M., Ramiro-Garcia, J., Davids, M., van den Bogert, B., Smidt, H., Smid, E.J., Boekhorst, J., Zoetendal, E.G., Schaap, P.J., and Kleerebezem, M. (2013). A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. BMC Genomics *14*, 530.

Liao, R., Duan, L., Lei, C., Pan, H., Ding, Y., Zhang, Q., Chen, D., Shen, B., Yu, Y., and Liu, W. (2009). Thiopeptide biosynthesis featuring ribosomally synthesized precursor peptides and conserved posttranslational modifications. Chem. Biol. *16*, 141–147.

Long, S.R. (2001). Genes and signals in the rhizobium-legume symbiosis. Plant Physiol. *125*, 69–72.

Malcolmson, S.J., Young, T.S., Ruby, J.G., Skewes-Cox, P., and Walsh, C.T. (2013). The posttranslational modification cascade to the thiopeptide berninamycin generates linear forms and altered macrocyclic scaffolds. Proc. Natl. Acad. Sci. USA *110*, 8483–8488.

Mazmanian, S.K., Liu, C.H., Tzianabos, A.O., and Kasper, D.L. (2005). An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. Cell *122*, 107–118.

Mazmanian, S.K., Round, J.L., and Kasper, D.L. (2008). A microbial symbiosis factor prevents intestinal inflammatory disease. Nature *453*, 620–625.

McInerney, B.V., Gregson, R.P., Lacey, M.J., Akhurst, R.J., Lyons, G.R., Rhodes, S.H., Smith, D.R., Engelhardt, L.M., and White, A.H. (1991). Biologically active metabolites from Xenorhabdus spp., Part 1. Dithiolopyrrolone derivatives with antibiotic activity. J. Nat. Prod. *54*, 774–784.

Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E., and Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. *39* (Web Server issue), W339–W346.

Nelson, K.E., Weinstock, G.M., Highlander, S.K., Worley, K.C., Creasy, H.H., Wortman, J.R., Rusch, D.B., Mitreva, M., Sodergren, E., Chinwalla, A.T., et al.; Human Microbiome Jumpstart Reference Strains Consortium (2010). A catalog of reference genomes from the human microbiome. Science *328*, 994–999.

Nougayrède, J.P., Homburg, S., Taieb, F., Boury, M., Brzuszkiewicz, E., Gottschalk, G., Buchrieser, C., Hacker, J., Dobrindt, U., and Oswald, E. (2006). Escherichia coli induces DNA double-strand breaks in eukaryotic cells. Science *313*, 848–851.

Oh, D.C., Poulsen, M., Currie, C.R., and Clardy, J. (2009). Dentigerumycin: a bacterial mediator of an ant-fungus symbiosis. Nat. Chem. Biol. *5*, 391–393.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al.; MetaHIT Consortium (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65.

Ravel, J., Gajer, P., Abdo, Z., Schneider, G.M., Koenig, S.S., McCulle, S.L., Karlebach, S., Gorle, R., Russell, J., Tacket, C.O., et al. (2011). Vaginal microbiome of reproductive-age women. Proc. Natl. Acad. Sci. USA *108* (*Suppl 1*), 4680–4687.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. Nat. Methods *9*, 811–814.

Smillie, C.S., Smith, M.B., Friedman, J., Cordero, O.X., David, L.A., and Alm, E.J. (2011). Ecology drives a global network of gene exchange connecting the human microbiome. Nature *480*, 241–244.

Sonnenburg, E.D., Zheng, H., Joglekar, P., Higginbottom, S.K., Firbank, S.J., Bolam, D.N., and Sonnenburg, J.L. (2010). Specificity of polysaccharide use in intestinal bacteroides species determines diet-induced microbiota alterations. Cell *141*, 1241–1252.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., et al. (2009). A core gut microbiome in obese and lean twins. Nature *457*, 480–484.

Tzianabos, A.O., Onderdonk, A.B., Rosner, B., Cisneros, R.L., and Kasper, D.L. (1993). Structural features of polysaccharides that induce intra-abdominal abscesses. Science *262*, 416–419.

Wieland Brown, L.C., Acker, M.G., Clardy, J., Walsh, C.T., and Fischbach, M.A. (2009). Thirteen posttranslational modifications convert a 14-residue peptide into the antibiotic thiocillin. Proc. Natl. Acad. Sci. USA *106*, 2549–2553.

Wieland Brown, L.C., Penaranda, C., Kashyap, P.C., Williams, B.B., Clardy, J., Kronenberg, M., Sonnenburg, J.L., Comstock, L.E., Bluestone, J.A., and Fischbach, M.A. (2013). Production of α-galactosylceramide by a prominent member of the human gut microbiota. PLoS Biol. *11*, e1001610.

Wyatt, M.A., Wang, W., Roux, C.M., Beasley, F.C., Heinrichs, D.E., Dunman, P.M., and Magarvey, N.A. (2010). Staphylococcus aureus nonribosomal peptide secondary metabolites regulate virulence. Science *329*, 294–296.

Yoshimoto, S., Loo, T.M., Atarashi, K., Kanda, H., Sato, S., Oyadomari, S., Iwakura, Y., Oshima, K., Morita, H., Hattori, M., et al. (2013). Obesity-induced gut microbial metabolite promotes liver cancer through senescence secretome. Nature *499*, 97–101.

Yu, Y., Duan, L., Zhang, Q., Liao, R., Ding, Y., Pan, H., Wendt-Pienkowski, E., Tang, G., Shen, B., and Liu, W. (2009). Nosiheptide biosynthesis featuring a unique indole side ring formation on the characteristic thiopeptide framework. ACS Chem. Biol. *4*, 855–864.

Zou, Y., Yin, H., Kong, D., Deng, Z., and Lin, S. (2013). A trans-acting ketoreductase in biosynthesis of a symmetric polyketide dimer SIA7248. ChemBioChem *14*, 679–683.