


Spring 5-15-2017

The Effects of the Gut Microbiota on the Host Chromatin Landscape

Nicholas Semenkovich
Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/art_sci_etds

 Part of the [Biology Commons](#), [Genetics Commons](#), and the [Medicine and Health Sciences Commons](#)

Recommended Citation

Semenkovich, Nicholas, "The Effects of the Gut Microbiota on the Host Chromatin Landscape" (2017). *Arts & Sciences Electronic Theses and Dissertations*. 1145.
https://openscholarship.wustl.edu/art_sci_etds/1145

This Dissertation is brought to you for free and open access by the Arts & Sciences at Washington University Open Scholarship. It has been accepted for inclusion in Arts & Sciences Electronic Theses and Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Molecular Genetics and Genomics

Dissertation Examination Committee:

Jeffrey I. Gordon, Chair

Barak Cohen

Gautam Dantas

Todd Druley

Daniel Goldberg

Ting Wang

The Effects of the Gut Microbiota on the Host Chromatin Landscape

by

Nicholas Paul Semenkovich

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2017
St. Louis, Missouri

© 2017, Nicholas Paul Semenkovich

Table of Contents

List of Figures	iv
List of Tables.....	v
Acknowledgments.....	vi
Abstract	vii

Chapter 1

Introduction

Introduction.....	2
Epigenetic interactions between the gut microbiota and its host.....	2
Potential mechanisms by which the gut influences the host epigenome	4
Overview of the thesis	6
References.....	8
Figure legend	10
Figure	11

Chapter 2

The impact of the gut microbiota on enhancer accessibility in gut intraepithelial lymphocytes

Abstract.....	14
Significance.....	14
Results.....	16
Experimental Design and Approach to Analysis.	16
The Super-Enhancer Landscapes of IELs and Circulating T Cells, Independent of Colonization Status.	17
Contrasting the Chromatin Landscapes of TCR $\alpha\beta^+$ vs. TCR $\gamma\delta^+$ IELs, Independent of Colonization Status.	19
Chromatin Landscape in TCR $\alpha\beta^+$ IELs from CONV-R vs. GF Mice.....	20
Chromatin Landscape in TCR $\gamma\delta^+$ IELs from CONV-R vs. GF Mice.	22
Analysis of Enhancers in IELs Purified from CONV-D Mice.....	23
Transcription Factor Regulatory Circuitry.....	24

Discussion.....	26
References.....	29
Materials and methods.....	35
Acknowledgments.....	35
Figure legends.....	36
Figures.....	38
List of tables.....	40
Supplemental materials.....	44
SI Materials and methods.....	45
Supplemental figure legends.....	54
Supplemental figures.....	58

Chapter 3

Future Directions

Future Directions.....	66
References.....	69

Appendices

Appendix A.....	72
Appendix B.....	88

List of Figures

Figure 1.1:	An ATAC-seq analysis overview.....	11
Figure 2.1:	The effects of colonization on enhancer populations in IELs.....	38
Figure 2.2:	Analysis of colonization-associated TF circuitry.....	39
Figure 2.1.1:	FACS sorting strategies and reproducibility of ATAC-seq datasets.....	58
Figure 2.1.2:	Overview of the RIESLING pipeline.	59
Figure 2.1.3:	The enhancer landscape of TCR $\alpha\beta$ + and TCR $\gamma\delta$ + IELs and peripheral CD4+ and CD8+ T cells, independent of colonization status.	60
Figure 2.1.4:	Predicted GeNets communities from both IEL lineages.....	61
Figure 2.1.5:	Enhancer populations distinguish IEL lineages.....	62
Figure 2.1.6:	Comparison of gut microbial community structure in CONV-R and CONV-D mice.....	63
Figure 2.1.7:	Track graphs comparing normalized median ATAC-seq signals obtained from analysis of TCR $\alpha\beta$ +IELs purified from CONV-R and GF mice.	64

List of Tables

- Table 2.1: Dataset S1. Overview of samples and ATAC-seq datasets.
- Table 2.2: Dataset S2. Super-enhancers identified in all populations, independent of colonization status, rank ordered by signal intensity.
- Table 2.3: Dataset S3. Pathway analysis within the GeNets-predicted communities from the top IEL super-enhancers.
- Table 2.4: Dataset S4. Lineage-specific enhancers independent of colonization status.
- Table 2.5: Dataset S5. Pathway analysis of lineage-specific differentially accessible enhancers independent of colonization status.
- Table 2.6: Dataset S6. Pathway analysis of genes near predicted enhancers in IEL populations, independent of colonization status.
- Table 2.7: Dataset S7. Pathway analysis of genes near predicted enhancers in CD4+ and CD8+ T-cell populations, independent of colonization status.
- Table 2.8: Dataset S8. Differentially accessible regions of chromatin identified comparing GF to CONV-R mice.
- Table 2.9: Dataset S9. Pathway analysis within the GeNets-predicted communities from the genes adjacent to differentially accessible enhancers identified in IELs purified from CONV-R vs. GF mice.
- Table 2.10: Dataset S10. GREAT analysis applied to differentially accessible enhancers and their adjacent genes identified in IELs from CONV-R vs. GF mice.
- Table 2.11: Dataset S11. Enhancers classified as microbiota-responsive, based on whether their accessibility changes in the GF vs. CONV-D and CONV-R groups.
- Table 2.12: Dataset S12. Analysis of TF circuitry in GF and CONV-R IEL enhancers.

Acknowledgments

Spend enough time with Jeff, and you'll hear one of his signature phrases, including my personal favorite: "If you want to go fast, go alone. If you want to go far, go together." (Perhaps only the latter half is true in the sciences.) In my own training, I owe many people thanks for their insights and thoughtfulness. My parents Janice and Clay, my amazing sister Katherine, and my brilliant wife Tara — have been endlessly supportive during my research & pre-clinical training.

None of this work would've been possible without the generous guidance and mentorship from Jeff and this lab. Joe Planer spent many late nights with me, developing strategies for cell isolation, staining, and sorting. Andy Kau and Philip Ahern always offered great FACS guidance. Jess Hoisington-Lopez was a wizard of sequencing technologies. Outside of Jeff's lab, my good friend Charles Lin provided early insights into ATAC-seq and enhancer dynamics.

My thesis committee members have provided great scientific discussion throughout my projects. I'd like to especially thank a former member, Bob Heuckeroth, for his feedback and provocative questions. This work was also advanced by the continuous support from the Medical Scientist Training Program. I owe special gratitude to Dr. Wayne Yokoyama, Dr. Vicky Fraser, & Brian Sullivan, for their knowledge of information technology.

None of this work would've been possible without the support I received throughout my graduate years by a T32 pre-doctoral training fellowship from the National Human Genome Research Institute (T32HG000045/HG/NHGRI), the MSTP program (with support from NIH grant GM007200) and from the Bill and Melinda Gates Foundation.

Nicholas P. Semenkovich

Washington University in St. Louis

May 2017

Abstract

The Effects of the Gut Microbiota on the Host Chromatin Landscape

by

Nicholas Paul Semenkovich

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Genetics and Genomics

Washington University in St. Louis, 2017

Professor Jeffrey I. Gordon, Chair

The human gut microbiota is home to tens of trillions of microbes belonging to all three domains of life. The structure and expressed functions of this community have myriad effects on host physiology, metabolism, and immune function. My studies focused on a facet of host-microbial interactions and mutualism that has not been explored to a significant degree in part because of the absence of suitable tools: namely, if, when, and how the gut microbiota produces durable effects on host biology through its impact on the epigenome. To address this area, I turned to gnotobiotic mice and developed a variety of experimental, methodological, and computational approaches to characterize the chromatin landscape of various host cell populations. I selected small populations of T cells likely to be exposed to the gut microbiota and its products, principally TCR $\alpha\beta$ ⁺ and TCR $\gamma\delta$ ⁺ small intestinal intraepithelial lymphocytes (IELs). I also chose to study circulating CD4⁺ and CD8⁺ T cells, with the advantage that these populations can be isolated from donors without using highly invasive techniques. I designed a series of approaches to enrich, isolate, and purify each of these cell populations from single animals, and to subsequently compare their chromatin landscape within and across mice using a recently described Assay for Transposase-Accessible Chromatin with high throughput sequencing (ATAC-seq), focusing on enhancer and super-enhancer loci within the mouse genome. These analyses revealed a conserved set of super-enhancer loci between $\alpha\beta$ and $\gamma\delta$ IELs, including super-enhancers near genes responsible for phospholipid binding and T cell

receptor signaling. In comparing C57BL/6J male mice reared under germ-free (GF) conditions to age- and sex-matched conventionally raised (CONV-R) mice (i.e., animals that acquired microbes from their environment beginning at birth), I was able to directly assess the impact of colonization on chromatin ‘state’ in these different purified cell populations and to identify colonization-dependent effects in IELs on enhancers associated with genes involved in a number of metabolic and signaling pathways. I then compared the results to GF mice that had been colonized following the end of the weaning period with an intact cecal microbiota from a CONV-R C57BL/6J donor. The resulting conventionalized (CONV-D) animals allowed me to identify modifications to host chromatin landscape that are ‘induced’ following the suckling-weaning transition and, in doing so, ascertain whether there were developmental windows that could constrain the durable effects of the microbiota on chromatin accessibility. In doing so, I observed changes in chromatin accessibility with colonization that may reveal a functional maturation of IEL populations that is related to the timing of exposure to the microbiota during postnatal development. My thesis project involved development of an elaborate, multi-faceted computational pipeline for the analysis of these novel, large datasets, including the prediction and characterization of putative enhancers and super-enhancers, proximally associated genes, and metabolic pathways influenced by those elements. As a whole, this work defines the impact of gut microbial colonization on host chromatin landscape and provides an analytical toolkit for further studies.

Chapter 1

Introduction

Introduction

The gut microbiota and host can be thought of as a ‘holobiont,’ where both host and microbes coexist and interact in a complex range of relationships, ranging from symbiotic and mutually beneficial (e.g., the biosynthesis of vitamins and co-factors) to harmful or parasitic. This spectrum of interactions has been present as both host and bacteria have evolved over millennia, and has resulted in numerous systems — largely immunologic — that regulate and influence these relationships. While some of these regulatory systems have been described over the past few decades, one possibility that remains largely unexplored is that host-microbiota interactions are influenced by epigenetic factors — that is, the microbiota and its products may alter the host chromatin landscape to subsequently potentiate and perpetuate phenotypes. Some critical questions in this field include: Do periods of nutritional unavailability and dysbiosis potentiate longer-term impacts in the host via epigenetic mechanisms? Are host-chromatin modifications durable (does the host keep an epigenetic ‘memory’ of prior exposure to a given microbial community configuration or configurations)? Is there a critical temporal window of exposure, where microbial signals must be received by the host before they are able to alter the chromatin landscape in a permanent way? These key questions are what drove my exploration and interest throughout this thesis.

Epigenetic interactions between the gut microbiota and its host

Several studies have presented a range of evidence for epigenetic interactions between the gut microbiota and its host. There are a multitude of analytical techniques available to characterize the epigenome (reviewed in Winter et al. 2015). In light of their strengths and weaknesses, it is important to consider each report in the context of the methods applied.

A number of studies have looked at changes in DNA methylation in host tissues, and its association with microbiota composition. Given high input material requirements of many methylation-based techniques (Tsompana & Buck 2014), some of these studies necessarily focus on heterogeneous host cell populations, where a large amount of cellular material is available — but mixed input populations can complicate interpretation of results. Some of the earliest studies of

microbiota-epigenetic interactions focused on the immune system, using microarray-based assays for methylation. Kellermayer et al. (2011), interrogated the methylation patterns of the colonic mucosa in wild type C57BL/6 mice, and in a Tlr2(-/-) mouse model (of interest given its associations with IBD and intestinal barrier function (Cario 2008)), using a custom microarray. They determined that, compared to WT mice, Tlr2(-/-) mice showed significant methylation differences around a number of genes associated with inflammatory pathways (e.g., Alanyl Aminopeptidase [Anpep], and annexin A8 [Anxa8]); these changes were associated with a decrease in the relative abundance of Firmicutes and increases in the representation of members of Proteobacteria, Bacteroidetes, and Actinobacteria. However, their study was not designed to separate microbiota-induced changes from the effects of the Tlr2 knockout alone.

In another example of a methylation-based approach, (Kumar et al. 2014) looked at whole blood methylation patterns across pregnant women, segregated based on whether their fecal microbiota appeared to be dominated by Bacteroidetes and Proteobacteria, versus Firmicutes. In women whose microbiota harbored Firmicutes as the dominant phylum (as compared to a Bacteroidetes/Proteobacteria dominated group), they identified differential methylation patterns around genes associated with increased risk for obesity and the metabolic syndrome. Other studies have focused on dietary contributions to host methylation, including Schaible et al. 2011, where female mice were placed on a methyl-donor diet (a diet supplemented with folic acid, B12, betaine, and choline) during pregnancy. The offspring of these conventionally-raised mice were observed to have alterations in DNA methylation patterns (associated with a variety of loci, including *Ppara*), an increased susceptibility to dextran sodium sulfate-induced colitis, and structural changes in the microbiota (including an increased proportional representation of *Clostridia* and a decreased representation of *Lactobacilli*); these differences did not last beyond 90 days.

Numerous associational studies have also been performed in humans, highlighting epigenetic and microbial factors in a range of (generally gastrointestinal) diseases. Based on observations that *Fusobacterium* was detected frequently in biopsies and excised tumors of patients with colorectal cancer (Castellarin et al. 2012), Tahara et al. (2014) characterized intestinal methylation

differences in patients with colorectal cancer, stratified by the presence of *Fusobacterium* in tumor sections. They discovered that tumors enriched for *Fusobacterium* had CpG island methylator phenotypes (CIMP), and an increase in microsatellite instability.

A small number of recent studies have tried to more directly define links between the microbiota and epigenetic changes, leveraging sterile ‘germ free’ (GF) mice while profiling epigenetic features. Obata et al. (2014) identified a specific DNA methylation adapter (*Uhrf1*) that was upregulated in intestinal Treg cells upon conventionalization of GF mice. Upon deletion of this locus in Tregs hypomethylation was observed in the promoter region of *Cdkn1a*, resulting in cell-cycle arrest and severe colitis. Camp et al. (2014) performed ChIP-seq and RNA-seq on intestinal epithelial cells (IECs), in GF and CONV-R mice. Their study determined that transcriptional changes were not associated with changes in ileal or colonic chromatin accessibility within epithelial cell populations. Yu et al. (2015) used whole-genome bisulfite sequencing to characterize methylation differences in GF versus conventionally raised (CONV-R) C57BL/6J mice. Focusing on *Lgr5*⁺ intestinal stem cells (ISCs), the group identified an increase in ISC methylation at CpG islands during the suckling period that were correlated with changes in transcription of certain glycosylation genes. They also performed a time-series analysis of GF vs CONV-R mice, which identified a reduction in methylation at numerous sites in GF mice and that these alterations were correlated with gene expression changes in at least two loci (*B4galnt1* and *Phosphol1*). They observed a rescue of this phenotype (an increase in methylation and associated changes in gene expression) in small cohort of GF mice (n=3) that were subsequently colonized at 25 days of life, and sacrificed at 100 days (a 75-day colonization period).

Potential mechanisms by which the gut influences the host epigenome

In part through the studies above, a number of mechanisms have been proposed by which the microbiota may influence the host epigenome (Alenghat & Artis 2014). One recurring theme in the literature is that elaborated microbial products and metabolites can directly influence the host epigenome. These metabolites, their influences, and proposed mechanisms, range from straight-

forward (e.g., methyl donors that provide methyl groups for epigenetic modifications), to complex microbial processing of food components that subsequently signal host cells.

The most fundamental proposed mechanism for gut-host epigenetic signaling is through the simple availability of methyl donors. A variety of gut-derived products serve as methyl donors or cofactors for base-pair methylation and the remodeling of chromatin, including folate and a variety of B vitamins (B2, B6, and B12), methionine, choline, and betaine (Anderson et al. 2012). These are all components of one-carbon metabolism pathways that contribute to the availability of S-adenosylmethionine (SAM), and transporters for many of these methyl donors are expressed within the colonic epithelium (Said et al. 2000; Visentin et al. 2014; Albersen et al. 2013). SAM itself is utilized by methyltransferases in the maintenance and establishment of DNA methylation. A large body of work has observed associations between dietary intake of these methyl donors and host DNA methylation patterns, including the studies described above (reviewed in Anderson et al. 2012). Notably, the microbiota can also produce or modify some of these methyl donors directly, including folate and a variety of B vitamins (Jeffery & O'Toole 2013).

Beyond direct contributions to the methyl donor pool, diet and the microbiota may contribute to more nuanced epigenetic signaling through metabolites, including short chain fatty acids (SCFAs). SCFAs are elaborated by commensal bacteria through their anaerobic fermentation of plant polysaccharides, and include propionate, acetate, and butyrate. Beyond being used as critical energy source by the host (Donohoe et al. 2011), these metabolites can signal through GPCRs, and act as histone deacetylase inhibitors (HDACs) (Vinolo et al. 2011). Through these mechanisms, commensal bacteria may influence host phenotypes: Alenghat et al. (2013) demonstrated that the absence of HDAC3 in IECs resulted in broad dysregulation in gene expression and increased sensitivity in a DSS-induced colitis model, a phenotype that was abrogated in GF mice. Additionally, a number of models have identified specific signaling mechanisms with immunologic impacts, for example, Arpaia et al. (2013) demonstrated one route through which butyrate enhanced extrathymic differentiation of Tregs.

Overview of the thesis

Based on this background, I set out to explore the potential contributions of the microbiota to the host chromatin landscape, using a mouse model. I chose to analyze this in the most controllable, reproducible approach, using sterile ('germ-free') mice and comparing them to their conventionally raised (CONV-R) counterparts. To better understand potential 'critical window' effects, I also included a 'conventionalized' arm, consisting of mice that were raised germ-free, and then subsequently colonized ('conventionalized', CONV-D) following the three-week suckling-weaning transition.

I turned to a nascent technology when this thesis began to profile chromatin accessibility across multiple cell populations in these mice. ATAC-seq (Assay for Transposase-Accessible Chromatin using Sequencing (Buenrostro et al. 2013)) is a transposase-based technique similar to DNase-seq. In ATAC-seq, the Tn5 transposase is pre-loaded with Illumina sequencing adapters and incubated with isolated nuclei. The transposase can then insert these sequencing adapters into regions of chromatin that are 'open' (areas of uncondensed, accessible chromatin). The resulting fragments of DNA can be amplified and sequenced, which reveals a genome-wide map of the open chromatin landscape (see Figure 1).

I focused on cell lineages potentially exposed to the microbiota and its products (and cells that could drive responses to microbial colonization) and selected four cell populations that met those criteria: circulating peripheral lymphocytes (both CD4+ and CD8+ T-cells), and intestinal intraepithelial lymphocytes (TCR $\alpha\beta$ + and TCR $\gamma\delta$ + T-cells).

I developed a protocol to rapidly and simultaneously isolate these populations from single mice, to ensure I could generate data without pooling samples across mice. Within these cell populations, I focused my analyses on enhancers — regions of chromatin that can drive the expression of genes (including non-classical elements like lncRNAs) via the recruitment of activating factors, and characterized the enhancer landscape of these cells for both colonization dependent and independent features.

The computational challenges I faced were great: assuming ATAC-seq data could be generated, no pipeline existed for its analysis and processing. Additionally, existing pipelines for related types of data (e.g., DNase-seq) tended to be fragile and poorly documented, which largely precluded their adaptation to ATAC-seq data. Additionally, the computational resources for super-enhancer prediction were designed for different data types (ChIP-seq) and differential accessibility analyses tended to be performed on an ad-hoc basis. I developed a flexible, portable, and documented computational pipeline for the analysis of ATAC-seq data, including its pre-processing, enhancer and super-enhancer characterization, and differential accessibility prediction, with the goal of establishing a permanent toolkit for other researchers, as the use of ATAC-seq and DNase-seq grows.

References

- Albersen, M. et al., 2013. The intestine plays a substantial role in human vitamin B6 metabolism: a Caco-2 cell model. *PloS One*, 8(1), p.e54113.
- Alenghat, T. et al., 2013. Histone deacetylase 3 coordinates commensal-bacteria-dependent intestinal homeostasis. *Nature*, 504(7478), pp.153–7.
- Alenghat, T. & Artis, D., 2014. Epigenomic regulation of host-microbiota interactions. *Trends in Immunology*, 35(11), pp.518–25.
- Anderson, O.S., Sant, K.E. & Dolinoy, D.C., 2012. Nutrition and epigenetics: an interplay of dietary methyl donors, one-carbon metabolism and DNA methylation. *The Journal of Nutritional Biochemistry*, 23(8), pp.853–9.
- Arpaia, N. et al., 2013. Metabolites produced by commensal bacteria promote peripheral regulatory T-cell generation. *Nature*, 504(7480), pp.451–5.
- Buenrostro, J.D. et al., 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12), pp.1213–8.
- Camp, J.G. et al., 2014. Microbiota modulate transcription in the intestinal epithelium without remodeling the accessible chromatin landscape. *Genome Research*, 24(9), pp.1504–16.
- Cario, E., 2008. Barrier-protective function of intestinal epithelial Toll-like receptor 2. *Mucosal Immunology*, 1 Suppl 1, pp.S62–6.
- Castellarin, M. et al., 2012. *Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma. *Genome Research*, 22(2), pp.299–306.
- Donohoe, D.R. et al., 2011. The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon. *Cell Metabolism*, 13(5), pp.517–26.
- Jeffery, I.B. & O’Toole, P.W., 2013. Diet-microbiota interactions and their implications for healthy living. *Nutrients*, 5(1), pp.234–52.

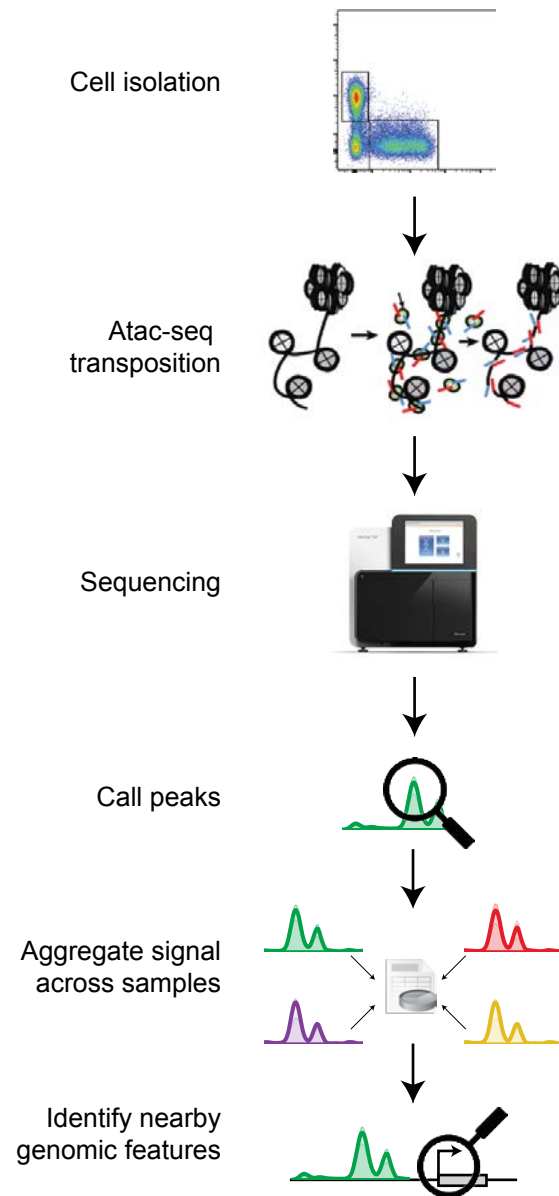
- Kellermayer, R. et al., 2011. Colonic mucosal DNA methylation, immune response, and microbiome patterns in Toll-like receptor 2-knockout mice. *FASEB*, 25(5), pp.1449–60.
- Kumar, H. et al., 2014. Gut microbiota as an epigenetic regulator: pilot study based on whole-genome methylation analysis. *mBio*, 5(6), pp.e02113–14.
- Obata, Y. et al., 2014. The epigenetic regulator Uhrf1 facilitates the proliferation and maturation of colonic regulatory T cells. *Nature Immunology*, 15(6), pp.571–9.
- Said, H.M. et al., 2000. Riboflavin uptake by human-derived colonic epithelial NCM460 cells. *American Journal of Physiology, Cell Physiology*, 278(2), pp.C270–6.
- Schaible, T.D. et al., 2011. Maternal methyl-donor supplementation induces prolonged murine offspring colitis susceptibility in association with mucosal epigenetic and microbiomic changes. *Human Molecular Genetics*, 20(9), pp.1687–96.
- Tahara, T. et al., 2014. Fusobacterium in colonic flora and molecular features of colorectal carcinoma. *Cancer Research*, 74(5), pp.1311–8.
- Tsompana, M. & Buck, M.J., 2014. Chromatin accessibility: a window into the genome. *Epigenetics & Chromatin*, 7(1), p.33.
- Vinolo, M.A.R. et al., 2011. Regulation of Inflammation by Short Chain Fatty Acids. *Nutrients*, 3(12), pp.858–876.
- Visentin, M. et al., 2014. The intestinal absorption of folates. *Annual Review of Physiology*, 76, pp.251–74.
- Winter, D.R., Jung, S. & Amit, I., 2015. Making the case for chromatin profiling: a new tool to investigate the immune-regulatory landscape. *Nature Reviews Immunology*, 15(9), pp.585–94.
- Yu, D.-H. et al., 2015. Postnatal epigenetic regulation of intestinal stem cells requires DNA methylation and is guided by the microbiome. *Genome Biology*, 16(1), p.211.

Figure legend

Figure 1. A broad overview of an ATAC-seq analysis is shown, where cells are first isolated and nuclei are prepared. Then, nuclei are incubated with the Tn5 transposase, and subsequently sequenced. Next, areas of high signal compared to background (‘peaks’) are called, and aggregated across samples, followed by the identification of nearby genomic features. The transposase graphic is adapted from Buenrostro et al. 2013.

Figure

Figure 1.



Chapter 2

The impact of the gut microbiota on enhancer accessibility in gut intraepithelial lymphocytes

**The impact of the gut microbiota on enhancer accessibility
in gut intraepithelial lymphocytes**

Nicholas P. Semenkovich^{1,2}, Joseph D. Planer^{1,2}, Philip P. Ahern^{1,2}, Nicholas W. Griffin^{1,2}, Charles Y. Lin³, Jeffrey I. Gordon^{1,2*}

¹Center for Genome Sciences and Systems Biology, ²Center for Gut Microbiome and Nutrition Research, Washington University School of Medicine, St. Louis, MO 63110, USA

³Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA

*To whom correspondence should be addressed: e-mail jgordon@wustl.edu

Abstract

The gut microbiota impacts many aspects of host biology including immune function. One hypothesis is that microbial communities induce epigenetic changes with accompanying alterations in chromatin accessibility, providing a mechanism that allows a community to have sustained host effects even in the face of its structural or functional variation. We used Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) to define chromatin accessibility in predicted enhancer regions of intestinal $\alpha\beta^+$ and $\gamma\delta^+$ intraepithelial lymphocytes purified from germ-free mice, their conventionally raised (CONV-R) counterparts, and mice reared germ free and then colonized with CONV-R gut microbiota at the end of the suckling–weaning transition. Characterizing genes adjacent to traditional enhancers and super-enhancers revealed signaling networks, metabolic pathways, and enhancer-associated transcription factors affected by the microbiota. Our results support the notion that epigenetic modifications help define microbial community-affiliated functional features of host immune cell lineages.

Significance

Comparing germ-free mice with those colonized at birth or later provides a way to determine how gut microbial community exposure affects the chromatin landscape of cells along the gut or at remote sites, ascertain how alterations in chromatin accessibility are correlated with functional features of different lineages, and determine whether there is a critical window of exposure when microbial signals must be received to alter the landscape durably. Genome-wide analysis of chromatin accessibility in intraepithelial lymphocytes and circulating T cells purified from gnotobiotic mice revealed enhancers and flanking genes involved in signaling and metabolic pathways that are sensitive to colonization status. Colonization does not fundamentally alter lineage-specific cis-regulatory landscapes but induces quantitative changes in the accessibility of preestablished enhancer elements.

Human gut microbial communities interact with their hosts in ways that affect physiology, metabolism, and immune function. The underlying mechanisms are subjects of intense investigation. One hypothesis is that the gut microbiota induces epigenetic changes in host cells, altering chromatin accessibility and the subsequent poising of genes for expression. Chromatin modifications have the potential to outlast a given microbiota configuration and endow the host with a “memory” of microbial exposure (1). This hypothesis has been investigated in a limited number of studies. DNase-seq applied to intestinal epithelial cells identified few differences in the chromatin landscape between germ-free (GF) and conventionally raised (CONVR) mice (2). Another study, using bisulfite sequencing, analyzed pooled populations of Lgr5⁺ intestinal stem cells and found colonization dependent methylation of numerous CpG loci (3). More recently, iChIP-IVT (Indexing first Chromatin Immunoprecipitation-In Vitro Transcription) was used to assess chromatin state in all three subsets of innate lymphoid cells harvested from CONV-R and antibiotic-treated mice. The results identified thousands of H3K4me2 regions that were sensitive to antibiotic treatment, with a number of these regions losing their specificity for different innate lymphoid cell subsets (4).

In the present study, we use ATAC-seq (Assay for Transposase Accessible Chromatin with high-throughput sequencing) to examine this hypothesis. ATAC-seq uses a Tn5 transposase preloaded with sequencing adaptors; this approach enables efficient in vitro transposition of the adaptors into regions of accessible (“open”) chromatin (5) which correlate with primed or active enhancers. ATAC-seq is agnostic to the type of epigenetic modification and, importantly, allows analyses to be performed on small populations of cells. Our focus on the effects of the gut microbiota on enhancers was based on the supposition that identifying these regions, their neighboring genes, and the signaling and metabolic pathways to which these genes belong would provide mechanistic insights about how the gut microbial community influences the biological properties of immune cell populations. We targeted two populations that are in intimate association with the gut microbiota and its products: T-cell receptor (TCR) $\alpha\beta^+$ and TCR $\gamma\delta^+$ intraepithelial lymphocytes (IELs). IELs participate in immune surveillance, immune tolerance, wound repair, mainte-

nance of gut barrier function, and protection from infectious agents (6). Given these roles, it is not surprising that considerable attention has been directed to how the microbiota shapes the IEL compartment (7, 8), although the relationship between the microbiota and IELs is less well characterized than other components of the immune system. We also examined peripheral CD4+ and CD8+ T cells, reasoning that they may have transient exposures to products of the gut microbiota, can be sampled more readily than IELs, and would allow us to delineate extraintestinal epigenetic effects of the microbiota.

Results

Experimental Design and Approach to Analysis.

Three groups of 8- to 10-wk-old male C57BL/6 mice were studied: a group reared under sterile conditions (GF mice), a CONV-R group exposed to microbes present in their environment from the time of birth, and a group initially maintained GF until the suckling–weaning transition at 3 wk of age and then colonized with a cecal microbiota from a CONV-R donor (conventionalized mice, CONV-D) (n = 8 mice per treatment group). The relatively low numbers of cells required to perform ATAC-seq allowed us to study different cell populations from single animals without needing to pool samples (Fig. S1A) and to identify predicted enhancers within a given cell lineage, including those whose accessibility is altered by microbial exposure during postnatal development.

To analyze the large datasets generated (Dataset S1), we developed an open-source, Python-based computational pipeline named RIESLING (Rapid Identification of Enhancers Linked to Nearby Genes). RIESLING was inspired by the ROSE (Rank Ordering of Super-Enhancers) software designed for ChIP-seq (9, 10) and developed as a toolkit to standardize analysis of datasets generated from ATAC-seq. RIESLING is designed to enable de novo prediction of putative enhancers and characterization of their differential accessibility (Fig. S2). Briefly, sequencing data generated following ATAC-seq transposition are used to call “peaks” (open regions of chromatin with a read density statistically higher than background) identified using a Poisson model adapted

from standard approaches used for ChIP-seq and DNase-seq analyses. To predict putative enhancers, peaks and their underlying raw read depth are aggregated based on one of three approaches: (i) unstitched (peaks and their underlying raw read depth); (ii) fixed stitching, in which peaks within a certain distance of each other [typically a 12.5-kb window (10)] are merged together to represent one broad region of interest; or (iii) dynamic stitching, in which a window is slowly increased from 0 kb (unstitched) to a maximum of 15 kb, in adjustable increments of 500 bp, and the number of enhancer clusters identified per window is recorded. The rate of growth (the first derivative) of this growing window is computed, and the point at which the rate of growth of called enhancers decreases (the inflection point) is used to merge nearby peaks together. (See SI Materials and Methods for further discussion of the rationale for using these three approaches in various facets of our analyses).

The Super-Enhancer Landscapes of IELs and Circulating T Cells, Independent of Colonization Status.

We identified putative enhancers shared by the two types of purified IEL populations ($\alpha\beta^+$ and $\gamma\delta^+$ T cells) and those shared by the two types of purified peripheral lymphocyte populations ($CD4^+$ and $CD8^+$ T cells), independent of colonization status (GF \cup CONV-R \cup CONV-D; Fig. S1B). Data generated from these purified cell populations were highly reproducible (Fig. S1C). The different enhancer populations were reflected in a hierarchical clustering that clearly stratified by cell type (Fig. S3A). We divided these putative enhancers into two groups: a set of “traditional” enhancers, and a subpopulation of super-enhancers. Super-enhancers are a subclass of enhancers that are stronger (more occupied by transcriptional activators) and can span longer regions of DNA than traditional enhancers (9, 10). Super-enhancers can drive cell lineagespecific functions and play critical roles in disease pathogenesis (11, 12). We identified super-enhancers by first combining adjacent peaks within a 12.5-kb window (the fixed stitching approach described above), ranking and graphing these merged enhancers by their normalized read depth (ATAC-seq signal), and discriminating traditional enhancers from super-enhancers by determining where the slope of

a tangent line fitted to their graph reached one (10). The resulting graph fit an exponential curve in which the largest ATAC-seq signal occurred at a small minority of enhancers (Fig. S3B and Dataset S2).

Super-enhancers, although comprising <5% of the aggregate IEL enhancer population (4.4% in TCR $\alpha\beta^+$ IELs and 4.9% in TCR $\gamma\delta^+$ IELs), accounted for a disproportionate share of the aggregate ATAC-seq signal (35%) compared with the traditional enhancer group; super-enhancers also spanned broader regions of the genome than did traditional enhancers [$29,447 \pm 440$ vs. $3,988 \pm 31$ bp (mean \pm SEM)]. These patterns fit well with other studies, which show similar distributions of signal and enhancer width when applying ATAC-seq and complementary techniques, including p300 and Mediator ChIP-seq, in different tissues (9, 12, 13). The colonization-independent super-enhancer populations in TCR $\alpha\beta^+$ and TCR $\gamma\delta^+$ IELs were highly similar, with a Jaccard similarity index of 83% (Fig. S3A). The top-ranked super-enhancer in both IEL populations is adjacent to *Prkch* (Fig. S3B), a nonclassical Ca^{2+} -insensitive η -type PKC that promotes T-cell activation and cytoskeletal polarization (14, 15).

Applying the same analytic approach to peripheral circulating T cells revealed that super-enhancers comprised 5.7% of the population but 37.7% of the aggregate ATAC-seq signal [values comparable to those reported in mice and humans (9, 12, 13, 16)]. The super-enhancer associated with *Bach2* (basic leucine zipper transcription factor 2) produced the top-ranked ATAC-seq signal in CD4⁺ T cells (Fig. S3 C and D and Dataset S2). This finding was notable, given a previous report that used p300 ChIP-seq to identify *Bach2* as the strongest super-enhancer in CD4⁺ T cells, and as a critical negative regulator of pathogenic T-cell effector differentiation and as a positive regulator of regulatory T-cell induction (12, 17).

We performed a network analysis of genes located in the proximity of the top 250 super-enhancers (ranked by ATAC-seq signal) from the combined population of TCR $\alpha\beta^+$ and TCR $\gamma\delta^+$ IELs, independent of colonization status, using GeNets (SI Materials and Methods). GeNets defines “communities” as sets of genes with a high degree of connectivity based on a random for-

est classifier trained on datasets of established protein–protein interactions. This network analysis clustered super-enhancer–associated genes into eight distinct communities that display functions that are lineage-defining components of IELs (e.g., TCR signaling; see Dataset S3 for a complete list of components and their assigned functions). GeNets also identified significant connections among these eight communities (Fig. S4).

Contrasting the Chromatin Landscapes of TCR $\alpha\beta^+$ vs. TCR $\gamma\delta^+$ IELs, Independent of Colonization Status.

We applied a negative-binomial model to the dataset of enhancers described above to identify lineage-specific features of the TCR $\alpha\beta^+$ and TCR $\gamma\delta^+$ chromatin landscapes, independent of colonization status. We were wary of applying this model only to the collection of peaks stitched within the 12.5-kb window initially applied to the datasets: Although appropriate for defining strong super-enhancers, this fixed window could, potentially, drown out signals from differentially accessible but narrower enhancer regions. Understanding that the width of enhancers can vary across different cell types and tissues, we used the dynamic stitching approach to merge peaks, with the goal of adapting to variable enhancer widths (Fig. S5A). Applying a binomial model to these dynamically stitched peaks yielded 4,704 predicted enhancers demonstrating differential accessibility (13.2% of all dynamically stitched IEL enhancers at $P < 0.05$ by Benjamini–Hochberg-adjusted Wald test). Filtering to a more stringent cutoff ($P < 5 \times 10^{-5}$ by Benjamini–Hochberg-adjusted Wald test) yielded 598 predicted enhancers with increased accessibility in TCR $\alpha\beta^+$ compared with TCR $\gamma\delta^+$ IELs and 295 with increased accessibility in TCR $\gamma\delta^+$ vs. TCR $\alpha\beta^+$ IELs (Fig. S5B). (For all differentially accessible enhancers observed at $P < 0.05$, see Dataset S4 A and B. For an analysis in peripheral T cells, see Dataset S4 C and D).

We performed a pathway analysis on the collection of statistically significant enhancers using GREAT (Genomic Regions Enrichment of Annotations Tool) (18), first focusing on regions with increased accessibility in the TCR $\alpha\beta^+$ IEL population and comparing them to the set of all enhancers in both IEL lineages independent of colonization status. We used GREAT because it can

consider multiple genes in proximity to a given enhancer when calculating pathway enrichment (as opposed to a GeNets-based analysis, which incorporates only the most proximal gene). The results revealed significant enrichment ($P < 0.05$; hypergeometric test) of enhancer-adjacent genes belonging to numerous pathways in the MSigDB (Molecular Signatures Database) encompassing Reactome, KEGG (Kyoto Encyclopedia of Genes and Genomes), other metabolic databases, and GO (Gene Ontology). These pathways are rank ordered based on their statistical significance in Dataset S5A (see SI Results for further discussion).

Chromatin Landscape in TCR $\alpha\beta^+$ IELs from CONV-R vs. GF Mice.

Comparing the TCR $\alpha\beta^+$ IEL chromatin landscape in CONV-R vs. GF animals using our dynamic stitching approach and binomial model yielded 7,137 predicted enhancers with colonization-associated differences in their accessibility ($P < 0.05$, Benjamini–Hochberg-adjusted Wald test) with 1,362 surviving more stringent filtering ($P < 5 \times 10^{-5}$). Of these, 497 were significantly more accessible in TCR $\alpha\beta^+$ IELs purified from CONV-R mice than in TCR $\alpha\beta^+$ IELs from GF animals, and 865 enhancers met this more stringent cutoff for increased accessibility in GF animals (Fig. S5 C and E and Dataset S8A). Fig. 1A shows that IEL enhancers cluster first by colonization history and then by cell type, in contrast to circulating peripheral T cells in which cell type has a larger discriminatory effect.

We used GeNets to identify potential interactions between genes located near colonization-associated enhancers. The resulting GeNets communities were enriched for (i) members of the PDGF signaling pathway whose associated enhancers were more accessible in TCR $\alpha\beta^+$ IELs from GF mice (community members *Grb10*, *Pik3cb*, *Pik3r2*, *Ptpn11*, and *Ptprj*; see Dataset S9A) and (ii) members of JAK-STAT [community members *Cblb* (Cbl proto-oncogene B), *Il12rb2* (IL-12 receptor β 2), *Il23r* (IL-23 receptor), *Il7r* (IL-7 receptor), and *Prlr* (Prolactin receptor)] and TCR (community members *Cblb*, *Nck2*, and *Prkcc*) signaling, and leukocyte transendothelial migration [*Cxcr4* (C-X-C chemokine receptor type 4), *Itga4* (Integrin subunit alpha 4), and *Ptk2* (protein

tyrosine kinase 2)] pathways whose associated enhancers were more accessible in CONV-R TCR $\alpha\beta^+$ IELs (Fig. 1B and Dataset S9B).

A GREAT-based analysis of all enhancers with statistically significant ($P < 0.05$) increased accessibility in CONV-R vs. GF TCR $\alpha\beta^+$ IELs disclosed significant overrepresentation of pathways involved in bile acid metabolism, propionate metabolism, and sulfur amino acid metabolism (Dataset S10A). Members of the microbiota direct a complex set of metabolic transformations of bile acids (19) which can influence intestinal inflammation through multiple mechanisms. Propionate, a product of microbial fermentation of dietary glycans, is a histone-deacetylase inhibitor (20) and has immunomodulatory effects in the gut (21). Sulfur-containing amino acids are coupled to methyl donor availability via their interconversion to S-adenosylmethionine. Moreover, gut inflammation induces elaboration of homocysteine from monocytes in the lamina propria; increased levels of this amino acid are associated with inflammatory bowel diseases (22).

Several genes in the GO “Regulation of lymphocyte activation” pathway with well-described roles in the accumulation of IELs were associated with enhancers having greater chromatin accessibility following colonization. These include *ahr*, encoding the aryl hydrocarbon receptor (AhR), which controls accumulation of CD8 $\alpha\alpha^+$ TCR β^+ IELs (23); *id2*, which is required for accumulation of CD8 $\alpha\beta^+$ IELs (24); *il15ra* (IL15 receptor α), which encodes a component of the IL-15 receptor; and *il7r*, which specifies a component of the IL-7 receptor. IL-15 and IL-7 are important cytokines controlling the accumulation and function of CD8 $\alpha\alpha^+$ IELs (25–28). Additionally, members of two pathways, KEGG “Cell adhesion molecules” (35 genes) and KEGG “ECM–receptor interaction” (25 genes), were also associated with enhancers exhibiting increased accessibility in CONV-R compared to GF TCR $\alpha\beta^+$ IELs (Dataset S10A); changes in chromatin state affecting enhancers positioned next to genes in these pathways may reflect broad changes in TCR $\alpha\beta^+$ interactions with the intestinal epithelium upon colonization.

Chromatin Landscape in TCR $\gamma\delta^+$ IELs from CONV-R vs. GF Mice.

We identified 4,579 differentially accessible enhancers in TCR $\gamma\delta^+$ IELs recovered from CONV-R and from GF animals ($P < 0.05$, Benjamini–Hochberg-adjusted Wald test). Filtering these datasets to $P < 5 \times 10^{-5}$ yielded 165 putative enhancers with significantly increased accessibility and 511 with significantly decreased accessibility (Fig. S5D and Dataset S8B). GeNets predicted a number of communities of interacting genes positioned near topranked colonization-responsive enhancers, including genes involved in TGF- β receptor signaling and adherens junction pathways; their associated enhancers exhibited increased accessibility in CONV-R animals (Dataset S9 C and D).

GREAT-based analysis using the entire collection of statistically significant enhancers with colonization-associated increases in their accessibility disclosed a significant enrichment of genes in KEGG “Cell adhesion molecules” and “ECM–receptor rinteractions” pathways (Dataset S10D). A similar enrichment of these two pathways was also observed in TCR $\alpha\beta^+$ IELs (Dataset S10A) and operationally defines a shared response of these two IEL lineages to the presence of a gut microbiota. Enhancers neighboring several genes in GO categories involved in leukocyte and lymphocyte activation and cytokine interactions also exhibited significant increases in accessibility with colonization in TCR $\gamma\delta^+$ IELs (Dataset S10D); these genes *include* *Cd28*, *Cxcr4*, *Il15ra*, *Il2*, *Il10*, *Cd40lg*, and *Ccr7*, some of which are associated with the activation of T cells and have been implicated in shaping the IEL compartment. As noted above, IL-15 responsiveness has been shown to be key for the accumulation of CD8 $\alpha\alpha^+$ TCR β^+ cells and also for TCR $\gamma\delta^+$ cells (25–28).

ion of cytokine production” were associated with enhancers with decreased chromatin accessibility in TCR $\gamma\delta^+$ IELs from CONV-R compared to TCR $\gamma\delta^+$ IELs from GF mice; they include *tigit*, a cell-intrinsic (29) and cell-extrinsic (30) inhibitor of effector T-cell function, *Il10*, which encodes an anti-inflammatory cytokine (31), and *furin*, which regulates levels of effector T-cell activation (32). The reduced accessibility of enhancers associated with these genes may reflect a homeostatic response to the increased activation of these cells upon exposure to microbial stimuli that is designed to maintain the delicate balance between pathogenic and protective immu-

nity in the intestine. The top-ranked KEGG pathways composed of genes proximal to enhancers that manifest significant decreases in their accessibility in CONV-R TCR $\gamma\delta^+$ IELs as compared with GF TCR $\gamma\delta^+$ IELs are presented in Dataset S10E (e.g., 23 genes involved in KEGG “Inositol phosphate metabolism”).

Comparing CD4⁺ and CD8⁺ T cells isolated from the peripheral circulation of GF and CONV-R mice revealed markedly fewer differentially accessible enhancers than in the IEL populations (n = 35 and 26 in the CD4⁺ and CD8⁺ populations, respectively; P < 0.05 Benjamini–Hochberg-adjusted Wald test; Dataset S8 C and D). Five enhancer-associated genes (*Cblb*, *Ic11*, *Lrrc8c*, *Mid1*, and *Nme7*) were shared between the two T-cell populations. All demonstrated increased accessibility in both lineages in CONV-R compared with GF mice. Many of the differentially accessible peripheral T-cell enhancers were also responsive to colonization in the IEL lineages (31% of CD4⁺ and 46% of CD8⁺ T-cell enhancers, respectively; compare Dataset S8 A and B with Dataset S8 C and D).

Analysis of Enhancers in IELs Purified from CONV-D Mice.

We expanded our analyses to the group of CONV-D mice that had been reared under GF conditions during the first three postnatal weeks and then were colonized with the cecal microbiota of an 8-wk-old C57BL/6J CONV-R animal and were killed 5 wk later (see Fig. S6 for the results of a bacterial 16S rRNA-based analysis of fecal samples collected at the time mice were killed). We binned enhancers based on differences in their accessibility in IELs recovered from CONV-D mice compared with IELs recovered from GF mice and in IELs recovered from CONV-D mice compared with IELs recovered from CONV-R mice. A set of microbiota-responsive enhancers was identified; these were defined as regions in CONV-D mice whose accessibility was significantly different compared with regions in GF mice but was not significantly different from regions in CONV-R animals.

The top-ranked microbiota-responsive enhancers in TCR $\alpha\beta^+$ IELs (n = 172) were located in the neighborhood of genes with broad roles in transcriptional regulation and cell growth, in-

cluding *Tshz1*, *Celf2*, and *Nav2* (Dataset S11A). Given the limited size of the dataset of these enhancers, we applied an alternative approach to identify pathways whose genes are associated with microbiota-responsive enhancers. Ingenuity Pathway Analysis (IPA) (SI Materials and Methods) revealed that the top-ranked pathway in TCR $\alpha\beta^+$ IELs is involved in the late stages of heparan sulfate synthesis. In the case of TCR $\gamma\delta^+$ IELs, the top-ranked pathways associated with enhancers classified as microbiota-responsive ($n = 141$) were linked to sphingosine-1-phosphate (S1P) and ceramide signaling (Dataset S11B). Enhancers located near *Slpr1*, *Slpr2*, and *Slpr3* contributed to these enrichments. One role of S1P is to direct the trafficking of IELs from the thymus to the intestine; global hematopoietic deficiency of *Slpr1* reduces the ability of TCR $\gamma\delta^+$ cells to populate the small intestine (33). The representation of these pathways among the microbiota-responsive group of enhancers may reflect a component of the codevelopment of the functional state of TCR $\gamma\delta^+$ IELs and the gut microbial community during an important malleable stage of their relationship.

Transcription Factor Regulatory Circuitry.

Understanding that the accessibility of cis-regulatory enhancer elements is mediated by the recruitment of transcription factors (TFs) and that TFs and enhancers interact in complex feedback loops, we performed a de novo reconstruction of TF regulatory circuitry (SI Materials and Methods) (34). In any cell type, a small number of the ~ 500 expressed TFs critically define cell identity; these master TFs are often associated with super-enhancer elements (35). To investigate the effects of the gut microbiota on the TF circuitry regulating cell identity, we modeled TF–enhancer interactions (TF_i) for all TFs associated with dynamically stitched super-enhancers. For each TF_i, we computed the inward connectivity (“IN degree”) to determine how the TF_i is regulated by itself and by other TFs at its own enhancer(s). We then computed the outward connectivity (“OUT degree”) to quantify how a TF_i regulates enhancers of genes encoding other TFs (Fig. 2A). We predicted enriched TF binding sites within ATAC-seq–accessible regions for all super-enhancers located proximal to TFs in TCR $\alpha\beta^+$ and TCR $\gamma\delta^+$ IELs from GF and CONV-R mice and then

computed the IN and OUT degrees for these super-enhancer-associated TFs. Super-enhancer-associated TFs were included in subsequent analyses if their unit-normalized total degree (IN + OUT) was greater than 0.75, indicating that the TF is in the top quartile of regulatory influence. For each IEL population, the interactions of all super-enhancer-associated TFs with a normalized total degree >0.75 are shown in Fig. 2B.

There was a significant concordance for TF regulatory connectivity across IELs, independent of colonization status, at 24 TF binding-site motifs (Dataset S12). Both TCR $\alpha\beta^+$ and $\gamma\delta^+$ IELs were enriched for a number of TF motifs, including ETS1/ETS2- binding sites. The ETS family plays a crucial role in T-cell development and can induce chromatin remodeling during early T-cell differentiation (36). There was also strong enrichment for RUNX1 and RUNX3; members of this family of TFs have been identified as regulators of critical TCR $\alpha\beta^+$ and TCR $\gamma\delta^+$ pathways (37, 38). A recent assessment of chromatin state in innate lymphoid cells also demonstrated that highly active enhancers show enrichment for ETS and RUNX motifs, suggesting potential commonality in developmental and functional pathways within IEL populations of the small intestine (39). *Ahr*, encoding the aryl hydrocarbon receptor, also was found to have high regulatory connectivity in both GF and CONV-R groups. This observation aligns well with its defined roles in TCR $\gamma\delta^+$ cells and in CD8 $\alpha\alpha^+$ TCR β^+ accumulation, and with our observation that this receptor is colonization-responsive in TCR $\alpha\beta^+$ IELs (23).

We also identified a regulatory network of TF motifs that were sensitive to colonization status. We limited our comparisons to GFvs. CONV-R IELs because a much more limited set of differentially accessible enhancers was identified in CONV-D IELs after applying our high-stringency cutoff. Ten TF binding-site motifs exhibited a high total degree of regulatory connectivity in both TCR $\alpha\beta^+$ and TCR $\gamma\delta^+$ IELs isolated from GF mice compared with those isolated from CONV-R mice (Fig. 2C and Dataset S12). They included the TFs IRF1 and STAT3, which have been identified as critical regulators of IEL maturation and the antiapoptotic activity of IL-15 in IELs (40). Although the multiple STAT TFs identified in the circuitry share similar binding motifs, and thus similar OUT degrees, we are able to infer differences in their regulation by examining

their IN degree, which predicts how the genes encoding the TFs themselves are regulated. Additionally, for the *Stat3* gene, there is an increase in enhancer accessibility in TCR $\alpha\beta^+$ IELs from GF mice compared with TCR $\alpha\beta^+$ IELs from CONV-R mice that is not present for *Stat1* (Fig. S7). The *Stat3* enhancer contains large numbers of predicted binding sites for the JUN/FOS and MAFK bZIP TFs that are not present in the *Stat1* gene (Fig. S7). Taken together, these data suggest an increase in the regulatory activity of Stat3 in IELs purified from GF mice compared with those from CONV-R animals.

In the reciprocal comparison, three TF binding-site motifs manifested a high degree of regulatory connectivity in TCR $\alpha\beta^+$ and TCR $\gamma\delta^+$ IELs recovered from CONV-R mice compared with those recovered from GF animals: GATA3, a critical regulator of T-cell function and master TF for Th2 helper cells (41); KLF3, which has broad regulatory roles, including intestinal IgA production; and REL, which plays a key role in driving both Th1 (42) and Th17 (43, 44) responses to pathogens (Dataset S12 C and D). These observations suggest a microbiota-regulated role for these TFs in the establishment and maturation of these IEL lineages.

Discussion

The location and cellular features of IELs are well described. However, the molecular mechanisms underlying their capacity to exert a rapid and diverse set of effector functions in response to stimulation have been largely undefined. One key stimulus occurs during colonization of the gut with a microbiota. Here, we used ATAC-seq and the RIESLING suite of software tools we developed to conduct a genome-wide search for enhancers and pathways enriched in genes adjacent to these enhancers. Our results demonstrate that this approach can clearly distinguish and classify cis-regulatory landscapes of distinct immune lineages. We find that gut colonization status does not fundamentally alter lineage-specific cis-regulatory landscapes but rather induces quantitative changes in the accessibility of these preestablished enhancer elements (Fig. S5E). Our findings differ from a previous DNase-seq study in which the authors concluded that the transcriptional state was modified by colonization with a gut microbiota from a specified pathogen-free mouse donor in

a manner that appeared to be independent of chromatin remodeling (2). That study used a heterogeneous population of intestinal epithelial cells. One possible explanation for the differences in our conclusions is that signals from enhancers with differential accessibility may be drowned out in a highly heterogeneous population of starting cells. We were able to identify quantitative differences in chromatin accessibility in purified cell lineages using ATAC-seq, which requires much smaller numbers of cells than DNase-seq. Our findings also are consistent with the quantitative changes observed in a recently reported iChIPVT analysis of innate lymphoid cells harvested from CONV-R animals that had or had not been treated with antibiotics (see the Introduction and ref. 4).

It is tempting to speculate that gut epithelial cell lineages that are in direct contact with the microbiota are poised to exploit a preexisting chromatin landscape in ways that allow them to respond rapidly to microbial stimulation. Although increases in chromatin accessibility correlate with transcriptional activity, these measurements are limited in their ability to characterize more complex modulations of the chromatin state. For instance, enhancers can be poised or active (45), and the recruitment of chromatin regulators at preexisting accessible sites can dramatically modulate transcriptional responses (46–48). Nonresponsive enhancers, which fail to exhibit alterations in chromatin accessibility after delayed exposure to a microbiota, could reflect (i) the existence of a critical developmental window beyond which chromatin modification is not possible or is unlikely; (ii) the need for a minimal length of time of microbial exposure, irrespective of when the microbiota is first introduced, that was not satisfied under our experimental conditions (5 wk of exposure starting at the end of postnatal week 3); or (iii) differences in the structural and functional configuration of the microbiota that we observed between CONV-R and CONV-D animals.

Our results provide an estimation of the magnitude of the gut microbiota's influence on the epigenetic state of IELs and define the enhancers and pathways impacted by colonization. Future efforts to understand how the functional epigenetic states of IELs are modified by specific members of the gut microbiota, the nature of the diet being consumed, and the period in postnatal development when these cells first encounter gut microbes will be important to develop a mechanistic understanding of how these interactions contribute to protective immunity. These questions

are relevant not only to normal host development but also to disorders in which various facets of development are impaired, including in children with undernutrition who have disrupted maturation of their gut microbiota (49, 50).

References

1. Olszak T, et al. (2012) Microbial exposure during early life has persistent effects on natural killer T cell function. *Science* 336(6080):489–493
2. Camp JG, et al. (2014) Microbiota modulate transcription in the intestinal epithelium without remodeling the accessible chromatin landscape. *Genome Res* 24(9):1504–1516
3. Yu D-H, et al. (2015) Postnatal epigenetic regulation of intestinal stem cells requires DNA methylation and is guided by the microbiome. *Genome Biol* 16(1):211
4. Gury-BenAri M, et al. (2016) The spectrum and regulatory landscape of intestinal innate lymphoid cells are shaped by the microbiome. *Cell* 166(5):1231–1246.e13
5. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNAbinding proteins and nucleosome position. *Nat Methods* 10(12):1213–1218
6. Cheroutre H, Lambolez F, Mucida D (2011) The light and dark sides of intestinal intraepithelial lymphocytes. *Nat Rev Immunol* 11(7):445–456
7. Ismail AS, et al. (2011) Gammadelta intraepithelial lymphocytes are essential mediators of hostmicrobial homeostasis at the intestinal mucosal surface. *Proc Natl Acad Sci USA* 108(21): 8743–8748
8. Ismail AS, Behrendt CL, Hooper LV (2009) Reciprocal interactions between commensal bacteria and gamma delta intraepithelial lymphocytes during mucosal injury. *J Immunol* 182(5):3047–3054
9. Whyte WA, et al. (2013) Master transcription factors and mediator establish superenhancers at key cell identity genes. *Cell* 153(2):307–319
10. Lovén J, et al. (2013) Selective inhibition of tumor oncogenes by disruption of superenhancers. *Cell* 153(2):320–334

11. Hnisz D, et al. (2013) Super-enhancers in the control of cell identity and disease. *Cell* 155(4):934–947
12. Vahedi G, et al. (2015) Super-enhancers delineate disease-associated regulatory nodes in T cells. *Nature* 520(7548):558–562
13. Qu K, et al. (2015) Individuality and variation of personal regulomes in primary human T cells. *Cell Syst* 1(1):51–61
14. Quann EJ, Liu X, Altan-Bonnet G, Huse M (2011) A cascade of protein kinase C isozymes promotes cytoskeletal polarization in T cells. *Nat Immunol* 12(7):647–654
15. Fu G, et al. (2011) Protein kinase C η is required for T cell activation and homeostatic proliferation. *Sci Signal* 4(202):ra84
16. Khan A, Zhang X (2016) dbSUPER: A database of super-enhancers in mouse and human genome. *Nucleic Acids Res* 44(D1):D164–D171
17. Roychoudhuri R, et al. (2013) BACH2 represses effector programs to stabilize T(reg)-mediated immune homeostasis. *Nature* 498(7455):506–510
18. McLean CY, et al. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* 28(5):495–501
19. Sayin SI, et al. (2013) Gut microbiota regulates bile acid metabolism by reducing the levels of tauro-beta-muricholic acid, a naturally occurring FXR antagonist. *Cell Metab* 17(2):225–235
20. Sweet MJ, Shakespear MR, Kamal NA, Fairlie DP (2012) HDAC inhibitors: Modulating leukocyte differentiation, survival, proliferation and inflammation. *Immunol Cell Biol* 90(1):14–22
21. Smith PM, et al. (2013) The microbial metabolites, short-chain fatty acids, regulate colonic treg cell homeostasis. *Science* 341(6145):569–573

22. Danese S, et al. (2005) Homocysteine triggers mucosal microvascular activation in inflammatory bowel disease. *Am J Gastroenterol* 100(4):886–895
23. Li Y, et al. (2011) Exogenous stimuli maintain intraepithelial lymphocytes via aryl hydrocarbon receptor activation. *Cell* 147(3):629–640
24. Kim J-K, Takeuchi M, Yokota Y (2004) Impairment of intestinal intraepithelial lymphocytes in Id2 deficient mice. *Gut* 53(4):480–486
25. Lodolce JP, et al. (1998) IL-15 receptor maintains lymphoid homeostasis by supporting lymphocyte homing and proliferation. *Immunity* 9(5):669–676
26. Lai YG, et al. (2008) IL-15 does not affect IEL development in the thymus but regulates homeostasis of putative precursors and mature CD8 alpha alpha+ IELs in the intestine. *J Immunol* 180(6):3757–3765
27. Isakov D, Dzutsev A, Berzofsky JA, Belyakov IM (2011) Lack of IL-7 and IL-15 signaling affects interferon- γ production by, more than survival of, small intestinal intraepithelial memory CD8+ T cells. *Eur J Immunol* 41(12):3513–3528
28. Fujihashi K, McGhee JR, Yamamoto M, Peschon JJ, Kiyono H (1997) An interleukin-7 internet for intestinal intraepithelial T cell development: Knockout of ligand or receptor reveal differences in the immunodeficient state. *Eur J Immunol* 27(9):2133–2138
29. Joller N, et al. (2011) Cutting edge: TIGIT has T cell-intrinsic inhibitory functions. *Immunol* 186(3):1338–1342
30. Joller N, et al. (2014) Treg cells expressing the coinhibitory molecule TIGIT selectively inhibit proinflammatory Th1 and Th17 cell responses. *Immunity* 40(4):569–581
31. Moore KW, de Waal Malefyt R, Coffman RL, O’Garra A (2001) Interleukin-10 and the interleukin-10 receptor. *Annu Rev Immunol* 19(1):683–765
32. Pesu M, et al. (2008) T-cell-expressed proprotein convertase furin is essential for maintenance of peripheral immune tolerance. *Nature* 455(7210):246–250

33. Odumade OA, Weinreich MA, Jameson SC, Hogquist KA (2010) Krüppel-like factor 2 regulates trafficking and homeostasis of gammadelta T cells. *J Immunol* 184(11):6060–6066
34. Lin CY, et al. (2016) Active medulloblastoma enhancers reveal subgroup-specific cellular origins. *Nature* 530(7588):57–62
35. Lee TI, Young RA (2013) Transcriptional regulation and its misregulation in disease. *Cell* 152(6):1237–1251
36. Cauchy P, et al. (2016) Dynamic recruitment of Ets1 to both nucleosome-occupied and -depleted enhancer regions mediates a transcriptional program switch during early T-cell differentiation. *Nucleic Acids Res* 44(8):3567–3585
37. Tani-Ichi S, Satake M, Ikuta K (2011) The pre-TCR signal induces transcriptional silencing of the TCR γ locus by reducing the recruitment of STAT5 and Runx to transcriptional enhancers. *Int Immunol* 23(9):553–563
38. Reis BS, Rogoz A, Costa-Pinto FA, Taniuchi I, Mucida D (2013) Mutual expression of the transcription factors Runx3 and ThPOK regulates intestinal CD4⁺ T cell immunity. *Nat Immunol* 14(3):271–280
39. Koues OI, et al. (2016) Distinct gene regulatory pathways for human innate versus adaptive lymphoid cells. *Cell* 165(5):1134–1146
40. Malamut G, et al. (2010) IL-15 triggers an antiapoptotic pathway in human intraepithelial lymphocytes that is a potential new target in celiac disease-associated inflammation and lymphomagenesis. *J Clin Invest* 120(6):2131–2143
41. Zheng W, Flavell RA (1997) The transcription factor GATA-3 is necessary and sufficient for Th2 cytokine gene expression in CD4 T cells. *Cell* 89(4):587–596
42. Hilliard BA, et al. (2002) Critical roles of c-Rel in autoimmune inflammation and helper T cell differentiation. *J Clin Invest* 110(6):843–850

43. Ruan Q, et al. (2011) The Th17 immune response is controlled by the Rel-ROR γ -ROR γ T transcriptional axis. *J Exp Med* 208(11):2321–2333
44. Chen G, et al. (2011) The NF- κ B transcription factor c-Rel is required for Th17 effector cell development in experimental autoimmune encephalomyelitis. *J Immunol* 187(9):4483–4491
45. Creighton MP, et al. (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA* 107(50):21931–21936
46. Wang E, et al. (2015) The transcriptional cofactor TRIM33 prevents apoptosis in B lymphoblastic leukemia by deactivating a single enhancer. *eLife* 4(4):e06377
47. Brown JD, et al. (2014) NF- κ B directs dynamic super enhancer formation in inflammation and atherogenesis. *Mol Cell* 56(2):219–231
48. Dixon JR, et al. (2015) Chromatin architecture reorganization during stem cell differentiation. *Nature* 518(7539):331–336
49. Subramanian S, et al. (2014) Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature* 510(7505):417–421
50. Blanton LV, et al. (2016) Gut bacteria that prevent growth impairments transmitted by microbiota from malnourished children. *Science* 351(6275):aad3311
51. Meyer-Hoffert U, et al. (2008) Identification of heparin/heparan sulfate interacting protein as a major broad-spectrum antimicrobial protein in lung and small intestine. *ASEB J* 22(7):2427–2434
52. Garner OB, Yamaguchi Y, Esko JD, Videm V (2008) Small changes in lymphocyte development and activation in mice through tissue-specific alteration of heparan sulphate. *Immunology* 125(3):420–429
53. Long ET, Baker S, Oliveira V, Sawitzki B, Wood KJ (2010) Alpha-1,2-mannosidase and hence N-glycosylation are required for regulatory T cell migration and allograft tolerance in mice. *PLoS One* 5(1):e8894

54. Gebuhr I, et al. (2011) Differential expression and function of α -mannosidase I in stimulated naive and memory CD4⁺ T cells. *J Immunother* 34(5):428–437
55. Mallick-Wood CA, et al. (1996) Disruption of epithelial gamma delta T cell repertoires by mutation of the Syk tyrosine kinase. *Proc Natl Acad Sci USA* 93(18):9704–9709
56. Puddington L, Olson S, Lefrançois L (1994) Interactions between stem cell factor and c-Kit are required for intestinal immune system homeostasis. *Immunity* 1(9):733–739
57. DeSantis TZ (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72(7):5069–5072
58. Caporaso JG, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7(5):335–336
59. Lara-Astiaso D, et al. (2014) Chromatin state dynamics during blood formation. *Science* 345(6199):943–9
60. Zhang Y, et al. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9(9):R137
61. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15(12):550
62. Stamatoyannopoulos JA, et al.; Mouse ENCODE Consortium (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* 13(8):418
63. Grant CE, Bailey TL, Noble WS (2011) FIMO: Scanning for occurrences of a given motif. *Bioinformatics* 27(7):1017–1018
64. Matys V, et al. (2006) TRANSFAC(R) and its module TRANSCompel(R): Transcriptional gene regulation in eukaryotes. *Nucl Acids Res* 34(suppl_1):D108–110
65. Consortium TEP, et al.; ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74

Materials and methods

Mouse experiments were performed using protocols approved by the Washington University Animal Studies Committee. A detailed description of protocols used for (i) gnotobiotic animal husbandry, (ii) 16S rRNA-based characterization of bacterial taxa present in the fecal microbiota of CONV-R and CONV-D mice, (iii) purification of small intestinal IELs and circulating CD4/CD8+ T cells, and (iv) ATAC-seq (including transposition and sequencing protocols; the architecture of the RIESLING computational pipeline; the approach for identifying enhancers that exhibit significant differences in their accessibility in the different immune cell lineages as a function of colonization status; data normalization; pathway analyses; and characterization of TF regulatory circuitry) is provided in SI Materials and Methods.

Acknowledgments

We thank Jessica Hoisington-Lopez for assistance with DNA sequencing, David O'Donnell and Maria Karlsson for help with gnotobiotic mouse husbandry, and Drew Hughes for valuable insights about the analytical approach. This work was supported in part by NIH Grants DK30292, DK70977, and DK078669. N.P.S. is supported by NIH Grant T32HG000045. N.P.S. and J.D.P. are members of the Washington University Medical Scientist Training Program supported by NIH Grant GM007200. P.P.A. is the recipient of a Sir Henry Wellcome Postdoctoral Fellowship (096100).

Figure legends

Fig. 1.

The effects of colonization on enhancer populations in IELs. (A) Unsupervised clustering of dynamically stitched enhancers identified in all four immune cell populations purified from GF, CONV-R, and CONV-D mice demonstrates that these enhancers not only stratify by cell type but also stratify TCR $\alpha\beta^+$ and TCR $\gamma\delta^+$ IEL lineages based on their history of microbial exposure. The areas boxed in red show that IELs from GF mice or from CONV-D mice cluster separately from IELs recovered from CONV-R mice exposed to microbes beginning at birth. Jaccard distance metric scores between enhancer regions from a given cell type are shown in blue shading where a value of 1.0 reflects complete overlap. Clustering was performed on the largest dynamically stitched enhancers using the super-enhancer cutoffs described in the text. (B) Heatmap demonstrating the ATAC-seq signal from components of the “JAK-STAT,” and “leukocyte transendothelial migration” signaling pathways whose enhancers demonstrate increased accessibility in CONV-R TCR $\alpha\beta^+$ IELs. These members were identified as overrepresented in a GeNets-based analysis (Dataset S9B).

Fig. 2.

Analysis of colonization-associated TF circuitry. (A) Overview of IN/OUT degree. In this example, the IN degree is the number of TFs (TF-A, TF-B, and TF-C) that regulate the expression of a gene encoding a TF (TF-A). The OUT degree is the number of enhancers (E1, E2, E3, E4) that contain binding-site motifs for TF-A. (B) Regulatory connectivity of enhancer-associated TFs in TCR $\alpha\beta^+$ and TCR $\gamma\delta^+$ IELs purified from the indicated groups of mice. For each IEL population, the interactions of all enhancer-associated TFs with a normalized total (IN+OUT) degree >0.75 are shown. TFs with a normalized total degree of at least 0.75 in either GF or CONV-R cells are highlighted in red or blue, respectively. (C) The total IN+OUT degree for all enhancer-associated TFs with a normalized total degree >0.75 are hierarchically clustered using a Euclidean distance metric allowing categorization of groups of TFs with similar regulatory patterns. TFs are color

coded by their degree of connectivity in each colonization group: blue- and red-colored TFs demonstrated a high degree of connectivity in CONV-R and GF IELs, respectively. TFs colored black displayed high connectivity in both IEL populations, and those in brown displayed a mixed pattern of connectivity.

Figures

Figure 1.

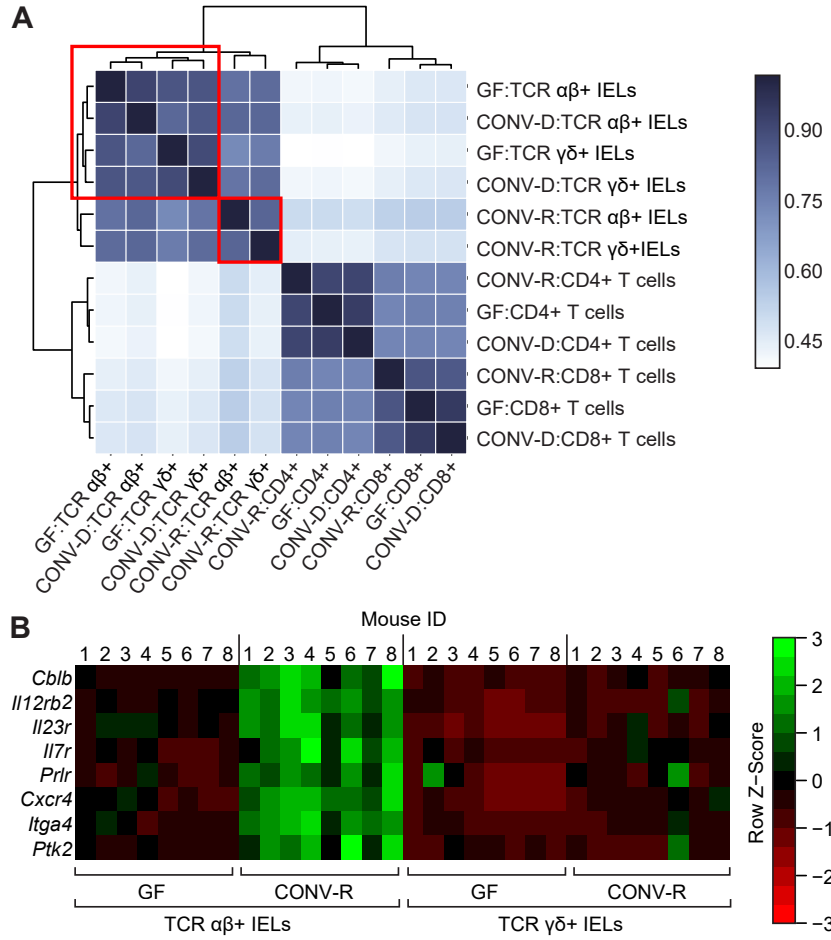
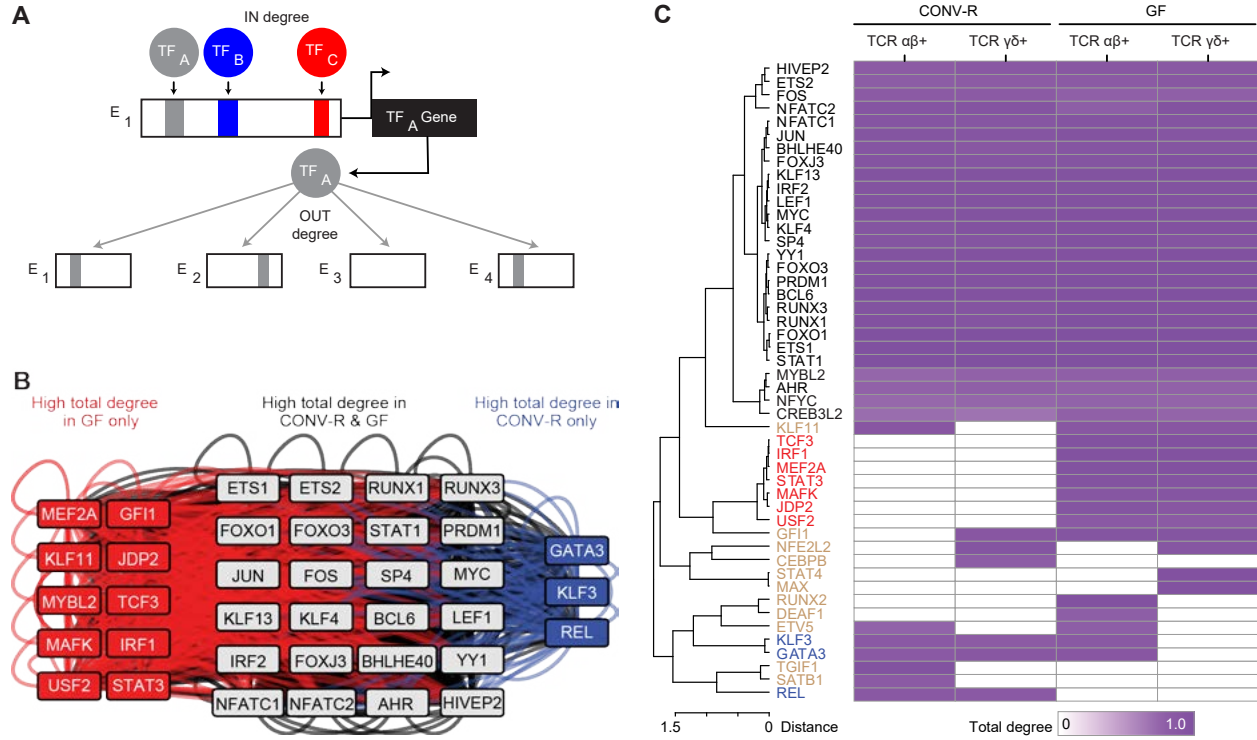


Figure 2.



List of tables

Dataset S1. Overview of samples and ATAC-seq datasets

A summary of all samples in the study (including read counts), stratified by cell type and colonization status.

Dataset S2. Super-enhancers identified in all populations, independent of colonization status, rank ordered by signal intensity This dataset encompasses all the super-enhancers presented in Fig. S1B. Super-enhancers were identified using the RIESLING pipeline. For each predicted super-enhancer, chromosome location, predicted boundaries, signal intensity, and most proximal gene (independent of orientation) are listed. (A) TCR $\alpha\beta$ + IELs. (B) TCR $\gamma\delta$ + IELs. (C) Both IEL populations. (D) CD4+ T cells only. (E) CD8+ T cells only. (F) Both T-cell populations (CD4+ U CD8+).

Dataset S3. Pathway analysis within the GeNets-predicted communities from the top IEL super-enhancers

The GeNets analysis identified eight predicted communities of interconnected super-enhancers (applied to the top 250 super-enhancers from the combined population of TCR $\alpha\beta$ + and TCR $\gamma\delta$ + IELs independent of host colonization status). Pathway analyses applied to these communities identified a number of enriched pathways, described here by community.

Dataset S4. Lineage-specific enhancers independent of colonization status

Differentially accessible enhancers were identified using binomial modeling via DESeq2. Statistical significance, based on a Benjamini–Hochberg-adjusted Wald test, reflects the consistency of differences in enhancer accessibility in each cell lineage across all animals regardless of their membership in the GF, CONVR, or CONV-D groups (colonization independent). (A) Enhancers more accessible in the TCR $\alpha\beta$ + IELs, rank ordered based on the statistical significance. (B) Enhancers more accessible in the TCR $\gamma\delta$ + IELs. (C) Enhancers more accessible in CD4+ T cells. (D) Enhancers more accessible in CD8+ T cells.

Dataset S5. Pathway analysis of lineage-specific differentially accessible enhancers independent of colonization status

Pathway analyses were applied to genes adjacent to enhancers described in Dataset S4. Analyses used GO Molecular Function, Biological Process, and Cellular Component and MSigDB Pathways. Pathways are rank ordered based on the statistical significance of their enrichment (hypergeometric and binomial tests). (A–D) Pathways with increased accessibility in TCR $\alpha\beta^+$ IELs (A), TCR $\gamma\delta^+$ IELs (B), CD4⁺ T cells (C), or CD8⁺ T cells (D).

Dataset S6. Pathway analysis of genes near predicted enhancers in IEL populations, independent of colonization status

Pathway analyses were performed comparing genes proximal to identified super-enhancers with the whole mouse genome as a background and comparing genes proximal to traditional enhancers with the whole mouse genome. A third comparison, designed to highlight potential functional differences between super-enhancers and traditional enhancers, involved a direct comparison of genes proximal to super-enhancers with genes proximal to all enhancers [i.e., the set of traditional and super-enhancers instead of the mouse genome (mm10)]. Pathways from GO and MSigDB are rank ordered based on the statistical significance of their enrichment (hypergeometric and binomial tests). (A) IEL (combined TCR $\alpha\beta^+$ and $\gamma\delta^+$) super-enhancers vs. mm10. (B) IEL traditional enhancers vs. mm10. (C) IEL super-enhancers vs. traditional enhancers. (D) TCR $\alpha\beta^+$ super-enhancers vs. mm10. (E) TCR $\gamma\delta^+$ super-enhancers vs. mm10. (F) TCR $\alpha\beta^+$ traditional enhancers vs. mm10. (G) TCR $\gamma\delta^+$ traditional enhancers vs. mm10. (H) TCR $\alpha\beta^+$ super-enhancers vs. traditional enhancers. (I) TCR $\gamma\delta^+$ super-enhancers vs. traditional enhancers.

Dataset S7. Pathway analysis of genes near predicted enhancers in CD4⁺ and CD8⁺ T-cell populations, independent of colonization status

Pathway analyses were performed as described in Dataset S6. Pathways are rank ordered based on the statistical significance of their enrichment (using hypergeometric and binomial tests). (A) Peripheral T-cell (combined CD4⁺ and CD8⁺) super-enhancers vs. mm10. (B) Peripheral T-cell

traditional enhancers vs. mm10. (C) Peripheral T-cell super-enhancers vs. traditional enhancers. (D) CD4⁺ super-enhancers vs. mm10. (E) CD8⁺ super-enhancers vs. mm10. (F) CD4⁺ traditional enhancers vs. mm10. (G) CD8⁺ traditional enhancers vs. mm10. (H) CD4⁺ super-enhancers vs. traditional enhancers. (I) CD8⁺ super-enhancers vs. traditional enhancers.

Dataset S8. Differentially accessible regions of chromatin identified comparing GF to CONV-R mice

Included here are statistically significant ($P < 0.05$, Benjamini–Hochberg-adjusted Wald test) differentially accessible enhancers and associated genomic features identified in TCR $\alpha\beta^+$ IELs (A), TCR $\gamma\delta^+$ IELs (B), CD4⁺ T cells (C), and CD8⁺ T cells (D).

Dataset S9. Pathway analysis within the GeNets-predicted communities from the genes adjacent to differentially accessible enhancers identified in IELs purified from CONV-R vs. GF mice

Also included are inferred candidate genes (and their pathways) that displayed connectivity and known interactions similar to those we observed. (A) Increased accessibility in GF TCR $\alpha\beta^+$ IELs. (B) Increased accessibility in CONV-R TCR $\alpha\beta^+$ IELs. (C) Increased accessibility in GF TCR $\gamma\delta^+$ IELs. (D) Increased accessibility in CONV-R TCR $\gamma\delta^+$ IELs.

Dataset S10. GREAT analysis applied to differentially accessible enhancers and their adjacent genes identified in IELs from CONV-R vs. GF mice

Included are statistically significant results (hypergeometric test $P < 0.05$) from the GO consortium and MSigDB. (A–C) Within TCR $\alpha\beta^+$ IELs, pathways enriched for genes proximal to enhancers demonstrating increased accessibility (A) and decreased accessibility (B), and independent of accessibility (C). (D–F) Within TCR $\gamma\delta^+$ IELs, pathways enriched for genes proximal to enhancers with increased accessibility (D), decreased accessibility (E), and independent of accessibility (F).

Dataset S11. Enhancers classified as microbiota-responsive, based on whether their accessibility changes in the GF vs. CONV-D and CONV-R groups

Included are statistically significant ($P < 0.05$, Benjamini–Hochberg-adjusted Wald test) differentially accessible enhancers and their most adjacent genes, classified as microbiota-responsive in TCR $\alpha\beta^+$ IELs (A) and in TCR $\gamma\delta^+$ IELs (B). Fold-changes and statistics reflect the comparison between GF vs. CONV-R groups.

Dataset S12. Analysis of TF circuitry in GF and CONV-R IEL enhancers

TF circuitry based on IN/OUT degree was analyzed using dynamically stitched enhancers. TFs shown in boldface have a high degree of connectivity and also are displayed in Fig. 2B. Those with an asterisk are also colonization specific. (A) GF TCR $\alpha\beta^+$ IELs. (B) GF TCR $\gamma\delta^+$ IELs. (C) CONV-R TCR $\alpha\beta^+$ IELs. (D) CONV-R TCR $\gamma\delta^+$ IELs.

Supplemental materials

SI Results

We also contrasted the chromatin landscapes of TCR $\alpha\beta^+$ vs. TCR $\gamma\delta^+$ IELs, independent of colonization status. Our GREAT-based analysis revealed that the most statistically significant enrichment for MSigDB pathways in TCR $\alpha\beta^+$ IELs was “Glycosaminoglycan biosynthesis– heparin sulfate” (10 genes) (Dataset S5A). Heparan sulfate-binding proteins have been reported to have antimicrobial effects in both gut and lung (51). The two most significantly enriched GO Molecular Functions were “UDP-xylosyltransferase activity” and “Protein xylosyltransferase activity,” which were driven by multiple enhancers around three genes (*Gxytl2*, *Xylt1*, and *Xylt2*) and which contribute to glycosaminoglycan biosynthesis, including heparan sulfate. T cells with a deficiency in heparan sulfate biosynthetic enzymes are hyperresponsive to low-level activation (52).

The most statistically significant MSigDB pathway in TCR $\gamma\delta^+$ IELs was N-glycan biosynthesis (based on eight distinct genes adjacent to enhancers), including a number of mannosidases (Dataset S5B). N-glycosylation and mannosidase activity are important for T-cell migration in an allograft model (53) and serve an important inhibitory role in preventing T-cell activation (54). The most statistically significant GO Molecular Function was “Receptor signaling protein tyrosine kinase activity,” driven by multiple enhancers around three genes, *ErbB4*, *Kit*, and *Syk*; the latter two genes have been identified as factors affecting TCR $\gamma\delta^+$ survival in the intestine (55, 56).

The rank order of pathway enrichment in these two IEL lineages differs from that observed in CD4⁺ and CD8⁺ T cells (compare Dataset S5 C and D with Dataset S5 A and B). For parallel GREAT analyses comparing enhancers with the whole mouse genome (mm10), see Datasets S6 (IELs) and S7 (peripheral T cells).

SI Materials and methods

Animals. Adult (8- to-10-wk-old) GF male C57BL/6J mice were maintained either in flexible gnotobiotic isolators (for the GF and CONV-D groups; Class Biologically Clean Ltd.) or in a specific pathogen-free barrier facility (CONV-R group), under a strict 12-h light cycle (light on at 0600, off at 1800). Pups were weaned onto a sterilized low-fat, plant polysaccharide-rich chow (B&K Universal; Product 7378000) that was provided ad libitum. All mice were housed in cages containing indigestible cellulosic bedding (Aspen wood shavings; NEPCO). CONV-D mice were colonized at 3 wks of age by gavage of cecal contents harvested from an adult male CONV-R C57BL/6J donor. Gavage was performed using 200 μ L of a slurry of cecal contents homogenized in 5 mL sterile PBS, with the remaining mixture spread on fur.

All mice were killed in the morning, and all four immune cell populations were isolated in parallel from up to four mice simultaneously. Where possible, subsets of mice from different treatment groups (e.g., two GF mice and two CONV-R mice) were killed at the same time to reduce technical and day-to-day biases associated with the purification of cell populations and preparation of nuclear DNA for ATAC-seq.

Bacterial 16S rRNA Analysis.

16S rRNA PCR was performed on fecal samples collected from CONV-R, CONV-D, and (as a quality control) GF mice. Fecal samples were collected at the time mice were killed, frozen immediately in liquid nitrogen, and stored at -80°C until analysis. DNA was isolated from samples by bead beating and subsequent extraction in 210 μ L of 20% SDS, 500 μ L buffer A (200 mM NaCl, 200 mM Trizma base, 20 mM EDTA), and 500 μ L phenol:chloroform:isoamyl alcohol (at a ratio of 25:24:1). Isolated DNA was purified using QIAquick columns (Qiagen) followed by elution in 70 μ L Tris-EDTA. Purified DNA was quantified using the Quant-iT dsDNA assay (broad range; Invitrogen) and was normalized to 1 ng/mL. PCR was performed using phased, barcoded primers, and amplicons were generated from variable region 4 (V4) of bacterial 16S rRNA genes. Fecal samples from the GF mice failed to produce detectable levels of amplicons. Amplicons were subsequently

sequenced from CONV-D and CONV-R mice using an Illumina MiSeq (paired-end 250-nt reads). Reads were de-multiplexed and clustered based on 97% sequence identity to the Greengenes database for operational taxonomic units (OTUs) (57). Novel OTUs from the remaining sequences were grouped using the UCLUST method [using QIIME version 1.9 (58)], and taxonomy was assigned using RDP 2.11 trained on a custom database aggregating the Greengenes “Isolated named strains 16s collection” and the National Center for Biotechnology Information (NCBI) taxonomy database.

A phylogenetic tree of the OTUs’ representative V4-16S rRNA sequences was constructed and used to calculate Faith’s phylogenetic diversity as well as weighted and unweighted UniFrac distances. The number of observed OTUs was also calculated for each sample (rarefied to 5,000 reads), and binary Jaccard dissimilarities were calculated for each pair of samples. ANOVAs were used to test for differences in the means of alpha-diversity metrics (phylogenetic diversity, Shannon’s diversity index, observed OTUs) between CONV-D and CONV-R mice, and permutational multivariate analyses of variance (PERMANOVA) were used to test for differences in community structure (unweighted UniFrac distances, weighted UniFrac distances, and binary Jaccard dissimilarities). Indicator species analysis was used to identify specific OTUs that were more strongly associated with CONV-D or CONV-R mice. Before this analysis, OTUs that were not detected in at least three CONV-D or three CONV-R mice or that had mean relative abundances less than 0.1% in both groups were removed from the dataset. Significance was assessed by permutation tests using 10,000 randomizations, and P values were corrected for false discovery using the Benjamini–Hochberg method.

Isolation of cell populations.

Small intestinal IELs. When mice were killed, the small intestines were excised, placed in HBSS lacking calcium and magnesium at 4 °C, rinsed gently with ice-cold HBSS to remove luminal contents, dissected along their cephalocaudal axis, and minced into 3- to 5-mm fragments. These fragments were agitated gently for 15 min (37 °C) in 15 mL of HBSS containing EDTA (5 mM),

and the suspension was passed through a sterile 100- μ m mesh nylon filter (Falcon/Corning; no. 352360). This step was repeated, and the filtrates were combined and centrifuged for 5 min at 500 \times g. Supernatants were discarded, and cell pellets were resuspended in 10 mL of 40% Percoll/RPMI (RPMI Medium 1640, Gibco; Percoll no.17-0891-01, GE Life Sciences). The suspension was subsequently layered over 5 mL of 80% Percoll/RPMI, and the material was centrifuged for 20 min at 650 \times g (18 $^{\circ}$ C) with the brake off. Cells at the resulting 40%/80% Percoll interface were collected, washed once with 50 mL of RPMI, and isolated again by centrifugation (5 min at 500 \times g). The resulting supernatant was discarded, and cell pellets were transferred to a round-bottomed 96-well plate for Fc receptor blocking and antibody staining as described below.

Circulating CD4/CD8+ T cells. CD4+ and CD8+ T cells were isolated by collecting \sim 500 μ L of whole blood at the time mice were killed in EDTA-containing tubes (BD Microtainer; no. 365973). Each sample was gently mixed with \sim 3 mL of cold HBSS and layered on top of Ficoll 1.084 (GE Life Sciences; no. 17-5446-02). Samples were spun at 400 \times g for 30 min at 18 $^{\circ}$ C with no brake. After centrifugation, the upper plasma layer was removed, and the entire mononuclear layer was transferred to a 50-mL conical tube (Falcon; no. 352098) wetted with a minimal volume of cold HBSS. Additional cold HBSS was added to the maximum volume of the tube. Tubes then were spun at 500 \times g for 10 min (4 $^{\circ}$ C), and the supernatants were discarded. The pelleted peripheral blood mononuclear cells (PBMCs) were resuspended in 50 μ L of HBSS and were transferred to a round-bottomed 96- well plate for blocking and staining, as described below.

Both cell preparations (IELs, and PBMCs) were incubated on ice for 20 min with anti-CD16/CD32 (clone 2.4G2; BD) to prevent nonspecific staining during subsequent steps. Cells were washed with 200 μ L of ice-cold PBS-BSA, centrifuged at 500 \times g for 5 min (4 $^{\circ}$ C), and the supernatant was discarded. Cell pellets were resuspended in the appropriate antibody staining mix. For IELs, this mix included anti-TCR $\gamma\delta$ phycoerythrin (clone eBioGL3; eBioscience), anti-CD45.2 allophycocyanin (APC)-eFluor780 (clone 104; BD), anti-TCR β V450 (clone H57-597; BD), and antiB220 V500 (clone RA3-6B2; BD). For PBMCs, this mix was composed of anti-

CD8a FITC (clone 53-6.7; BD), anti-CD4 APC (clone RM4-5; eBioscience), anti-TCR β V450 (clone H57-597; BD), and anti-B220 V500 (clone RA3-6B2; BD).

Live cells were identified by exclusion of the DNA-binding dye 7-aminoactinomycin D (7-AAD) (Life Technologies) added to the staining mix at a dilution of 1:1,000. After incubation with these reagents in the dark for 20 min on ice, cells were washed with 150 μ L of PBS-BSA, centrifuged at $500 \times g$ for 5 min (4 °C), and the supernatant was discarded. Cells were kept in the dark, on ice, until flow cytometry and cell sorting.

Flow cytometry and cell sorting were performed with a FACSAria III instrument (BD Biosciences) in a laminar flow biocontainment hood (BioProtect IV Safety Cabinet; Baker Co.), using the manufacturer's protocol for aseptic sorting. Sheath fluid consisted of autoclaved PBS and was prepared no more than 1 day before sorting. Cells were sorted into sterile 1.5-mL tubes in a water-cooled jacket maintained at 4 °C. After sorting, cell purity was validated in a post sort; i.e., for every sorted population, 10% of the sorted volume was transferred to a new sterile tube and analyzed by flow cytometry (see Fig. S1A for the gating strategies).

ATAC-seq.

Transposition and sequencing. Purified, sorted cells ($\leq 50,000$ cells per cell type) were prepared for ATAC-seq transposition following the original published protocol (5), incorporating modifications and optimizations (59). Purified cells from FACS were first centrifuged at $500 \times g$ for 10 min (4 °C), and the supernatants were discarded. Cell pellets were rinsed with 150 μ L icecold PBS and centrifuged at $500 \times g$ for 10 min (4 °C). The supernatants were discarded, and cell pellets were resuspended in 25 μ L of ice-cold lysis buffer [10 mM Tris·HCl (pH 7.4), 10 mM NaCl, 3 mM MgCl₂, 0.1% (vol/vol) IGEPAL CA-630 (Sigma no. I8896)], followed by centrifugation at $500 \times g$ for 15 min (4 °C). Supernatants were discarded, and the pellets were kept on ice while being gently resuspended in 25 μ L of ATAC-seq transposase reaction mixture [12.5 μ L 2 \times Illumina Tagment DNA (TD) buffer; 10.5 μ L nuclease-free water; 2.0 μ L Tn5 transposase (Illumina/Nextera no. FC-121-1030)]. Samples were incubated at 37 °C for 1 h and then were placed immediately on

ice, and transposed DNA was isolated using a standard PCR purification kit (QIAGEN MinElute no. 28004).

Isolated DNA samples were transferred to a 96-well PCR plate and amplified for nine cycles using unique 8-bp indexes per sample (Dataset S1). Cycling conditions were as described in the original ATAC-seq protocol (5). Following this first amplification, samples were enriched for smaller fragments using solid-phase reversible immobilization (SPRI) beads. Briefly, after PCR amplification, samples were mixed with SPRI beads (Beckman Coulter/Agencourt AMPure XP; no. A63881) at an SPRI to DNA ratio of 0.5 (25 μ L SPRI beads). After a 5-min incubation, the PCR plate was transferred to a magnet stand for separation, and the supernatant was transferred to a new 96-well plate. Additional SPRI beads were added to an SPRI to DNA ratio of 1.2 (65 μ L of SPRI beads). Samples were mixed at room temperature for an additional 5 min and then were transferred to a magnet stand for bead separation. The resulting supernatant was discarded, and the magnet-immobilized beads were rinsed twice with 80% ethanol. The ethanol-washed beads were airdried following the manufacturer's instructions. DNA was subsequently eluted in 20 μ L of 10 mM Tris·HCl, pH 8.0 (Integrated DNA Technologies; no.11-05-01-13). The size-selected DNA was subsequently amplified via a second nine-cycle PCR using the same unique 8-bp indexes per sample and the cycling conditions described in the original ATAC-seq protocol (5). Amplicons were purified using a standard kit (QIAGEN MinElute no. 28004).

The amplified, uniquely indexed samples were combined and normalized before pooling and then were sequenced to a target of 25 million reads per sample. Pools were run on an Illumina NextSeq 75-cycle high-output flow cell (paired-end, $2 \times 40+8$ bp index) and HiSeq high-output flowcell (paired-end, $2 \times 50+8$ bp index). To minimize group effects, samples were chosen so that each sequencing pool included mice from different experimental groups. Samples that were under-sequenced were repooled for subsequent runs.

Computational pipeline. The RIESLING pipeline was developed to process and analyze ATAC-seq data. RIESLING is a Pythonbased, open-source (license: MIT) pipeline, available at

[https:// github.com/GordonLab/riesling-pipeline](https://github.com/GordonLab/riesling-pipeline). The pipeline was designed for de novo prediction of putative enhancer and superenhancer regions from either single or multiple samples, with the option of including background or control samples (if available) for comparison. The full pipeline, detailed in Fig. S2, consists of mapping paired-end sequence data to host genomes (using Bowtie 2), mitochondrial read filtering (a nuisance in ATACseq, given the open nature of mitochondrial DNA), removal of identical PCR duplicate reads, quality filtering (using adjustable parameters, with a default mapping quality ≥ 10), and hotspot/ blacklist masking (from the ENCODE and modENCODE consortia, with varying availability depending on the host organism, or from internally generated experimental hotspots).

Filtered data were used to perform multistage enhancer prediction, beginning with peak calling, in which regions with signal higher than background (peaks) were identified using a standard Poisson modeling technique from ChIP-seq [user adjustable, with a default Model-based Analysis of ChIP-Seq (MACS) (60) and a significance cutoff of $P < 1 \times 10^{-9}$]. A window around known transcription start sites was subsequently masked because of their generally low signal-to-noise ratio in ATAC-seq data (user adjustable, with a default masking of ± 2.5 kb), and the raw, sequenced read depth underlying the identified peaks was computed and summarized using bamliquidator (<https://github.com/bradnerlab/pipeline>). Peaks were additionally grouped using one of three strategies to merge adjacent peaks: (i) unstitched (only the raw peaks without merging); (ii) fixed stitching, in which peaks within a user-definable window (generally 12.5 kb) are merged to represent one potential large super-enhancer; or (iii) dynamic stitching, in which the rate of change of peaks is computed within a window, and the window size is slowly increased to determine an inflection point. Using these three different approaches to aggregating peaks, predicted enhancers were graphed using the aggregate signal underlying each enhancer region. Super-enhancers can be stratified from traditional enhancers by drawing a tangent line to this graph and placing the cutoff where the slope of this line equals one (this distinction is somewhat arbitrary; see refs. 10 and 11).

Differential accessibility analysis. Differentially accessible regions of chromatin were identified and classified by applying a negative-binomial model (DESeq2) (61) to counts data out-

put from the RIESLING pipeline. The RIESLING pipeline was used to preprocess (filter, blacklist, de-replicate reads, call peaks, and so forth) data from each sample, compute the raw read depth, and then generate a count matrix (read depth per locus in each sample) for input to DESeq2. (Users also can input these data to alternative analytical packages, including edgeR and baySeq.) Peaks were selected for inclusion in the counts matrix if they were identified as statistically significant (based on the MACS Poisson model described above) in any input sample; any overlapping peaks were merged. Because adjacent peaks are likely to represent the same enhancer regions, RIESLING automatically expands peaks to include nearby peaks within 50-bp, 500-bp, and 1,000-bp windows (in addition to the unmerged raw data). DESeq2 is then invoked with the default parameters and run at both the genome position level (mm10) and at gene level for comparison (binning peaks and their associated read depth by proximal gene). For classifying microbiota-responsive enhancers, we considered groups to be significantly different with $P < 0.05$.

Normalization. Numerous normalization approaches were compared to determine the most appropriate choice. For enhancer and super-enhancer characterizations (nondifferential analyses), we used bamliquidator to produce normalized scores of the ATAC-seq signal. This approach takes each called peak, combined with the underlying raw sequencing data, and computes the raw read counts under each peak, divided by the total sequencing depth and the peak width.

For differential accessibility analyses, we compared results obtained from DESeq2 when using raw counts data, counts data aggregated by nearby peaks, or counts data following bamliquidator normalization to sequencing parameters (which may violate assumptions of DESeq2's median normalization model). Comparing these three options revealed that providing raw counts data to DESeq2 generated the most conservative estimates of differential accessibility (i.e., the fewest number of differentially accessible regions); therefore we selected that method for our analyses.

DESeq2's standard normalization approach uses counts data from all peaks and applies a median-of-ratios method described in ref. 61 to predict unchanging (pseudoreference) regions. We also considered guiding this median normalization by providing predictions of unchanged (consti-

tively accessible or condensed) chromatin regions, i.e., de novo regions chosen from overlapping peaks from our input data and available peaks from public ENCODE datasets. Our de novo peaks were derived from those shared across groups ($GF \cap \text{CONV-R} \cap \text{CONV-D}$) and in intersections of cell types ($CD4+ \cap CD8+$, $\text{TCR } \alpha\beta+ \cap \text{TCR } \gamma\delta+$ IELs \cap PBMC) and at varying overlap percentages (10, 50, or 90% bp overlap). In these permutations, DESeq2 performed either comparably to or worse than (i.e., less conservatively, calling more regions as differentially accessible) simply applying the standard median-of-ratios method to all the input counts data. Using the available mouse ENCODE data (62) resulted in normalization effects similar to those seen using all raw counts data, perhaps because the original mouse ENCODE data (i) were generated from heterogeneous tissues not directly comparable to our datasets and (ii) used ChIP-seq. In light of these alternatives, the data in our study reflect the standard DESeq2 median-of-ratios method applied to the raw input counts data. Pathway analyses. GeNets (<https://apps.broadinstitute.org/genets>) was used to identify clusters of functionally associated genes and to group genes by community. The GeNets random forest classifier was based on the GeNets Metanetwork 1.0 training database, which encompasses InWeb3, ConsensusPathDB, and known drug–target interactions. If multiple enhancers mapped to a single gene, they were coalesced into one gene for GeNets analysis. Analyses and community predictions were restricted to the top 250 enhancer-associated genes ranked by ATAC-seq signal (for super-enhancers) or by P value (when comparing two different colonization states). GeNets community enrichment significance is based on a Bonferroni-adjusted hypergeometric test. The GeNets empirical P values for network connectivity are based on gene network density. The number of edges (interactions between genes) out of the total number of possible edges (the density) is computed for a given input network as well as for randomly sampled sets of genes; the density is considered significant if it is greater than that of 95% of the randomly sampled gene sets.

Gene overrepresentation analyses were performed GREAT (18). GREAT 3.0 was run on predicted enhancer locations (provided from RIESLING) and using the mm10 mouse genome annotation. For identifying proximal genes, the configured parameters of “basal plus extension” were used (with parameters left at their default values for gene regulatory domain definitions). Back-

ground regions were provided where applicable (to remove potential biases from over- or under-represented regions from our original peak calling), or the whole mm10 genome was used where appropriate. DESeq2 differential accessibility data were filtered at $P < 0.05$ for use by GREAT. Additional pathway and overrepresentation analyses were performed using IPA (QIAGEN). IPA was provided log ratios, P values, and the NCBI ID of the nearest gene for each predicted enhancer. Analyses were run using only the IPA data from “Mouse restricted to Experimentally Observed data from Epithelial Cells and Immune Cells.”

Analysis of TF Regulatory Circuitry. Network analysis was performed using the Coltron software package (34) available at <https://pypi.python.org/pypi/coltron>. In each IEL sample, we identified all dynamically stitched super-enhancer regions associated with TFs. These super-enhancer-associated TFs became the cohort of nodes for the network analysis. Edges were defined as predicted TF-binding sites within ATAC-accessible regions found inside super-enhancers. Thus, for any given TF_i, connectivity could be described as how it is regulated itself (IN degree) and how it regulates the super-enhancers of other TF nodes in the network. IN degree and OUT degree are normalized from 0 to 1 where 0 indicates that no TF-binding sites are predicted, and 1 indicates that all TFs have a predicted binding site. Multiple TF-binding sites in a super-enhancer from the same factor are considered a single edge in the degree calculations. Enriched TF-binding sites were predicted using the FIMO algorithm (63) trained on TF position weight matrices defined in the TRANSFAC database (64). A false-discovery rate cutoff of 0.01 was used to identify enriched TF-binding sites.

Supplemental figure legends

Fig. S1.

FACS sorting strategies and reproducibility of ATAC-seq datasets. (A) Gating strategies for both IEL populations. Live TCR $\alpha\beta^+$ and TCR $\gamma\delta^+$ IELs from the small intestine were sorted simultaneously. (B) Venn diagrams showing the total number of putative enhancers within and between the four cell types studied. (C) ATAC-seq read counts per predicted enhancers identified from each cell preparation were normalized using a negative binomial model (via DESeq2). Pairwise comparisons were then performed within a given treatment group. Pearson correlation values (mean \pm SEM) are plotted for each group.

Fig. S2.

Overview of the RIESLING pipeline. The RIESLING pipeline was developed to characterize enhancer regions in ATAC-seq data rapidly and to generate output for processing with additional tools (e.g., DESeq2, edgeR, baySeq, and others) to identify differentially accessible regions of chromatin. (A) Input to this pipeline is de-multiplexed paired-end sequencing reads. (B–D) Reads are mapped to a host genome (via bowtie2) (B) and then are filtered for mitochondrial reads and ENCODE blacklists (65) (C) and de-duplicated (D). (E) Peaks are called using a user-selectable peak caller (60). (F) The preprocessed data undergo signal aggregation using bamliquidator. (G and H) RIESLING internally filters transcriptional start sites (TSS) and promoters (G) and then predicts enhancer regions using one of three peak stitching approaches: unstitched (raw peaks); dynamic stitching, based on the rate of change of enhancer prediction vs. a growing window; or fixed stitching in which peaks within a user-selectable parameter (generally 12.5 kb) are merged (H). (I) These regions are then ranked by signal (normalized read depth), and super-enhancer (SE) regions are separated from traditional enhancers. (J and K) RIESLING calls enhancer-proximal genes based on user-specified parameters (J) and prepares a counts matrix for differential accessibility analysis (K). (L–N) For differential accessibility analysis, the output of RIESLING is used as input to DESeq2 to generate quality control graphs (L) and differential accessibility analyses

(M), including optional pathway and overrepresentation analyses (N). For additional details of the pipeline and source code availability, see SI Materials and Methods, Computational pipeline.

Fig. S3.

The enhancer landscape of TCR $\alpha\beta^+$ and TCR $\gamma\delta^+$ IELs and peripheral CD4⁺ and CD8⁺ T cells, independent of colonization status. (A) Hierarchical clustering of enhancers identified in ATAC-seq datasets shows clear separation of IEL populations from circulating T cells. Clustering was performed on Jaccard distance metric scores between peaks from a given cell type (shown in blue shading where a value of 1.0 reflects complete overlap). (Left) All enhancers. (Right) Super-enhancers. (B) Distribution of enhancer signals in TCR $\alpha\beta^+$ IELs and TCR $\gamma\delta^+$ IELs, demonstrating a hockeystick asymmetric distribution. Super-enhancers were stratified from traditional enhancers by the point on the exponential curve where the slope approaches 1. The vertical dashed line represents the cutoff between traditional and super-enhancers. A subset of super-enhancers and their adjacent genes is highlighted. (C) Distribution of enhancer signals in CD4⁺ and CD8⁺ T cells. Super-enhancers with the top-ranked signal intensity are annotated with their most proximal gene. (D) The Bach2 super-enhancer overlaid on normalized sequencing data (reads per kilobase per million reads, within a 50-bp bin) from the TCR $\alpha\beta^+$ and TCR $\gamma\delta^+$ IEL populations. Genomic information for Bach2 and its neighbors are from mm10 RefSeq.

Fig. S4.

Predicted GeNets communities from both IEL lineages. (A) A network and community clustering analysis performed using GeNets, applied to the top 250 super-enhancer-associated genes shared by TCR $\alpha\beta^+$ and TCR $\gamma\delta^+$ IELs, independent of colonization status. Nodes represent genes, and edges (lines) represent gene-gene interactions. Note that there is significant connectivity between networks ($P < 0.005$; permutation test of nodes). See Datasets S6 and S7 for an alternative GREAT-based analysis of pathways enriched in the various purified cell populations independent of colonization status. (B) Each individual community is expanded here for both TCR $\alpha\beta^+$ and TCR $\gamma\delta^+$ IELs, with each node representing a super-enhancer-associated gene and each edge representing a

gene–gene interaction (based on a network including ConsensusPathDB and InWeb3; see SI Materials and Methods, Pathway analyses). A description of each gene and analyses of overrepresented pathways in each community are available in Dataset S3.

Fig. S5.

Enhancer populations distinguish IEL lineages. (A) The dynamic stitching model from RIESLING applied to the IEL population (independent of colonization status), demonstrating a window of 2.5 kb for grouping together nearby predicted enhancers. (B) Enhancers distinguishing lineages independent of host colonization status. (C) Enhancers distinguishing TCR $\alpha\beta^+$ IELs as a function of colonization status. (D) Enhancers distinguishing TCR $\gamma\delta^+$ IELs as a function of colonization status. (E) Histograms representing the distribution of fold change (\log_2) in dynamically stitched enhancers identified as statistically significant to our stringent threshold ($P < 5 \times 10^{-5}$) in GF vs. CONV-R TCR $\alpha\beta^+$ and TCR $\gamma\delta^+$ IELs. (F) A histogram of fold change (\log_2) of dynamically stitched enhancers identified as statistically significant ($P < 5 \times 10^{-5}$) in CD4⁺ vs. CD8⁺ T cells.

Fig. S6.

Comparison of gut microbial community structure in CONV-R and CONV-D mice. Bacterial 16S rRNA analysis was performed on fecal samples obtained when 8-wk-old mice were killed. CONV-D animals were colonized at 3 wks of age. The fecal microbiota of CONV-D and CONV-R animals had similar alpha-diversity with no significant differences detected in Faith's phylogenetic diversity ($P = 0.395$, ANOVA), Shannon's diversity ($P = 0.301$, ANOVA), or the number of observed 97% ID OTUs that survived our initial filtering and rarefaction ($P = 0.646$, ANOVA). (A) Heatmap of the relative abundances of the most strongly associated OTUs for the CONV-D and the CONV-R groups of mice, as judged by the differences in their indicator values. Relative abundance is calculated as the base 10 log of the number of reads (plus one) per 5,000 reads. (B) Phylogenetic tree showing evolutionary relationships between the V4-16S rRNA sequences of the significant indicator OTUs. OTUs in red were positively associated with CONV-R mice, and OTUs in blue were positively associated with CONV-D mice. An approximately-maximum-likelihood phyloge-

netic tree was constructed using the FastTree method implemented in QIIME. PERMANOVA revealed significant differences in the phylogenetic composition and structure of the fecal microbiota of CONV-D compared with CONV-R mice based on a number of metrics [unweighted UniFrac distances ($P = 0.001$, $R^2 = 0.422$); weighted UniFrac distances ($P = 0.001$, $R^2 = 0.571$); and binary Jaccard dissimilarities ($P = 0.001$, $R^2 = 0.298$)].

Fig. S7.

Track graphs comparing normalized median ATAC-seq signals obtained from analysis of TCR $\alpha\beta^+$ IELs purified from CONV-R and GF mice. Tracks represent the normalized signal at enhancers near Stat1 and Stat3 and demonstrate the increased signal at Stat3 in GF TCR $\alpha\beta^+$ IELs compared with CONV-R TCR $\alpha\beta^+$ IELs ($P < 0.05$, Benjamini–Hochberg-adjusted Wald test). Binding motifs for a subset of TFs implicated in the IEL TF regulatory circuitry are annotated; outlined in red are the JUN/FOS and MAFK bZIP TFs that have binding sites in Stat3 but not in Stat1. Both TCR $\alpha\beta^+$ and TCR $\gamma\delta^+$ IELs show enrichment in GF mice compared with CONV-R mice at a second (intronic) enhancer ($P < 0.05$, Benjamini–Hochberg-adjusted Wald test). No statistically significant differences were noted between ATAC-seq signals from enhancers in Stat1 when comparing the IEL populations from GF and CONV-R animals. For a more global view of differential accessibility, see Fig. S5.

Supplemental figures

Figure S1.

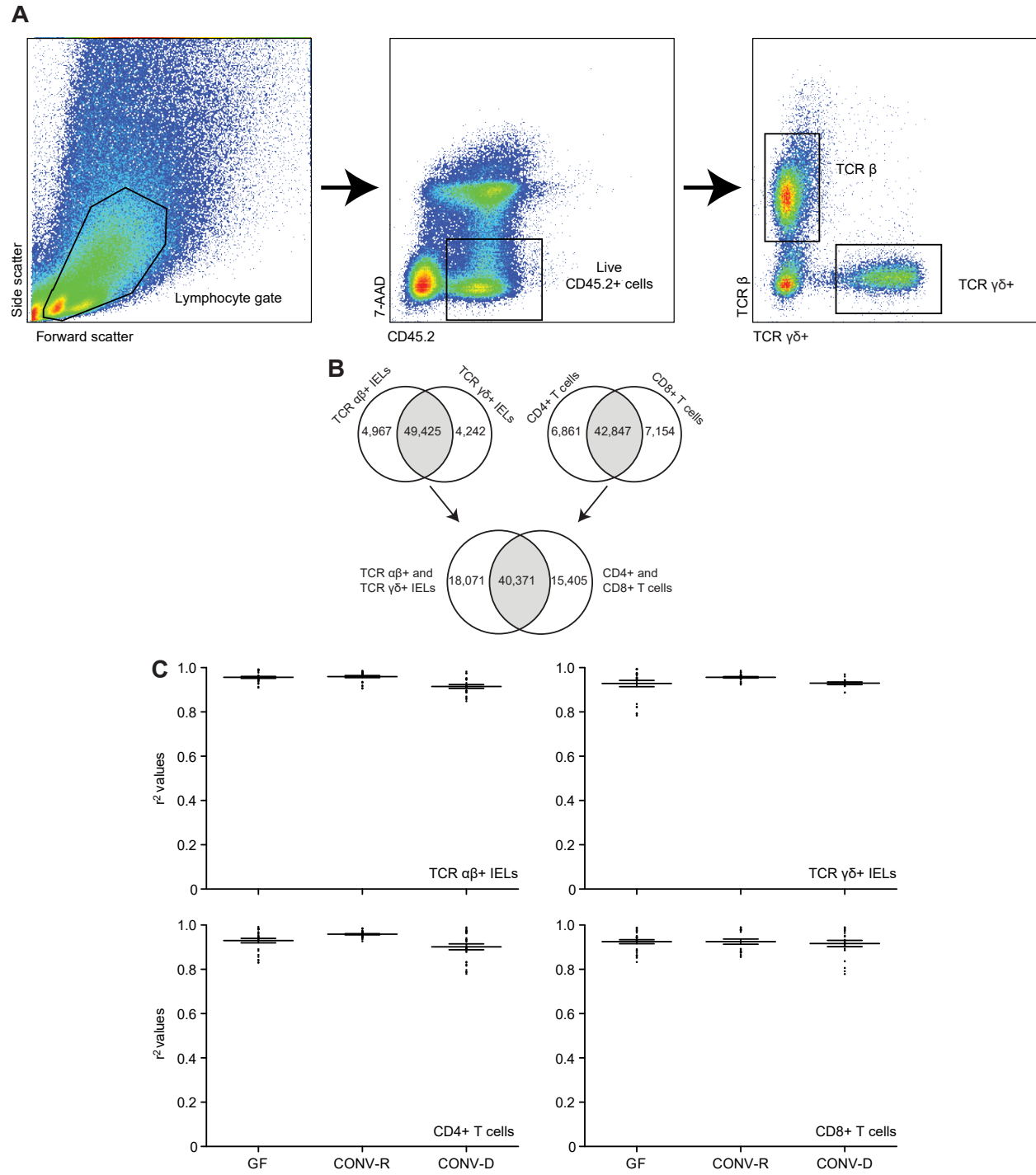


Figure S2.

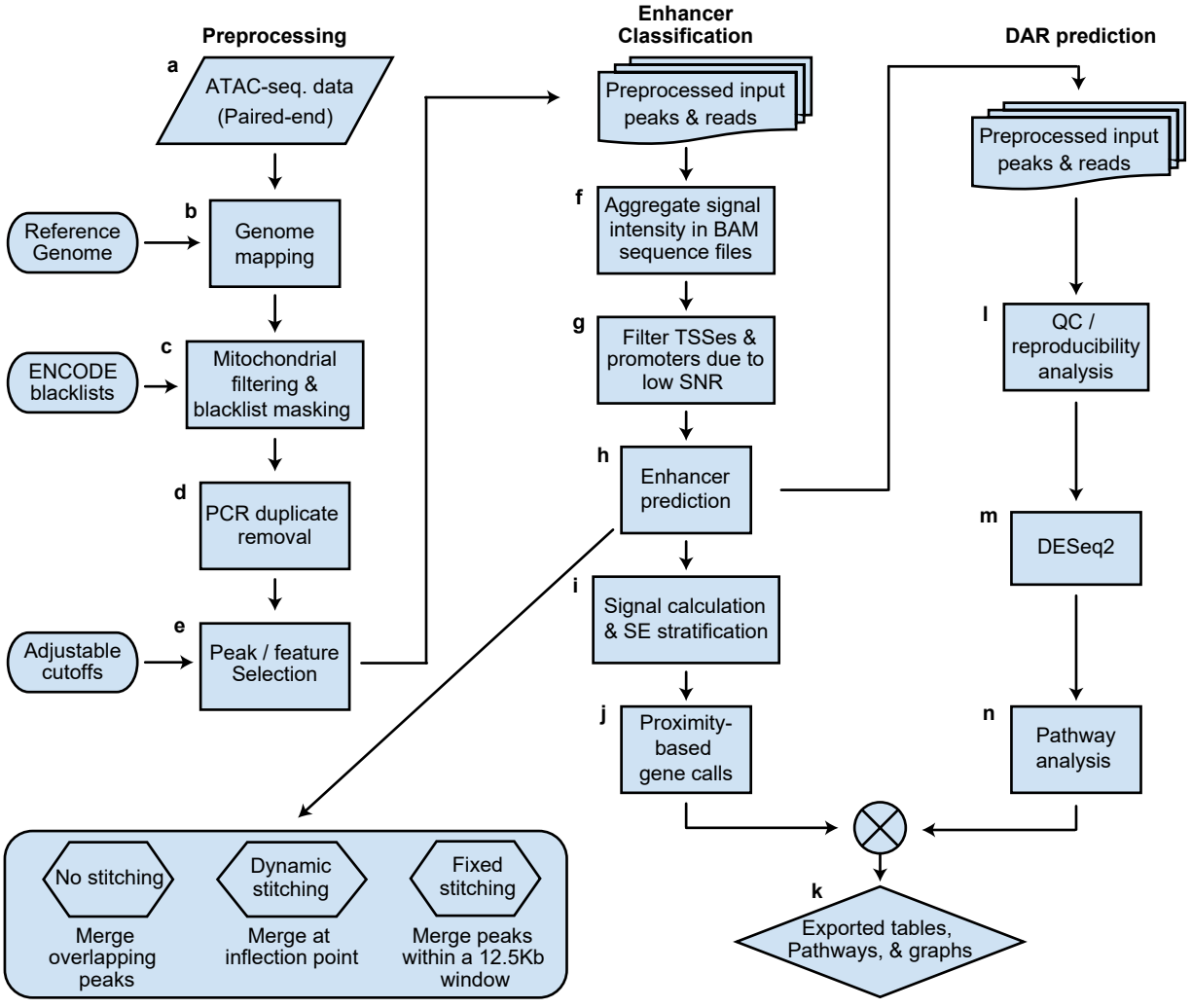


Figure S3.

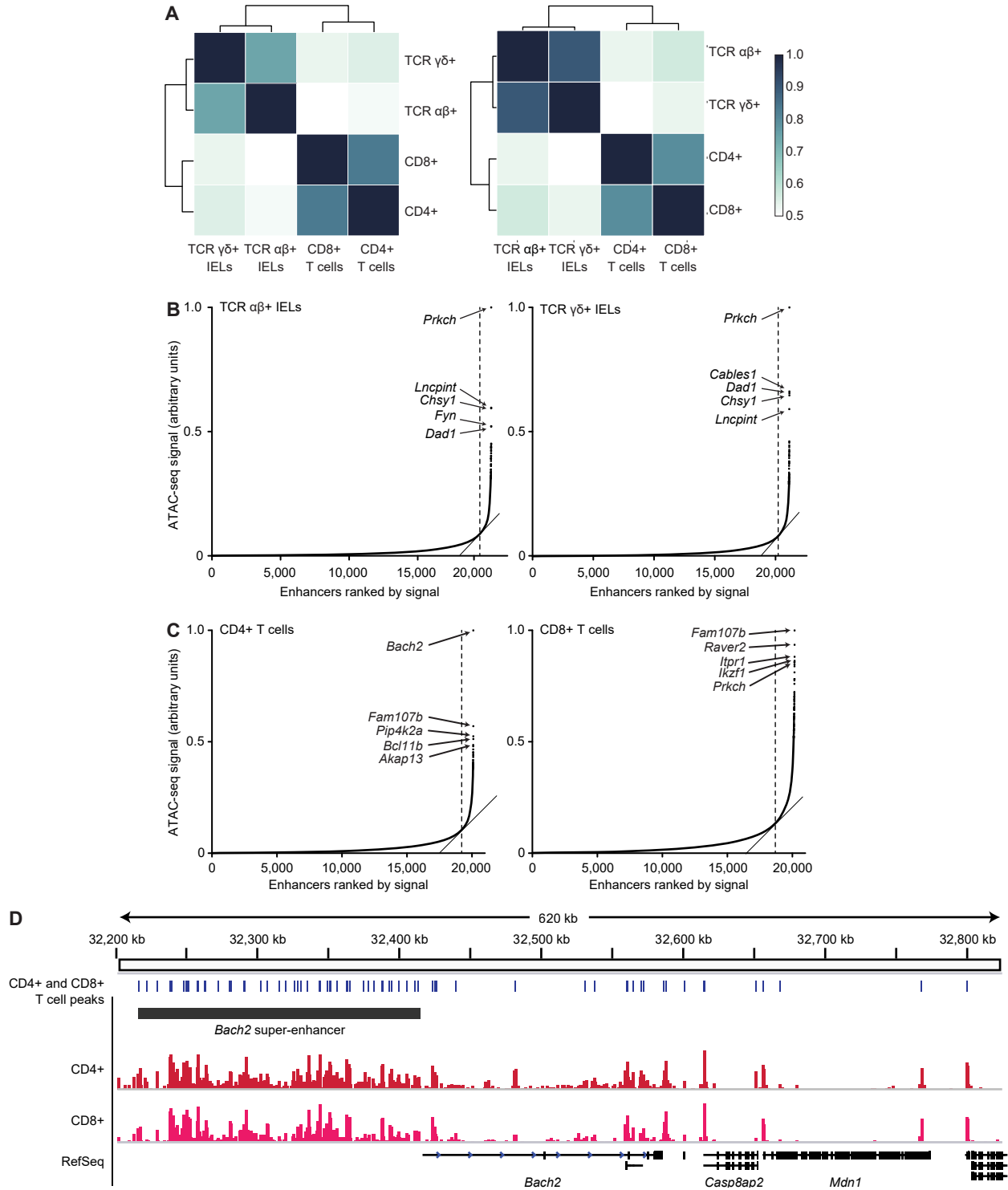


Figure S4.

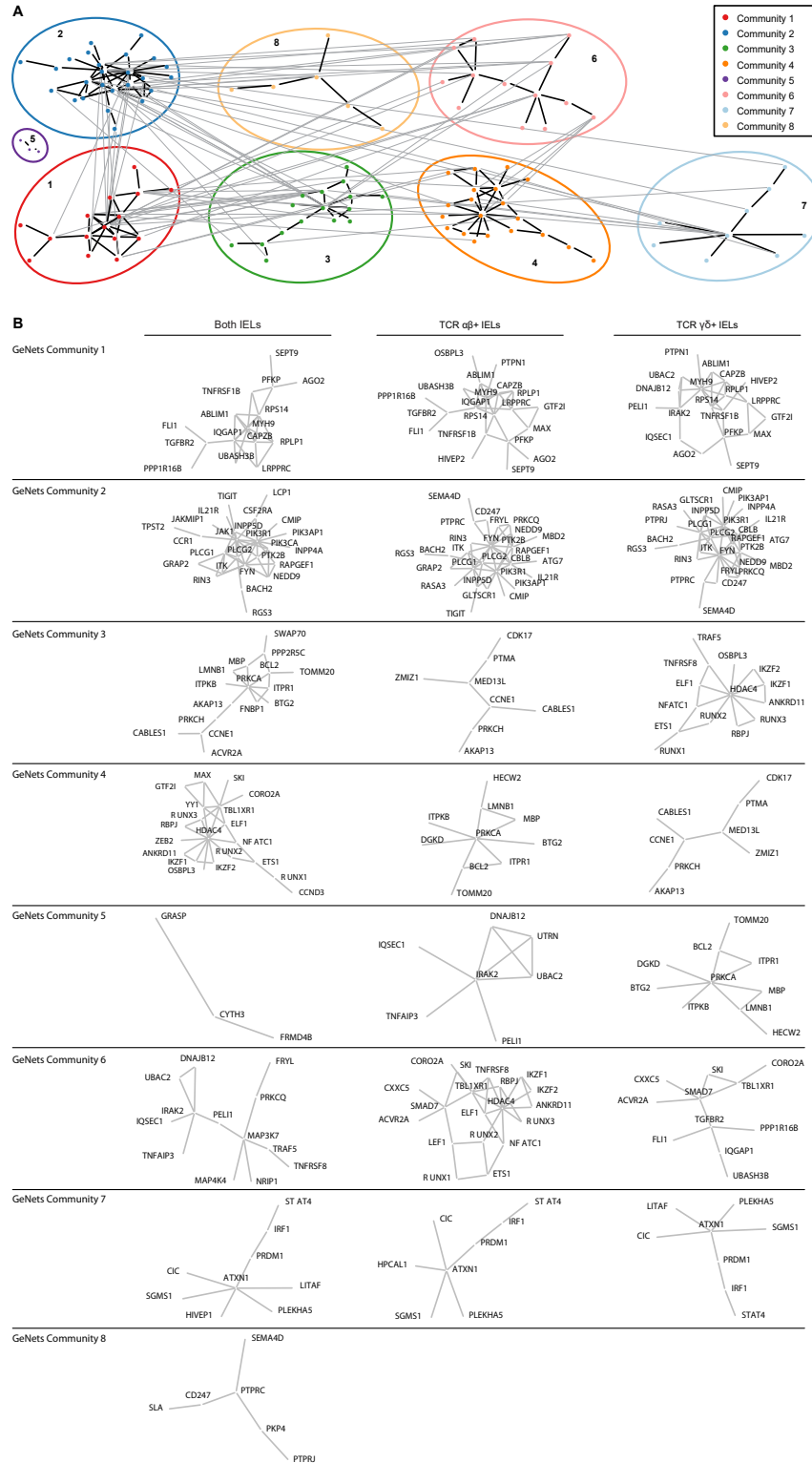


Figure S5.

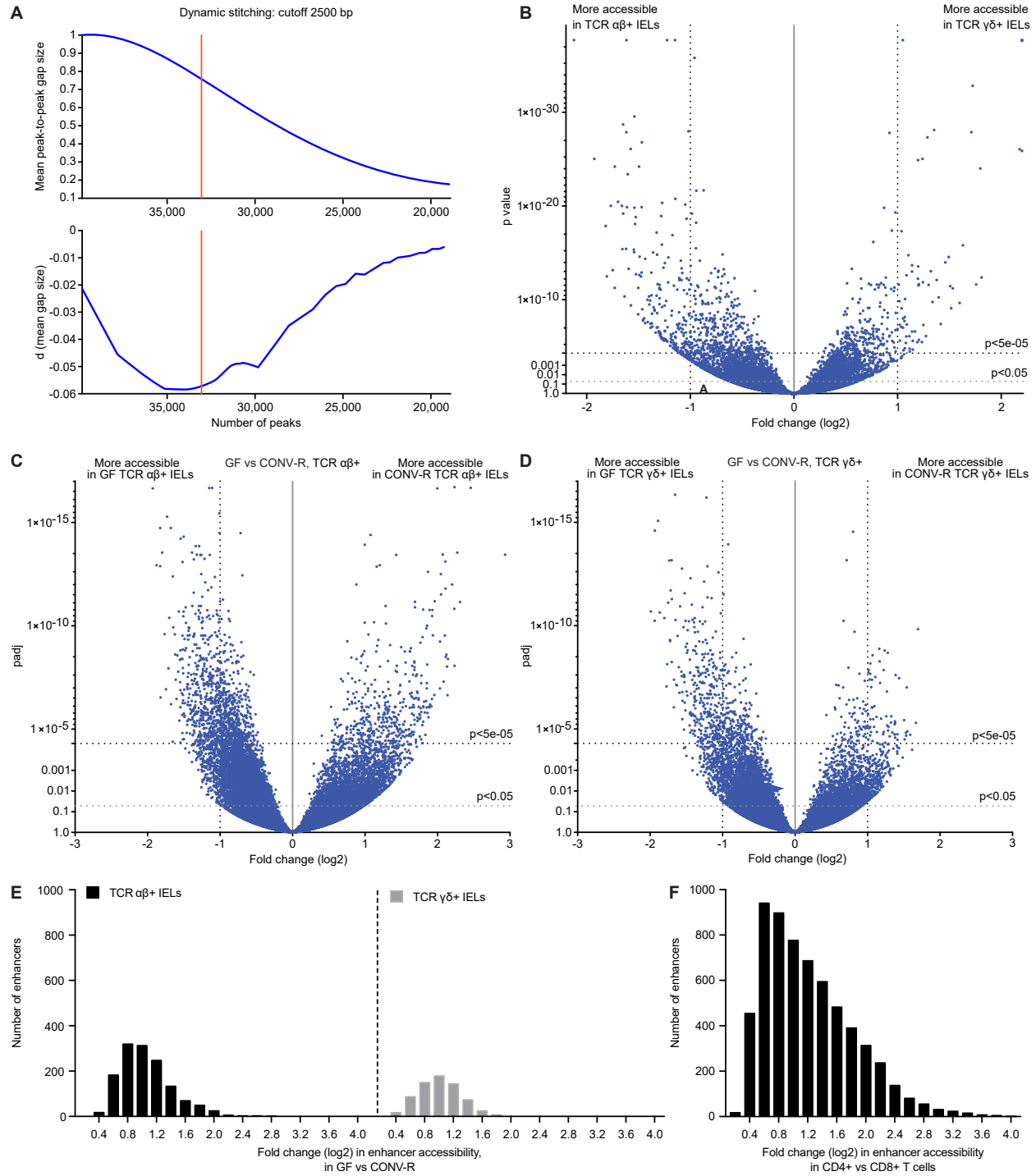
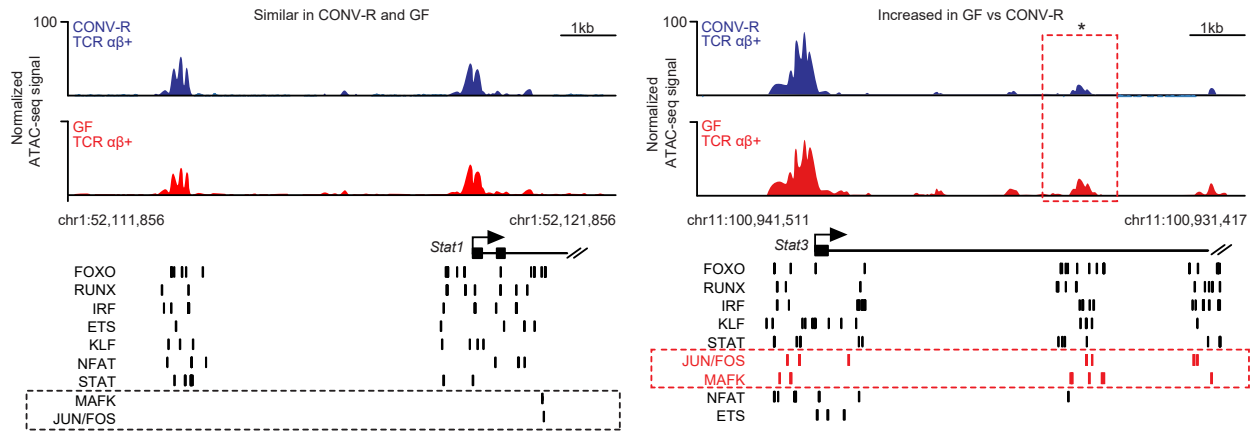


Figure S7.



Chapter 3

Future Directions

Future Directions

This thesis was enabled by the bleeding edge of technology. At the start of these experiments, and during an unpublished proof-of-concept experiment, ATAC-seq was nascent. It was unclear if the technology could be adapted from its original paper (which looked primarily at human CD4+ T cells) to subpopulations of mouse immune cells, and used to call putative enhancer loci. My studies suggest one potential mechanism by which the microbiota induces durable effects in its host. Several follow-up studies are warranted.

First, although ATAC-seq is in part empowered by its histone-agnostic approach to chromatin characterization, detailed data on individual histone modifications would be very valuable in determining which specific modifications may underlie changes in chromatin accessibility. At the start of this project, this wasn't feasible with such small cell populations, though some recent technologies (iChIP (Lara-Astiaso et al. 2014)) may be of value when applied to these and/or other populations. iChIP is an index-first modification to ChIP-seq that could potentially assay numerous histone marks in small populations of cells. Unfortunately, in a small collaboration with my colleague David Russler-Germain, we were unable to reproduce this technique. Since then, other techniques have been published aiming to enable ChIP-seq and related techniques using small numbers of cells, including Mint-ChIP (a multiplexed ChIP-seq-based approach employing linear amplification (van Galen et al. 2015)), scM&T-seq (parallel, single-cell parallel sequencing of the transcriptome and methylome (Angermueller et al. 2016)), and single-cell versions of DNase-seq and ATAC-seq (Jin et al. 2015; Buenrostro et al. 2015; Cusanovich et al. 2015). An alternative approach would be to look even more globally at chromatin architecture. One novel method (ChromATin, Linhoff et al. 2015) takes this global view, and provides 3D chromatin architecture in large and structurally complex tissues. This technique, which uses high-resolution imaging combined with fixation and immunostaining, could potentially be applied across the layers of the small intestine, where marked differences in chromatin landscape may exist between the same cell type localized at the junction between the epithelium and underlying lamina propria, or deeper within

the submucosa (e.g., tissue-resident macrophages or neurons traversing the intestinal layers may show markedly different chromatin compaction or localization of specific histones).

Secondly, this work would be enhanced by exploring potential pharmacologic or dietary components that influence enzymes and pathways responsible for microbiota-mediated host chromatin remodeling and histone modification. Recent work has suggested that high dietary fat intake can induce significant changes in hepatic chromatin landscape (assayed by FAIRE-seq; a common formaldehyde cross-linking technique that is similar to DNase-seq in 4–6 week-old C57BL/6J and DBA/2J mice fed a standard 10% fat/kcal vs 30% fat/kcal diet for eight weeks) and related transcriptional pathways (via hepatic RNA-seq) (Leung et al. 2014).

It would be fascinating to apply this approach to the study of populations suffering from malnutrition. The literature on epigenetic impacts of malnutrition is complex, and largely focused on either single nutrient deficiencies, or on broad studies of methylation. For example, one recent study identified significant perturbations (generally hypomethylation) in the mouse sperm methylome following parental undernutrition (Radford et al. 2014). In one of the few studies of undernourished humans, whole peripheral blood showed differential methylation at numerous loci — including some linked to the insulin receptor — in individuals with prenatal exposure to the Dutch Hunger Winter (at the end of World War II) (Tobi et al. 2014). It could be extremely informative to apply the approaches outlined in this thesis to the gnotobiotic mouse model described in the Appendix, where animals were colonized with culture collections generated from the fecal microbiota of Bangladeshi children with severe acute malnutrition or with healthy growth phenotypes, fed a diet representative of those consumed by the donors. In these studies, weight loss phenotypes with associated alterations host energy metabolism were dependent on the presence of an enterotoxigenic strain of *Bacteroides fragilis* (ETBF). The weight loss phenotype could be prevented by the presence of a non-toxigenic strain, was diet-dependent, and could be transmitted across generations. We used ATAC-seq to characterize the chromatin landscape in splenic CD4⁺ and CD8⁺ T cells harvested from these mice. Although we noted few statistically significant changes between mice harboring the stunted donors ETBF(+),NTBF(-) culture collection versus the healthy donors

NTBF(+),ETBF(-) culture collection, our work could be limited by our choice of splenic T cells, which was restricted by sample availability. A logical follow-up series of experiments would be to generate ATAC-seq profiles from IELs or other cell populations exposed more directly to the microbiota and its products (e.g., hepatic Kupffer cells).

Other intergenerational models of malnutrition are also ripe for investigation, including one of a collaborator who has identified that maternal undernutrition in mice can result in metabolic disease and insulin resistance in offspring (reviewed in Patti 2013). It is possible that this intergenerational phenotype reflects microbial contributions to the host, either during *in utero* development or from postnatal colonization.

Finally, it will be of value for future researchers to expand the computational platform and associated visualization tools developed in this thesis. Since this study began, the broad field of epigenetic profiling has expanded to include more expression data, frequently by parallel RNA-seq of samples. Integration of RNA-seq data into the RIESLING pipeline for enhancer prediction could be helpful, however not without potential complications. RNA recovered from cells after their purification may not accurately reflect the features of gene expression *in vivo*. Notably, the state of chromatin can be more indicative of cell identity and cell state than gene expression data (Stergachis et al. 2013), and chromatin can encode functional capabilities that are not reflected in RNA (Samstein et al. 2012; John et al. 2011). Additionally, as the epigenetic field has expanded, many studies now focus on even smaller cell populations, including the use of single-cell ATAC-seq. Expanding this software toolkit to analyze single-cell experiments could be of great value to future research, and would help standardize what can be otherwise piecemeal analytical approaches.

References

- Angermueller, C. et al., 2016. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*.
- Buenrostro, J.D. et al., 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561), pp.486–90.
- Cusanovich, D.A. et al., 2015. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237), pp.910–4.
- Jin, W. et al., 2015. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature*, 528(7580), pp.142–6.
- John, S. et al., 2011. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics*, 43(3), pp.264–8.
- Lara-Astiaso, D. et al., 2014. Chromatin state dynamics during blood formation. *Science*, 345(6199), pp.943–9.
- Leung, A. et al., 2014. Open chromatin profiling in mice livers reveals unique chromatin variations induced by high fat diet. *The Journal of Biological Chemistry*, 289(34), pp.23557–67.
- Linhoff, M.W., Garg, S.K. & Mandel, G., 2015. A High-Resolution Imaging Approach to Investigate Chromatin Architecture in Complex Tissues. *Cell*, 163(1), pp.246–255.
- Patti, M.-E., 2013. Intergenerational programming of metabolic disease: evidence from human populations and experimental animal models. *Cellular and Molecular Life Sciences*, 70(9), pp.1597–608.
- Radford, E.J. et al., 2014. In utero effects. In utero undernourishment perturbs the adult sperm methylome and intergenerational metabolism. *Science*, 345(6198).
- Samstein, R.M. et al., 2012. Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. *Cell*, 151(1), pp.153–66.

Stergachis, A.B. et al., 2013. Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell*, 154(4), pp.888–903.

Tobi, E.W. et al., 2014. DNA methylation signatures link prenatal famine exposure to growth and metabolism. *Nature Communications*, 5, p.5592.

van Galen, P. et al., 2015. A Multiplexed System for Quantitative Comparisons of Chromatin Landscapes. *Molecular Cell*, 61(1), pp.170–80.

Appendices

Appendix A

Vitas E. Wagner, Neelendu Dey, Janaki Guruge, Ansel Hsiao, Philip P. Ahern, Nicholas P. Semenkovich, Laura V. Blanton, Jiye Cheng, Thaddeus S. Stappenbeck, Olga Ilkayeva, Christopher Newgard, William Petri, Rashidul Haque, Tahmeed Ahmed, & Jeffrey I. Gordon.

Effects of a gut pathobiont in a gnotobiotic mouse model of childhood undernutrition.

Sci Transl Med. 2016 Nov 23;8(366):366ra164.

MICROBIOME

Effects of a gut pathobiont in a gnotobiotic mouse model of childhood undernutrition

Vitas E. Wagner,^{1,2*} Neelendu Dey,^{1,2,3*} Janaki Guruge,^{1,2} Ansel Hsiao,^{1,2} Philip P. Ahern,^{1,2} Nicholas P. Semenkovich,^{1,2} Laura V. Blanton,^{1,2} Jiye Cheng,^{1,2} Nicholas Griffin,^{1,2} Thaddeus S. Stappenbeck,⁴ Olga Ilkayeva,⁵ Christopher B. Newgard,^{5,6,7,8} William Petri,⁹ Rashidul Haque,¹⁰ Tahmeed Ahmed,¹⁰ Jeffrey I. Gordon^{1,2†}

2016 © The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science.

To model how interactions among enteropathogens and gut microbial community members contribute to undernutrition, we colonized gnotobiotic mice fed representative Bangladeshi diets with sequenced bacterial strains cultured from the fecal microbiota of two 24-month-old Bangladeshi children: one healthy and the other underweight. The undernourished donor's bacterial collection contained an enterotoxigenic *Bacteroides fragilis* strain (ETBF), whereas the healthy donor's bacterial collection contained two nontoxigenic strains of *B. fragilis* (NTBF). Analyses of mice harboring either the unmanipulated culture collections or systematically manipulated versions revealed that ETBF was causally related to weight loss in the context of its native community but not when introduced into the healthy donor's community. This phenotype was transmissible from the dams to their offspring and was associated with derangements in host energy metabolism manifested by impaired tricarboxylic acid cycle activity and decreased acyl-coenzyme A utilization. NTBF reduced ETBF's expression of its enterotoxin and mitigated the effects of ETBF on the transcriptomes of other healthy donor community members. These results illustrate how intraspecific (ETBF-NTBF) and interspecific interactions influence the effects of harboring *B. fragilis*.

INTRODUCTION

Undernutrition is the leading cause of childhood mortality worldwide; it is not due to food insecurity alone but rather reflects a complex and incompletely understood set of interactions involving a variety of factors that operate within and across generations (1–4). Among these factors is the gut microbiota. A recent culture-independent study of fecal microbiota samples collected monthly from members of a Bangladeshi birth cohort with healthy growth phenotypes identified age-associated changes in the representation of bacterial species during the first 2 years of postnatal life (5). These age-discriminatory species together define a developmental program for the microbiota that is shared across biologically unrelated Bangladeshi infants and children (5). Children with moderate or severe undernutrition exhibit disruptions in this program, resulting in microbial communities that appear younger than those of chronologically age-matched healthy individuals (5, 6). Clinical studies have provided evidence that microbiota immaturity in children with severe acute undernutrition is not repaired when they are fed therapeutic food supplements (5). Transplantation of fecal microbiota from children with healthy growth phenotypes, or immature microbiota from chronologically age-matched undernourished children, into germfree

mice fed diets similar to those consumed by the human donors has provided preclinical evidence for a causal role of the gut community in disease pathogenesis (6–8). Young mouse recipients of undernourished donors' microbiota exhibit growth faltering, manifested in part by a reduced rate of lean body mass gain compared to recipients of healthy donors' microbiota. These differences are not associated with differences in food consumption (6). Adult mouse recipients of undernourished donor's microbiota exhibit a weight loss (wasting) phenotype (7, 8).

A large enteropathogen burden is associated with childhood undernutrition (9, 10). However, the effects of microbiota configuration on enteropathogen invasion and burden, or how community context influences the expressed properties of organisms classified as enteropathogens in children at risk for or already showing undernutrition, remain poorly understood. Identifying gut community configurations that are resistant to invasion or accommodate organisms currently classified as enteropathogens without supporting their ability to produce deleterious host effects has potential diagnostic and therapeutic implications.

Bacteroides fragilis provides a model for examining how community context, including intraspecific interactions involving different strains of a given species and interspecific interactions involving different species, affects the properties of an enteropathogen and its effects on the host. Enterotoxigenic *B. fragilis* (ETBF) strains contain one of three alleles of *B. fragilis* toxin (*bft-1*, *bft-2*, and *bft-3*) in their 6-kb pathogenicity island (BfPAI). The *bft* locus encodes a zinc-dependent metalloprotease, fragilysin, that perturbs gut barrier function through cleavage of E-cadherin at epithelial cell adherence junctions (11). ETBF strains cause diarrhea in children residing in Bangladesh (12, 13). Nontoxigenic *B. fragilis* strains (NTBF) lack *bft*. Thus, *B. fragilis* can be viewed as a potentially pathogenic symbiont (pathobiont) on the basis of the presence or absence of this gene, or more broadly, on the basis of its community context.

Here, we describe a gnotobiotic mouse model for analyzing how intraspecific and interspecific interactions determine the effects of ETBF. We started with fecal samples collected from Bangladeshi

¹Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63110, USA. ²Center for Gut Microbiome and Nutrition Research, Washington University School of Medicine, St. Louis, MO 63110, USA. ³Department of Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA. ⁴Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110, USA. ⁵Sarah W. Stedman Nutrition and Metabolism Center, Duke University Medical Center, Durham, NC 27710, USA. ⁶Duke Molecular Physiology Institute, Duke University Medical Center, Durham, NC 27710, USA. ⁷Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, NC 27710, USA. ⁸Department of Medicine, Duke University Medical Center, Durham, NC 27710, USA. ⁹Departments of Medicine, Microbiology, and Pathology, University of Virginia School of Medicine, Charlottesville, VA 22908, USA. ¹⁰International Centre for Diarrhoeal Disease Research, Bangladesh, Dhaka 1212, Bangladesh.

*These authors contributed equally to this work.

†Corresponding author. Email: jgordon@wustl.edu

children who were participants in a previously completed birth cohort study (14). Using anthropometric scores to define healthy growth versus stunting and wasting (15), and a polymerase chain reaction (PCR)-based assay that targeted *bft*, we selected two chronologically age-matched individuals: one markedly stunted, underweight, and ETBF-positive, and the other with a healthy growth phenotype who was ETBF-negative but NTBF-positive. We transplanted intact uncultured fecal microbiota samples, collected from these two children at 24 months of age, into adult germfree mice. These gnotobiotic mice were fed cooked diets containing ingredients embodying those consumed by the population from which the microbiota donors were selected. Finding that the severely stunted and underweight but not the healthy donor's microbiota transmitted a weight loss phenotype to recipient animals and subsequently to their offspring, we performed follow-up transplant studies using clonally arrayed collections of sequenced anaerobic bacterial strains cultured from the donors' fecal samples. These culture collections allowed us to dissect microbe-microbe and host-microbe interactions by testing the effects of several types of manipulations: (i) removing the ETBF strain from the stunted donor's cultured community, (ii) introducing the ETBF strain into the healthy donor's culture collection together with or in lieu of its own NTBF strains, and (iii) introducing the NTBF strain into the stunted donor's culture collection together with or in lieu of its own ETBF strain. Microbiota community structure and gene expression, microbial and host metabolism, and host weight and immune phenotypes were characterized as a function of these alterations.

RESULTS

Uncultured fecal gut microbiota from an underweight donor confers weight loss on gnotobiotic mice

We used anthropometric data collected from members of a birth cohort study (14) of 100 children living in Mirpur thana in Dhaka, Bangladesh, to define whether they were healthy or undernourished (table S1). Those with height-for-age *z* scores (HAZ) greater than or equal to -2 were classified as "healthy," whereas those with scores less than or equal to -3 were deemed severely stunted. At 18 months, 30 and 25 children satisfied these criteria for healthy and severely stunted, respectively, whereas at 24 months, 27 and 20 children received these designations; the remaining children were classified as moderately stunted (HAZ between -2 and -3). A PCR-based screen for ETBF targeting all three fragilysin gene subtypes (14) was performed using DNA isolated from fecal samples that had been collected from these children at 18 and 24 months of age. The results revealed that ETBF was variably present between individuals and within a given individual over time, with a total of 25% of 18-month-old and 14% of 24-month-old children having a positive test (table S1). In this small cohort, ETBF carriage was not significantly correlated with indices of linear or ponderal growth [HAZ, weight-for-age *z* score (WAZ), and weight-for-height *z* score (WHZ) measured at 12 and 24 months of age ($P = 0.8$ and $P = 0.4$, $P = 0.7$ and $P = 0.2$, and $P = 0.5$ and $P = 0.2$, respectively; two-tailed Student's *t* test)]. We combined anthropometric and PCR data to select fecal samples collected at 24 months from two children: (i) a healthy individual (child ID 7114 in table S1) with a HAZ score of -0.71 , a WAZ score of -1.49 , and a WHZ score of -1.62 who was ETBF-negative at the two time points tested, and (ii) a severely stunted and moderately underweight individual (child ID 7004) with a HAZ score of -3.02 , a WAZ score of -2.51 , and a WHZ score of -1.34 who was ETBF-positive at both time points. Of the 35 individuals with a positive ETBF test at

either time point, only this stunted/underweight child was positive at both 18 and 24 months of age. Fecal samples obtained from members of this singleton birth cohort were screened for parasites using microscopic methods (5); neither of the two donors tested positive (see Materials and Methods for details).

To define the effects of diet and these two children's gut microbiota on host biology, we generated three representative versions (embodiments) of the diets consumed by the population represented by the donors. To do so, we determined the relative daily caloric contributions of various selected ingredient types, based on a study by Arsenault and coworkers (16). Selection of specific food items as representative of each ingredient type was based on consumption incidence surveys tabulated by Islam *et al.* (17), and the results were incorporated into a database consisting of 54 food ingredients. We filtered this database to remove items consumed by $<20\%$ of households and categorized each of the remaining 39 items (see Materials and Methods for additional details). From the resulting diet ingredient matrix, we randomly sampled (without replacement) one item each from cereals, pulse vegetables, roots/tubers, leafy vegetables, fruits, and fish, plus three nonleafy vegetables, to populate three separate diet lists. Using the U.S. Department of Agriculture National Nutrient Database for Standard References (18), we determined the caloric information for each ingredient and subsequently calculated proportions required to match the predetermined contributions of each ingredient type. Food items were cooked in a manner intended to simulate Bangladeshi practices, and the resulting three embodiments of a Bangladeshi diet were sterilized by irradiation. This approach allowed us to generate several representative Bangladeshi diets that were not dominated by the idiosyncrasies of a single individual's diet or by our own biases. The composition and results of nutritional analysis of the three diet embodiments are described in table S2 (A and B). The nutritional requirements of mice and children are compared in table S2C.

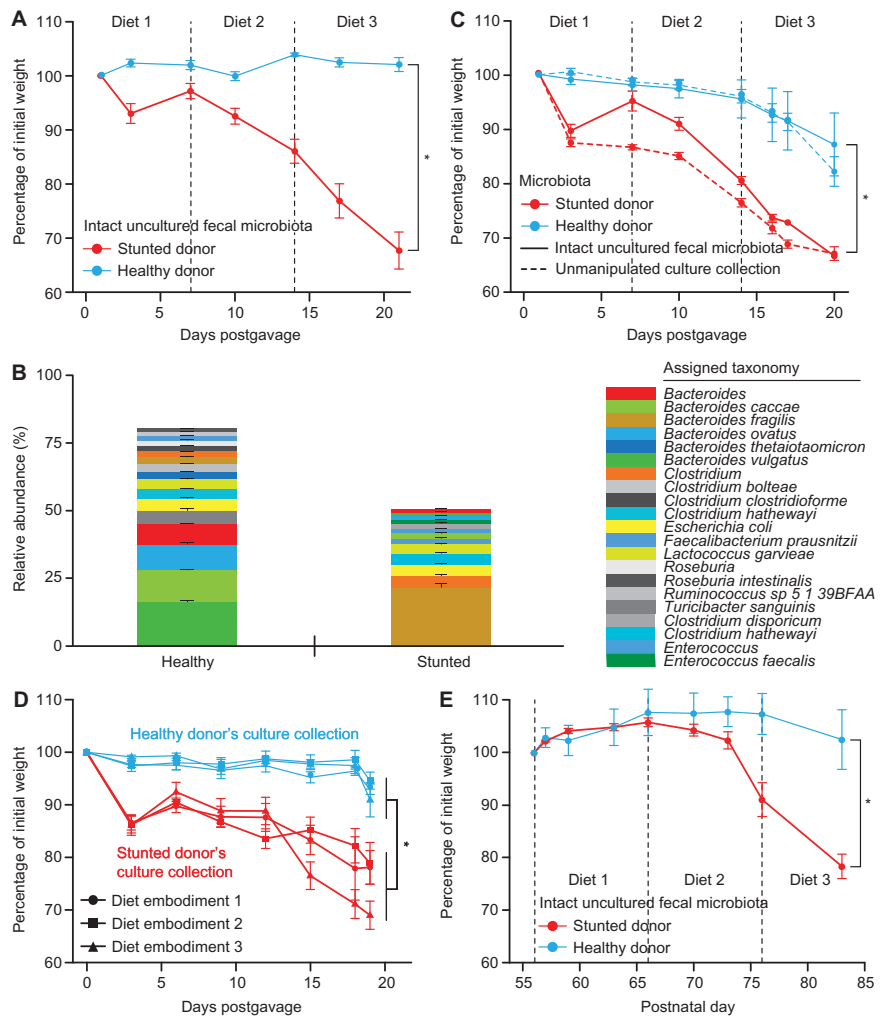
The results of a 12-year survey of demographic variations in the nutritional status of 16,278 Bangladeshi children found no significant sex differences in WHZ, WAZ, or HAZ (19). Therefore, in these and subsequent experiments, we eliminated gender as an experimental variable and only studied male mice. We gavaged separate groups of 8- to 9-week-old germfree C57BL/6 mice with the intact uncultured fecal microbiota samples obtained from the healthy or stunted/underweight Bangladeshi donors (two independent experiments; $n = 4$ singly caged mice per donor microbiota per experiment; see fig. S1A for study design). Fecal microbiota transplantation occurred 2 days after mice had been switched from an irradiated, nutritionally complete, low-fat/high-plant polysaccharide (LF/HPP) mouse chow that they had received since weaning to the first of the three embodiments of the Bangladeshi diet. Animals were subsequently fed, ad libitum, embodiment 1 for 1 week, followed by embodiment 2 for 1 week, and finally embodiment 3 for 1 week, with frequent sampling of their fecal microbiota during the course of each diet. Sequencing PCR amplicons generated from variable region 4 (V4) of bacterial 16S ribosomal RNA (rRNA) genes present in the donor's fecal sample and in fecal samples collected over time from recipient gnotobiotic mice (table S3) provided an *in vivo* assay of colonization efficiency for each human donor sample. 16S rRNA sequencing reads were grouped into operational taxonomic units (OTUs) on the basis of a threshold of $\geq 97\%$ nucleotide sequence identity (97% ID). The results revealed that at the conclusion of the experiment, $65.8 \pm 2.5\%$ (mean \pm SEM) of OTUs in the stunted/underweight donor's fecal microbiota sample and $68.4 \pm 8.8\%$ (mean \pm SEM) of the OTUs in the healthy donor's microbiota were detectable in

recipient mice (that is, each OTU had a relative abundance of $\geq 0.1\%$ in $\geq 1\%$ of fecal samples obtained from the animals).

Although gnotobiotic animals colonized with the healthy donor's intact uncultured fecal microbiota maintained weight, recipients of the severely stunted/underweight donor's intact uncultured fecal microbiota exhibited progressive and significant weight loss ($P < 0.005$, paired two-tailed Student's *t* test, comparison of final versus initial weights between the two treatment groups; Fig. 1A). In contrast to mice colonized with the healthy donor's microbiota, those that received the stunted donor's microbiota exhibited statistically significant weight loss at 10 days postgavage (dpg), during consumption of diet embodiment 2. Weight loss in this group worsened progressively, reaching

$31 \pm 6\%$ (mean \pm SEM) of original starting weight by 21 dpg ($P < 0.001$, two-tailed Student's *t* test, comparison of final weights; Fig. 1A); in a linear mixed-effects model, both dpg and the interaction between microbiota and dpg were significant factors affecting weight throughout the experiment ($P < 1 \times 10^{-7}$ for each). Food consumption was not different between the two treatment groups as their weight phenotypes diverged. The relative abundance of *B. fragilis*, defined by V4-16S rRNA analysis of fecal samples obtained at the time of killing, was significantly greater in mice colonized with the stunted/underweight donor's microbiota than in mice colonized with the healthy donor's microbiota ($P = 1.9 \times 10^{-6}$, two-tailed Student's *t* test; Fig. 1B).

Fig. 1. Intact uncultured human fecal microbiota and derived culture collections from healthy and undernourished Bangladeshi children transmit discordant weight phenotypes to gnotobiotic mice. (A) Germfree male C57BL/6 mice (8 to 9 weeks old) ($n = 8$ per treatment group) gavaged with intact uncultured fecal microbiota from Bangladeshi donors were fed a sequence of three embodiments of a representative Bangladeshi diet consumed by members of the donor population. See fig. S1A for experimental design. Mean weights (\pm SEM) as a function of dpg are shown as percentages of weights immediately before fecal microbiota transplantation. (B) Efficiency of capture of bacterial OTUs present in the donor's intact uncultured fecal samples in gnotobiotic mice. Mean relative abundances (\pm SEM) of 97% ID OTUs representing $\geq 1\%$ of the total fecal microbial communities in recipient animals. Results are based on V4-16S rRNA data sets and summarized at the species level (or genus when species could not be determined). OTUs present at lower abundances are not shown and account for the proportion not represented in each stacked barplot. (C) Transplantation of culture collections (dashed lines) generated from the fecal microbiota of the healthy or stunted/underweight donors recapitulated the discordant weight phenotype seen with the corresponding intact uncultured microbiota (solid lines) ($n = 6$ mice per treatment group, mean weights \pm SEM plotted). $*P < 0.05$ (paired two-tailed Student's *t* test and linear mixed-effects model, as above). (D) The weight-loss phenotype observed in recipients of the stunted/underweight donor's culture collection is not significantly different between the three Bangladeshi diet embodiments tested ($P > 0.05$; two-tailed Student's *t* test). Mean weights (\pm SEM) are plotted as a function of dpg ($n = 6$ mice per culture collection per diet embodiment). Significant weight differences were seen between mice colonized with the healthy donor's compared to the stunted/underweight donor's culture collection in the context of all three embodiments of the Bangladeshi diet. $*P < 0.05$ (paired two-tailed Student's *t* test and linear mixed effects model). (E) Intergenerational transmission of discordant weight phenotypes. See fig. S1C for experimental design. Mean weights (\pm SEM) of offspring of female gnotobiotic mice colonized with the indicated donors' microbiota ($n = 3$ to 4 mice per treatment group) are plotted as a function of age. Animals were switched from a nutrient sufficient LF/HPP mouse chow to embodiments of the Bangladeshi diets beginning on postnatal day 56. $*P < 0.05$ (tested by both paired two-tailed Student's *t* test comparing weights at killing and linear mixed effects model assessing interaction of weight, dpg, and microbiota through the experiment). The efficiency of intergenerational transmission of 97% ID OTUs was $96 \pm 1.8\%$ and $88 \pm 2.3\%$ (mean \pm SEM) for the healthy and stunted/underweight donor's microbiota, respectively (defined at the time of killing).



Bacterial culture collections from donor fecal microbiota transmit contrasting weight phenotypes

We next cultured bacterial strains from the healthy and stunted/underweight donors' fecal samples (20, 21). Each collection of cultured strains was clonally arrayed in multiwell plates so each well contained a monoculture of a given bacterial isolate (20). Each culture collection consisted of organisms that had coexisted in the donor's gut and thus were the products of the donor's history of environmental exposures to various microbial reservoirs (including those of family members and various enteropathogens endemic to the Mirpur *thana*), as well as the selective pressures and evolutionary events placed on and operating within their microbiota (for example, immune, antibiotic, dietary, and horizontal gene transfer). Individual isolates in the clonally arrayed culture collection were grouped into "strains" if they shared an overall level of nucleotide sequence identity of >96% across their assembled draft genomes (21). On the basis of this criterion and the results of sequencing amplicons generated from the isolates' 16S rRNA genes, we determined that the healthy and stunted donors' culture collections contained 53 and 37 strains, respectively. Only one strain was shared between the two culture collections: *Bifidobacterium breve* hVEW9 [see table S4 for a list of all isolates in the culture collection derived from the stunted/underweight child and (21) for details of the healthy donor's culture collection]. The two *B. fragilis* strains present in the healthy donor's culture collection (hVEW46 and hVEW47) lacked a BfPAI and were therefore classified as NTBF. The stunted donor's collection contained a single *B. fragilis* strain (mVEW4) with a *bft-3* allele. ETBF strains of this type are globally distributed but most common in Southeast Asia (22) (see table S5 for a comparison of the functions encoded by genes in the genomes of these ETBF and NTBF strains and the reference *B. fragilis* type strain ATCC 25285).

To ascertain whether the contrasting weight phenotypes conferred by the two intact uncultured fecal microbiota samples could be transmitted by the strains captured in their derivative culture collections, we colonized 8-week-old adult male germfree C57BL/6 mice with all members of either of these two culture collections ($n = 6$ singly caged mice per collection; all mice receiving a given culture collection were maintained in a single gnotobiotic isolator). As a reference control for this experiment, and to compare results between this and the previous experiment, we colonized mice with the corresponding intact uncultured fecal microbiota samples, housing these mice in separate isolators from those used for the culture collection transplants. All mice were fed three embodiments of the Bangladeshi diet (1 week per diet) in the same order described for the previous experiment. As with the intact uncultured microbiota, the corresponding culture collections transmitted discordant weight phenotypes to recipient animals ($P < 0.002$, two-tailed Student's *t* test, comparison of final weights; Fig. 1C). Moreover, the weight phenotypes (change in body weight over time as a percentage of initial weight before gavage) observed with each intact uncultured fecal microbiota and the corresponding derivative culture collection were not significantly different ($P > 0.05$ for both microbiota donors, two-tailed Student's *t* test; Fig. 1C). The difference in weight phenotypes first became statistically significant between the two groups of mice midway through consumption of diet embodiment 2, continued to increase with diet embodiment 3 (Fig. 1C), and again were not attributable to differences in food consumption.

Effect of diet.

To test whether the weight loss phenotype was sensitive or robust to diet embodiment type, we gavaged the two clonally arrayed bacterial culture

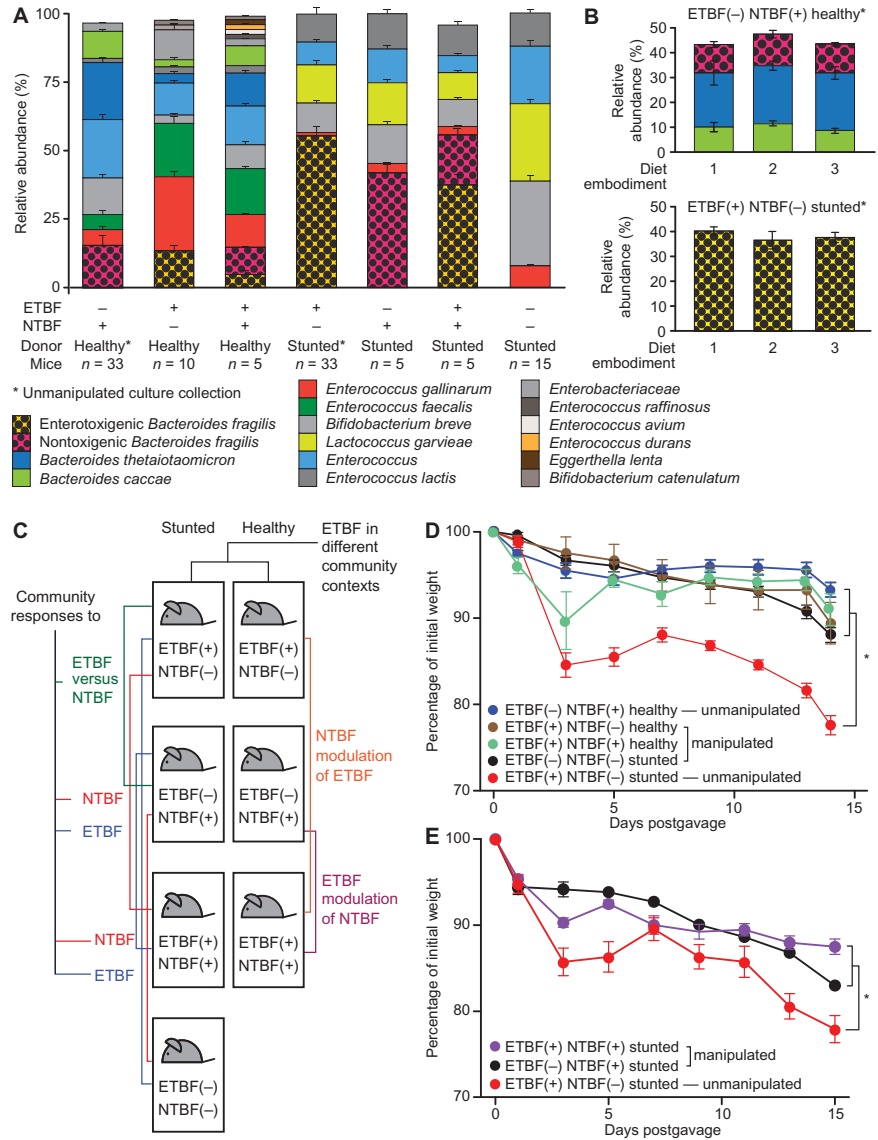
collections into separate groups of 8-week-old adult male germfree C57BL/6 mice who were monotonously fed Bangladeshi diet embodiment 1, 2, or 3 for 3 weeks ($n = 6$ singly caged recipient mice per culture collection per diet embodiment; fig. S1B). The discordant weight phenotype observed previously was preserved irrespective of the Bangladeshi diet embodiment consumed ($P < 0.01$, two-tailed Student's *t* test, comparison of final weights of mice regardless of diet embodiment consumed; $n = 18$ mice per culture collection; Fig. 1D). Moreover, no significant differences in weights were noted between groups of mice colonized with the same culture collection but fed different diet embodiments ($P > 0.05$ for embodiments 1 versus 2, 1 versus 3, and 2 versus 3; two-tailed Student's *t* test, comparison of final weights; Fig. 1D).

Transmission of strains was assessed by short-read shotgun sequencing of DNA isolated from fecal samples collected at the end of the experiment. This method, known as community profiling by sequencing (COPRO-Seq) (21), maps reads onto the draft genome assemblies of community members. At the depth of sequencing used [$354,352 \pm 23,216$ (mean \pm SEM), 50-nucleotide (nt) unidirectional reads/fecal DNA sample], we could reliably detect strains whose relative abundance is $\geq 0.1\%$. COPRO-Seq demonstrated that transplantation of the culture collections was efficient and reproducible, with $98.1 \pm 0.6\%$ and $94.5 \pm 1.6\%$ (mean \pm SEM) of strains in the collections derived from the healthy and stunted donors, respectively, appearing in recipient animals. The relative abundance of ETBF in the fecal microbiota of mice containing the stunted/underweight donor's culture collection was significantly greater than the cumulative relative abundance of the two NTBF strains in recipients of the healthy culture collection irrespective of the diet embodiment consumed ($75.5 \pm 4.1\%$ versus $17.0 \pm 4.4\%$; $P = 2.8 \times 10^{-9}$, two-tailed Student's *t* test; Fig. 2A). The relative abundances of the ETBF strain in recipients of the stunted/underweight donor's culture collection, the two NTBF strains in the healthy donor's collection, and all other *Bacteroides* species did not differ significantly between diet embodiments [$P > 0.2$ for all *Bacteroides*, one-way analysis of variance (ANOVA); Fig. 2B].

Intergenerational transmission of weight phenotypes.

To assess whether this weight loss phenotype was transmissible across generations of mice, two C57BL/6 male mice from the transplant experiment, one containing the stunted/underweight donor's culture collection and the other containing the healthy donor's collection, were switched to and subsequently maintained on an irradiated nutritionally enhanced mouse breeder chow from 21 to 48 dpg, at which time they were each cohoused with two germfree 6-week-old female mice that had received breeder chow since weaning. Seven days after cohousing, each male mouse was withdrawn from each mating trio, and the female mice were subsequently maintained on breeder chow throughout their pregnancy and as their pups completed the suckling period (fig. S1C). Male pups ($n = 3$ to 4 per litter) were then weaned onto an irradiated, nutritionally sufficient, LF/HPP chow, until they were 9 weeks old, at which time they were switched to the Bangladeshi diets (10 days per diet; same order of sequential presentation of the embodiments as before). Mice born to mothers colonized with either of these arrayed culture collections experienced identical weight gain profiles while consuming the LF/HPP diet ($P = 0.9$, two-tailed Student's *t* test; table S6). However, once they were transitioned to the sequence of three Bangladeshi diet embodiments (consumed from postnatal days 56 to 86), mice born to mothers harboring a stunted/underweight donor's microbial community exhibited significantly greater weight loss ($P = 0.03$, two-tailed Student's *t* test comparing

Fig. 2. ETBF is necessary but not sufficient to produce weight loss in recipient gnotobiotic mice. (A) Gut microbial community composition, defined by COPRO-Seq, in mice colonized with either of the two unmanipulated culture collections or the derived manipulated versions. Mean values for relative abundances \pm SEM are plotted using aggregate data generated from fecal samples collected from mice colonized with a given community. Taxa present at abundances lower than 1% are not represented in the stacked barplots. (B) The proportional representation of *Bacteroides* taxa in unmanipulated culture collections installed in gnotobiotic mice does not differ significantly as a function of the diet embodiments animals were fed. Means \pm SEM for data generated from feces are shown ($n = 5$ to 6 per group; one-way ANOVA). (C) Schematic illustrating the different groups of gnotobiotic mice generated by manipulating the presence/absence of ETBF and NTBF within the stunted/underweight or healthy donors' culture collections and the questions addressed by the indicated comparisons. (D) Removal of ETBF prevents weight loss in mice colonized with the stunted/underweight donor's culture collection. In contrast, addition of ETBF with the simultaneous removal of NTBF does not significantly affect weight in mice colonized with the culture collection derived from the healthy child ($n = 5$ to 6 mice per treatment group). Means \pm SEM are plotted. $*P < 0.05$ (paired two-tailed Student's *t* test and linear mixed effects model as above). (E) Addition of NTBF to the stunted/underweight donor's culture collection ameliorates ETBF-associated weight loss in gnotobiotic mice fed embodiment 2 of a representative Bangladeshi diet ($n = 6$ mice per treatment group). Means \pm SEM are plotted. $*P < 0.05$ (paired two-tailed Student's *t* test and linear mixed effects model as above).



weights at killing). The total relative abundance of the two NTBF strains in fecal samples obtained from recipients of the healthy donor's culture collection was $4.2 \pm 0.7\%$ at the conclusion of the LF/HPP diet period and $4.6 \pm 0.9\%$ at the conclusion of the Bangladeshi diet embodiment sequence, whereas the relative abundance of ETBF at these two time points was $34.3 \pm 4.2\%$ and $50.0 \pm 0.7\%$, respectively, in mice colonized with the stunted/underweight donor's culture collection.

An independent intergenerational transfer experiment was performed, in this case using the donors' intact uncultured fecal microbiota. The efficiency of ETBF and NTBF transmission from mothers to pups was 100%. As with the culture collections, there was diet-dependent transmission of the discordant weight loss phenotype (Fig. 1E; compare with Fig. 1A).

Microbial community context determines the effects of ETBF on community members and host

To establish whether ETBF is necessary and sufficient to cause marked weight loss in multiple community contexts, we performed a series of manipulations that involved removing the ETBF strain from the stunted/underweight donor's culture collection and adding it to the healthy donor's culture collection, with or without subtraction of its two NTBF strains (Fig. 2C). These manipulations allowed us to characterize (i) the role of community context in determining ETBF pathogenicity, (ii) the community/host responses to ETBF, (iii) the ability of NTBF to modulate ETBF effects, and (iv) the effects of ETBF on NTBF. Recipient C57BL/6 male mice in each of the different treatment groups were 8 to 9 weeks old at the time of colonization; all were placed on diet embodiment 2 for 2 days before

gavage and subsequently maintained on this diet for 14 days until they were killed ($n = 5$ singly caged animals per treatment group, maintained in separate gnotobiotic isolators). Fecal samples were collected at the time points described in fig. S1D.

Weight phenotypes.

Removal of the ETBF strain from the stunted/underweight donor's culture collection prevented the transmissible weight loss phenotype (Fig. 2D; $P = 5.9 \times 10^{-8}$, two-tailed Student's t test, comparison of weights at killing). However, addition of the ETBF strain to the healthy donor's culture collection did not produce significant weight loss, regardless of whether the NTBF strains were present or absent ($P = 0.3$ and $P = 0.2$, respectively, two-tailed Student's t test, comparison of weights at killing; Fig. 2D). On the basis of these findings, we concluded that whether ETBF produces weight loss (cachexia) is dependent on microbial community context.

COPRO-Seq analysis of the fecal microbiota of recipients of the unmanipulated ETBF(-) NTBF(+) healthy donor's culture collection revealed that it contained the two NTBF strains [total relative abundance of $14.5 \pm 3.0\%$ (mean \pm SEM), with *B. fragilis* hVEW46 and *B. fragilis* hVEW47 comprising 1.1 and 13.5%, respectively], two other *Bacteroides* (*B. thetaiotaomicron* and *B. caccae*), plus *Bifidobacterium breve* and *Enterococcus*. The relative abundance of *B. fragilis* was not significantly different between mice harboring the transplanted unmanipulated healthy donor's culture collection and its two manipulated ETBF(+) NTBF(-) and ETBF(+) NTBF(+) versions ($P > 0.5$, two-tailed Student's t test; Fig. 2A). (The term "unmanipulated" indicates that all bacterial isolates that comprise a culture collection were pooled before transplantation, whereas "manipulated" refers to the inclusion and/or exclusion of *B. fragilis* strains as part of the gavaged consortium.) The fecal microbiota of recipients of the unmanipulated stunted donor's culture collection was dominated by ETBF (relative abundance, $62.3 \pm 4.0\%$). Removal of ETBF led to significant increases in the relative abundances of *B. breve*, another *Bifidobacterium* strain, *Enterococcus lactis*, and *Enterococcus gallinarum* ($P < 0.02$, two-tailed Student's t test; Fig. 2A).

To determine whether NTBF alone is sufficient to protect mice from ETBF's cachectic effects, we colonized three groups of C57BL/6 male gnotobiotic mice, each with a different version of the stunted donor's culture collection: the unmanipulated culture collection containing ETBF alone or one of two manipulated versions, one with NTBF alone, and the other with both ETBF and NTBF strains. Mice were placed on diet embodiment 2 for 2 days before gavage and maintained on this diet for 2 weeks until killed ($n = 6$ animals per treatment group, all singly caged; one treatment group per gnotobiotic isolator; fig. S1D). We observed a significant difference in weight phenotypes between mice colonized with the unmanipulated undernourished donor's ETBF(+) NTBF(-) culture collection compared to the manipulated ETBF(-) NTBF(+) version ($P = 0.01$, one-tailed Student's t test; Fig. 2E). Addition of NTBF [yielding the ETBF(+) NTBF(+) community] markedly ameliorated the weight loss phenotype ($P = 0.0004$ for weights at killing compared to mice with the unmanipulated community, one-tailed Student's t test; Fig. 2E). Follow-up COPRO-Seq analysis revealed that the relative abundances of ETBF at the conclusion of the experiment were $38.9 \pm 3.9\%$ and $39.0 \pm 3.5\%$ when animals were colonized with and without NTBF, respectively. Thus, NTBF does not appear to mediate its effects by reducing the fractional representation of ETBF in the community. However, ETBF appears to reduce the relative abundance of NTBF, which constituted $41.8 \pm 3.2\%$ of the total community when ETBF was absent but only $19.2 \pm 2.6\%$ when ETBF was present ($P = 0.04$, one-tailed Student's t test).

The effects of intraspecific interactions on microbial gene expression.

We performed microbial RNA sequencing (RNA-seq) of cecal contents harvested at killing to characterize the transcriptomes of members of the unmanipulated and manipulated versions of the healthy and stunted communities. Our goal was to assess (i) the effects of intraspecific competition (NTBF on ETBF and vice versa) in the healthy and stunted community contexts, (ii) the effects of the cultured stunted/underweight versus healthy donor community on ETBF, and (iii) the effects of cocolonization with ETBF on other bacterial members (including other *Bacteroides*). ETBF genes with significant differential expression attributable to the presence or absence of NTBF, in both healthy and stunted community contexts, are listed in table S8 (B and F). Conversely, NTBF genes with significant differential expression attributable to the presence or absence of ETBF, in both healthy and stunted community contexts, are highlighted in table S8 (C and E).

Fragipain is a cysteine protease that activates fragilysin by removing its autoinhibitory prodomain. In mouse models of colitis, host proteases can also serve this function, but fragipain is required for sepsis to occur (23, 24). In the presence of NTBF, ETBF expression of fragilysin (*bft-3*) in the cecal metatranscriptome of mice harboring the manipulated ETBF(+) NTBF(+) healthy donor's community was significantly decreased compared to the manipulated version of the community where ETBF, but not NTBF, was present (39-fold, based on normalized transcript counts; $P = 0.002$, one-tailed Student's t test). Fragipain expression was also significantly reduced (14.2-fold; $P = 0.0005$, one-tailed Student's t test) (table S8B). In the context of the stunted community, the reduction in *bft-3* expression associated with introducing NTBF was considerably more modest (5.9-fold; $P = 0.09$, one-tailed Student's t test), whereas fragipain expression was not significantly different between the two treatment groups ($P > 0.5$, one-tailed Student's t test; table S8).

When we abrogated fragilysin (*bft-3*) expression through insertional mutagenesis (fig. S2), the mutant $\Delta bft-3$ strain grew robustly in vitro. However, when germfree mice were gavaged with a manipulated version of the stunted donor's culture collection containing this isogenic strain with a disrupted *bft-3* locus substituted for the wild-type ETBF strain, we observed no detectable colonization of the mutant ($n = 5$ mice fed diet embodiment 2 for 14 days); the number of COPRO-Seq reads mapping to the mutant $\Delta bft-3$ strain was no greater than background, and a PCR assay that used *B. fragilis*-specific *bft* primers was negative. However, these results led us to conclude that this locus functions as an important colonization factor for this particular ETBF strain in this community context. However, these experiments did not allow us to directly address the hypothesis that attenuation of *bft-3* expression produced by inclusion of NTBF in the stunted community contributed to the observed mitigation of weight loss.

Looking beyond the effects of intraspecific interactions on *bft-3* expression, we compared the cecal metatranscriptomes of gnotobiotic mice colonized with the unmanipulated NTBF(+) ETBF(-) healthy donor's culture collection versus mice harboring the two manipulated versions where ETBF was added, with or without removal of the two NTBF strains. The results revealed that ETBF in the absence of NTBF produced significant alterations in the expression of a number of transcripts related to various features of stress responses in several community members [*Enterococcus faecalis*, *E. gallinarum*, *B. breve*, and two members of *Enterobacteriaceae*; differentially expressed genes identified using the Robinson and Smyth exact negative binomial test (25), with Bonferroni correction for multiple hypotheses] (Fig. 3). Both *rpoS*, which is a key general stress response sigma factor that positively

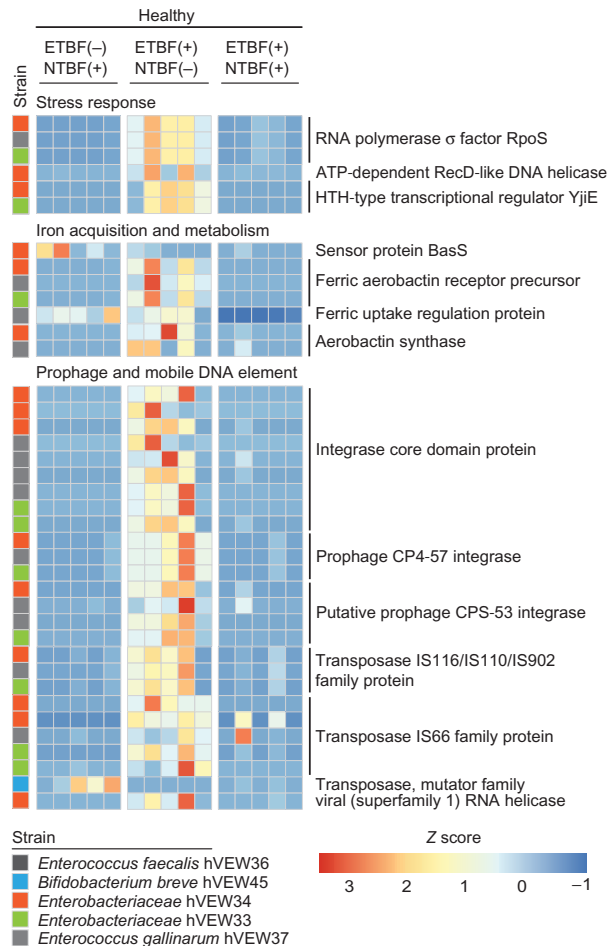


Fig. 3. The effects of intraspecific NTBF-ETBF interactions on the community metatranscriptome. Adult mice were colonized with the indicated unmanipulated and manipulated versions of the healthy donor's culture collection. All treatment groups were monotonously fed diet embodiment 2. Cecal contents were collected at the time of killing 14 days after initial colonization, and gene expression in the community was analyzed by microbial RNA-Seq. Each column represents data from an individual mouse. Each row represents the levels of a given transcript, normalized across that row. Addition of ETBF to and removal of NTBF from the healthy donor's culture collection (middle set of columns) produced an increase in expression of the indicated genes in strains whose identity is denoted by the color code on the left, compared to their expression in the unmanipulated ETBF(-) NTBF(+) version (left set of columns) or the manipulated version where the NTBF strains were retained when ETBF was added (right set of columns). UniProt-based annotations are shown on the right. ATP, adenosine triphosphate; HTH, helix-turn-helix.

controls expression of genes involved in transport of carbon sources and iron acquisition, and *recD*, which is involved in DNA repair, exhibited significant increases in their expression in the setting of ETBF without NTBF ($P < 0.05$). Several genes involved in the acquisition and metabolism of iron were either up-regulated in the presence of ETBF (for example, ferric aerobactin receptor, ferric uptake regulation protein, and aerobactin synthase) or repressed (for example, an *Enterobacteriaceae* strain hVEW34 homolog of the *Escherichia coli* BasSR system component BasS, which is normally induced under high-iron conditions)

(26). ETBF's effect on expression of these latter genes was mitigated when NTBF was present (Fig. 3), highlighting the importance of iron in intraspecific and interspecific interactions in the healthy donor's consortium of transplanted cultured bacterial strains. In contrast, the presence or absence of ETBF or NTBF did not evoke significant changes in the expression of these or other genes involved in iron metabolism in the context of the stunted/underweight donor's community. Numerous genes related to prophage and mobile DNA element biology were also expressed at significantly higher levels by healthy community members when ETBF was present in the absence of NTBF ($P < 0.05$; Fig. 3). Prophage activation occurs in response to stress. Some studies have postulated that phage induction can "shuffle" community structure to favor an increased proportion of pathobionts (27).

Studies in gnotobiotic mice have shown that signaling by members of the human gut microbiota involving the quorum sensing molecule, autoinducer-2 (AI-2), can alter virulence factor expression in enteropathogens (28) and have linked AI-2 signaling to modulation of the levels of *Bacteroidetes* in the gut (29). LuxQ is involved in the detection of AI-2. In the context of the healthy community, expression of three of the four *luxQ* homologs in the ETBF genome was decreased when NTBF was present [$\log_2(\text{fold change})$ of -2.8 , -4.4 , and -9.5 , $P < 0.005$, exact negative binomial test; table S8B]. Comparing the mice colonized with the unmanipulated ETBF(-) NTBF(+) and manipulated ETBF(+) NTBF(+) versions of the healthy donor's culture collection revealed differential regulation of five other *luxQ* transcripts encoded by *Bacteroides* members (three in *B. thetaiotaomicron* hVEW3, and two in *B. caccae* hVEW51; table S8D). In the context of the stunted donor's community, the presence of ETBF had no significant effects on *lux* gene expression in NTBF or any other community members, nor did the presence of NTBF have any effect on *lux* expression in ETBF (table S8, E and F). Together, these results illustrate the importance of community context in determining the transcriptional effects of intraspecific (and interspecific) interactions involving ETBF.

Metabolism.

The metabolic effects of manipulating the representation of ETBF and NTBF in the healthy and stunted donor's communities were studied by targeted mass spectrometry (MS) of tissue samples obtained from mice in the fed state (table S7). Quantifying amino acids, organic acids, acyl-carnitines, and acyl-CoAs in livers obtained from animals colonized with either of the two unmanipulated culture collections disclosed that compared to mice harboring the healthy donor's ETBF(-) NTBF(+) culture collection, those colonized with the stunted donor's ETBF(+) NTBF(-) culture collection had higher concentrations of propionyl-CoA and isovaleryl-CoA [by-products of oxidation of branched-chain and other amino acids; $P < 0.05$, false discovery rate (FDR)-adjusted two-tailed Student's *t* test; Fig. 4A], and lower concentrations of acetyl-CoA ($P = 0.07$) and its cognate metabolite acetyl carnitine (that is, C2 acylcarnitine; $P = 0.001$; Fig. 4B). Mirroring these trends, cecal contents harvested at the time of killing from mice harboring the stunted donor's unmanipulated culture collection contained higher concentrations of branched-chain amino acids ($P = 0.066$ for isoleucine/leucine and $P = 0.1$ for valine) and lower concentrations of acetyl-carnitine ($P = 0.067$). However, these trends were not observed in skeletal muscle.

Acetyl-CoA and acetyl-carnitine are generated by glucose, fatty acid, or amino acid metabolism, with a primary contribution from glucose metabolism in the fed state. A deficit in these metabolite pools suggests impaired flux of glucose to acetyl-CoA in mice with the stunted compared to healthy donor's unmanipulated culture collection. The decline in acetyl-CoA and acetyl-carnitine levels occurred in concert with significant

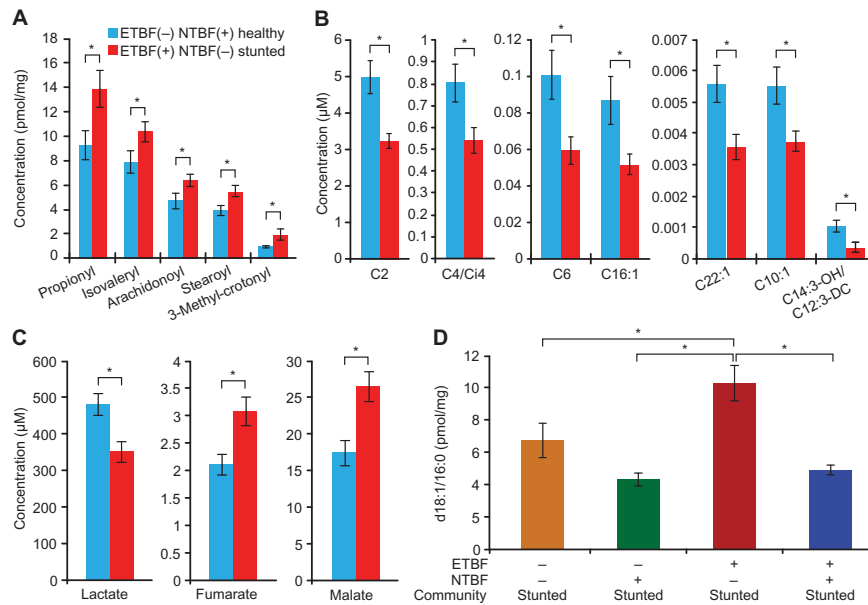


Fig. 4. Host metabolic abnormalities associated with the stunted donor's community. Mice were colonized with the indicated culture collections and fed diet embodiment 2. Animals were killed 21 days after colonization, and metabolites in their livers were quantified by mass spectrometry ($n = 5$ to 6 mice per treatment group). Means \pm SEM are plotted. Altered concentrations of acyl-CoAs (A), acylcarnitines (B), and organic acids (C) in the livers of mice colonized with the unmanipulated ETBF(+) NTBF(-) stunted donor's culture collection versus animals harboring the unmanipulated ETBF(-) NTBF(+) healthy donor's culture collection provide evidence for metabolic dysfunction involving the TCA cycle and mitochondrial fatty acid oxidation in the former group. (D) ETBF in the setting of the unmanipulated stunted donor's community is associated with greater hepatic concentrations of the d18:1/16:0 ceramide known to inhibit mitochondrial oxidative metabolism compared to manipulated versions of the culture collection where ETBF has been removed with or without replacement by NTBF or when NTBF and ETBF are both present. $*P < 0.05$ (two-tailed Student's t test).

increases in hepatic levels of the late tricarboxylic acid (TCA) cycle intermediates malate and fumarate ($P < 0.005$ compared to mice with the healthy donor's culture collection; Fig. 4C), as well as a trend toward increased succinate ($P = 0.1$). These patterns were again mirrored in the cecum but not in skeletal muscle. Early-stage TCA cycle intermediates citrate and α -ketoglutarate were not different across these groups. Conversely, and consistent with impaired glucose availability or metabolism, lactate concentrations were significantly lower in the livers of mice colonized with the stunted donor's culture collection ($P = 0.002$; Fig. 4C), with an associated decrease in cecal lactate concentrations ($P = 0.0005$; table S7B). Hepatic pyruvate also showed a downtrend ($P = 0.16$; table S7B).

Combining the acyl-CoA and organic acid data, this metabolic profile is consistent with a scenario in which mice colonized with the unmanipulated ETBF-containing stunted donor's culture collection have decreased capacity for metabolism of glucose to acetyl-CoA. It appears that these animals attempt to compensate by an increase in amino acid metabolism that is sufficient to produce enough acetyl-CoA to maintain normal citrate and α -ketoglutarate levels. However, use of these fuels also creates substrates that lead to production of distal TCA cycle intermediates (anaplerosis), for example, via propionyl-CoA conversion to succinate, and aspartate conversion to oxaloacetate, fumarate, and malate.

Comparing mice colonized with the unmanipulated stunted donor's culture collection versus the manipulated ETBF(-) NTBF(-) version of the stunted donor's culture collection revealed that removal of ETBF

was not sufficient to reverse the metabolic abnormalities described above. ETBF removal produced a significant reduction in hepatic levels of d18:1/16:0 ceramide, a known inhibitor of mitochondrial electron transport complex IV ($P = 0.02$; Fig. 4D) and consistent with decreased oxidative metabolism. Ceramides are involved in key intracellular stress response pathways, including those associated with immunoinflammatory responses. This effect was strain-specific: NTBF alone did not produce the same result in mice colonized with the manipulated ETBF(-) NTBF(+) version of the stunted donor's culture collection. Moreover, the ETBF-associated elevation in d18:1/16:0 was prevented by the presence of NTBF [$P = 0.007$, FDR-adjusted two-tailed Student's t test, comparison of mice colonized with the unmanipulated stunted community versus the manipulated ETBF(+) NTBF(+) community].

Gut barrier/immune function.

One view of the effects of *B. fragilis* enterotoxin is that it perturbs gut barrier function, increasing permeability via cleavage of E-cadherin (11), triggering the release of proinflammatory cytokines including tumor necrosis factor- α (TNF- α) (30), and augmenting β -catenin signaling (31). Treatment of chow-fed, conventionally raised mice with antibiotics followed by introduction of wild-type NTBF, NTBF with an engineered active *bft* allele, wild-type ETBF with an active *bft* allele, or ETBF with a $\Delta bft1$ allele revealed that under these conditions, *bft* was necessary to produce colitis (32). Small-scale human studies have demonstrated a greater representation of ETBF in individuals with active inflammatory bowel disease (33) and colorectal cancer (34). Strains harboring *bft-3* have been reported to generate less biologically active toxin than strains having *bft-1* or *bft-2* alleles (22).

Analyses of hematoxylin and eosin-stained sections of small intestine and colon obtained from mice consuming the representative Bangladeshi diet and colonized with either of the two unmanipulated culture collections or two manipulated versions [ETBF(-) NTBF(-) stunted or ETBF(+) NTBF(+) healthy] revealed no hallmarks of inflammation, such as an overabundance of neutrophils in the gut mucosa, crypt hyperproliferation or loss, or loss/damage of villi in the small intestine ($n = 5$ mice examined per treatment group).

We performed a follow-up flow cytometry analysis of immune cell populations in the colon, and mesenteric lymph nodes of four groups of mice: those colonized with the stunted donor's culture collection with or without ETBF and those harboring the unmanipulated healthy donor's culture collection or its manipulated ETBF(+) NTBF(-) version. Consistent with a previous study of specified pathogen-free mice treated with antibiotics to boost colonization with an ETBF isolate (34), colonic T helper 17 (T_H17) cell populations (TCR- β^+ CD4⁺CD8a⁻IL-17A⁺) were significantly increased in animals colonized with manipulated ETBF(+) NTBF(-) healthy and unmanipulated ETBF(+)

NTBF(–) stunted culture collections compared to the unmanipulated ETBF(–) NTBF(+) healthy and manipulated ETBF(–) NTBF(–) stunted culture collections ($P < 0.001$ and $P < 0.02$, respectively, Student's t test; fig. S3, A and B). The presence of ETBF was not associated with significant differences in the size of the anti-inflammatory FoxP3⁺ regulatory T cell population in mesenteric lymph nodes in either the stunted/underweight or healthy community contexts ($P > 0.05$, Student's t test; fig. S3C).

B. fragilis polysaccharide A (PSA) is the product of one of the organism's capsular polysaccharide synthesis (*CPS*) loci. A number of studies have shown that PSA functions as an immunomodulatory factor mitigating inflammatory responses in the gut via its effects on reducing interleukin-17 (IL-17) and inducing FoxP3⁺ regulatory T cells (35, 36). Microbial RNA-seq disclosed that the expression of genes in this *CPS* locus (37) was not significantly different in the ETBF and NTBF strains as a function of community context (unmanipulated versions of the healthy or stunted donor consortia) or engineered intraspecific interactions (table S8B).

We expanded our analysis to examine the effects of intraspecific interactions on serum cytokine profiles in three groups of mice: those harboring the unmanipulated stunted donor's culture collection and those colonized with the manipulated ETBF(+) NTBF(+) and ETBF(–) NTBF(+) versions. ETBF(+) NTBF(–) mice had significantly higher levels of the proinflammatory cytokines IL-17A, TNF- α , and interferon- γ compared to ETBF(–) NTBF(+) mice ($P = 0.02$, $P = 0.03$, and $P = 0.03$, respectively; one-tailed Student's t test) (fig. S3D). However, these effects were not reversed with NTBF cocolonization ($P > 0.05$ for all comparisons) despite NTBF's protective effect with respect to weight loss, leading us to postulate that (i) the weight loss phenotype was a reflection, at least in part, of the combined effects of disturbances in central metabolism and subtle perturbations in gut barrier function, and (ii) other members of the undernourished donor's transplanted culture collection, besides ETBF, were contributors to these abnormalities.

DISCUSSION

The question of what determines the effects of a large enteropathogen burden in children at risk for undernutrition, or with already overt disease, is rooted in the definition of "pathogen"—both conceptually and operationally. Koch's postulates invoke the requirement for an isolated candidate pathogen to produce a disease when introduced into a host species, without necessarily considering the microbial ecology of the body habitat that the organism invades and establishes itself. The approach described in this study, involving generation of sequenced, clonally arrayed culture collections from the fecal microbiota of healthy and stunted/underweight Bangladeshi children and subsequent introduction of these consortia, with or without addition or subtraction of ETBF and NTBF strains, into germfree mice fed embodiments of the diets consumed by the microbiota donors, allowed us to characterize how *B. fragilis* functions as a pathobiont in a microbial community context-dependent manner.

A topic of interest to those studying the pathogenesis of undernutrition is the relative effects of enteropathogen burden in nondiarrheal fecal samples and the presence or absence of environmental enteric dysfunction, an enigmatic disorder associated with populations where sanitation is poor and enteropathogen burden is high (38). Although ETBF itself was not correlated with stunting in our small study cohort of 100 children, additional studies in larger populations coupled with quantitative PCR assays for other enteropathogens are needed to ascertain the contributions of ETBF to disease. Results from our gnoto-

biotic mouse studies emphasize that microbiota community context also needs to be considered when understanding the potential effects of this and other enteropathogens (39). The approach described in this report should help shed light on this issue. Constructing communities from culture collections captures a donor's history of microbial exposures in their places of residency, as well as the evolutionary events that have shaped the genetic features of the gut strains they harbor. Culture collections generated from healthy and undernourished donors from this or other populations can be introduced into young rapidly growing animals (either through direct gavage or maternal transmission), or into adult animals, to better understand how intraspecific and interspecific interactions with one or more enteropathogens affect host biology in specified and/or systematically manipulated dietary and microbial community settings. A limitation of the present study is that the microbiota dissection was limited to two donors from a small cohort of 100 children. An obvious next step is to address the generalizability of our findings; this can be accomplished in a variety of ways, including analyses of the effects of introducing different ETBF and NTBF strains recovered from other Bangladeshi donors into the existing culture collections, as well as culture collections generated from individuals representing other populations.

An intriguing difference transmitted by the healthy and undernourished donor's culture collections involves the host metabolic phenotype. On the basis of our findings, we propose that the limited rate of acetyl-CoA production in mice receiving the stunted donor's culture collection cannot keep pace with the increased influx of substrates that replenish TCA cycle intermediates (anaplerosis), explaining the increase in distal TCA cycle substrates. At a more global level, increased oxidation of amino acids to maintain a minimal acetyl-CoA pool may divert amino acids away from protein synthesis, contributing to a wasting phenotype. Furthermore, concentrations of amino acids were modestly but consistently lower in the sera of mice colonized with the culture collection from the stunted compared to the healthy donor. This is consistent with use of amino acids for anaplerosis, thereby contributing to lower amino acid supplies for host growth and metabolism.

Some bacterial members of the gut microbiota contain the complete enzymatic apparatus for executing the TCA reaction sequence. The finding that there is marked inhibition of the distal TCA cycle in harvested cecal contents and liver but not skeletal muscle leads to the following testable hypotheses: (i) interactions between the undernourished donor's microbiota and dietary components yield products that are present at high-enough levels to inhibit the TCA cycle in community members; (ii) these products are transported from the gut by the portal circulation in sufficient quantities to inhibit the TCA cycle in hepatocytes; (iii) either first-pass metabolism of these products by the liver or other processes limit their levels in the systemic circulation so that their effects on the TCA cycle are not observed in muscle; and (iv) TCA cycle inhibition negatively affects the ability to harvest energy from an already impoverished diet.

In our previous study of Malawian twins discordant for severe acute malnutrition, ¹H nuclear magnetic resonance analyses of urine obtained from adult gnotobiotic mice colonized with intact uncultured microbiota indicated that TCA cycle inhibition was a feature of animals harboring the undernourished co-twin's microbiota and consuming a macronutrient- and micronutrient-deficient prototypic Malawian diet. These animals but not their counterparts that had been colonized with the healthy co-twin's microbiota exhibited a pronounced diet-dependent weight loss phenotype that did not occur in the context of a macronutrient- and micronutrient-sufficient diet (7). Together, these findings suggest

that metabolic studies of serum, urine, and feces obtained from children before, during, and after treatment for their undernutrition should include quantitative assessment of analytes linked to TCA cycle activity.

MATERIALS AND METHODS

Human study design

Samples were obtained from an already completed observational birth cohort study (*Field Studies of Human Immunity to Amebiasis in Bangladesh*) (14). The study was conducted using protocols for obtaining informed consent, clinical samples, and clinical metadata that were approved by institutional review boards from the International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b) (study ID number 2007-041), the University of Virginia, Charlottesville (study ID number 7563), and Washington University in St. Louis (study ID number 201111065). Fecal specimens used in this study were covered by a material transfer agreement between icddr,b and Washington University in St. Louis. Fecal samples from 100 of the 147 children enrolled in this study were used for the analysis described in the current report; this number was not based on a previous power calculation derived from knowledge of ETBF carriage rate but rather represented samples from individuals who had been surveyed during the second year of postnatal life.

bft PCR assay

DNA isolated from fecal samples collected at 18 and 24 months of age from members of the birth cohort was used for a *bft* PCR assay using primer pairs 5'-GAACCTAAAACGGTATATGT-3' (GBF-201) and 5'-GTTGTAGACATCCCACTGGC-3' (GBF-210) (40), OneTaq 2× Master Mix (New England Biolabs) and the following thermocycling conditions: after initial melting at 95°C for 30 s, 30 cycles of 95°C for 30 s, 53°C for 30 s, and 68°C for 30 s, with a final annealing time of 7 min at 68°C.

Testing fecal samples for parasites

A clinical microscopy-based screen for *Entamoeba histolytica*, *Entamoeba dispar*, *Blastocystis hominis*, *Trichomonas hominis*, coccidian-like bodies, *Giardia lamblia*, *Ascaris lumbricoides*, *Trichuris trichiura*, *Ancylostoma duodenale*, *Necator americanus*, *Hymenolepis nana*, *Endolimax nana*, *Iodamoeba butschlii*, and *Chilomastix mesnili* was also performed on these fecal samples, as reported in (5).

Preparation of human fecal samples for transplantation to germfree mice

Aliquots (1 g) of previously frozen fecal samples were resuspended under anaerobic conditions (77% N₂, 20% CO, and 3% H₂) in 15 ml of gut microbiota medium (GMM) (20) and homogenized at a setting of “high” in a sterilized blender (Waring). Homogenates were clarified by passage through 100- μ m-pore-diameter nylon filters (BD Falcon). Five milliliters of sterile 2-mm-diameter glass beads was added, and remaining cell clumps were disrupted by vortexing (four on/off cycles, each for 30 s). A final filtration through a 40- μ m-pore-diameter nylon filter (BD Falcon) was performed before storage in GMM containing glycerol [final concentration of 15% (v/v)] at -80°C in 2-ml amber glass vials with a crimp-top butyl septum (Wheaton Scientific).

Preparation of Bangladeshi diet embodiments

Diet embodiments with ingredient and nutritional contents described in table S2 were prepared in 15-kg batches. Tilapia fish fillets (Whole Foods Markets) were simmered for 30 min in a 20-liter stainless steel

pot over a Corning hot plate (temperature set at “4”). Fruits and vegetables were combined and simmered in a separate pot for 45 min on a Corning hot plate (temperature set at “4”). Parboiled rice (Delta Star) was cooked in a rice cooker (KRUPS). Lentils were simmered for 90 min (temperature set at “2”). All ingredients for a given diet embodiment were combined in a vertical cutter-mixer (Robot Coupe Model R23) and pureed for 5 min. Aliquots were cooled in large plastic containers for 12 hours at 4°C. Aliquots (500 g) were then vacuum-sealed in 8 × 10-inch plastic bags (Uline). The vacuum-sealed food paste was placed in a second 8 × 10-inch plastic bag (as an added barrier against incidental contamination), and the contents of the packages were sterilized by irradiation (20 to 50 kGy) within 24 hours of food production (Steris Co.). Sterility was verified by culturing samples of the irradiated diet in brain heart infusion (BHI) medium for 5 days at 37°C under anaerobic and aerobic conditions (21). In addition, *B. subtilis* spore strips that had been included along with the food during the irradiation were cultured under the same conditions. Final nutritional profiles were obtained for samples of each embodiment (Nestlé Purina Analytical Laboratories, St. Louis, MO).

Clonally arrayed bacterial culture collections

Collections of cultured anaerobic bacterial strains were generated from frozen fecal samples according to previously published methods (20, 21). Cultures were arrayed in 384-well plates in Coy Chambers under strict anaerobic conditions (77% N₂, 20% CO, and 3% H₂) using a Precision XS liquid handling robot (BioTek). Bacteria isolates occupying each well were first grouped into 100% ID OTUs on the basis of the results of V4-16S rRNA amplicon sequencing. Most OTUs were observed more than once across an arrayed library. Next, three to four isolates representing each OTU were picked robotically from the 384-well arrays and struck out individually onto eight-well agar plates containing GMM (20). DNA from isolates was subjected to whole-genome shotgun sequencing using an Illumina HiSeq 2000 instrument (101-nt paired-end reads) or a MiSeq machine (150-nt paired-end reads). Sequences were assembled using MIRA (41) version 4.02 (parameters: “-NW:cnfs = warn, -NW:cmml = no, -GE:not = 4”; template_size = “150 500 autorefine”). Coverage of isolate genomes from the severely stunted/underweight donor’s culture collection was 38.6 ± 5.0-fold (mean ± SEM) with an N50 contig length of 28.5 ± 3.5 kb (mean ± SEM) (table S3B). The corresponding values for members of the healthy donor’s collection are described in (21). Genes were annotated using Prokka (v1.10) (42). Predicted genes in each isolate’s genome assemblies were mapped to KEGG pathways and assigned KEGG Ortholog groups by querying the KEGG reference database (release 72.1) (BLAST 2.2.29+, blastp *E* value ≤ 10⁻¹⁰, single best hit defined by *E* value and bit score). Putative virulence factors were identified using the Virulence Factor Database [(43); BLAST hits with *E* value < 10⁻¹⁷ bearing UniProt database annotations].

Full-length 16S rRNA gene amplicons were generated from isolates using primers 8F and 1391R. Isolates sharing ≥99% nucleotide sequence identity in their 16S rRNA genes and ≥96% sequence identity throughout their genomes as determined by NUCmer (44) were defined as representing a unique strain. Full-length 16S rRNA sequences were used to define taxonomy [Ribosomal Database Project (RDP) version 2.4 classifier (45)]. Clonally arrayed, sequenced culture collections were stored in GMM/15% glycerol at 80°C in 96-well plates (TPP Tissue Culture Test Plates).

Generation of a *bft-3:pGERM* mutant in the *ETBF* strain cultured from the stunted donor’s microbiota.

Disruption of the *bft-3* gene was accomplished using the *Bacteroides* suicide vector pGERM. A 380-base pair (bp) fragment of the 5' end of

bft-3 was introduced into Bam HI–digested pGERM using the Gibson Assembly Cloning Kit (New England Biolabs). Primers to create the PCR product for the Gibson reaction were created using the NEBuilder Assembly Tool (v1.7.2). Assembled constructs were introduced by electroporation into kanamycin-resistant, electrocompetent *E. coli* DH5- α RK231 (46). Transformants were plated on LB agar containing isopropyl- β -D-thiogalactopyranoside/X-gal (Sigma-Aldrich) and ampicillin (100 μ g/ml). Individual colonies were picked and grown in LB medium plus ampicillin; successful transformation was verified by PCR. The pGERM-transformed *E. coli* DH5- α RK231 and the ETBF strain (isolate mB11 of strain *B. fragilis* mVEW4) were conjugated by filter-mating on BHI-blood agar plates incubated under anaerobic conditions at 37°C for 24 hours. Bacterial cells from the resulting lawn were collected by scraping and plated on BHI agar containing erythromycin/gentamicin (25 and 100 μ g/ml, respectively). Plates were incubated under anaerobic conditions at 37°C for 48 hours. Picked colonies were sequenced to confirm successful pGERM insertion into *bft-3*, and positive colonies were stored in tryptone yeast glucose (TYG) containing 15% (v/v) glycerol.

The mutant and wild-type strains were cultured at 37°C under anaerobic conditions in TYG liquid medium to mid-log phase. Cells were then harvested, and RNA was extracted using TRIzol reagent (Life Technologies). To confirm that *bft* expression had been disrupted, complementary DNA (cDNA) was generated from the extracted RNA (see RNA-seq section below) and used as a template for a PCR-based assay that used ReddyMix PCR Master Mix (Thermo Fisher Scientific), a forward primer that spans the vector insertion site (5'-ATGAA-GAATGTAAAGTTACTTTTAATGCTAGGAACCG-3'), a reverse primer 3' to the crossover site (5'-CTCCACTTTGTACTTTATAC-TACTGAATATGCTTG-3'), and the following cycling conditions: 94°C for 10 min followed by 30 cycles of 94°C for 30 s, 56°C for 30 s, 72°C for 60 s, with a final extension of 72°C for 5 min.

Assembling consortia of cultured bacterial strains for gavage

Archived 96-well culture collection plates were thawed in an anaerobic chamber and aliquots of each well inoculated into 600 μ l of GMM in 96-deep-well culture plates (Thermo Fisher Scientific). Cultures were grown anaerobically to stationary phase at 37°C, with growth measured at OD₆₀₀ (optical density at 600 nm). Equal amounts of selected isolates (100 μ l) were pooled using the Precision XS liquid handling robot. For manipulations involving addition of NTBF, both strains of the nontoxigenic strain were used. For manipulations involving subtraction of the single ETBF strain or both NTBF strains, the strains were simply excluded from the pooling step. Final pools were mixed 1:1 with sterile phosphate-buffered saline (PBS) plus glycerol (to achieve a final concentration of 15%) and aliquots (1 ml) were placed in 2-ml butyl septum-stoppered amber glass vials (Wheaton) for storage or gavage into germfree animals.

Gnotobiotic mouse experiments

All mouse experiments were performed using protocols approved by Washington University Animal Studies Committee. Before the initiation of experiments, germfree adult male C57BL/6 mice were maintained in plastic flexible film gnotobiotic isolators under a strict 12-hour light cycle and fed an autoclaved LF/HPP chow ad libitum (B&K Universal, diet 7378000). Different treatment groups (that is, mice gavaged with different microbial communities) were maintained in separate gnotobiotic isolators. All animals studied were included in subsequent analyses. Male mice were age- and weight- matched before gavage of human

donor microbiota. Investigators were not blinded to either diet or microbiota treatments.

Feeding Bangladeshi diet embodiments was initiated 2 days before colonization with either uncultured fecal microbiota samples or pools generated from arrayed bacterial culture collections. Diets were extruded as a paste from their plastic bags into food trays. Fresh 30-g aliquots of the paste were provided to each singly caged animal per day. For experiments where diet embodiments were changed, autoclaved Aspen hardwood laboratory bedding (NEPCO) was replaced at the start of each embodiment transition to minimize carryover of ingredients from the preceding diet embodiment exposure.

For the intergenerational transmission experiments, male pups ($n = 3$ to 4 for each intact donor microbiota and $n = 5$ for each culture collection), representing the combined litter from trio matings were weaned onto LF/HPP chow and transitioned to the Bangladeshi diet embodiments at 8 weeks of age (embodiments 1→2→3; 10 days per diet phase).

Characterizing microbial community composition and gene expression

Multiplex sequencing of bacterial 16S rRNA PCR amplicons.

Genomic DNA was extracted from mouse fecal pellets using a phenol-chloroform and beat-beating protocol (45). Bar-coded primers 515F and 806R were used to generate PCR amplicons covering the V4 region of bacterial 16S rRNA genes present in the samples. Multiplex sequencing of pooled amplicons with sample-specific bar codes was performed using an Illumina MiSeq instrument (250-bp paired-end reads). Reads were trimmed in silico to 200 bases to retain the highest-quality base calls, assembled with FLASH (v1.2.11) and demultiplexed in QIIME (v1.8.0) (21). The 16S rRNA sequence data sets were first analyzed using open-reference OTU picking (97% ID OTUs) (uclost-ref against the Greengenes reference database). Sequences (98.7% of the 26,419,497) that passed QIIME's default quality filters were successfully clustered with a reference sequence of $\geq 97\%$ identity. For all subsequent analyses, OTU tables were rarefied to 4000 reads per sample and filtered to retain only those OTUs whose relative abundances in fecal microbiota were $\geq 0.1\%$ in $\geq 1\%$ of all samples sequenced. Taxonomies were assigned using RDP classifier 2.4 trained on the manually curated Greengenes database "Isolated named strains 16S."

COPRO-Seq of fecal samples collected from mice colonized with unmanipulated and manipulated culture collections.

Unidirectional reads (75 nt) were generated from fecal DNA samples using an Illumina MiSeq instrument and trimmed to 50 nt to preserve the highest-quality reads before mapping onto the genomes of culture community members. The analytic pipeline for COPRO-Seq is described in (21) and uses software available at <https://github.com/nmnculty/COPRO-Seq>.

Microbial RNA-seq.

Procedures for performing microbial RNA-seq are detailed in previous publications (21). Aliquots of cecal contents (~50 mg) that had been collected at killing (14 dpv) and stored at -80°C were suspended in 500 μ l of extraction buffer (200 mM NaCl and 20 mM EDTA), 210 μ l of 20% SDS, 500 μ l of phenol/chloroform/isoamyl alcohol (pH 4.5, 125:24:1; Ambion/Life Technologies), and 150 μ l of acid-washed glass beads (212- to 300- μ m diameter; Sigma-Aldrich). Cells were lysed by mechanical disruption using a bead beater (maximum setting, 5 min at room temperature; BioSpec Products), followed by phenol/chloroform/isoamyl alcohol extraction and isopropanol precipitation on ice. After treatment with ribonuclease (RNase)-free TURBO-DNase (Ambion/Life

Technologies), MEGAclear columns (Life Technologies) were used to remove 5S rRNA and transfer RNAs. A second deoxyribonuclease (DNase) treatment (Baseline-Zero DNase, Epicentre/Illumina) was performed before a second MEGAclear purification followed by rRNA depletion (Ribo-Zero rRNA removal kits, Epicentre/Illumina). cDNA was synthesized (SuperScript II, Invitrogen), and second-strand synthesis was performed with RNase H, *E. coli* DNA polymerase, and *E. coli* DNA ligase (all from New England Biolabs). cDNA libraries were prepared by shearing samples using a Bioruptor Pico sonicator (Diagenode), then size-selecting 150- to 200-bp fragments for blunting, A-tailing, and ligating sample-specific barcoded sequencing adapters, and finally PCR enrichment. cDNA libraries were pooled for multiplex sequencing using an Illumina NextSeq instrument (13.4 ± 1.1 million unidirectional 75-nt reads/cecal RNA sample), with the exception of the sample IDs ETBF.expt.112-116 (table S3D), which were sequenced on the MiSeq platform (6.6 ± 0.7 million unidirectional 75-nt reads per sample). Identification of differentially expressed genes was performed in R (version 3.2.3) using the Robinson and Smyth exact negative binomial test with Bonferroni correction (edgeR package, version 3.10.2) as described previously (21).

Mass spectrometry

Tissues previously stored at -80°C were weighed while frozen and immediately homogenized in a solution of 50% aqueous acetonitrile/0.3% formic acid (50 mg of tissue per milliliter solution; procedure done at 4°C using an IKA T25 Ultra-Turrax High-Speed Homogenizer at maximum setting for 30 to 45 s). Amino acids, acylcarnitines, organic acids, acyl-CoAs, and ceramides were analyzed using stable isotope dilution techniques. Amino acid and acylcarnitine measurements were made by flow injection tandem MS using sample preparation methods described previously (47, 48). Data were acquired using a Waters Acquity UPLC system (Waters) equipped with a TQ (triple quadrupole) detector and a data system controlled by the MassLynx 4.1 operating system (Waters). Organic acids were quantified (49) using a Trace Ultra Gas Chromatograph coupled to ISQ MS operating under Xcalibur 2.2 (Thermo Fisher Scientific). Acyl-CoAs were extracted and purified as described previously (50), and analyzed by flow injection analysis using positive electrospray ionization on a Xevo TQ-S, TQ mass spectrometer (Waters). Heptadecanoyl CoA was used as an internal standard. Ceramides were extracted (51) and analyzed by flow injection tandem MS using a Xevo TQS spectrometer (Waters).

Histologic analysis

Intestines were harvested and placed in 10% neutral-buffered formalin for 3 hours at room temperature, washed in 70% ethanol, prepared as “Swiss rolls,” paraffin-embedded, and 5- μ m-thick sections were cut. Hematoxylin and eosin–stained sections were evaluated for alterations in crypt number (evaluated by crypt density), crypt depth, villus height, epithelial proliferation [measured by M-phase cells per crypt ($n = 100$ crypts per specimen evaluated) and neutrophils per crypt unit ($n = 100$ crypt units evaluated per specimen)]. All slides were blinded before quantification.

Immune cell isolation and characterization

Cells from the spleen, mesenteric lymph node, and lamina propria of the colon were isolated as described previously (52), with two minor modifications: (i) Hanks’ Balanced Salt Solution (HBSS) was used in place of Dulbecco’s PBS, and (ii) collagenase type VIII (Sigma-Aldrich) was used at a concentration of 0.1 mg/ml for isolation of colonic lamina propria cells.

Intracellular cytokine staining was performed on cell suspensions from mesenteric lymph nodes and colonic lamina propria after restimulation in round-bottom 96-well plates in complete RPMI containing β -mercaptoethanol (0.05 mM; Sigma-Aldrich) supplemented with phorbol 12-myristate 13-acetate (50 ng/ml; Sigma-Aldrich), ionomycin (750 ng/ml; Sigma-Aldrich), and brefeldin A (eBioscience; diluted according to the manufacturer’s recommendations). Cells were restimulated at 37°C for 3 hours in a humidified CO₂ tissue culture incubator. Cells in which FoxP3 was assessed were not restimulated.

Cell surface and intracellular staining was performed as described previously (52) with the following two modifications: (i) cells were washed free of excess protein from restimulation medium, or from the HBSS/0.1% bovine serum albumin (BSA) (w/v) buffer, using HBSS without BSA, and (ii) the incubation containing anti-CD16/CD32 [Fc block (2.4G2); BD Pharmingen] also included Live/Dead Fixable Aqua Dead Cell Stain (Life Technologies) to allow identification and removal of dead cells during the analysis that followed fixation and permeabilization. Cells were stained with the following antibodies: CD4 eFluor 450 (GK1.5, eBioscience), CD4 allophycocyanin (RM4-5, BD Pharmingen), T cell receptor β (TCR- β) Alexa Fluor 488 (H57-597, BioLegend), CD8a allophycocyanin (53-6.7, BioLegend), FoxP3 eFluor 450 (FJK-16s, eBioscience), and IL-17A phycoerythrin (TC11-18H10.1, BioLegend). Appropriate isotype control antibodies were purchased from the same vendors that supplied antibodies targeting molecules of interest. Cells were analyzed with an Aria III flow cytometer (BD Biosciences), and data were analyzed with FlowJo software (Tree Star; version 7.6.1).

Measurements of serum cytokine levels

Blood was collected via retro-orbital phlebotomy at the time of killing, and derived serum samples were analyzed, alongside standards, using the Bio-Plex Pro Mouse Cytokine 23-plex Assay (Bio-Rad) following the manufacturer’s instructions. Technical duplicates were performed for each sample.

Statistical analyses

Routine statistical analyses were performed in R (version 3.1.2). $P < 0.05$ was considered to be statistically significant. Specific statistical tests are noted in the figure legends and throughout the text. Bonferroni correction for multiple hypotheses was applied to identify genes that exhibited significant differences in their expression between treatment groups.

SUPPLEMENTARY MATERIALS

www.sciencetranslationalmedicine.org/cgi/content/full/8/366/366ra164/DC1

Fig. S1. Experimental designs.

Fig. S2. Generation of a Δbft -3 strain of ETBF.

Fig. S3. Assessing the effects of ETBF on gut barrier/immune function.

Table S1. Summary of anthropometric scores and results of PCR assays for *bft* in the fecal microbiota of members of a birth cohort of 100 Bangladeshi children.

Table S2. Composition and nutritional content of Bangladeshi diet embodiments, and nutritional requirements of mice and children.

Table S3. V4-16S rRNA, COPRO-Seq, microbial RNA-seq, and cultured bacterial strain genome sequencing data sets.

Table S4. Bacterial culture collection generated from the fecal microbiota of a severely stunted Bangladeshi child.

Table S5. Genome features of isolates in the clonally arrayed bacterial culture collection generated from the fecal microbiota of a severely stunted 2-year-old Bangladeshi child.

Table S6. Weights of all individual mice as a function of diet, microbiota, and dpg.

Table S7. MS-based quantitation of metabolites in host tissues, serum, and cecum obtained from mice in the different treatment groups.

Table S8. Effects of intraspecific and interspecific interactions on ETBF, NTBf, and other *Bacteroides* gene expression profiles in healthy or stunted donor community contexts.

REFERENCES AND NOTES

- P. Christian, L. C. Mullany, K. M. Hurley, J. Katz, R. E. Black, Nutrition and maternal, neonatal, and child health. *Semin. Perinatol.* **39**, 361–372 (2015).
- R. E. Black, C. G. Victora, S. P. Walker, Z. A. Bhutta, P. Christian, M. de Onis, M. Ezzati, S. Grantham-McGregor, J. Katz, R. Martorell, R. Uauy; Maternal and Child Nutrition Study Group, Maternal and child undernutrition and overweight in low-income and middle-income countries. *Lancet* **382**, 427–451 (2013).
- R. Martorell, A. Zongrone, Intergenerational influences on child growth and undernutrition. *Paediatr. Perinat. Epidemiol.* **26** (suppl. 1), 302–314 (2012).
- C. G. Victora, L. Adair, C. Fall, P. C. Hallal, R. Martorell, L. Richter, H. S. Sachdev; Maternal and Child Undernutrition Study Group, Maternal and Child Undernutrition: Consequences for adult health and human capital. *Lancet* **371**, 340–357 (2008).
- S. Subramanian, S. Huq, T. Yatsunenko, R. Haque, M. Mahfuz, M. A. Alam, A. Benezra, J. DeStefano, M. F. Meier, B. D. Muegge, M. J. Barratt, L. G. VanArendonk, Q. Zhang, M. A. Province, W. A. Petri Jr., T. Ahmed, J. I. Gordon, Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature* **510**, 417–421 (2014).
- L. V. Blanton, M. R. Charbonneau, T. Salihi, M. J. Barratt, S. Venkatesh, O. Ilkaveya, S. Subramanian, M. J. Manary, I. Trehan, J. M. Jorgensen, Y.-m. Fan, B. Henrissat, S. A. Leyn, D. A. Rodionov, A. L. Osterman, K. M. Maleta, C. B. Newgard, P. Ashorn, K. G. Dewey, J. I. Gordon, Gut bacteria that prevent growth impairments transmitted by microbiota from malnourished children. *Science* **351**, aad3311 (2016).
- M. I. Smith, T. Yatsunenko, M. J. Manary, I. Trehan, R. Mkakosya, J. Cheng, A. L. Kau, S. S. Rich, P. Concannon, J. C. Mychaleckyj, J. Liu, E. Houpt, J. V. Li, E. Holmes, J. Nicholson, D. Knights, L. K. Ursell, R. Knight, J. J. Gordon, Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science* **339**, 548–554 (2013).
- A. L. Kau, J. D. Planer, J. Liu, S. Rao, T. Yatsunenko, I. Trehan, M. J. Manary, T.-C. Liu, T. S. Stappenbeck, K. M. Maleta, P. Ashorn, K. G. Dewey, E. R. Houpt, C.-S. Hsieh, J. I. Gordon, Functional characterization of IgA-targeted bacterial taxa from undernourished Malawian children that produce diet-dependent enteropathy. *Sci. Transl. Med.* **7**, 276ra24 (2015).
- R. L. Guerrant, R. B. Oriá, S. R. Moore, M. O. B. Oriá, A. A. M. Lima, Malnutrition as an enteric infectious disease with long-term effects on child development. *Nutr. Rev.* **66**, 487–505 (2008).
- G. Lee, M. Paredes Olortegui, P. Peñataro Yori, R. E. Black, L. Caulfield, C. Banda Chavez, E. Hall, W. K. Pan, R. Meza, M. Kosek, Effects of Shigella-, Campylobacter- and ETEC-associated diarrhea on childhood growth. *Pediatr. Infect. Dis. J.* **33**, 1004–1009 (2014).
- C. L. Sears, A. L. Geis, F. Housseau, *Bacteroides fragilis* subverts mucosal biology: From symbiont to colon carcinogenesis. *J. Clin. Invest.* **124**, 4166–4172 (2014).
- P. Pathela, K. Z. Hasan, E. Roy, K. Alam, F. Huq, A. K. Siddique, R. B. Sack, Enterotoxigenic *Bacteroides fragilis*-associated diarrhea in children 0–2 years of age in rural Bangladesh. *J. Infect. Dis.* **191**, 1245–1252 (2005).
- C. L. Sears, S. Islam, A. Saha, M. Arjumand, N. H. Alam, A. S. G. Faruque, M. A. Salam, J. Shin, D. Hecht, A. Weintraub, R. B. Sack, F. Qadri, Association of enterotoxigenic *Bacteroides fragilis* infection with inflammatory diarrhea. *Clin. Infect. Dis.* **47**, 797–803 (2008).
- D. Mondal, J. Minak, M. Alam, Y. Liu, J. Dai, P. Korpe, L. Liu, R. Haque, W. A. Petri Jr., Contribution of enteric infection, altered intestinal barrier function, and maternal malnutrition to infant malnutrition in Bangladesh. *Clin. Infect. Dis.* **54**, 185–192 (2012).
- World Health Organization (WHO), WHO Child Growth Standards: Growth velocity based on weight, length and head circumference: Methods and development. WHO, (2009) www.who.int/childgrowth/standards/velocity/technical_report/en/
- J. E. Arsenault, E. A. Yakes, M. B. Hossain, M. M. Islam, T. Ahmed, C. Hotz, B. Lewis, A. S. Rahman, K. M. Jamil, K. H. Brown, The current high prevalence of dietary zinc inadequacy among children and women in rural Bangladesh could be substantially ameliorated by zinc biofortification of rice. *J. Nutr.* **140**, 1683–1690 (2010).
- S. N. Islam, M. N. I. Khan, M. Akhtaruzzaman, A Food Composition Database for Bangladesh with Special Reference to Selected Ethnic Foods (INFS-NFPCSP-FAO, Dhaka, Bangladesh, 2010); www.nfpcsp.org/agridrupal/content/food-composition-database-bangladesh-special-reference-selected-ethnic-foods.
- US Department of Agriculture (USDA), Agricultural Research Service, Nutrient Data Laboratory, USDA National Nutrient Database for Standard Reference, Release 28, (2015); www.ars.usda.gov/nea/bhnrc/ndl.
- M. Mohsena, R. Goto, C. G. N. Mascie-Taylor, Socioeconomic and demographic variation in nutritional status of under-five Bangladeshi children and trend over the twelve-year period 1996–2007. *J. Biosocial Sci.*, 1–17 (2016).
- A. L. Goodman, G. Kallstrom, J. J. Faith, A. Reyes, A. Moore, G. Dantas, J. I. Gordon, Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 6252–6257 (2011).
- N. Dey, V. E. Wagner, L. V. Blanton, J. Cheng, L. Fontana, R. Haque, T. Ahmed, J. I. Gordon, Regulators of gut motility revealed by a gnotobiotic model of diet-microbiome interactions related to travel. *Cell* **163**, 95–107 (2015).
- A. S. Scotto d'Abusco, M. Del Grosso, S. Censini, A. Covacci, A. Pantosti, The alleles of the *bft* gene are distributed differently among enterotoxigenic *Bacteroides fragilis* strains from human sources and can be present in double copies. *J. Clin. Microbiol.* **38**, 607–612 (2000).
- V. M. Choi, J. Herrou, A. L. Hecht, W. P. Teoh, J. R. Turner, S. Crosson, J. Bubeck Wardenburg, Activation of *Bacteroides fragilis* toxin by a novel bacterial protease contributes to anaerobic sepsis in mice. *Nat. Med.* **22**, 563–567 (2016).
- J. Herrou, V. M. Choi, J. Bubeck Wardenburg, S. Crosson, Activation mechanism of the *Bacteroides fragilis* cysteine peptidase, fragipain. *Biochemistry* **55**, 4077–4084 (2016).
- M. D. Robinson, G. K. Smyth, Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9**, 321–332 (2008).
- J. Hagiwara, T. Yamashino, T. Mizuno, A genome-wide view of the *Escherichia coli* BasS–BasR two-component system implicated in iron-responses. *Biosci. Biotechnol. Biochem.* **68**, 1758–1767 (2004).
- L. Selva, D. Viana, G. Regev-Yochay, K. Trzcinski, J. M. Corpa, I. Lasa, R. P. Novick, J. R. Penadés, Killing niche competitors by remote-control bacteriophage induction. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 1234–1238 (2009).
- A. Hsiao, A. M. S. Ahmed, S. Subramanian, N. W. Griffin, L. L. Drewry, W. A. Petri, R. Haque, T. Ahmed, J. I. Gordon, Members of the human gut microbiota involved in recovery from *Vibrio cholerae* infection. *Nature* **515**, 423–426 (2014).
- J. A. Thompson, R. A. Oliveira, A. Djukovic, C. Ubeda, K. B. Xavier, Manipulation of the quorum sensing signal AI-2 affects the antibiotic-treated gut microbiota. *Cell Rep.* **10**, 1861–1871 (2015).
- J. M. Kim, S. J. Cho, Y.-K. Oh, H.-Y. Jung, Y.-J. Kim, Nuclear factor-kappa B activation pathway in intestinal epithelial cells is a major regulator of chemokine gene expression and neutrophil migration induced by *Bacteroides fragilis* enterotoxin. *Clin. Exp. Immunol.* **130**, 59–66 (2002).
- S. Wu, P. J. Morin, D. Maouyo, C. L. Sears, *Bacteroides fragilis* enterotoxin induces c-Myc expression and cellular proliferation. *Gastroenterology* **124**, 392–400 (2003).
- K.-J. Rhee, S. Wu, X. Wu, D. L. Huso, B. Karim, A. A. Franco, S. Rabizadeh, J. E. Golub, L. E. Mathews, J. Shin, R. B. Sartor, D. Golenbock, A. R. Hamad, C. M. Gan, F. Housseau, C. L. Sears, Induction of persistent colitis by a human commensal, enterotoxigenic *Bacteroides fragilis*, in wild-type C57BL/6 mice. *Infect. Immun.* **77**, 1708–1718 (2009).
- T. P. Prindiville, R. A. Sheikh, S. H. Cohen, Y. J. Tang, M. C. Cantrell, J. Silva, *Bacteroides fragilis* enterotoxin gene sequences in patients with inflammatory bowel disease. *Emerg. Infect. Dis.* **6**, 171–174 (2000).
- S. Wu, K.-J. Rhee, E. Albesiano, S. Rabizadeh, X. Wu, H.-R. Yen, D. L. Huso, F. L. Brancati, E. Wick, F. McAllister, F. Housseau, D. M. Pardoll, C. L. Sears, A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. *Nat. Med.* **15**, 1016–1022 (2009).
- S. K. Mazmanian, J. L. Round, D. L. Kasper, A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature* **453**, 620–625 (2008).
- J. L. Round, S. K. Mazmanian, Inducible Foxp3⁺ regulatory T-cell development by a commensal bacterium of the intestinal microbiota. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 12204–12209 (2010).
- M. J. Coyne, A. O. Tzianabos, B. C. Mallory, V. J. Carey, D. L. Kasper, L. E. Comstock, Polysaccharide biosynthesis locus required for virulence of *Bacteroides fragilis*. *Infect. Immun.* **69**, 4342–4350 (2001).
- L. V. Blanton, M. J. Barratt, M. R. Charbonneau, T. Ahmed, J. I. Gordon, Childhood undernutrition, the gut microbiota, and microbiota-directed therapeutics. *Science* **352**, 1533 (2016).
- A. J. Bäuml, V. Sperandio, Interactions between the microbiota and pathogenic bacteria in the gut. *Nature* **535**, 85–93 (2016).
- N. Kato, C. Liu, H. Kato, K. Watanabe, H. Nakamura, N. Iwai, K. Ueno, Prevalence of enterotoxigenic *Bacteroides fragilis* in children with diarrhea in Japan. *J. Clin. Microbiol.* **37**, 801–803 (1999).
- B. Chevreux, T. Wetter, S. Suhai, Computer Science and Biology, in *Proceedings of the German Conference on Bioinformatics*, Hannover, 4 to 6 October 1999 (GCB, 1999), pp. 45–46.
- T. Seemann, Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
- L. Chen, Z. Xiong, L. Sun, J. Yang, Q. Jin, VFDB 2012 update: Toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res.* **40**, D641–D645 (2012).
- S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, S. L. Salzberg, Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
- V. K. Ridaura, J. J. Faith, F. E. Rey, J. Cheng, A. E. Duncan, A. L. Kau, N. W. Griffin, V. Lombard, B. Henrissat, J. R. Bain, M. J. Muehlbauer, O. Ilkaveya, C. F. Semenkovich, K. Funai, D. K. Hayashi, B. J. Lyle, M. C. Martini, L. K. Ursell, J. C. Clemente, W. V. Treuren, W. A. Walters, R. Knight, C. B. Newgard, A. C. Heath, J. I. Gordon, Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* **341**, 1241214 (2013).
- N. B. Shoemaker, C. Getty, J. F. Gardner, A. A. Salyers, Tn4351 transposes in *Bacteroides* spp. and mediates the integration of plasmid R751 into the *Bacteroides* chromosome. *J. Bacteriol.* **165**, 929–936 (1986).

47. J. An, D. M. Muoio, M. Shiota, Y. Fujimoto, G. W. Cline, G. I. Shulman, T. R. Koves, R. Stevens, D. Millington, C. B. Newgard, Hepatic expression of malonyl-CoA decarboxylase reverses muscle, liver and whole-animal insulin resistance. *Nat. Med.* **10**, 268–274 (2004).
48. C. T. Ferrara, P. Wang, E. C. Neto, R. D. Stevens, J. R. Bain, B. R. Wenner, O. R. Ilkayeva, M. P. Keller, D. A. Blasiolo, C. Kendzioriski, B. S. Yandell, C. B. Newgard, A. D. Attie, Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling. *PLoS Genet.* **4**, e1000034 (2008).
49. M. V. Jensen, J. W. Joseph, O. Ilkayeva, S. Burgess, D. Lu, S. M. Ronnebaum, M. Odegaard, T. C. Becker, A. D. Sherry, C. B. Newgard, Compensatory responses to pyruvate carboxylase suppression in islet β -cells. Preservation of glucose-stimulated insulin secretion. *J. Biol. Chem.* **281**, 22342–22351 (2006).
50. M. Monetti, M. C. Levin, M. J. Watt, M. P. Sajan, S. Marmor, B. K. Hubbard, R. D. Stevens, J. R. Bain, C. B. Newgard, R. V. Farese Sr., A. L. Hevener, R. V. Farese Jr., Dissociation of hepatic steatosis and insulin resistance in mice overexpressing DGAT in the liver. *Cell Metab.* **6**, 69–78 (2007).
51. A. H. Merrill Jr., M. C. Sullards, J. C. Allegood, S. Kelly, E. Wang, Sphingolipidomics: High-throughput, structure-specific, and quantitative analysis of sphingolipids by liquid chromatography tandem mass spectrometry. *Methods* **36**, 207–224 (2005).
52. J. J. Faith, P. P. Ahern, V. K. Ridaura, J. Cheng, J. I. Gordon, Identifying gut microbe-host phenotype relationships using combinatorial communities in gnotobiotic mice. *Sci. Transl. Med.* **6**, 220ra11 (2014).

Acknowledgments: We thank D. O'Donnell, M. Karlsson, S. Wagoner, and J. Serugo for assistance with gnotobiotic mouse husbandry; M. Meier, S. Deng, and J. Hoisington-Lopez for superb technical assistance; and M. Charbonneau and M. Barratt for providing valuable insights during the course of this study. **Funding:** This study was supported by the Bill & Melinda Gates Foundation and the NIH (DK30292). N.D. is the recipient of a Young Investigator

Grant for Probiotics Research from the Global Probiotics Council. P.P.A. is the recipient of a Sir Henry Wellcome Postdoctoral Fellowship from the Wellcome Trust (096100). **Author contributions:** V.E.W., N.D., and J.I.G. designed the experiments. V.E.W. and N.G. designed the representative Bangladeshi diet embodiments. R.H., T.A., and W.P. directed the clinical study design, enrollment, plus clinical data and sample collection for the Bangladeshi birth cohort. V.E.W. and N.D. performed gnotobiotic mouse experiments and generated 16S rRNA, COPRO-Seq, microbial RNA-seq, and bacterial isolate genome sequencing data sets. V.E.W. and J.G. produced culture collections from the fecal microbiota of the healthy and stunted Bangladeshi children. J.C., O.I., and L.V.B. conducted MS-based studies of biospecimens obtained from gnotobiotic mice. P.P.A. performed FACS of immune cell populations. N.P.S. and N.D. measured serum cytokine levels. N.D., V.E.W., A.H., T.S.S., J.C., P.P.A., N.G., C.B.N., and J.I.G. analyzed the data. N.D., V.E.W., and J.I.G. wrote the paper. **Competing interests:** J.I.G. is a cofounder of Matatu Inc, a company characterizing the role of diet-by-microbiota interactions in animal health. The other authors declare that they have no competing interests. **Data and materials availability:** 16S rRNA, COPRO-Seq, and microbial RNA-seq data sets, plus whole-genome shotgun sequencing data sets from cultured bacterial strains are available through the European Nucleotide Archive (study accession number PRJEB9703).

Submitted 30 June 2016
Accepted 4 November 2016
Published 23 November 2016
10.1126/scitranslmed.aah4669

Citation: V. E. Wagner, N. Dey, J. Guruge, A. Hsiao, P. P. Ahern, N. P. Semenkovich, L. V. Blanton, J. Cheng, N. Griffin, T. S. Stappenbeck, O. Ilkayeva, C. B. Newgard, W. Petri, R. Haque, T. Ahmed, J. I. Gordon, Effects of a gut pathobiont in a gnotobiotic mouse model of childhood undernutrition. *Sci. Transl. Med.* **8**, 366ra164 (2016).



Effects of a gut pathobiont in a gnotobiotic mouse model of childhood undernutrition

Vitas E. Wagner, Neelendu Dey, Janaki Guruge, Ansel Hsiao, Philip P. Ahern, Nicholas P. Semenkovich, Laura V. Blanton, Jiye Cheng, Nicholas Griffin, Thaddeus S. Stappenbeck, Olga Ilkayeva, Christopher B. Newgard, William Petri, Rashidul Haque, Tahmeed Ahmed and Jeffrey I. Gordon (November 23, 2016)
Science Translational Medicine 8 (366), 366ra164. [doi: 10.1126/scitranslmed.aah4669]

Editor's Summary

Neighbors matter

A big unanswered question is what determines the effects of enteropathogen burden in children who are undernourished or at risk for undernutrition. In a new study, Wagner and colleagues introduce collections of sequenced gut bacterial strains cultured from healthy or underweight Bangladeshi children into germfree mice fed diets resembling those consumed by the children. The gut bacterial strains were transplanted with or without nontoxigenic or enterotoxigenic *Bacteroides fragilis* strains. Addition of enterotoxigenic *B. fragilis* induced cachexia in the transplanted mice, and altered gene expression and metabolic activity of the transplanted bacterial strains. These effects were mitigated by cocolonization with nontoxigenic *B. fragilis*, illustrating the influence of intra- and interspecies interactions in determining the impact of an enteropathogen on its host.

The following resources related to this article are available online at <http://stm.sciencemag.org>.
This information is current as of March 16, 2017.

- | | |
|-------------------------------|--|
| Article Tools | Visit the online version of this article to access the personalization and article tools:
http://stm.sciencemag.org/content/8/366/366ra164 |
| Supplemental Materials | " <i>Supplementary Materials</i> "
http://stm.sciencemag.org/content/suppl/2016/11/21/8.366.366ra164.DC1 |
| Permissions | Obtain information about reproducing this article:
http://www.sciencemag.org/about/permissions.dtl |

Science Translational Medicine (print ISSN 1946-6234; online ISSN 1946-6242) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005. Copyright 2017 by the American Association for the Advancement of Science; all rights reserved. The title *Science Translational Medicine* is a registered trademark of AAAS.

Appendix B

Alejandro Reyes, Nicholas P. Semenkovich, Katrine Whiteson, Forest Rohwer & Jeffrey I. Gordon.

Going viral: next-generation sequencing applied to phage populations in the human gut.

Nature Reviews Microbiology 10, 607-617 (September 2012)

REVIEWS

Going viral: next-generation sequencing applied to phage populations in the human gut

Alejandro Reyes¹, Nicholas P. Semenkovich¹, Katrine Whiteson², Forest Rohwer² and Jeffrey I. Gordon¹

Abstract | Over the past decade, researchers have begun to characterize viral diversity using metagenomic methods. These studies have shown that viruses, the majority of which infect bacteria, are probably the most genetically diverse components of the biosphere. Here, we briefly review the incipient rise of a phage biology renaissance, which has been catalysed by advances in next-generation sequencing. We explore how work characterizing phage diversity and lifestyles in the human gut is changing our view of ourselves as supra-organisms. Finally, we discuss how a renewed appreciation of phage dynamics may yield new applications for phage therapies designed to manipulate the structure and functions of our gut microbiomes.

CRISPRs

(Clustered regularly interspaced short palindromic repeats). Widespread genetic systems in bacteria and archaea, consisting of multiple copies of palindromic repeats flanking short spacers of viral or plasmid origin. CRISPR elements provide acquired resistance to foreign DNA.

From Alfred Hershey and Martha Chase's studies indicating that DNA is the genetic material¹, to Francis Crick and Sydney Brenner's experiment establishing the triplet nature of the genetic code², viruses that infect bacteria (referred to below as phages) have helped define the fundamental components of modern biology. Most of the tools for early molecular biology arose from the work of phage biologists³. The first genomes sequenced were from phages and other viruses, and the first comparisons of multiple genomes were carried out on phages of *Lactobacillus* and *Mycobacterium* spp. These early studies showed that there is extensive diversity in essentially every phage community. It is now clear that viruses are the most diverse and uncharacterized components of the major ecosystems on Earth⁴ and have intricate roles in ecosystem function that go far beyond simple predator-prey dynamics⁵ (BOX 1).

The clinical world has become increasingly interested in phage-based therapeutics because of the increased prevalence of antibiotic-resistant bacteria⁶. The idea of using phages as therapeutic tools is not new. Félix d'Herelle, the co-discoverer of phages, recognized their potential medical applications nearly a century ago⁷, and his first phage therapies were tested as early as 1919 (REF. 8). However, our then-rudimentary understanding of the composition and dynamic operations of the human microbiome, our lack of knowledge of phage biology, and poor quality

control during phage production made this therapeutic approach unreliable⁹.

In this Review, we detail recent advances in the rising field of viral metagenomics, with an emphasis on our current views of the phage communities associated with the human gut. We do not discuss many of the mechanisms of resistance to phages, such as CRISPRs (clustered regularly interspaced short palindromic repeats), or the extremely large field of eukaryotic virus discovery, which is being propelled by metagenomics, as these topics have been reviewed elsewhere¹⁰⁻¹⁴. Instead, we focus exclusively on viruses that infect bacteria. There are many challenges facing phage metagenomics at the present time, several of which are discussed below; some of the technical improvements that are required to overcome these challenges are outlined in BOX 2.

Methods for viral metagenomics

The introduction of small subunit (SSU) rRNA — that is, 16S rRNA — as a reliable bacterial and archaeal phylogenetic marker¹⁵⁻¹⁷ allowed remarkable insights to be gained about the diversity and dynamics of microbial communities. Phylogenetic-marker 'envy' rapidly 'infected' the psyche of phage biologists: they did not have an SSU rRNA equivalent, and there was (and still is) no conserved protein or gene enabling a similar characterization of all or even the majority of phages present in a sample. Efforts to characterize phage diversity focused

¹Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, Missouri 63108, USA.

²Department of Biology, San Diego State University, San Diego, California 92182, USA.

Correspondence to J.I.G.
e-mail: jgordon@wustl.edu
doi:10.1038/nrmicro2853
Published online 6 August 2012

Box 1 | Phage–bacterial host dynamics: lessons learned from environmental ecosystems

Most of life on Earth exists as bacteria and archaea in oceans, sediments, land and the deep biosphere¹¹². In the early 1980s, researchers became aware that there are literally millions of actively growing microorganisms per millilitre of sea water, and efforts to characterize the impact of phage life cycles on planetary-scale biogeochemistry were initiated^{113–116}. As an example of how dramatic the effects of phages can be, consider the widespread cyanobacterial clades *Prochlorococcus* and *Synechococcus*. The species in these two genera of unicellular algae carry out about half of the primary production in the world's oceans; these bacteria are also infected with cyanophages, and variations in cyanophage and bacterial concentrations are tied to daily and seasonal cycles¹¹⁵. The isolation and characterization of phages infecting *Synechococcus* spp. (in the 1990s) and *Prochlorococcus* spp. (in the 2000s) revealed not only that cyanophages are widespread, often infecting 40–50% of cyanobacteria, but also that they kill 10–50% of their hosts daily^{114,115}, rapidly driving the diversification of their hosts as these bacteria co-evolve resistance¹¹⁷, and simultaneously driving carbon into a dissolved form when bacterial cells lyse. When levels of available resources are low, cyanophages can enter a lysogenic state. A growing list of genes that are important for bacterial host metabolism and function have been found in marine phages, including photosystem genes that can increase the host's photosynthetic output and maintain energy production during infection so that phages can eventually lyse their host cells^{51,118–120}. Lysogeny may be an important lifestyle under a number of suboptimal conditions, including when host abundance or nutrient abundance is low¹²¹.

Bacteria can use temperate phages to enable invasion of new habitats. When part of the bacterial population is sacrificed through phage lysis, the released phages will target competitors but allow bacterial kin harbouring the prophages to survive, as they are resistant to attack owing to a process called superinfection exclusion^{122–125}.

The concept that a virus can have a beneficial effect beyond that experienced directly by its host cell is not novel. Three-way symbioses have been well described in macro- and micro-ecosystems. For example, a symbiotic bacterium that inhabits the pea aphid protects the aphid from a wasp that can otherwise lay eggs in the aphid haemocoel; this protection is conferred by a phage-encoded toxin expressed by the bacterium^{126,127}. Drought, heat and cold tolerance are conferred to plants through viruses^{122,128}. In Yellowstone National Park, USA, a fungal endophyte infecting panic grass confers thermal tolerance, allowing the grass to grow in hot geothermal soils; however, the fungus is not heat tolerant without the virus that infects it¹²⁹. In a phage-related example, there is a reported case of phage-associated corynetoxin synthesis in the bacterium *Rathayibacter toxicus* (formerly *Clavibacter toxicus*), which colonizes ryegrass plants; the toxin makes the grass toxic to grazing animals such as sheep¹³⁰.

Virus-like particles

(VLPs). Particles that can be recovered from microbial communities using physical separation methods such as density gradient ultracentrifugation and/or filtration. Purified VLPs have physical characteristics that resemble those of viruses, although their capacity for infection has to be subsequently defined.

Virotype

A taxonomic classification that is typically based on a selected percentage identity threshold among viral reads, rather than on phylogenetic markers.

Multiple displacement amplification

A method for exponential isothermal amplification of a DNA template using ϕ 29 DNA polymerase and random primers. Exponential amplification is achieved by attachment of the polymerase to newly elongated fragments, coupled with the strong displacement activity of the enzyme on extension.

Deep biosphere

The deepest oceanic regions in which life is supported.

on partially conserved fragments of phage genes such as those encoding polymerases. However, this method was useful only within certain viral families^{18–20}. Horizontal gene transfer further complicates the use of marker genes in phages. For example, most members of the order Caudovirales have identifiable conserved genes, including those encoding terminases, portal proteins and capsid proteins, but horizontal transfer and recombination events generate extensive genome mosaicism that challenges phylogenetic characterization²¹. The arrival of next-generation sequencing, together with methods for purifying virus-like particles (VLPs), set the stage for defining viral diversity using shotgun sequencing.

Purification of VLPs. Although viruses outnumber microbial cells 10:1 in most environments, viral DNA represents 2–5% of the total DNA in a microbial community^{22–24}. For this reason, it is often desirable to separate viruses from microbial cells. If the sample volume is large and the viral density low (such as in ocean environments), tangential flow filtration can be used to recover and concentrate VLPs. For solid samples with a high viral density, such as faeces, a common approach is to resuspend the material in an osmotically neutral buffer, followed by one or more steps designed to remove large particles, including undigested or partially digested food fragments and microbial cells, by caesium chloride density gradient ultracentrifugation²⁵ and subsequent filtration (FIG. 1). This procedure has been successfully applied to faecal material that has been stored at -80°C for several years, without any pre-processing of the sample,

indicating that VLP structures are stable under conditions of freezing and thawing²².

Amplification of VLP-derived DNA. After VLPs are purified, non-encapsulated free nucleic acids are removed by treatment with DNase and RNase, and VLP-derived nucleic acids are then isolated. The methods chosen determine the purity of the DNA and RNA, influence the DNA and RNA yields, and represent a selection step that can bias the subsequent interpretation of virotype abundance and viral community diversity²⁶. Unfortunately, the yield of DNA following extraction of nucleic acids from purified VLPs is often below the required minimum for sequencing. Therefore, a range of amplification methods have been developed, such as random amplified shotgun library (RASL)²⁷ and linker-amplified shotgun library (LASL)²⁸, among others^{29–32}. Caveats concerning random PCR amplification of viral DNA include an uneven coverage of viral genomes, the limitation of this approach to double-stranded DNA (dsDNA) templates, and an inherent bias attributable to the exponential amplification of mixed templates. Another common method uses the phage-derived ϕ 29 polymerase for multiple displacement amplification (MDA)³³. MDA takes advantage of the high processivity of this DNA polymerase ($>70,000$ nucleotides (nt) per association–dissociation cycle) and its strong strand displacement capability, which together permit the amplification of complete viral genomes. The result is a fast method that can efficiently amplify minute amounts of both single-stranded DNA (ssDNA) and dsDNA. Although the method is fast, it is not without flaws, including the overamplification of small circular ssDNA

Box 2 | Technical challenges in viral metagenomics

- New and better tools for the recovery of virus-like particles (VLPs) from small amounts of starting microbial community biomass, and methods for less biased amplification of extracted nucleic acids before shotgun sequencing.
- Improved methods for deep-draft assemblies of full-length viral genomes. Particularly problematic are the ends of phage genomes, which can be blocked or permuted or can have hairpins.
- The automation of methods (for example, MaxiPhi) for carrying out comparative metagenomics and estimating α -diversity, β -diversity and γ -diversity, in order to describe the pan-virome in a given environment.
- New and better tools for defining the host specificity of known and novel phages from either assembled genomes or VLP-derived short-read sequences, and for identifying the determinants of microbial host cell range.
- Improved methods of experimentally and computationally assigning functions to 'conserved' viral genes with no known functions (shedding light on the 'genetic dark matter' that is represented by conserved hypothetical genes).
- Better *in vitro* and *in vivo* models for determining phage–bacterial host dynamics and the impact of these interactions on energy availability and niche partitioning in a microbiota.
- Experimental and computational methods, and related visualization tools, for efficient analyses of temporal variation in model microbial communities, and for measuring the effects of perturbations.
- Models for predicting the cost or benefit of having prophage present in a candidate probiotic species; for example, weighing invasiveness and persistence in a targeted microbiota.
- Methods to test the utility of phages to directly perturb a targeted microbiota in ways that facilitate invasion by a probiotic species or species consortium.

viruses³⁴ and the potential formation of chimaeras^{35,36}. Procedures for avoiding some of these limitations continue to be developed^{37,38}, including a novel transposon-based method for rapidly generating DNA libraries from small quantities of dsDNA³⁹. RNA viruses can be sequenced by reverse transcription followed by application of the protocols described above. Alternatively, whole-transcriptome amplification approaches can be used⁴⁰.

Sequencing strategies. Although sequencing costs are falling at an astonishingly rapid rate as newer technologies offer higher degrees of parallelism (that is, greater numbers of reads per run, and multiplex sequencing using sample-specific DNA bar-codes), read length matters⁴¹. When characterizing a viral community from which most of the sequences are novel and enriched in regions of low-complexity repeats, obtaining the longest possible reads will aid accurate assembly and taxonomic assignment^{41,42}. The earliest next-generation sequencing analyses were powered by the 454 GS20 instrument (from Roche) with ~100 nt reads. Advances in pyrosequencing technology, including today's 454 FLX+ machine, have produced average read lengths that exceed 800 nt^{22,23,43–46}. Most recently, total microbial-community DNAs have been subjected to deep shotgun sequencing with more highly parallel Illumina instruments; analyses of the resulting metagenomic data sets have shown that the percentage of reads with similarity to known viral sequences is generally less than 0.01%^{30,47,48}. This low value is in part due to the short read length (≤ 100 nt). However, the percentage of similarity increases when VLPs are purified³². Other studies have obtained better assignments by assembling the short reads^{49,50}.

In summary, researchers engaged in viral metagenomic studies have tended to opt for technologies that prioritize long read lengths over those that offer short read lengths but higher throughput. However, as the high-throughput platforms approach 150 nt read lengths

and 250 million reads per lane (such as the Illumina HiSeq 1500 and 2500 instruments), and as the cost per read falls, we will undoubtedly see a rapid transition to these types of sequencers — so long as improvements in assembly algorithms keep pace (FIG. 1).

Computational approaches for characterizing sequenced viromes. To address the question of viral community composition, shotgun metagenomic sequences are typically compared to individual viral genomes. Although public sequence databases have expanded considerably — from 500 viral genomes in 2007 to more than 3,000 full viral genomes in 2012 — the number of deposited genomes is far less than the expected number of virotypes present in 100 litres of sea water⁵¹. Compounding this problem, existing databases include few viral proteins in their training sets, meaning that many taxonomic assignments are based on proteins that have been transferred from a virus to a microbial host or that are present in prophages and are described as part of a microbial genome. Databases with a particular focus on viruses are under development and include A Classification of Mobile Genetic Elements (ACLAME)⁵² and Phage SEED⁵³. Novel data analysis pipelines are also being constructed to improve the accuracy and efficiency of homology searches (see REFS 54–59, and Phage SEED and Virome).

When taxonomic and functional assignments have been made for a given sample, a viral community profile can be created that characterizes the diversity present in that sample. Multidimensional reduction methods such as principal component analysis (PCA) and hierarchical clustering have been used to visualize similarities among viral communities, and methods such as supervised learning can help to identify discriminatory features.

Given that most of the available viral metagenomic data lack similarity to entries in databases, similarity-independent methods have been developed to better understand viral community structure. One example,

Prophages

Temperate phages in a host-incorporated state.

 α -diversity

Diversity, whether defined using taxonomic or functional characteristics, within a particular locale (habitat) at a particular moment in time.

 β -diversity

Diversity measured between samples or locales at a particular moment in time or over time.

 γ -diversity

A combination of alpha and beta diversity.

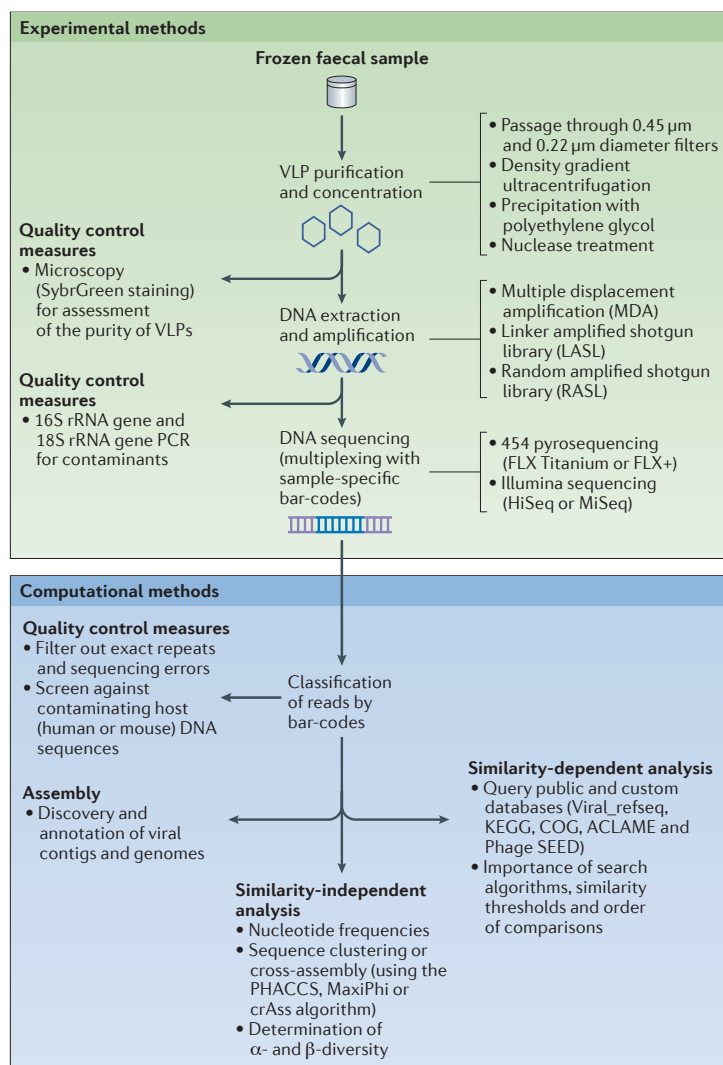


Figure 1 | Experimental and computational methods for the characterization of the phage populations present in the human gut microbiota. See main text for further details. VLP, virus-like particle.

Pan-genome

The global gene repertoire of a microbial species; defined by sequencing the genomes of isolates of that species obtained from a single or multiple habitats.

Bacterial phylotypes

Taxonomic classifications that are based on phylogenetic markers, classically the 16S rRNA gene. Isolates can be arbitrarily assigned to a species-level phylotype if they share ≥97% sequence identity among their 16S rRNA genes.

PHACCS (Phage Communities from Contig Spectrum), was designed to quantify virotypes^{60,61} on the basis of the assumption that if a virotype is present in high abundance in a VLP sample, it is more likely to be assembled into a large contig. Moreover, we can posit that if assembly of a single sample data set allows prediction of community structure and diversity, then pooling two samples together and carrying out a cross-assembly analysis could determine the inter-sample diversity (by, for example, using MaxiPhi⁶²). Another alternative for identifying shared viruses among different samples comes from crAss, an algorithm that allows for the simultaneous cross-assembly of all the samples in a data set as opposed to the pairwise assemblies used in MaxiPhi. As more tools are developed, special attention

should be given to the assembly parameters in order to prevent mixed assemblies and chimeras between viral genomes (FIG. 1).

Phages in the human gut

The gut provides an exciting place to characterize the role of phages in community assembly and dynamics. Assembly of the human gut microbiota begins at birth, with evolution towards an adult-like configuration during the first 3 years of life⁶³. The importance of early environmental exposures is emphasized by the fact that the overall phylogenetic composition of the gut microbiota is not significantly more similar in adult monozygotic twins than in dizygotic twins, and the fact that family members have a higher degree of similarity in the gut microbiota than unrelated individuals living in different households. These patterns are robust to different cultural traditions, and the observations about mono- versus dizygotic twins apply to children as well as teenagers and adults^{63,64}. Microorganisms in this densely populated ecosystem are engaged in a constant fight for nutrients and survival. Peristalsis moves an ephemeral menu of dietary components along the cephalocaudal axis of the intestine, and the microbial inhabitants face the omnipresent threat of washout from the gut 'bioreactor'. Maintaining a foothold in this ecosystem depends not only on physical interactions with the perpetually renewing mucous layer and partially digested food particles, but also on functional interactions with other community members. Preserving functional redundancy contributes to community resilience, and horizontal gene transfer provides an opportunity to constantly modify the pan-genome of a given species-level bacterial phylotype. Each adult appears to harbour a persistent collection of one hundred, or at most a few hundred, species in their intestines, although strain level diversity is great^{24,65,66}. Although the proportional representation of taxa changes as the community responds to various environmental perturbations, intrapersonal variation in species content is considerably less than interpersonal differences^{63,64,67,68}.

As our knowledge of inter- and intrapersonal variations in the microbiota has increased, a lagging question has been the role of phages in shaping the properties of these microbial communities. Although a number of individual phages have been extensively characterized, providing an important genomic context against which metagenomic data can be interpreted, recently more attention has been given to phage dynamics at the microbial community level. In 2003, the first report of a human-associated gut virome was published; it described the results of shotgun Sanger sequencing of DNA from VLPs isolated from a faecal sample obtained from a single healthy adult. The identifiable fraction of the virome was dominated by phages, including temperate phages. This report estimated that there were 1,200 different virotypes in the single sample analysed, with the majority assigned to the *Siphoviridae* family²⁸. *Siphoviridae* and temperate phages have subsequently been reported to be the most abundant identifiable viruses in other sampled faecal viromes, followed by members of the family *Podoviridae*^{22,23,69}.

The prominence of members of the family *Microviridae* in adult human gut microbiota was initially dismissed as an artefact of the MDA method, which has a preference for ssDNA. However, a novel branch of the family *Microviridae* has been identified recently, from prophages present in the genomes of *Bacteroides* and *Prevotella* spp.⁷⁰. These two genera are prominently represented in the microbiota of adult human populations living in a number of diverse geographical areas^{63,71}. Another study characterizing *Microviridae* members from healthy human donors also clustered these novel viruses with prophages from *Bacteroides* and *Prevotella* spp.⁶⁹. These analyses suggest that *Microviridae* is an important viral family in the human gut, and also that what was previously considered to be an exclusively lytic phage can in fact integrate into bacterial hosts in an environment that encourages a temperate (lysogenic) virus–host lifestyle (see FIG. 2 and below).

Marine environments can contain millions of different virotypes in a single sample⁵¹. By comparison, none of the human faecal samples characterized thus far has had more than 1,500 virotypes. Moreover, the ratio of virotypes to species-level bacterial phylotypes is 10:1

in the ocean but closer to 1:1 in the gut²². Microscopy counts have further validated these estimated ratios, demonstrating an average of 10^8 – 10^9 VLPs per gram of faeces compared with $\sim 10^9$ bacterial cells per gram⁶⁹. These findings also support the notion that phages exhibit a more temperate lifestyle in the gut, in contrast to the active kill-the-winner virus–bacterial host dynamic manifest in marine environments.

The temperate lifestyle of phages observed in the gut, along with bacterial CRISPR elements involved in conferring immunity to infection with foreign DNA (including phages), has facilitated bioinformatic efforts to discover new phages in this body habitat. Recently, data sets obtained from deep shotgun sequencing of human faecal-community DNA²⁴ were used to extract CRISPR spacers present in gut bacterial genomes⁷². These spacers were then used to query data sets of shotgun sequencing reads from faecal VLP-derived DNA^{22,23}, and contigs assembled from total faecal-community DNA²⁴. This approach allowed a large collection of these previously unassigned contigs to be designated as viral and assigned to potential bacterial hosts. This study also led to an appreciation of the wide distribution of novel phages across human gut communities.

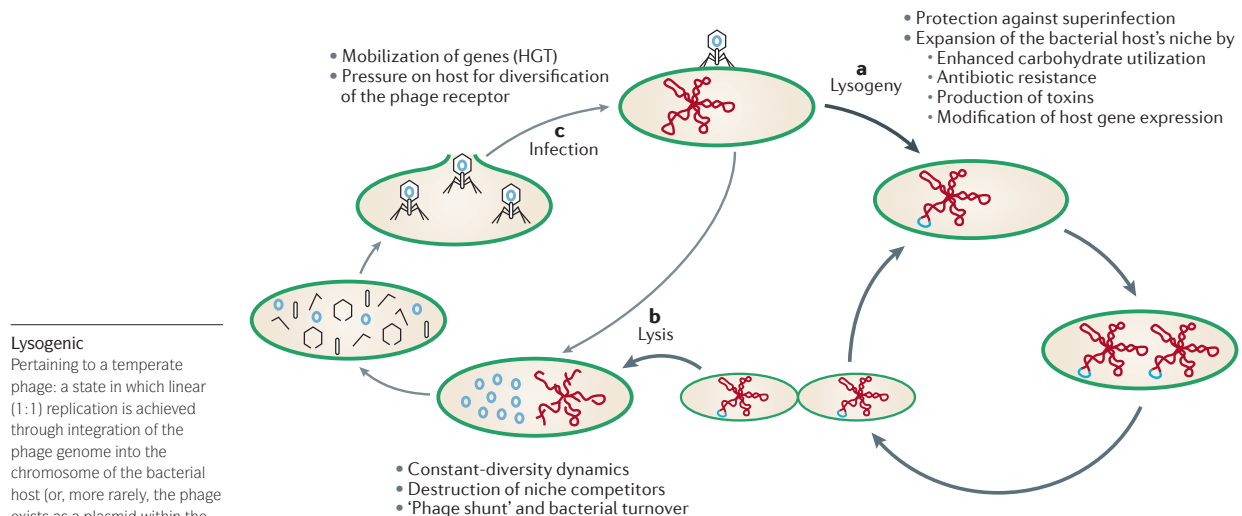


Figure 2 | Potential consequences of a temperate phage life cycle in the human gut. Metagenomic studies of viruses suggest that a temperate lifestyle is dominant in the distal human gut, in contrast to the lytic lifestyle observed in open oceans. This temperate lifestyle can have benefits for the phage and the bacterial host, and can alter phage–host dynamics. Integration as a prophage (part **a**) protects the host from superinfection, effectively 'immunizing' the bacterial host against infection from the same or a closely related phage. Furthermore, the genes encoded by the phage genome may expand the niche of the bacterial host by enabling metabolism of new nutrient sources (for example, carbohydrates), providing antibiotic resistance, conveying virulence factors or altering host gene expression. This temperate (lysogenic) life cycle allows phage expansion in a 1:1 ratio with the bacterial host. If the prophage conveys increased fitness to its bacterial host, there will be an increased prevalence of the host and phage in the microbiota. Induction of a lytic cycle (part **b**) can follow a lysogenic state and can be triggered by environmental stress. As a consequence, bacterial turnover is accelerated and energy utilization is optimized through a 'phage shunt', in which the debris remaining after lysis is used as a nutrient source by the surviving bacterial population. Furthermore, a bacterial subpopulation that undergoes lytic induction sweeps away other sensitive species and increases the niche for survivors (that is, bacteria that already have the specific phage integrated into their genome). Periodic induction of prophages can also lead to a constant-diversity dynamic¹³⁹, which helps maintain community structure and functional efficiency. Novel infections or infections of novel bacterial hosts by phages (part **c**) bring the benefit of horizontally transferred genes and create selective pressure on the hosts for diversification of their phage receptors, which are often involved in carbohydrate utilization. HGT, horizontal gene transfer.

Temporal variation. To date, only three reported metagenomic studies of the human gut virome have characterized temporal variation^{22,23,73}. One of these studies used VLPs that were isolated from frozen faecal samples collected from four adult female monozygotic twin pairs and their mothers at three time points over a 12-month period²². VLP-derived virome data sets were compared to data sets of both sequenced bacterial 16S rRNA genes, and shotgun reads from total faecal-community DNA generated from the same faecal samples used to prepare the VLPs. The results disclosed that the viromes of co-twins and their mothers exhibited a significantly greater degree of interpersonal variation than did the corresponding bacterial communities. Despite the marked interpersonal variation in gut viromes and their encoded gene functions, intrapersonal diversity was extremely low: >95% of virotypes were retained over the period surveyed, and DNA viromes were dominated by a few temperate phages that exhibited remarkable genetic stability (>99% sequence conservation). These observations suggest that a temperate viral lifestyle is more prevalent in the distal intestine than a kill-the-winner dynamic (see FIG. 2).

Another study of temporal variation involved adults who were subjected to a defined diet for a period of 8 days²³. During this time, both bacterial and viral faecal communities changed in a comparable manner. Importantly, interpersonal variation at the late time points was reduced among individuals consuming the same diet, suggesting that diet has an important effect in shaping both bacterial and viral components of the microbiota.

A third study examining temporal variation characterized the DNA virome of a 1-week-old healthy infant and used DNA microarrays to compare relative viral abundances in the faecal microbiota between postnatal week one and two⁷³. The results showed that the viral population changes drastically in the early stages of human life: more than half of the virotypes that were present at week one were undetectable by week two. Although these findings contrast with the stability of the DNA viromes in healthy adult faeces, they are consistent with the dynamic and rapid nature of assembly of the infant bacterial microbiota^{63,74}.

Functions encoded in phage genomes. There are a number of examples of known phage-encoded host fitness factors in gut bacteria (for example, the phage λ lipoprotein Bor and outer-membrane protein Lom, and the phage 933W Shiga-like toxin 2 (SLT2; also known as Stx2 (REFS 75,76)), but most of these appear to be virulence determinants of one kind or another. When comparing purified VLP-derived viromes and faecal microbiomes in the study of monozygotic twins, phages have been shown to exhibit enrichment for genes involved in anaerobic nucleotide synthesis, as well as cell wall biosynthesis and degradation²². Other distinctive features of phage genomes include genes that can alter bacterial phage-receptors and prevent superinfection⁷⁷. Interestingly, many phage receptors may be involved in carbohydrate transport and utilization. In an

environment such as the gut, where carbohydrate utilization is an important fitness factor, mobilization of these bacterial genes by phages could endow bacterial hosts with benefits (FIG. 2). There are probably a great number of bona fide metabolic and other fitness factors that have yet to be characterized which are encoded by phages.

Intriguingly, new evidence suggests that carbohydrate-binding components of the human gut virome can change at an extremely high rate. A recent metagenomic study examining VLPs purified from faecal samples collected from 12 humans identified 51 hypervariable loci — areas with mutation rates that are much higher than those for the rest of the viral genomes⁵⁰. Structural predictions for proteins putatively encoded by these regions revealed little homology to known folds; however, some of these proteins have similarity to immunoglobulin superfamily domains, and others to C-type lectin domains that participate in carbohydrate binding. Moreover, these loci appear to be specifically targeted for mutation by a reverse transcriptase-based mechanism, suggesting that a crucial functional advantage is provided by these hypervariable loci. It is tempting to speculate that these loci confer a selective advantage to phages, enabling immune evasion through immunoglobulin A (IgA) binding, or improving the chances of infecting a host cell in the rapidly changing conditions of the gut through adaptable binding to relevant environmental materials or bacterial surface receptors. There is a well-documented precedent for this scenario: hypervariable loci confer a fitness advantage in phages that infect *Bordetella* spp., as they allow tropism switching in the phage receptor-binding protein⁷⁸. The hypothesis that these loci allow a phage to bind IgA or environmental ligands is speculative and needs experimental validation; non-receptor structural phage proteins have been found to contain Ig-like domains, which may aid in host binding by weakly interacting with the cell surface⁷⁹.

RNA virome. The RNA virome of two healthy adults has been characterized by purifying and sequencing the RNA from VLPs⁸⁰. In this study, most RNA viruses appeared to be consumed together with food. Indeed, a pepper-associated virus (pepper mild mottle virus) constituted more than 80% of the identifiable gut viruses. The only animal-infecting RNA virus observed was a picobirnavirus that had previously been found in the faeces of healthy individuals as well as in the faeces of patients with diarrhoea; this virus has not been associated with any particular disease to date.

Comparative studies of other mammals

Comparative studies of different mammalian species represent a source of information about the effects of environmental factors, including diet^{81,82}, and various host factors on phage diversity in the gastrointestinal tract. Extensive surveys have been conducted of viruses associated with different mammalian species, with the aim of finding potential sources of zoonoses, uncovering the aetiology of animal diseases and identifying common mammalian viruses¹³. These studies, which

Box 3 | Characterizing the eukaryotic virome in the gut of healthy individuals

The eukaryotic virome can be considered from at least three different perspectives: viruses associated with the eukaryotic component of the gut microbiota, viruses associated with various human cell populations exposed to this microbiota, and viruses associated with ingested food. Metagenomic studies of faecal microbial communities from healthy individuals indicate that these communities are dominated by phages rather than eukaryotic DNA viruses^{22,23,28,50,69,73}. Eukaryotic RNA viruses are abundant but appear to be largely derived from food⁸⁰.

Our view of the eukaryotic virome in the gut comes largely from metagenomic studies of patients with various gastrointestinal diseases^{131–137}. These studies have identified known enteric viruses (adenoviruses, rotaviruses, enteroviruses and noroviruses), and novel members of the Bocavirus, Picobirnavirus and Cosavirus genera and of the family *Anelloviridae* that are potential human pathogens, as well as novel viruses that may be related to diet (members of the genus Gyrovirus and of the families *Nodaviridae*, *Dicistroviridae*, *Virgaviridae* and *Partitiviridae*). The high prevalence of single-stranded eukaryotic DNA viruses in metagenomic data sets has led to a new perspective on the potential importance and diversity of these small viruses¹³⁸. Although these viruses were initially identified in symptomatic individuals, they have also been identified, at a similar prevalence, in asymptomatic contacts of individuals with disease. The broad representation of these eukaryotic viruses in the human gut as well as in other body habitats has prompted a call to consider the functional significance of the human eukaryotic virome¹². The almost ubiquitous presence of human viruses that are not phages and have not been associated with any disease indicates that viruses, especially those acquired in early childhood, might be essential for proper immune system development and that particular host genotypes or immunological constraints might cause normally benign viruses to induce disease states.

include surveys of mammals occupying distinct habitats^{29,31,32,83–88}, identified viruses from certain families in the human gut virome, further underscoring the prevalence and long-standing nature of the evolved virus–mammalian host relationship (BOX 3). In all these studies, ssDNA viruses were found to be ubiquitous and accompanied in some cases by positive-sense ssRNA enteric viruses^{43,87,88}.

An early survey of coliphages in cows, pigs and humans⁸⁹ showed that these phages are present in titres of up to 10^7 VLPs per gram of faeces and that temperate coliphages are the most common. Interestingly, humans and pigs (omnivores with simple guts) had higher counts of temperate coliphages than cows (herbivores with foregut fermentation chambers). In an independent study, estimates of phage diversity from bovine rumen fluid⁹⁰ suggested that up to 28,000 different virotypes can be present in titres as high as 10^9 VLPs per millilitre of sample, hinting at a strikingly higher viral diversity and abundance in the cow gut than in the human gut. In contrast to the large interpersonal variation observed in human gut viromes^{22,23}, this latter study demonstrated a high degree of similarity between the phage communities of co-habiting animals on a similar diet. Metagenomic studies in horses⁸³ (herbivores with a hindgut fermentation chamber) revealed an intermediate level of phage diversity between that documented in herbivorous foregut-fermenting ruminants and omnivorous mammals with simple guts.

Together, these studies suggest that diet, gut physiology and potentially the transit time of food all have important roles in determining the life cycle and diversity of phages in the mammalian gut. Further dissection of these relationships requires manipulable, representative

Coliphages
Phages that infect coliform bacteria, in particular *Escherichia coli*.

and defined *in vivo* models. Moving in this direction, Maura and colleagues used mice to study the effects of a lytic enteric phage, observing stable long-term phage replication over 3 weeks⁹¹. As noted below, gnotobiotic mouse models might also be very informative.

Phage therapy

Much has already been written about the history, successes and failures of phage therapy. Most of the studies to date have focused on the use of lytic phages to destroy pathogenic bacteria^{92,93} (FIG. 3). Clinically oriented phage research began soon after the discovery of phages, when Félix d'Herelle used phages to treat bacillary dysentery in a number of human patients⁸. However, this optimistic start led to several misconceptions and missteps, both scientific and political, regarding the use of phages. d'Herelle incorrectly assumed that there was only one universally efficacious strain of lytic phage⁹⁴, although we now know that phages exhibit exquisite host cell specificity. In the 1930s, pharmaceutical companies began distributing enormous amounts of lytic phages as generic antibacterial therapies, but in part because of the perceived universality of phages, they had little knowledge of the components of these products.

In retrospect, we know that many of the commonly used phage preparations were destroyed by the organo-mercury preservatives added to the vials that contained the preparations, or were contaminated with bacterial exotoxins secreted by the cultures used to generate the phages⁹⁵. Inevitably, these problems, along with manufacturing inconsistencies (the supposedly standardized strains of phages would change from batch to batch), led to distrust among the medical and scientific community.

The recent resurgence of phages as possible therapeutic agents has been driven by a number of factors. The alarming prevalence of antibiotic-resistant strains of pathogenic bacteria, combined with the inexorable spread of antibiotic-degrading enzymes, such as New Delhi metallo- β -lactamase (NDM1), have led to calls for new therapeutic strategies⁹⁶. From a practical standpoint, efforts in antibiotic discovery have produced few novel compounds over the past decade⁹⁷. Phages are promising tools because they are easy to manufacture, have good host specificity and can be genetically manipulated. Moreover, resistance to phages may develop more slowly than resistance to antibiotics, although the reasons for this are multifaceted⁹⁸. Phage resistance can occur spontaneously in cultures (as frequently as 1 in 10^5 cells), but there can be fitness costs associated with resistance. By contrast, many forms of antibiotic resistance cannot occur spontaneously but instead require the introduction of a foreign DNA element. In many ways, addressing bacterial resistance is much easier with phages than with antibiotics because one can isolate different phages, or the phage may spontaneously mutate to overcome host resistance.

Perhaps a more interesting question, in the context of community dynamics and our growing understanding of the virome and microbiome, is whether we can produce more subtle phenotypic shifts in an ecological niche. Rather than using phages to destroy a single pathogenic

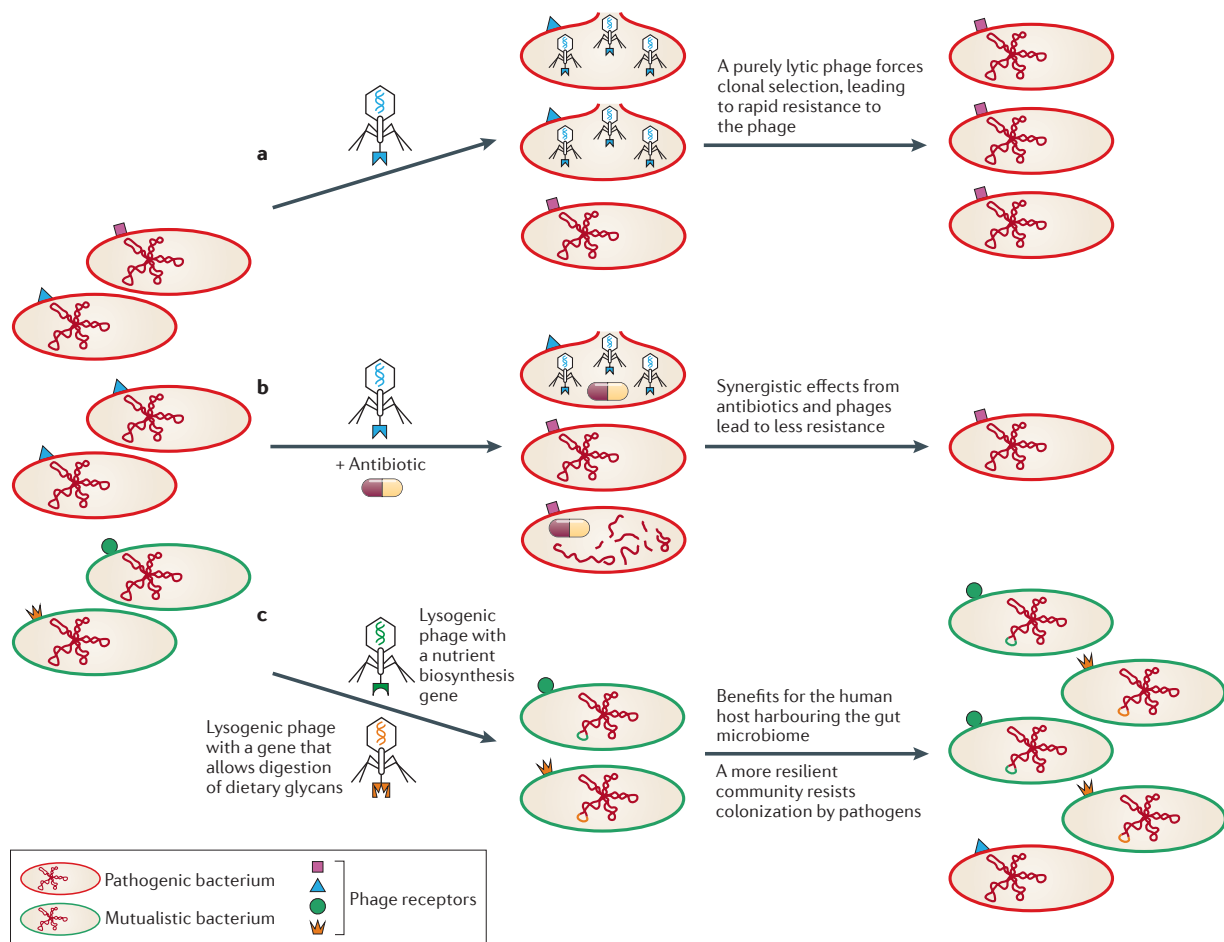


Figure 3 | Potential strategies for phage therapy. **a** | The traditional strategy for phage therapy has been to use a lytic phage that targets pathogenic bacteria. Although this approach is transiently useful, it can lead to rapid bacterial resistance owing to the resultant positive selection for subpopulations (clones) that are resistant to the lytic phage. Note that this is an antiquated approach to phage therapy and can be trivially improved by using multiple phages with non-overlapping host resistance patterns or by selecting for phage mutants that overcome host resistance. **b** | More recently, the synergistic relationships between phages and antibiotics have been exploited, with the use of lysogenic phages that do not kill the pathogen when used alone but instead decrease its survival when used in concert with antibiotics. An example is a phage that inhibits a DNA damage repair system (the SOS response), thus rendering the host bacterium exquisitely sensitive to quinolone class antibiotics¹⁴⁰. **c** | With our growing understanding of the human microbiome, it may be possible to take a more nuanced approach, selectively manipulating (that is, enhancing) microbial community functions or clearing the way for invasion by probiotic consortia. Strategies can be envisioned that benefit both microorganisms and their hosts; for example, introducing into phage genomes certain genes that are involved in nutrient biosynthesis (with direct benefits to the bacterial, and potentially also the human, host) or degradation (potentially stabilizing the abundance and niches of beneficial microorganisms, especially during times of acute stress).

member of a community, lysogenic phages could be introduced to promote a community structure that is beneficial to both the human host and microbial community members (FIG. 3). For example, we could expand the capacity of the gut microbiome to degrade dietary components⁹⁹. Similarly, phages could be used to introduce novel, beneficial traits to community members, such as those involving nutrient biosynthesis. In these cases, it may be difficult to introduce traits that are not purely beneficial to the bacterial host of a lysogenic phage, as the

energetic effects of synthesizing an unnecessary protein may impose a selection pressure.

Given the potential power and replicative nature of phages, a number of questions must be addressed before these viruses can be more widely adopted as therapeutics, including issues related to biocontainment⁹⁸. Although phages are frequently sold as viruses that ‘can only infect bacteria’, their safety has yet to be completely defined. The intravenous administration of phages (for example, in the case of bacterial sepsis) is particularly complex

given the immunogenicity of some preparations and the rapid clearance of phage particles by the reticuloendothelial system of the spleen¹⁰⁰. It is tempting to assume that other routes of administration, such as oral cocktails of phages to target the human gut microbiome, would not have such effects; however, phage DNA is detectable by PCR and FISH (fluorescence *in situ* hybridization) in the serum shortly after oral consumption of phages¹⁰¹. Other studies have provided evidence of *trans*-placental passage of phages¹⁰². There are data suggesting that enzymes transcribed from phage DNA can be expressed in mammalian cells¹⁰³; this finding has even led to attempts to use phages as gene therapy vectors^{104,105}.

Despite these concerns, we are exposed to millions of phages every day, including those from our own microbiota, without significant observable harm. Given this fact, it is interesting to consider the potential therapeutic use of phages in the context of current efforts to apply microbiome-directed therapies¹⁰⁶. Questions that can be asked include whether it is beneficial or detrimental for candidate probiotic bacterial taxa to possess or lack prophages, or whether phages should be deliberately administered coincidentally with or preceding the introduction of a probiotic consortium in order to help create a niche that promotes successful invasion and engraftment of the consortium.

Future directions

Little experimental work has been carried out on the ecology of phages *in vivo*. Germ-free mice and mice monocolonized with different strains of *Escherichia coli*, including strains that were isolated from children with diarrhoea, have been used to examine the replication of T4 and T7 phages^{107,108}. Models of the human gut microbiota using gnotobiotic mice may not only provide a better understanding of phage–bacterial host dynamics, but also represent a potentially valuable tool for establishing a preclinical pipeline designed to evaluate the feasibility of phage therapy. Recent work has shown that transplanting intact, uncultured human gut microbial communities (in the form of a faecal transplant) into gnotobiotic mice is efficient, and captures the majority of the microbial diversity and microbiome-encoded functions that are present in the human donor's

community^{109,110}. Mice with replicated human gut microbiomes can be fed diets resembling those of the human donor to explore diet–microbiome–phage interactions. An additional refinement to this approach is to culture and sequence collections of bacteria (some members of which will contain prophages) from a given donor's faecal sample and transplant these collections into recipient mice¹¹⁰. Various perturbations can be applied to the gnotobiotic mice harbouring these microbiota (such as changes in diet or manipulation of the immune system), and the effects of these perturbations on phage–bacterium dynamics can be studied over time under highly controlled conditions.

Yet another envisioned approach is to take cultured, sequenced members of the human gut microbiota and use these to assemble defined communities in formerly germ-free mice. Preparing VLPs from human faecal samples and introducing them into these mice would allow investigators to directly determine the bacterial–host specificity of VLP-associated phages. This approach could also allow researchers to assess the effects of the presence or absence of a prophage in community members, the effects of the diet administered to the animals and the contribution of phages to mammalian host physiology, including immune function. The impact of a staged phage attack on community structure and functions can also be defined in gnotobiotic animal models over time and as a function of location within the length of the gut, where the available nutrient and energy resources vary considerably.

These model systems should help us understand how phages influence metabolism in the gut and how we can use phages to manipulate human microbial communities. This will almost certainly demand new and more efficient methods for deliberately curing bacterial hosts of their prophages. It may also require the application of whole-genome transposon mutagenesis methods married to next-generation sequencing platforms¹¹¹ to identify the functional contributions of genes in a given prophage. Although the journey ahead will certainly be demanding, approaches are in hand (or can be envisioned) that will help propel the 'new age of phage' forward so that long-standing questions can be addressed and new insights can be obtained.

- Hershey, A. D. & Chase, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.* **36**, 39–56 (1952).
- Crick, F. H., Barnett, L., Brenner, S. & Watts-Tobin, R. J. General nature of the genetic code for proteins. *Nature* **192**, 1227–1232 (1961).
- Cairns, J., Stent, G. S. & Watson, J. D. *Phage and the Origins of Molecular Biology* (Cold Spring Harbor Laboratory Press, 1992).
- Mokili, J. L., Rohwer, F. & Dutilh, B. E. Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* **2**, 63–77 (2012).
- Breitbart, M. & Rohwer, F. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* **13**, 278–284 (2005).
- Fernandes, P. Antibacterial discovery and development—the failure of success? *Nature Biotech.* **24**, 1497–1503 (2006).
- d'Herelle, F. Sur un microbe invisible antagoniste des bacilles dysentériques. *C. R. Acad. Sci. Ser. D* **165**, 373–375 (1917).
- Sulakvelidze, A., Alavidze, Z. & Morris, J. G. Jr. Bacteriophage therapy. *Antimicrob. Agents Chemother.* **45**, 649–659 (2001).
- Levin, B. R. & Bull, J. J. Population and evolutionary dynamics of phage therapy. *Nature Rev. Microbiol.* **2**, 166–173 (2004).
- Marraffini, L. A. & Sontheimer, E. J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nature Rev. Genet.* **11**, 181–190 (2010).
- Horvath, P. & Barrangou, R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167–170 (2010).
- An overview of CRISPR-mediated defence mechanisms against phage attack.**
- Virgin, H. W., Wherry, E. J. & Ahmed, R. Redefining chronic viral infection. *Cell* **138**, 30–50 (2009).
- A discussion of the nuanced role of the immune system in chronic viral infections.**
- Delwart, E. Animal virus discovery: improving animal health, understanding zoonoses, and opportunities for vaccine development. *Curr. Opin. Virol.* **2**, 1–9 (2012).
- Haynes, M. & Rohwer, F. in *Metagenomics of the Human Body* (ed. Nelson, K. E.) 63–77 (Springer, 2011).
- Fox, G. E. *et al.* The phylogeny of prokaryotes. *Science* **209**, 457–463 (1980).
- Lane, D. J. *et al.* Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl Acad. Sci. USA* **82**, 6955–6959 (1985).
- Hugenholtz, P., Goebel, B. M. & Pace, N. R. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* **180**, 4765–4774 (1998).
- Culley, A. I., Lang, A. S. & Suttle, C. A. High diversity of unknown picorna-like viruses in the sea. *Nature* **424**, 1054–1057 (2003).
- Breitbart, M., Miyake, J. H. & Rohwer, F. Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol. Lett.* **236**, 249–256 (2004).
- Hambly, E. *et al.* A conserved genetic module that encodes the major virion components in both the coliphage T4 and the marine cyanophage S-PM2. *Proc. Natl Acad. Sci. USA* **98**, 11411–11416 (2001).

21. Casjens, S. R. Comparative genomics and evolution of the tailed-bacteriophages. *Curr. Opin. Microbiol.* **8**, 451–458 (2005).
22. Reyes, A. *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334–338 (2010).
The finding that the human faecal virome in healthy individuals is not highly shared between family members, and exhibits surprising stability over the course of 1 year.
23. Minot, S. *et al.* The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* **21**, 1616–1625 (2011).
A longitudinal study of the impact of controlled diet changes on the human gut virome.
24. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
25. Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nature Protoc.* **4**, 470–483 (2009).
A protocol for isolating VLPs for subsequent metagenomic characterization.
26. Willner, D. *et al.* Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. *Proc. Natl Acad. Sci. USA* **108** (Suppl. 1), 4547–4553 (2011).
27. Rohwer, F., Seguritan, V., Choi, D. H., Segall, A. M. & Azam, F. Production of shotgun libraries using random amplification. *BioTechniques* **31**, 108–112 (2011).
28. Breitbart, M. *et al.* Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* **185**, 6220–6223 (2003).
29. Shan, T. *et al.* The fecal virome of pigs on a high-density farm. *J. Virol.* **85**, 11697–11708 (2011).
30. Yozwiak, N. L. *et al.* Virus identification in unknown tropical febrile illness cases using deep sequencing. *PLoS Negl. Trop. Dis.* **6**, e1485 (2012).
31. Li, L. *et al.* Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses. *J. Virol.* **84**, 6955–6965 (2010).
32. Ge, X. *et al.* Metagenomic analysis of viruses from bat fecal samples reveals many novel viruses in insectivorous bats in china. *J. Virol.* **86**, 4620–4630 (2012).
33. Hutchison, C. A., Smith, H. O., Pfannkoch, C. & Venter, J. C. Cell-free cloning using ϕ 29 DNA polymerase. *Proc. Natl Acad. Sci. USA* **102**, 17332 (2005).
34. Kim, K. H. *et al.* Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Appl. Environ. Microbiol.* **74**, 5975–5985 (2008).
35. Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the multiple displacement amplification reaction. *BMC Biotechnol.* **7**, 19 (2007).
36. Kim, K. H. & Bae, J. W. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl. Environ. Microbiol.* **77**, 7663–7668 (2011).
37. Andrews-Pfannkoch, C., Fadrosch, D. W., Thorpe, J. & Williamson, S. J. Hydroxypatite-mediated separation of double-stranded DNA, single-stranded DNA, and RNA genomes from natural viral assemblages. *Appl. Environ. Microbiol.* **76**, 5039–5045 (2010).
38. Fadrosch, D. W., Andrews-Pfannkoch, C. & Williamson, S. J. Separation of single-stranded DNA, double-stranded DNA and RNA from an environmental viral community using hydroxypatite chromatography. *J. Vis. Exp.* **2011**, e3146 (2011).
39. Marine, R. *et al.* Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. *Appl. Environ. Microbiol.* **77**, 8071–8079 (2011).
40. Nakamura, S. *et al.* Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS ONE* **4**, e4219 (2009).
41. Wommack, K. E., Bhavsar, J. & Ravel, J. Metagenomics: read length matters. *Appl. Environ. Microbiol.* **74**, 1453–1463 (2008).
42. Bibby, K., Viau, E. & Peccia, J. Viral metagenome analysis to guide human pathogen monitoring in environmental samples. *Lett. Appl. Microbiol.* **52**, 386–392 (2011).
43. Ng, T. F. *et al.* Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PLoS ONE* **6**, e20579 (2011).
44. Pasic, L. *et al.* Metagenomic islands of hyperhalophiles: the case of *Salinibacter ruber*. *BMC Genomics* **10**, 570 (2009).
45. Vega Thurber, R. L. *et al.* Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral *Porites compressa*. *Proc. Natl Acad. Sci. USA* **105**, 18413–18418 (2008).
46. Dinsdale, E. A. *et al.* Functional metagenomic profiling of nine biomes. *Nature* **452**, 629–632 (2008).
47. Yang, J. *et al.* Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *J. Clin. Microbiol.* **49**, 3463–3469 (2011).
48. Xu, B. *et al.* Metagenomic analysis of fever, thrombocytopenia and leukopenia syndrome (FTLS) in Henan Province, China: discovery of a new bunyavirus. *PLoS Pathog.* **7**, e1002369 (2011).
49. Coetzee, B. *et al.* Deep sequencing analysis of viruses infecting grapevines: virome of a vineyard. *Virology* **400**, 157–163 (2010).
50. Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D. & Bushman, F. D. Hypervariable loci in the human gut virome. *Proc. Natl Acad. Sci. USA* **109**, 3962–3966 (2012).
The finding that hypervariable loci in the virome are predicted to encode Ig superfamily and C-type lectin folds.
51. Rohwer, F. & Thurber, R. V. Viruses manipulate the marine environment. *Nature* **459**, 207–212 (2009).
An overview of the interactions between marine viruses and their hosts.
52. Leplae, R., Lima-Mendez, G. & Toussaint, A. ACLAME: A Classification of Mobile genetic Elements, update 2010. *Nucleic Acids Res.* **38**, D57–D61 (2010).
53. Overbeek, R. *et al.* The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33**, 5691–5702 (2005).
54. Sharon, I. *et al.* Comparative metagenomics of microbial traits within oceanic viral communities. *ISME J.* **5**, 1178–1190 (2011).
55. Ghosh, T. S., Mohammed, M. H., Komanduri, D. & Mande, S. S. ProVIDE: a software tool for accurate estimation of viral diversity in metagenomic samples. *Bioinformatics* **6**, 91–94 (2011).
56. Lorenzi, H. A. *et al.* TheViral MetaGenome Annotation Pipeline(VMGAP): an automated tool for the functional annotation of viral metagenomic shotgun sequencing data. *Stand. Genomic Sci.* **4**, 418–429 (2011).
57. Meyer, F. *et al.* The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386 (2008).
58. Roux, S. *et al.* Metavir: a web server dedicated to virome analysis. *Bioinformatics* **27**, 3074–3075 (2011).
59. Sun, S. *et al.* Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res.* **39**, D546–D551 (2011).
60. Breitbart, M. *et al.* Genomic analysis of uncultured marine viral communities. *Proc. Natl Acad. Sci. USA* **99**, 14250–14255 (2002).
61. Angly, F. *et al.* PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* **6**, 41 (2005).
62. Angly, F. E. *et al.* The marine viromes of four oceanic regions. *PLoS Biol.* **4**, e368 (2006).
63. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
64. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484 (2009).
65. Turnbaugh, P. J. *et al.* Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc. Natl Acad. Sci. USA* **107**, 7503–7508 (2010).
66. Hansen, E. E. *et al.* Pan-genome of the dominant human gut-associated archaeon, *Methanobrevibacter smithii*, studied in twins. *Proc. Natl Acad. Sci. USA* **108** (Suppl. 1), 4599–4606 (2011).
67. Caporaso, J. G. *et al.* Moving pictures of the human microbiome. *Genome Biol.* **12**, R50 (2011).
68. Costello, E. K. *et al.* Bacterial community variation in human body habitats across space and time. *Science* **326**, 1694–1697 (2009).
69. Kim, M. S., Park, E. J., Roh, S. W. & Bae, J. W. Diversity and abundance of single-stranded DNA viruses in human feces. *Appl. Environ. Microbiol.* **77**, 8062–8070 (2011).
70. Krupovic, M. & Forterre, P. *Microviridae* goes temperate: microvirus-related proviruses reside in the genomes of *Bacteroidetes*. *PLoS ONE* **6**, e19893 (2011).
71. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
72. Stern, A., Mick, E., Tirosh, I., Sagy, O. & Sorek, R. CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res.* 25 Jun 2012 (doi:10.1101/gr.138297.112).
73. Breitbart, M. *et al.* Viral diversity and dynamics in an infant gut. *Res. Microbiol.* **159**, 367–373 (2008).
An article that describes the rapid assembly and unstable features of the gut virome following birth.
74. Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intestinal microbiota. *PLoS Biol.* **5**, e177 (2007).
75. Baroness, J. J. & Beckwith, J. A bacterial virulence determinant encoded by lysogenic coliphage λ . *Nature* **346**, 871–874 (1990).
76. Plunkett, G. 3rd, Rose, D. J., Durfee, T. J. & Blattner, F. R. Sequence of Shiga toxin 2 phage 933W from *Escherichia coli* O157:H7: Shiga toxin as a phage late-gene product. *J. Bacteriol.* **181**, 1767–1778 (1999).
77. Markine-Goriaynoff, N. *et al.* Glycosyltransferases encoded by viruses. *J. Gen. Virol.* **85**, 2741–2754 (2004).
78. Liu, M. *et al.* Reverse transcriptase-mediated tropism switching in *Bordetella* bacteriophage. *Science* **295**, 2091–2094 (2002).
79. Fraser, J. S., Yu, Z., Maxwell, K. L. & Davidson, A. R. Ig-like domains on bacteriophages: a tale of promiscuity and deceit. *J. Mol. Biol.* **359**, 496–507 (2006).
80. Zhang, T. *et al.* RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol.* **4**, e3 (2006).
81. Ley, R. E. *et al.* Evolution of mammals and their gut microbes. *Science* **320**, 1647–1651 (2008).
82. Muegge, B. D. *et al.* Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* **332**, 970–974 (2011).
83. Cann, A. J., Fandrich, S. E. & Heaphy, S. Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. *Virus Genes* **30**, 151–156 (2005).
84. Donaldson, E. F. *et al.* Metagenomic analysis of the viromes of three North American bat species: viral diversity among different bat species that share a common habitat. *J. Virol.* **84**, 13004–13018 (2010).
85. Blinkova, O. *et al.* Novel circular DNA viruses in stool samples of wild-living chimpanzees. *J. Gen. Virol.* **91**, 74–86 (2010).
86. Ng, T. F. *et al.* Metagenomic identification of a novel anellovirus in Pacific harbor seal (*Phoca vitulina richardsii*) lung samples and its detection in samples from multiple years. *J. Gen. Virol.* **92**, 1318–1323 (2011).
87. Phan, T. G. *et al.* The fecal viral flora of wild rodents. *PLoS Pathog.* **7**, e1002218 (2011).
88. van den Brand, J. M. *et al.* Metagenomic analysis of the viral flora of pine marten and European badger feces. *J. Virol.* **86**, 2360–2365 (2012).
89. Dhillon, T. S., Dhillon, E. K., Chau, H. C., Li, W. K. & Tsang, A. H. Studies on bacteriophage distribution: virulent and temperate bacteriophage content of mammalian feces. *Appl. Environ. Microbiol.* **32**, 68–74 (1976).
90. Berg Miller, M. E. *et al.* Phage–bacteria relationships and CRISPR elements revealed by a metagenomic survey of the rumen microbiome. *Environ. Microbiol.* **14**, 207–227 (2012).
91. Maura, D. *et al.* Intestinal colonization by enteroaggregative *Escherichia coli* supports long-term bacteriophage replication in mice. *Environ. Microbiol.* 28 Nov 2011 (doi:10.1111/j.1462-2920.2011.02644.x).
92. Fischetti, V. A., Nelson, D. & Schuch, R. Reinventing phage therapy: are the parts greater than the sum? *Nature Biotech.* **24**, 1508–1511 (2006).
93. Lu, T. K. & Koeris, M. S. The next generation of bacteriophage therapy. *Curr. Opin. Microbiol.* **14**, 524–531 (2011).
94. van Helvoort, T. The controversy between John H. Northrop and Max Delbrück on the formation of bacteriophage: bacterial synthesis or autonomous multiplication? *Ann. Sci.* **49**, 545–575 (1992).
95. Calendar, R. L. *The Bacteriophages* (Oxford Univ. Press, 2005).
96. Kumarasamy, K. K. *et al.* Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: a molecular, biological, and epidemiological study. *Lancet Infect. Dis.* **10**, 597–602 (2010).

97. Piddock, L. J. The crisis of no new antibiotics—what is the way forward? *Lancet Infect. Dis.* **12**, 249–253 (2012).
98. Clokie, M. R. J. & Kropinski, A. M. *Bacteriophages: Methods and Protocols* (Humana Press, 2009).
99. Hehemann, J. H. *et al.* Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**, 908–912 (2010).
100. Geier, M. R., Trigg, M. E. & Merrill, C. R. Fate of bacteriophage lambda in non-immune germ-free mice. *Nature* **246**, 221–223 (1973).
101. Schubbert, R., Renz, D., Schmitz, B. & Doerfler, W. Foreign (M13) DNA ingested by mice reaches peripheral leukocytes, spleen, and liver via the intestinal wall mucosa and can be covalently linked to mouse DNA. *Proc. Natl Acad. Sci. USA* **94**, 961–966 (1997).
102. Schubbert, R., Hohlweg, U., Renz, D. & Doerfler, W. On the fate of orally ingested foreign DNA in mice: chromosomal association and placental transmission to the fetus. *Mol. Gen. Genet.* **259**, 569–576 (1998).
103. Geier, M. R. & Merrill, C. R. Lambda phage transcription in human fibroblasts. *Virology* **47**, 638–643 (1972).
104. Barry, M. A., Dower, W. J. & Johnston, S. A. Toward cell-targeting gene therapy vectors: selection of cell-binding peptides from random peptide-presenting phage libraries. *Nature Med.* **2**, 299–305 (1996).
105. Dunn, I. S. Mammalian cell binding and transfection mediated by surface-modified bacteriophage lambda. *Biochimie* **78**, 856–861 (1996).
106. Silverman, M. S., Davis, I. & Pillai, D. R. Success of self-administered home fecal transplantation for chronic *Clostridium difficile* infection. *Clin. Gastroenterol. Hepatol.* **8**, 471–473 (2010).
107. Chibani-Chennoufi, S. *et al.* *In vitro* and *in vivo* bacteriolytic activities of *Escherichia coli* phages: implications for phage therapy. *Antimicrob. Agents Chemother.* **48**, 2558–2569 (2004).
108. Weiss, M. *et al.* *In vivo* replication of T4 and T7 bacteriophages in germ-free mice colonized with *Escherichia coli*. *Virology* **393**, 16–23 (2009).
109. Turnbaugh, P. J. *et al.* The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med.* **1**, 6ra14 (2009).
110. Goodman, A. L. *et al.* Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc. Natl Acad. Sci. USA* **108**, 6252–6257 (2011).
111. Goodman, A. L. *et al.* Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe* **6**, 279–289 (2009).
112. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proc. Natl Acad. Sci. USA* **95**, 6578–6583 (1998).
113. Bergh, O., Borsheim, K. Y., Bratbak, G. & Heldal, M. High abundance of viruses found in aquatic environments. *Nature* **340**, 467–468 (1989).
114. Clokie, M. R., Millard, A. D., Letarov, A. V. & Heaphy, S. Phages in nature. *Bacteriophage* **1**, 31–45 (2011).
115. Fuhrman, J. A. Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541–548 (1999).
116. Azam, F. *et al.* The ecological role of water column microbes in the sea. *Mar. Ecol. Prog. Ser.* **10**, 257–263 (1983).
117. Marston, M. F. *et al.* Rapid diversification of coevolving marine *Synechococcus* and a virus. *Proc. Natl Acad. Sci. USA* **109**, 4544–4549 (2012).
118. Mann, N. H., Cook, A., Millard, A., Bailey, S. & Clokie, M. Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature* **424**, 741 (2003).
119. Sullivan, M. B. *et al.* Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol.* **4**, e234 (2006).
120. Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. & Chisholm, S. W. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438**, 86–89 (2005).
121. Anderson, R. E., Brazelton, W. J. & Baross, J. A. Is the genetic landscape of the deep subsurface biosphere affected by viruses? *Front. Microbiol.* **2**, 219 (2011).
122. Roossinck, M. J. The good viruses: viral mutualistic symbioses. *Nature Rev. Microbiol.* **9**, 99–108 (2011). **An excellent outline of beneficial virus–host interactions in a variety of species.**
123. Roossinck, M. J. Changes in population dynamics in mutualistic versus pathogenic viruses. *Viruses* **3**, 12–19 (2011).
124. Brown, S. P., Le Chat, L., De Paepe, M. & Taddei, F. Ecology of microbial invasions: amplification allows virus carriers to invade more rapidly when rare. *Curr. Biol.* **16**, 2048–2052 (2006).
125. Brown, S. P., Inglis, R. F. & Taddei, F. Evolutionary ecology of microbial wars: within-host competition and (incidental) virulence. *Evol. Appl.* **2**, 32–39 (2009).
126. Moran, N. A., Degnan, P. H., Santos, S. R., Dunbar, H. E. & Ochman, H. The players in a mutualistic symbiosis: insects, bacteria, viruses, and virulence genes. *Proc. Natl Acad. Sci. USA* **102**, 16919–16926 (2005).
127. Oliver, K. M., Degnan, P. H., Hunter, M. S. & Moran, N. A. Bacteriophages encode factors required for protection in a symbiotic mutualism. *Science* **325**, 992–994 (2009).
128. Xu, P. *et al.* Virus infection improves drought tolerance. *New Phytol.* **180**, 911–921 (2008).
129. Marquez, L. M., Redman, R. S., Rodriguez, R. J. & Roossinck, M. J. A virus in a fungus in a plant: three-way symbiosis required for thermal tolerance. *Science* **315**, 513–515 (2007).
130. Ophel, K. M., Bird, A. F. & Kerr, A. Association of bacteriophage particles with toxin production by *Clavibacter toxicus*, the causal agent of annual ryegrass toxicity. *Phytopathology* **83**, 676–681 (1993).
131. Holtz, L. R., Finkbeiner, S. R., Kirkwood, C. D. & Wang, D. Identification of a novel picornavirus related to cosaviruses in a child with acute diarrhea. *Virology* **5**, 159 (2008).
132. Finkbeiner, S. R. *et al.* Metagenomic analysis of human diarrhea: viral detection and discovery. *PLoS Pathog.* **4**, e1000011 (2008).
133. Finkbeiner, S. R. *et al.* Human stool contains a previously unrecognized diversity of novel astroviruses. *Virology* **6**, 161 (2009).
134. Phan, T. G. *et al.* A third gyrovirus species in human feces. *J. Gen. Virol.* **93**, 1356–1361 (2012).
135. Kapoor, A. *et al.* Multiple novel astrovirus species in human stool. *J. Gen. Virol.* **90**, 2965–2972 (2009).
136. Kapoor, A. *et al.* Human bocaviruses are highly diverse, dispersed, recombination prone, and prevalent in enteric infections. *J. Infect. Dis.* **201**, 1633–1643 (2010).
137. Victoria, J. G. *et al.* Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *J. Virol.* **83**, 4642–4651 (2009).
138. Rosario, K., Duffy, S. & Breitbart, M. A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch. Virol.* **4**, Jul 2012 (doi:10.1007/s00705-012-1391-y).
139. Rodriguez-Valera, F. *et al.* Explaining microbial population genomics through phage predation. *Nature Rev. Microbiol.* **7**, 828–836 (2009). **A discussion of the consequences of phage predation on microbial substrain diversity, presented as a constant-diversity dynamics model.**
140. Lu, T. K. & Collins, J. J. Engineered bacteriophage targeting gene networks as adjuvants for antibiotic therapy. *Proc. Natl Acad. Sci. USA* **106**, 4629–4634 (2009).

Acknowledgements

Work from the authors' laboratories that is described in this Review was supported by the US National Institutes of Health (NIH) (grants DK78669, DK50292 and DK70977 to J.J.G. and grant GM095384 to F.L.R.) and by the Crohn's and Colitis Foundation of America. A.R. is the recipient of an International Fulbright Science and Technology Award. N.P.S. is a member of the Washington University Medical Scientist Training Program (MSTP), which is funded by NIH grant GM007200. Owing to space limitations, the authors were not able to cite many wonderful studies that are relevant to the topics covered.

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

Jeffrey Gordon's homepage: <http://gordonlab.wustl.edu>
 ACLAME: <http://aclame.ulb.ac.be>
 crAss: <http://edwards.sdsu.edu/crAss>
 Phage SEED: <http://www.phantom.org/PhageSeed/Phage.cgi>
 Virome: <http://virome.diagcomputing.org/view-home>

ALL LINKS ARE ACTIVE IN THE ONLINE PDF