The Role of Host-Associated Factors on Metazoan Microbiome Assembly

By

Andrew W. Brooks

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Human Genetics

May 31, 2019

Nashville, Tennessee

Committee:

John A. Capra, Ph.D. (Chair)

David Samuels, Ph.D.

Antonis Rokas, Ph.D.

Patrick Abbot, Ph.D.

Seth Bordenstein, Ph.D. (Advisor)

DEDICATION

To my parents, sister, grandmother, aunt, uncle, and cousins
for their constant love and support…

…to Alanna, my perpetually loving cheerleader,
and occasional adult supervision…

...to Kirby, Daniel, Jake, Saligram, Eric and all of my Nashville
friends who stood by me through thick and thin…

...to Seth, my committee, and the lab for
an amazing educational experience…

…and to Teddy, for responsibili-buddying through it all.

I would not be here today without your support,
and I cannot thank you enough for this incredible journey.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

Chapter 2: Phylosymbiosis

Support

We thank MR4 for providing mosquito eggs (contributed by Sandra Allan, Nora Besansky, Mustafa Dakeen, William Collins, Maureen Coetzee, William Reisen, William Brogdon, and MR4 Vector Activity). We would like to especially recognize the substantial efforts made by Michael Felder and Janet Crossland for evaluating the stock lines and collecting fecal samples from the *Peromyscus* Genetic Stock Center at the University of South Carolina, Dr. Sanford H. Feldman of the Center for Comparative Medicine at the University of Virginia for providing the pinworm DNA, and Theresa Bark for screening the *Peromyscus* stool for pinworm DNA. We thank Dr. M. Denise Dearing for supplying equipment and facilities to conduct the *Pero- myscus* functional experiments and Dr. Hopi Hoekstra for help with the *Peromyscus* host phylogeny. We would like to thank Ran Blekhman and Michael Burns for support in processing preliminary data and Sarah Bordenstein, Bojana Jovanovic, and Lisa Funkhouser for providing feedback on an earlier version of the manuscript.

Funding

- National Science Foundation - Division of Integrative Organismal Systems (grant number 1456778). Received by SRB.

  https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503623

- National Science Foundation - Division of Biological Infrastructure (grant number 1400456). Received by KDK.

  http://www.nsf.gov/div/index.jsp?div=DBI

- National Institute of Health - Predoctoral Training Grant (grant number 5T32GM080178). Received by AWB.

  https://researchtraining.nih.gov

- National Science Foundation - Division of Environmental Biology (grant number 1046149). Received by SRB.

  http://www.nsf.gov/div/index.jsp?div=DEB

- Rowland Institute at Harvard University Junior Fellowship to RMB.

Chapter 3: Virome Phylosymbiosis

Funding

- This work was supported by National Science Foundation award 1456778, National Institutes of Health R01 AI132581, and a Vanderbilt Microbiome

Trans-Institutional Initiative Award to SRB.

- AWB was supported by T32 National Institutes of Health training grants 6424T32GM08017810, 5T32GM08017809, and 5T32GM0817808.
- AM was supported by a Deutsche Forschungsgemeinschaft (DFG) postdoctoral fellowship (MI 2242/1-1).
- The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Chapter 4: American Gut

Support

We would like to thank Tony Capra, David Samuels, Patrick Abbot, Antonis Rokas, and other members of the Vanderbilt Genetics Institute and Bordenstein Lab for input. Thank you to the Minnesota Supercomputing Institute (MSI) at the University of Minnesota and the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University for providing resources that contributed to the research results reported within this paper. We also thank the American Society of Microbiology for supporting travel by AWB to present this work.

Funding

- National Institutes of Health (grant number 4T32GM08017810, 5T32GM08017809, and 5T32GM0817808). Supported through the Vanderbilt Genetics Institute. Received by AWB. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

  https://researchtraining.nih.gov/career/graduate

- Vanderbilt Office of Equity, Diversity and Inclusion. Received by AWB.

  https://www.vanderbilt.edu/equity-diversity-inclusion/

- Alfred P. Sloan Foundation Fellowship. Received by RB.

  https://sloan.org/fellowships/

- Vanderbilt Microbiome Initiative. Received by SRB.

  https://my.vanderbilt.edu/microbiome/

- National Institutes of Health / National Institute of General Medical Sciences (grant number NIH/NIGMS R35-GM128716). Received by RB.

  https://www.nigms.nih.gov/grants-and-funding

Chapter 5: Vanderbilt Microbiome Initiative

Support

We would like to thank the Vanderbilt Trans-Institutional Program (TIPs)

for fostering inter-school collaboration, providing funding, and supporting the resources and infrastructure necessary for the VMI clinical trial. A sincerest thank you to the Vanderbilt sequencing core (VANTAGE) and particularly Karen Beeri for support with metagenomic sequencing and human genotyping. We are grateful for support from the Vanderbilt Office of Equity, Diversity, and Inclusion in promoting and financing VMI-associated research. We are also grateful to the Vanderbilt Institute for Infection, Immunology, and Inflammation (VI4), John McLean, Chris Lopez, and particularly James Poland for providing mini-sabbatical funding and training to AWB in fecal metabolomics. Thank you to Matt Scholz for support with VMI outreach and training activities at Vanderbilt.

Funding

- Vanderbilt Trans-Institutional Partnership - Microbiome Initiative. Received by SRB.

  https://my.vanderbilt.edu/microbiome/

- Vanderbilt Institute for Infection, Immunology, and Inflammation Mini-Sabbatical Internal Fellowship. Received by AWB.

  https://www.vanderbilt.edu/equity-diversity-inclusion/

- Vanderbilt Office of Equity, Diversity and Inclusion. Received by AWB.

  https://www.vanderbilt.edu/equity-diversity-inclusion/

Chapter 6: Conclusion and Future Directions

Support

A sincerest thank you to Seth R. Bordenstein, Wendy W. Brooks, Edward J. van Opstal, and Alanna Salituro for feedback on the ethnicity editorial published in *Future Microbiology*.

LIST OF PUBLICATIONS BY CHAPTER

Chapter 2: Phylosymbiosis

- Phylosymbiosis: Relationships and Functional Effects of Microbial Communities across Host Evolutionary History.
    - **Brooks AW***, Kohl KD*, Brucker RM*, van Opstal EJ, Bordenstein SR. (*Co-first Authors). *PLOS Biology*. (2016).
    - http://dx.doi.org/10.1371/journal.pbio.2000225

Chapter 3: Virome Phylosymbiosis

- Finer Scale Phylosymbiosis: Insights from Insect Viromes.
    - Leigh B, Bordenstein SR, **Brooks AW**, Mikaelyan A, Bordenstein SR. *mSystems*. (2018).
    - https://doi.org/10.1128/mSystems.00131-18

Chapter 4: American Gut

- Gut Microbiota Diversity across Ethnicities in the United States.
    - **Brooks AW**, Priya S, Blekhman R, Bordenstein SR. *PLOS Biology*. (2018).

- https://doi.org/10.1371/journal.pbio.2006842

Chapter 6: Vanderbilt Microbiome Initiative

- How could ethnicity-associated microbiomes contribute to personalized therapies?
  - **Brooks AW.** *Future Microbiology*. (2019).
  - https://doi.org/10.1371/journal.pbio.2006842

Additional Publications at Vanderbilt

- Evolutionary Genetics of Cytoplasmic Incompatibility Genes cifA and cifB in Prophage WO of *Wolbachia*.
  - Lindsey ARI, Rice DW, Bordenstein SR, **Brooks AW**, Bordenstein SR, Newton ILG. *Genome Biology and Evolution*. (2018).
  - https://doi.org/10.1093/gbe/evy012
- Genetic Signatures for *Helicobacter pylori* Strains of West African Origin.
  - Bullock KK, Shaffer CL, **Brooks AW**, Secka O, Forsyth MH, McLain MS, Cover TL. *PLOS One*. (2017).
  - https://doi.org/10.1371/journal.pone.0188804

LIST OF FIGURES

Chapter 2: Phylosymbiosis

Chapter 3: Virome Phylosymbiosis

Chapter 4: American Gut

Chapter 5: Vanderbilt Microbiome Initiative

LIST OF TABLES

Chapter 2: Phylosymbiosis

Chapter 3: Virome Phylosymbiosis

Chapter 4: American Gut

Chapter 5: Vanderbilt Microbiome Initiative

LIST OF ABBREVIATIONS

Chapter 1: Introduction

ANOSIM	-	Analysis of Similarity
CCA	-	Canonical Correlation Analysis
FMT	-	Fecal Microbiome Transplant
GWAS	-	Genome-Wide Association Study
HGT	-	Horizontal Gene Transfer
HMO	-	Human Milk Oligosaccharide
IBD	-	Identity By Descent
IBS	-	Identity By State
NMDS	-	Non-metric Multidimensional Scaling
OTU	-	Operational Taxonomic Unit
PERMANOVA	-	Permutational Analysis of Variance
PCoA	-	Principle Coordinates of Analysis
PD	-	Phylogenetic Diversity
RNA	-	Ribonucleic Acid

SCFA	-	Short Chain Fatty Acid
SNP	-	Single Nucleotide Polymorphism
UNIFRAC	-	Unique Fraction (Beta Diversity Metric)

Chapter 2: Phylosymbiosis

ANOVA	-	Analysis of Variance
AVPR1A	-	Arginine Vasopressin Receptor 1A
COI	-	Cytochrome Oxidase
EPE	-	Expected Predicted Error
HSD	-	Honest Significant Difference
OTU	-	Operational Taxonomic Unit
PCoA	-	Principal Coordinates Analysis
RFC	-	Random Forest Classifier

Chapter 3: Virome Phylosymbiosis

COI	-	Cytochrome Oxidase I
HTH	-	Helix Turn Helix

ORF	-	Open Reading Frame
Pfam	-	Protein Family
UPGMA	-	Unweighted Paired Group Mean Arithmetic

Chapter 4: American Gut

AGP	-	American Gut Project
ANOSIM	-	Analysis Of Similarity
AUC	-	Area Under the Curve
A/U	-	Abundance/Ubiquity
BMI	-	Body Mass Index
eQTL	-	Expression Quantitative Trait Locus
FDR	-	False Discovery Rate
F_{ST}	-	Fixation Index
GWAS	-	Genome-Wide Association Studies
HMP	-	Human Microbiome Project
MAF	-	Minor Allele Frequency
OTU	-	Operational Taxonomic Unit

PERMANOVA	-	Permutational Multivariate Analysis of Variance
RF	-	Random Forest
ROC	-	Receiver Operating Characteristic
SMOTE	-	Synthetic Minority Oversampling Technique

Individuals

AM	-	Aram Mikaelyan
AWB	-	Andrew W. Brooks
KDK	-	Kevin D. Kohl
RB	-	Ran Blekhman
RMB	-	Robert M. Brucker
SRB	-	Seth R. Bordenstein

DATA AVAILABILITY

Chapter 2: Phylosymbiosis

- All sequencing and mapping files are available from the Dryad database repository:

  doi:10.5061/dryad.n3v49

- A GitHub repository contains custom analysis scripts and all of the necessary data for figure reconstruction (including BIOM Tables and Mapping files) for each clade are also publicly available:

  https://github.com/awbrooks19/phylosymbiosis

Chapter 3: Virome Phylosymbiosis

- Assembled contigs from each viral metagenome have been submitted to the WGS database of NCBI under BioProject PRJNA481165:

  https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA481165

- Additionally, each circular genome has been submitted to the NCBI-nr database under the accession numbers MK047638 to MK047643:

  https://www.ncbi.nlm.nih.gov/nuccore/MK047638

  https://www.ncbi.nlm.nih.gov/nuccore/MK047639

https://www.ncbi.nlm.nih.gov/nuccore/MK047640

https://www.ncbi.nlm.nih.gov/nuccore/MK047641

https://www.ncbi.nlm.nih.gov/nuccore/MK047642

https://www.ncbi.nlm.nih.gov/nuccore/MK047643

- Supplemental material for this article may be found at:

https://doi.org/10.1128/mSystems.00131-18.

Chapter 4: American Gut

- Code, scripts, and data underlying figures are publicly available from the GitHub repository:

https://github.com/awbrooks19/microbiota_and_ethnicity

- Individual metadata (age, sex, ethnicity...) for the Human Microbiome Project are held under restricted access available through dbGaP application [NCBI - dbGaP, Human Microbiome Project phs000228.v3]:

https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000228.v3.p1

Chapter 5: Vanderbilt Microbiome Initiative

- Information about the ongoing Vanderbilt Microbiome Initiative can be found:

https://my.vanderbilt.edu/microbiome/

CHAPTER I

Clinically Translating Ecological and Evolutionary Microbiome Assembly

Introduction

Defining Principles of Human Microbiome Assembly

Gut microbiomes, the genetic repertoire for millions to trillions of microbes residing throughout each metazoan's digestive tract, are shaped by their host and environment. Composed of up to 500 times the genetic diversity of the human genome and distributed across more than 1,000 microbial species, microbiomes result from ecological assembly of microbial communities known as microbiota[1-4]. The advent of high throughput genetic sequencing hastened the characterization of microbiota and microbiomes, revealing an under-appreciated diversity of microbial taxa, ecological compositions, and functional capabilities[1;4-6]. In these gastrointestinal communities, variation associates intrinsically with host physiology[7;8], genetics[9-13], metabolism[14;15], immunity[16-20], and evolutionary relatedness[12;21-25], as well as extrinsically with lifestyle[1;26-28], diet[4;29-34], environment[4;35;36], and sociality[37;38]. Many factors exerting influence on microbiome and microbiota composition likely drive the regular observation of higher variation between organisms than variation within an adult organism over

time, suggesting that personalized composition is a more stable phenomenon relative to the influence of intrinsic and extrinsic factors which vary across organisms[13;39-41]. Particular interest has been payed to microbiota associations with human diseases, where interpersonal variation likely plays important roles in disease risk and progression. Extrinsic microbiota control through manipulation of diet, lifestyle, and xenobiotics can rapidly change composition and functional compatibility with the host, making the microbiome a novel and promising target for health interventions[30;33;42-44]. However, composition specific to each individual means that interventions targeting the microbiota are not necessarily generalizable across a population, and defining microbiota-disease variation often shows limited reproducibility and contradictory results between studies. Characterizing microbial communities and their genetic capacity has been a rapid scientific breakthrough, but the ability to define "what varies" has far outpaced defining principles of "why". Therefore, reproducible principles need to be applied to the human microbiome in order to develop generalizable applications, and to distinguish deviations that indicate personalized interventions remain necessary.

Utilizing Ecological and Evolutionary Principles to Understand Human Health

A decade of characterizing microbiome and microbiota differences has laid

a foundation for applying principles developed over a century of ecological and evolutionary theory. Whether comparing closely related individuals or diverse host species, establishing reproducible patterns could help elucidate generalizable principles governing host and microbial roles in assembly, and inform shared mechanisms of compositional and functional breakdown across diseases and evolutionary history. However, clinical and translational studies are often disconnected from fundamental basic science research, resulting in pillars of knowledge united by underlying biology but disconnected in ideology, terminology, and methodology. To date no consensus has arisen in defining 'healthy' or 'dysbiotic' gut ecosystems, partially fueled by the realization that core taxa and functions can make up a minimal set of microbiota and microbiome composition when examined across populations[4;26;42;45;46]. In fact, differences in extrinsic lifestyle and intrinsic biology likely mean that defining these terms will be subjective to the individual and situationally dependent. This is not to say that individuality has precluded generalizations, with patterns emerging including reduction of ecological alpha diversity across a wide range of diseases[47;48], and the observation that microbiota ecological similarity parallels host evolutionary relatedness (a.k.a., phylosymbiosis) in a variety of animal systems and hominids[24;49-52]. Moving forward, questions elucidating ecological and evolutionary

roles in host-microbe symbioses can be investigated in clinical and translational studies, particularly if conscious consideration is given from the initial stages of study design. Likewise, the wealth of existing clinical studies warrant re-examination while considering that taxa and functions may lack reproducible associations because disease etiology affects the same mechanistic principle of community instability, but the resulting breakdown reflects individual intrinsic and extrinsic influences. The questions then arise, what ecological and evolutionary principles apply to microbiota and microbiomes, and more importantly how can they be tested in an empirical way?

Investigating these principles is a priority for many basic science researchers working in model systems, but generalization to the extent that principles can inform health interventions and evolutionary biology will require a body of evidence founded in clinical and basic science settings. Certainly, challenges and barriers exist to properly addressing the hypotheses underlying many ecological and evolutionary principles in humans, including limited control of extrinsic factors affecting the microbiome, the inability to use genetic engineering or interventions that lack a medical utility, a limited host genetic diversity relative to interspecific comparisons, and higher costs for recruiting and sampling enough individuals to be powered to detect patterns where they exist.

Challenges may also arise due to deviating approaches and motivations underlying basic and clinical research. Clinicians start with a factor of interest such as a disease, observe associated biological patterns such as changes in microbiome composition, investigate underlying mechanisms, and only after a large body of literature is compiled try to connect unifying principles. Basic researchers often begin with an ecological or evolutionary principle, predict biological mechanisms and patterns that fit the hypothesis, and finally in model systems compile examples of controlled intrinsic and extrinsic factors that associate with expected outcomes. For this reason, merely discussing principles and examples in basic science research is not a roadmap for clinicians. However, a barrier that should not exist is siloing of ecological and evolutionary principles into basic science merely by lacking communication of their utility and approachability.

A Framework Translating Ecological and Evolutionary Principles to Clinicians

Here the aim is to make the transfer of ecological and evolutionary concepts more salient to the broader clinical community by focusing on how they could be investigated and applied to understand disease. First, reviewing some of the fundamental eco-evolutionary measures already widely employed in

microbiome research will lay a groundwork for addressing hypotheses. Second, the patterns investigators look for in those measures that could be explained by higher eco-evolutionary principles will be discussed. Third, we will examine empirical examples in model systems that support how and why those patterns manifest, and consider which patterns may or may not be extensible to human studies. Fourth, we will discuss how clinical studies could incorporate these patterns to test if generalizable eco-evolutionary principles may influence or breakdown in disease etiology. Proposing approaches to investigate eco-evolutionary principles is an avenue for clinicians to incorporate such questions into their existing studies, and ultimately generalizing from model systems to humans could reveal novel applications of these principles to understanding human health and disease.

In the Light of Evolution

Phylosymbiosis: Host Evolutionary Impacts on Microbiome Ecology

Since multicellular metazoans arose hundreds of millions of years ago, they have been surrounded and inhabited by complex microbial communities. The same applies for millions of years of human and hominid evolution, and understanding the evolutionary context that has shaped human microbiomes

could prove critical as modernization leads to rapid shifts in human lifestyles that outpace rates of evolutionary adaptation. Efforts to elucidate factors shaping modern human microbiomes also reveal difficulties in disentangling the influence of covarying factors spanning diet, lifestyle, culture, geography, genetics, and others. Therefore, if host evolution influences microbiome assembly, confounding host intrinsic and extrinsic variation will need to be controlled or statistically assessed to quantify the effect host evolutionary divergence is playing. Since evolution acts constantly, one possible expectation is that evolutionarily changes affecting the microbiome would generally accrue consistently over time. **Phylosymbiosis** is the observation that host phylogenetic relationships correspond to microbiota or microbiome relationships as measured by beta diversity and depicted visually in dendrograms[24;51]. Conversely, stochastic assembly and microbial dispersal throughout the environment could result in variable microbiome composition that is not necessarily distinguishable by host species[53]. Observing phylosymbiosis is at first glance interpreted as due to host filtering or genetic effects through immune, metabolic, or physiological systems that interact with the microbiome. However, host filtering is too simplistic as phylosymbiosis could also arise by microbial adaptations for colonizing specific hosts at certain abundances, perhaps in an effort to increase their own replication

in a favorable host species.

While a diverse array of mechanisms could underlie the pattern, phylosymbiosis provides predictive hypotheses about the relationship between host evolution and microbiome ecology. If other factors are controlled, phylosymbiosis has two primary expectations: 1) genetically distinct hosts will have distinguishable microbiomes, and 2) microbiome beta-diversity relationships will parallel host evolutionary relationships. Thus, phylosymbiosis is observed as a phylogenetic signal on microbial community relationships. It is a pattern whose mechanisms remain to be studied in a wide array of hosts. Considering the relatively persistent accrual of genetic variation within species, each change has a probability of affecting the microbiome to varying degrees through allelic variation at a locus controlling host traits that interact or do not interact with the microbiome. While the magnitude of effect on the microbiome from a single variant could be negligible or enormous depending on the affected system, when averaged across thousands to millions of evolutionary changes, rates of host evolution could correlate with diverging microbiome ecologies. One utility of phylosymbiosis could be as a null hypothesis for the expected relationship between host evolution and microbiome ecology, allowing researchers to identify where community compositions are more stably maintained or diverge more

dramatically than expected between host species. While the pattern of phylosymbiosis more likely emerges through interspecific genetic divergence following the formation of reproductive, allopatric, or other barrier to shared microbial acquisition, the conceptual framework that accruing host genetic variation leads to ecological changes in the microbiome is generalizable across short and long timescales whether reproductive barriers are at play.

Quantifying Ecological and Evolutionary Microbiome Divergence

Functionally testing principles like phylosymbiosis requires measuring both microbiome ecological similarity and host evolutionary relatedness. Such measures are foundational to microbiome research and can be applied to a diverse array of hypotheses. There are far more tools used in microbiome research than can be discussed here, however reviewing these widely-used and fundamental measures provides a toolkit to understand (i) how measures can be used to test hypotheses about host-associated microbiome assembly, (ii) which principles and measures of phylosymbiosis could be extendable to humans, and (iii) what phylosymbiotic expectations predict for the human microbiome in an evolutionary context relative to our hominid and hominin relatives.

Ecological similarity can be assessed between the total compositions of two

microbial communities using **beta diversity**, a measure of ecological distance between samples. Beta diversity has been a traditional foundation of ecological studies over the past century, and is one of the most widely used measures in modern microbiome research. Depending on the metric used, beta diversity quantifies how similar or different two microbiomes are by comparing: 1) the unweighted presence / absence of unique observations (e.g. metric Binary Jaccard), 2) the weighted relative abundance of observations (e.g. metric Bray Curtis), 3) the phylogenetic diversity across observations (e.g. metric unweighted and weighted UNIFRAC)[54-56], 4) or some combination of all three.

An expectation under the first hypothesis of phylosymbiosis is that hosts with the same genetic background will have more similar microbial communities than genetically divergent hosts. This hypothesis could be evaluated if organisms of the same genetic background have lower ecological distances between their microbiomes than organisms of different genetic backgrounds. However, more advanced tests with statistical and nonparametric advantages have been developed to assess beta diversity distinguishability between categorical groups such as host species (ANOSIM[57;58]), across continuous variables (PERMANOVA[58]), correlation with other kinds of distance based outcomes (Mantel Correlation[59]), or supervised classification (Random Forest[60]), each with their own advantages[61]. In addition to

statistically evaluating hypotheses with beta diversity, methods for decomposing high dimensional distance matrices allow researchers to visually examine sample relationships in two or three dimension ordination plots (Principle Coordinates of Analysis, PCoA[62]; Non-metric Multidimensional Scaling, NMDS[63]; Canonical Correlation Analysis, CCA[64]), or in dendrograms with sample relationships as tree distances[65]. Commonly used beta diversity metrics each have their own nuances and caveats that must be considered, but together they provide a foundation to understand how microbiomes vary between organisms.

A similarly diverse set of measures quantify evolutionary divergence of hosts. Phylosymbiosis has focused on characterizing evolutionary divergence between closely- or distantly-related host species based on the genetic similarity at shared loci within their genome. The most widely used tool to assess host relatedness is to cluster genetic distances between samples into tree dendrograms known as **phylogenies**. Just as with beta diversity, there are a wide-variety of metrics and approaches to calculate genetic distances and cluster samples[66]. While phylogenies can be constructed with human genomes, finer measures of interspecific genetic divergence are more conducive to the degree of variation between humans lacking reproductive barriers. The advent of genotyping and sequencing abilities to characterize Single Nucleotide Polymorphisms (SNPs)

across the genome has allowed quantification of host genetic relatedness in microbiome studies using measures of: SNPs shared (Identity By State, IBS), SNPs shared due to common ancestry (Identity By Descent, IBD), distinguishing SNP variation across many loci (Population Structure)[67-69], how SNPs associate with covariates like height or disease status (Genome-Wide Association Studies, GWAS)[70;71], or the contribution to traits from additive SNP effects (Heritability)[72-74]. The second hypothesis of phylosymbiosis has been evaluated by looking for correlation between beta diversity and phylogenetic distances, or looking for concordance between microbiome dendrogram and host phylogenetic trees[23-25;52]. The first approach may be useful because direct distances are taken into account, but this also assumes equivalency in rate of change between the genome and microbiome which are governed by very different forces[23]. The second approach only utilizes the clustered interspecific relationships and therefore is less powerful; however, it reduces the assumption of equal rates of change[23]. Just as with beta diversity, there are many additional evolutionary measures of host genetic relatedness that can be applied in microbiome research, and each of these approaches have extensive nuance and caveats that must be considered. Regardless of the approach, tools to measure phylosymbiosis can reveal insights about individual microbial contributions to the observation of phylosymbiosis[23].

The combination of these basic ecological and evolutionary measures provide a foundation to assess the roles evolution may play in microbiome assembly, as well as a wide range of other hypotheses.

Evidence for Phylosymbiosis

What evidence exists in metazoan and model systems that host evolution shapes microbiomes, and how are these hypotheses of phylosymbiosis critically assessed using beta diversity and phylogenetic measures? The first hypothesis of phylosymbiosis predicts that intraspecific beta diversity distances will be lower than interspecific. Using the techniques discussed above, such a prediction would manifest as grouping of microbiomes by host species in ordination space and dendrogram trees. This is because the formation of reproductive barriers during speciation lead to evolutionary divergence, and therefore if each variant had an equal effect then microbiome composition would be expected to also continuously diverge. It should not be assumed however that each variant will have an effect on microbiome composition, and for variants that do affect microbiome assembly the magnitude of change could vary widely. Indeed, early microbiome studies across metazoans found that microbiomes were generally distinguishable across diverse host species[25;75]. A caveat of such broad interspecific examination is revealed

where beta diversity associates with host phylogeny, but also confounding factors like host diet, taxonomic order, and provenance[13]. Therefore, it is impossible to say if herbivores cluster away from carnivores because they are more closely related genetically, or if the dietary shifts correlated with evolutionary divergence are causal[13].

By limiting confounding effects like diet, environment, sex, endosymbionts, and age, phylosymbiosis has been observed within laboratory controlled clades of closely related host species of *Hydra*[49;76], *Nasonia*[24;51;77;78], *Peromyscus* deer mice[24], three families of mosquitoes[24], and *Drosophila*[24]. Outside of controlled lab conditions in natural populations the observation of phylosymbiosis is mixed, with the signal observed in clades of diverse mammals[23], coral[52], birds[79], and the skin microbiomes of grazing mammals[80] and fish[81], but not within clades of amphibians[82], *Drosophila*[83], and among more divergent birds[79], coral[52], and carnivorous mammals[23]. The difference in results between controlled lab and natural field studies usefully suggest conditions where predictions of phylosymbiosis may breakdown. One possibility is that the influence of varying intrinsic factors like age and sex, as well as extrinsic factors like diet and environment can play more influential roles in microbiome assembly than host evolution. Indeed, dietary differentiation was clearly a factor alongside host

phylogeny in explaining microbiome composition[23;75;81]. It could also be supposed that even in controlled environments, the influence of host evolution could be so subtle that stochastic factors like microbial dispersal limitation and community dynamics could obscure any signal of phylosymbiosis. These two possibilities raise the possibility that there may be a 'Goldilocks Zone' of evolutionary divergence where phylosymbiosis is most strongly observed, as under the first possibility hosts have become so diverged in factors like lifestyle or physiology that phylosymbiosis is obscured, or under the second where a lack of evolutionary divergence does not lead to interspecifically distinguishable communities. Indeed, the evolutionary age of divergence across species within a host clade strongly correlates with the degree of microbiome distinguishability measured by ANOSIM tests in controlled settings for clades diverging from 1mya to 100mya[24], but the correlation of host phylogeny with beta diversity is significantly overpowered by dietary correlation when examined across more divergent host species (i.e. >500mya)[23;79].

It is estimated that modern humans arose within the last 200k years, therefore does evidence exist that evolutionary changes in our genome have measurable effects on microbiome composition, particularly in such a short timeframe? This question leads to a third but albeit rare possibility for the

obfuscation of a phylosymbiotic signal, that a single evolutionary variant confers very little evolutionary divergence on its own, but if the variant has a very large effect on microbiome composition, it could disrupt phylosymbiotic expectations for microbiome relationships. In humans, the persistence into adulthood of lactase gene metabolism of lactose from milk arose at least three times across southern Europe and northern Africa, and the ability to utilize milk as a high energy food source has advantages that could lead to selection driven maintenance in the population[84;85]. Population geneticists have developed GWAS techniques to identify SNPs across the human genome that correlate with a phenotypic trait[71]. The technique has been adapted to measure which variants associate with changes in total microbiome composition measured by alpha and beta diversity, or the abundance of individual microbial taxa[9-11;13;86-90]. One of the first associations identified was between a SNP on chromosome 2 in the lactase gene and the abundance of *Bifidobacterium*, suggesting that variation in the ability to process lactose sugars could shape microbiome composition at least among individual taxa[9]. Considering the benefits such a high energy food source could confer, it is probable that there could be evolutionary selection on SNPs affecting the lactase gene, but also sociocultural selection to maintain lifestyles with sources of dietary milk. This therefore entangles host evolution and extrinsic

factors like herding and societal structures. Due to identifiable shifts in human microbiome composition associated with small evolutionary changes in the lactase gene over the last 50k years, the concept of unequal evolutionary contributions to microbiome composition spread across time could rapidly disrupt a signal of phylosymbiosis.

Under phylosymbiosis, the interspecific correlation between microbiome distinguishability and genetic divergence has a parallel in the observation that beta diversity across human individuals correlates with the degree of shared SNPs genome-wide[9]. This observation supports the potential utility of applying measures of human genetic relatedness to generating expectations about microbiome similarity across individuals. Measuring heritability often relies on twin studies, where the degree to which a phenotype is shared is compared between genetically identical monzygotic twins and dizygotic twins sharing ~50% of their parent's genetic material[72]. Applying twin studies of heritability to the microbiome means looking for differences in community composition across the most limited degrees of genetic divergence, yet it has been reproducibly observed that beta diversity and the abundance of some microbial taxa are significantly heritable[7;13;91;92]. Beyond microbiome divergence between twins, beta diversity distances between microbiomes significantly increased for parent-sibling

relationships, and further for microbiome similarity between unrelated individuals[7]. These examples highlight how the phylosymbiotic hypothesis that genetic divergence will lead to microbiome divergence may provide useful predictions about the relationship between host evolution and microbiome divergence both between and within species. They also demonstrate how phylosymbiosis may serve as a useful hypothesis to identify where intrinsic or extrinsic host variation, changes from neutral to selection driven evolutionary effects on the microbiome, or violations of assumptions about the gradual concordance of variation in both systems can obscure such predictions.

Human Microbiomes in a Phylosymbiotic Context

How can changes in human microbiome composition be understood relative to our closest hominid and hominin relatives? Across clades of humans and wild hominids the pattern of phylosymbiosis has been observed[24;50;93]. Divergence since the last common ancestor of humans and chimpanzees leads to the expectation that human microbiomes will be distinct, and indeed this is observed as greater interspecific versus interspecific beta diversity distances[24;50;93]. Relative to microbiome divergence between other hominids, however, human microbiomes are separated by larger interspecific distances than would be

expected from concordant change between the genome and microbiome[92]. Humans and hominids in captivity have lower microbiome **alpha diversity** than hominids living in the wild[50;94;95]. Alpha diversity is another widely used measure in microbiome research, which assigns a value to diversity within a microbiome based on the total community composition. Depending on the calculation metric used, alpha diversity is affected by community: 1) **richness** as the number of unique observations (e.g. metric observed OTUs), 2) **evenness** as the similarity of observation frequencies (e.g. metric Pielou's evenness), 3) phylogenetic **diversity** across observations (e.g. metric PD whole tree), or 4) a combination of richness, evenness, and diversity. While alpha and beta diversity are both calculated from total microbiome profiles, there are no inherent expectations about how alpha diversity will change across host evolutionary divergence. Still, differences in community alpha diversity will likely be reflected by increasing beta diversity distances between samples as the loss of alpha diversity reduces overlap of community composition.

Why do humans have lower gut alpha diversity compared to our hominid relatives, and what aspects of our changing lifestyles, diets, and genetics have played a role? Among hominin relatives, ancient Neanderthal oral microbiome compositions reveal significant beta diversity divergence between those that ate

primarily meat and those primarily subsisting on nut and plant diets[96]. Still, beta diversity distances between Neanderthals were dwarfed by distances between human and Neanderthal microbiomes, some of the latter which had dental microbiomes more similar to wild Chimpanzee[96]. One conclusion might be that relative to the dramatic lifestyles shifts in modern humans, ancient hominin shared more interspecific environmental and dietary similarity with hominid relatives, therefore leading to lower beta diversity distances. Across Europe, shifting from a hunter-gatherer lifestyle to farming and then industrial lifestyles corresponds with distinguishable changes in dental microbiomes and loss of alpha diversity in fossilized human remains[97]. Attributing causality is difficult as each historical period was also accompanied by shifts from complex to simple dietary carbohydrates, higher to lower overall dietary diversity, outdoor and rural to indoor and urban lifestyles, natural to more sterile dwellings, and many other factors.

One could critically argue that ancestral human and hominin microbiomes vary only for technical and not biological reasons, but shifts in lifestyle have parallels in modern human cultures from around the globe today. Some of the most profound differences in human gut microbiomes are observed between traditional hunter-gatherer or subsistence farming cultures and urbanized western

lifestyles, again accompanied by decreasing alpha diversity[28;98-100]. Western lifestyles also correlate with shifts in gut carbohydrate active enzymes abilities to process plant versus animal food source substrates, suggesting diet is playing roles in shaping microbiome metabolism across lifestyle gradients[99]. Microbiomes in four populations spanning traditional to urbanized lifestyles in the Himalayas associated with environmental factors like varying water sources and dwelling style[27]. Clearly many aspects of modernization in human culture may help to drive the observation of increased microbiome divergence outpacing what would be expected from evolutionary divergence alone. Even migration into more westernized cultures were accompanied by beta diversity divergence and alpha diversity reductions within individuals from two ethnic minorities as they immigrated from Thailand to the United States[101]. While diet and lifestyle appear to play important roles for these immigrants, effects covaried with unmeasured changes in habitation, stress associated with global migration, and new social interactions[101].

A gradient of beta diversity similarity from more traditional to westernized cultures around the globe today stratified microbiomes along the first principle component of PCoA plot, and this divergence in community composition correlates with the abundance of particular microbial taxa[99]. Along this gradient

there is a reproducible shift from taxonomic groups containing complex fiber degrading *Prevotella* in more traditional cultures, to *Bacteroides* which process simple sugars more readily and mucus degrading *Akkermansia* in westernized cultures[8;99;102]. The utilization of dietary metabolites is likely a key driver of taxonomic divergence across lifestyles, and this is reflected by differences in microbiome enzymatic capacities[99]. Ultimately divergence in microbiome composition across hominids, hominins, and lifestyle gradients in modern humans have likely been influenced by host genomic differences to varying degrees. However, the larger than expected degree of beta diversity divergence to hominid and hominin relatives and losses in alpha diversity at many stages of human advancement highlight how extrinsic environmental factors can confound relationships between host evolution and microbiome composition.

Microbiome divergence between humans and interspecific relatives and between intraspecific lifestyle shifts outpaces what would be expected from evolution, and this may inform modern human health[93]. Many common and chronic diseases associate with reduced alpha diversity, and share overlapping changes in microbiome taxonomic composition[103]. One hypothesis is that modern human microbiomes are shaped by very different extrinsic factors than what more slowly changing human genomes have evolved to contend with. This could lead to

ineffective host control of the microbiome through loss of immune or metabolic mechanisms shaped over millions of years. Widespread antibiotic use, more sterile built environments, reduced dietary diversity, and a range of other factors in modern human life could be responsible for lower alpha diversity[104-106]. These losses of community complexity could in turn allow openings for opportunistic pathogens like *Clostridium difficile* to take hold in the gut, and it has been shown that abundance reduction of Clostridiales precedes hospital acquired *C. difficile* infection[107]. Invasion of potentially pathogenic microbes is closely linked with host immunity, and it has been observed in mice that a disrupted gut microbiome allows invasion of oral microbes which in turn drive immune inflammation[108]. Invasion of oral microbes into the gut microbiome was significantly higher in humans with five diseases, which may reflect imbalances or 'dysbiosis' of community composition[108]. If lower alpha diversity is a precursor leading to disease, then the increased rates of many common and autoimmune diseases in westernized societies may be a reflection of the wide-spread reductions in alpha diversity that have occurred relative to hominids, hominins, and lifestyle modernization.

Roles of Reciprocal Host and Microbial Evolution in Shaping Communities

Composed of trillions of genomes from thousands of microbial taxa, the genetic components of the microbiome undergo evolutionary changes as well. The pressures acting on genetic variants in the microbiome may not be uniform across taxa however, and the intimacy of a microbe's interaction with a host could be a key determinant. Phylosymbiosis treats the total or portions of the microbiome as a measurable unit that may change between host species by a variety of mechanisms including host filtering, bacteria filtering, microbial evolution, and more. Within this context, host and microbial genomes can exert evolutionary selection pressures on each other, particularly if the association is maintained over multigenerational timescales. Some of the most intimate host-microbe interactions are revealed by comparing host and microbial evolutionary patterns of **cospeciation**, **codiversification**, and **cocladogenesis**. In these examples, evolutionary divergence patterns across host species are paralleled by evolutionary divergence of a particular microbial strain or species, which can be quantified as the similarity of evolutionary patterns between host and microbe phylogenies. These patterns suggest that a microbial taxon has been associated with a host or shared environment over evolutionary timescales, and that formation of reproductive barriers between host species are paralleled by separately evolving

microbial lineages. Coevolution and codivergence are most often observed in strict cases of host interactions with a microbe over many generations, and inter-generational transfer of microbes could be evolutionarily selected for in host and microbial genomes. Some of the most intimate examples of host-microbe codiversification result from the action of **coevolution**, where host and microbiome exert evolutionary pressures on the other's genome reciprocally over time. Two striking examples include aphids and their *Buchnera* endosymbionts where critical functions are lost from *Buchnera* genomes and instead fulfilled by the host[109;110], and bobtail squid which acquire *Vibrio fischerii* symbionts into a specialized organ that helps protect the host from predation[111;112]. Interestingly, microbes may not only be affected by host speciation events and subsequent evolutionary divergence, but they could also create reproductive barriers leading to host speciation[77]. Under strict lab conditions, such a pattern was observed in *Nasonia* wasp species where reproductively isolated species gained the ability to reproduce when reared germ free, and reproductive barriers returned when re-inoculated with native microbiomes[51;77]. While evidence for cospeciation of microbes with humans is tenuous[113], the evolutionary mechanisms selecting for inter-generational microbial transmission may still play important roles in human gut microbiome assembly.

Microbial Transmission Shapes Intimate Host-Microbe Associations

Definitions of heritability often imply host genomes affecting a phenotype, but it is important to remember that a microbe's genome could undergo selection for its own heritability if it experiences fitness or performance advantages in the host environment. In the case of *Buchnera* symbionts and their aphid hosts, genomic loss of microbial functions necessary to life has led *Buchnera* to become obligate to the aphid microbiome environment[109;110]. Without the aphid fulfilling key functions lost in the *Buchnera* genome, and reciprocally without vitamins provided by *Buchnera* to the aphid, neither would survive. This necessity has facilitated **vertical transmission** of *Buchnera* directly from aphid parent to offspring. Vertical transmission has led obligate symbionts of many arthropod species to become inviable without their host, and intertwining microbial fates with that of a host's lineage may be a key driver of codiversification patterns[114-120]. Cases of strict vertical transmission support a debated **hologenome** hypothesis, stating that in addition to selection acting on host and microbe as independent units of life, a level of selection can also act on the host's genome and microbiome as a single unit. Such a prediction makes sense in strict cases of coevolution like aphids and *Buchnera*, where host and microbial fates are directly intertwined.

Hologenome selection could also depend on how the host genome relates to a compilation of microbes, the total community composition, factors like community stability, or host-microbe metabolic complementarity. Still, extensibility of a hologenomic level of selection becomes less clear when host and microbial fates are not so closely intertwined, such as with *V. fischeri* and the bobtail squid[112]. While squid fitness can be linked to *V. fischeri* providing resistance to predation, both *V. fischeri* and the squid can survive in the ocean environment without the other. Squid acquire *V. fischeri* through **horizontal transmission** from the ocean environment on a daily basis, but horizontal transmission also encompasses microbial acquisition from other members of the same species that do not fall under strict parent-to-offspring vertical transmission. As *V. fischeri* can survive alone in the ocean its fate is not necessarily linked to acquisition by squid; however, squid have evolved a physiological organ devoted to **host filtering** that specifically selects for *V. fischeri*[112]. Through acquisition into the non-competitive squid-organ microbiome, *V. fischeri* gains advantages over other ocean microbes that could induce evolutionary selection for its acquisition by the squid. While the existential fate of host and microbe are not necessarily tied together in this example, the mutualistic benefits each provides the other may improve not only their own fitness, but also that of the hologenome.

By definition fitness depends on an organism's ability to produce offspring, and for this reason the act of propagating and raising young is under strong evolutionary pressures. In humans, the role of microbial transmission and maintenance from parent to offspring in early life is becoming increasingly recognized as critical for health as people age. The human vagina provides infants with their first exposure to complex microbial communities, and evidence abounds that vaginal community composition is controlled by host filtering[4;121;122]. Interestingly pH appears to be a key factor in vaginal filtering of microbial taxa, and reciprocally *Lactobacillus* and other dominant vaginal microbes may contribute to acidification through lactic acid production[123]. The apparent ubiquity of lactic acid production across vaginal taxa suggest similar evolutionary forces may be at play as in *V. fischeri* and squid, where host physiology and microbial contributions to host health could lead to a level of hologenomic complementarity. Maintaining a selective vaginal environment contributes to maternal performance through prevention of community imbalance and foreign microbial invasion that can lead to vaginosis[124]. Selective vaginal ecosystems also play roles in maternal fitness, both during initial childbirth and in long term health of the child. It has become widely recognized that inoculation with a healthy vaginal microbiome during childbirth helps to seed the long-term composition of infant microbiomes, and a

primary mechanism is to train children's immune systems through a process called **immune education**[37;125]. Adaptive immunity is developed throughout human's lifetimes with exposure to pathogens, but comparison of natural and c-section births highlight how important initial exposure is through dramatically higher rates of allergies and asthma when babies are not exposed to vaginal microbiomes at birth[126;127]. While vaginal inoculation at birth helps seed a baby's initial microbiome, replacement of maternal strains and species occurs quickly over the first few years of life which suggests that immune education and not long-term vertical inheritance may be a driving force of selection for transmission[125;128]. In addition to vertical inheritance of vaginal microbes, there is evidence that mothers contribute to horizontal acquisition of microbes in babies through breast feeding in multiple potential ways. First, it is believed that human breast milk is not sterile, and contribution of microbes through this early feeding source may have similar roles as vaginal communities in immune education[129]. It is unknown if genetic selection has shaped which microbes appear in breast milk, but a second way that breast milk may contribute to horizontal transmission of microbes is through conferred host filtering effects. Human milk oligosaccharides (HMOs) are formed from five monosaccharide sugar building blocks linked in over 100 identified combinations, far outnumbering the complexity observed in

other mammals[130-132]. These HMOs can be metabolized by the gut microbiome of babies, and the nutritional benefits provided may help select for certain combinations of beneficial microbes that regulate immune development[132]. An interesting proposition arises when the role of microbial evolution is considered in these cases, that microbial production of lactic acid in vaginal microbes and consumption of HMOs in infant guts could be conferred across strains and species by **horizontal gene transfer (HGT)**. Mobile genetic elements rich in metabolic capacities appear to be transferred between *Lactobacilli* in vaginal communities, and there may be a component of hologenomic selection at play because the conference correlates with exclusion of pathogenic *Gardnerella* species that would hurt the host[133]. In combination with selection on maternal fucosyltransferases that help microbes metabolize HMOs, results of the high prevalence of HGT in human gut microbiomes may help contribute to maternal horizontal transmission of *Lactobacillus* and *Bifidobacterium* to babies[134;135]. HGT has also been a key driver of antibiotic resistance spreading across diverse microbes in human microbiomes following decades of widespread antibiotic use, leading to a clear disconnect between selection pressures on host genomes and the microbiome[106;135]. The rise of antibiotic resistance highlights that hologenome complementarity can be superseded by selection on microbial fitness despite deleterious effects on the

host. Clearly evolution has shaped interplay between host genomes and microbiomes across millions of years, and there is distinct crossover between patterns identified in animal models which could have direct implications on human health. Principles like the hologenome and phylosymbiosis provide useful predictions about host-microbiome evolution, and the lack of ubiquity with which they are observed also allows the identification of alternative principles which can lead of breakdown of those expectations. Understanding which evolutionary principles apply in humans may allow researchers to correlate expected patterns, or lack thereof, with positive and negative health outcomes. The ecological forces dictating microbiome assembly, factors driving community dynamics through a host's lifetime, and roles of community composition and stability in human health will be considered next.

Ecological Building Blocks of Community Assembly

Ecological forces constantly shape the microbiome throughout an organism's lifetime, and provide mechanisms on which host and microbial evolutionary forces can act. A microbiome composes the genomic content for trillions of living microorganisms that represent the biotic component of a complex ecosystem. These organisms are constantly competing for non-living

abiotic components of that ecosystem necessary for life, including metabolites, minerals, vitamins, essential amino acids, and nucleotides. Such a complex ecosystem is analogous to the complexity of macro-ecologies such as forests or oceans, and many of the same measures developed over centuries of studying animals and their environments can be applied to microbiomes. Larger ecological principles dictating microbiome assembly also overlap those shaping macro-ecosystems, and their application to understanding community function, diversity, and stability can inform health over individuals' lifetimes and response to extrinsic perturbations. Function is a key determinant of ecological assembly, and characterizing functional contributions and requirements of individual microbes and the community as a whole can provide targets for interventions aimed at shaping the microbiome. It is with this in mind that an ecological framework will be established by discussing how function, competition, initial assembly, short- and long-term community dynamics, and stability play into human health.

Functional Diversity, Redundancy, and Complementarity among Microbes

The gut microbiome is among the most dense and diverse microbial ecosystems identified, and competition for resources means that microbes do not

live individually or independent of their host. With over a thousand-species identified throughout human digestive tracts, a multitude of diverse functional and metabolic roles help determine how microbes interact with one another[4]. These abilities are dictated by the genetic capacity within each microbial genome, and this content can vary widely even across strains of the same species[136]. Compared to the human genome with roughly 20 thousand human genes, metagenomic sequencing has identified more than 10 million unique microbial genes[6]. Therefore, **functional diversity** within the microbiome vastly outnumbers capacities found within the human genome. When considered across the breadth of microbial taxa which have evolved over billions of years however, it is interesting to observe that most healthy human gut microbiomes are dominated by two bacterial phyla, Bacteroides and Firmicutes[4]. The remaining bacterial taxa in the human gut are often predominantly Proteobacteria and Actinobacteria, many lineages of which are considered environmental and can exist extrinsic to host associated microbiomes[1;4]. Over 50 bacterial phyla have been identified to date, and the limited breadth of phyla in the human gut likely reflect the unique traits of such an environment. Genetically encoded functional capabilities such as anaerobic respiration in the oxygen deprived environment, adherence to epithelial and mucosal linings, and host filtering for certain traits may be key determinants

of microbial survival in the digestive tract. Selection on gut microbiome composition by the combined effects of the ecosystem may drive observations of **functional redundancy**, where the same advantageous traits that allow survival in the environment are maintained or even spread with HGT across many taxonomic lineages of gut microorganisms. A key goal of early studies like the Human Microbiome Project was to identify components of the **core microbiome**, defined as taxa or functional traits that are ubiquitously shared across most or all people in the population[4]. As more individuals have been sampled it has become clear that a core set of taxa are not maintained across microbiota, yet functions are more conserved across metagenomic profiles even when there is very little overlap of microbial strains or species[137]. With ecological measures applied to taxonomic or functional microbiome profiles, functional redundancy may be observed as lower inter-individual beta diversity distances and less alpha diversity variation between profiles of microbiome functions compared to taxonomy. In such a comparison however it is important to consider how methodological differences between metagenomic and 16S amplicon sequencing can affect diversity measures, particularly biases like depth of sequencing coverage and resolution at which features are defined (e.g. taxonomic species versus families, functional representation of particular enzyme versus collapsing counts into

metabolic pathways). Still, when limiting such variation by using taxa and functions annotated from the same metagenomic profiles, higher inter-individual correlation in genes was also observed compared to species[138]. Inter-individual correlation was lowest for metatranscriptomic profiles where more than half of transcripts were differentially expressed compared to their metagenomic abundance, suggesting observed functional redundancy in the metagenome represents the available genetic capacity and not necessarily functional activity[138]. Altering downstream products encoded by the microbiome occurs during transcription into RNA, translation into proteins, post-translational protein modification, enzymatic metabolism of those proteins, and a range of other mechanisms acting throughout those processes. Pairing metagenomic sequencing techniques with metabolomics is gaining wider popularity because it allows researchers to directly compare potential genetic capacities of the microbiome with actual functional consequences in abiotic metabolite targets and products. More holistic use of multi'omics techniques may reveal how gut microbiomes and abiotic factors function as an ecosystem, but a key challenge is distinguishing effects of microbiome versus host metabolism. This issue emerges because microbiomes and hosts metabolize many of the same molecules, and host-microbe and microbe-microbe **metabolic complementarity** has direct roles on

the gut ecosystem and host health. One example is microbial fermentation of complex dietary carbohydrates into short chain fatty acids (SCFAs). Resulting SCFAs from fermentation then become available as energy sources for: human colonocytes in the lower intestine (e.g. butyrate), gluconeogenesis in the human liver (e.g. propionate), and by other bacteria which is known as **cross-feeding** (e.g. acetate)[139]. Fermentation metabolizes what are called 'non-digestible carbohydrates' predominantly from dietary plant sources, yet such terminology seems flawed when considering they are only non-digestible without a complementary microbiome that catalyzes the first steps of digestion. Yet non-digestible may be the case for many people as fermentation into SCFA's is functionally redundant among a subset of Firmicutes, which may or may not be represented in each individuals' microbiome[140;141]. Much like tools for evolutionary comparison, functional capacity in the microbiome can be assessed through direct nucleotide alignments or creating hidden Markov models that learn important genetic signatures for functional gene families. Thus, metagenomic sequencing yields the genetic material that allows researchers to understand how capabilities like fermentation are distributed across the breadth of taxa in a microbiome, whether factors like HGT or convergent evolution could play into that functional redundancy, and how such functions are represented across

human populations.

Microbial Niche Specialization, Competition and Compatibility

How do functional abilities contribute to microbial survival and success within the gut ecosystem? Beyond identifying individual functions, measures leveraging total genomic capacities organisms have been developed to understand how well a microbe's total metabolic potential complements human metabolism[142]. Abiotic products of microbial metabolism that complement humans include essential vitamin B12, and the necessity of this nutrient could lead humans to shape an ecosystem favorable to vitamin B12 producing microbes that is analogous to aphid and *Buchnera* relationships. Through **niche specialization** microbes can take advantage of favorable conditions in the gut ecosystem, where environmental characteristics complement particular microbial traits leading to reduced **competition** with other microbes. The process of a microbe favorably occupying a niche could result from stochastic assembly of characteristics within a particular microenvironment and corresponding microbial traits, but it could also be directed by host control of the gut ecosystem or microbial construction of favorable conditions. Competition for abiotic factors may favor certain microbial traits, such as formation of a **metabolic niche**

through the bioavailability of essential metabolites and minerals that can only be utilized by a subset of taxa. Microbes compete fiercely for essential trace metals like zinc and iron, and humans are able to sequester these metals inside cells to control microbial access[143;144]. When pathogenic microbes take hold in the gut, the bioavailability of iron is key to their rapid growth, and a host's ability to limit this essential mineral creates a chokepoint to microbial over-proliferation known as "**nutritional immunity**[144]." Some viruses also compete for the iron metabolic niche to achieve the same ultimate goal of replication, but viruses favor acquisition of iron by human cells because they rely on the host's replication machinery for their own life cycle[145]. To create their own iron metabolic niches, different viruses have developed a series of tools including: targeting active iron transporters as receptors in host membranes to preferentially infect cells that are actively acquiring iron, disrupting signaling pathways that limit iron acquisition by host cells leading to increased uptake, and blocking iron efflux through degradation of ferroportin[145]. The availability of many metabolic niches within the microbiome depend on the extrinsic influence of host diet, with macronutrient, specific metabolite, and mineral content varying widely across food sources. As omnivores, human diets can vary widely, and microbiome composition varies widely between animal and plant, high fat and low fat, high protein versus high

carbohydrate, and traditional versus western diets[26;28;30]. Firmicutes that perform fermentation into SCFAs in the human gut seem to prosper with plant based diets for instance, and this makes sense because they leverage a metabolic niche through the specialized ability to metabolize complex plant carbohydrates[141;146;147]. Considering how diet can create (e.g. through additional plant carbohydrates) or eliminate (e.g. through low iron) metabolic niches could provide targets to control specific microbes in the gut. Limiting iron for instance has been shown to reduce or even ameliorate negative processes and health outcomes associated with human immunodeficiency virus, hepatitis B and C viruses, and human cytomegalovirus[145]. **Biogeographical niches** may also form in the gut through spatial exclusion of competition, either by creating local barriers that prevent entry of other microbes, or by leveraging traits that allows persistence within a particular selective microenvironment space. Regulating host mucus production to establish an exclusive mucosal biofilm for instance can create a barrier for other microbes and threats like antibiotics. Other microbes have developed ways to adhere to the gut epithelial lining that allows spatial persistence, countering the common microbial fate of colonic transit. Microbes may also leverage biogeographical niche microenvironments purely for their own benefit, and in ways that are pathogenic to the host. Habitation of colonic crypts bypasses host

"metabolic barriers" where colonocyte cells normally consume butyrate before it penetrates to basal crypt progenitor cells[148;149]. Microbial expansion into the less competitive base of crypts allows butyrate and other molecules to bypass the colonocyte "metabolic barrier", and has been linked to inflammation, delayed wound repair, and colorectal carcinogenesis[148;149]. If two microbes occupy the same metabolic and spatial niche, it is likely one will eventually displace the other through a process of **competitive exclusion**. The ability for microbes to colonize individual humans varies from person to person, which suggests interpersonal variation in niche specificity[148]. Alongside diet, human microbiome composition prior to fecal microbiome transplant (FMT) and probiotic inoculation is predictive, and likely deterministic, of which foreign microbes can take hold in the gut[148]. Following antibiotic treatment, reconstitution of a diverse human gut ecosystem seems to depend on the complexity of the inoculation, where simple probiotics reduced and diverse FMT improved reestablishment of community diversity compared to no innoculation[150;151]. This may suggest that filling a rich diversity of microbial niches leads to faster and more stable gut community assembly, when compared to less diverse inoculations like probiotics where microbes may partially and ineffectively fill many unoccupied niches. Under the theory of competitive exclusion, more available niches equate to more

opportunities for foreign and potentially pathogenic microbes to invade. Therefore, reductions in microbiome alpha diversity across westernized societies could allow opportunistic and potentially pathogenic microbes to overgrow by lacking direct competition, which in turn may be reflected in the correlation between alpha diversity and negative health outcomes. Going forward clinicians could consider the niche requirements of specific pathogens as targets to eliminate or fill with commensal microbes, and alpha diversity as a quantifier of a microbiomes capacity for competitive exclusion. As our understanding of microbial niche occupation and competitive exclusion grow, shaping the landscape of available niches may provide new ways to engineer ideal ecosystem homes for a diversity of 'healthy' microbes.

Priority Effects and Dispersal Limitation Shape Initial Microbiome Assembly

Functional capacity and niche preference help determine a microbe's survival in the gut ecosystem, but a range of community ecology principles capture or shape composition and diversity beyond single microbe interactions. As with vaginal inoculation and HMOs shaping human microbiomes early in life, which microbes colonize first can have long-term consequences on the microbiome and host health. **Priority effects** manifest when the sequential order

of microbial colonization shapes which microbes can subsequently colonize. Mechanisms of priority effects could embody many principles like competitive exclusion limiting niche access, functional diversity and redundancy controlling adaptability between niches, symbiotic cross-feeding creating new niches, and even indirect **apparent competition** for niches by shaping third-party predator-prey preferences or immune education to act on competitors. For example, *Haemophilus influenza* virus that causes flu may exert small changes on host immune education that allows a wider diversity of *Streptococcus pneumoniae* infection which causes pneumonia, yet the complexities of microbial coinfections and co-colonization experienced throughout an individual's life make such dependencies difficult to distinguish by merely looking for pairwise microbial co-occurence[152]. Examining human microbiomes temporally may help with such insights however, where patterns in colonization order may better reveal where underlying priority effects could be at play. In fly models, the likelihood of microbial colonization decreased with previous microbial inoculations, and spatial niche occupation appeared to be a factor in this competitive exclusion[153]. Still, a combination of stochastic dynamics and microbial dose in the inoculum seem to shape initial community assembly, while downstream assembly and community stability seemed to be shaped by priority effects stemming the 'lottery' of

successful initial colonizers[153]. A key contributor to colonization order is the '**exposome**' of microbes and metabolites that a host confronts in their environment on a daily basis[154;155]. Biogeographical partitioning of different microbes into physical regions is determined by **dispersal limitations**, essentially the distribution range of each microbial species. Clouds of microbes stochastically assembled in the air and on surfaces are ubiquitous to our world, but exposomes vary widely in composition across locations and time points[155]. Examination of wild mammal microbiomes showed that phylogeny and diet explained some microbiome variation, but was independent of a biogeographic signal that physically closer mammals had more microbiome similarity (i.e. lower beta diversity distances = more similar microbiomes)[156]. The role that living indoors has likely played in shaping human microbiomes is likely profound[93], and sterilized indoor air contains a very different exposome than the local outdoor environment[104;155;157;158]. Humans also seed the exposome with microbes, and one does not have to look farther than individuals' cell phones to see microbial transfer to the local environment[159]. Horizontal transfer of diverse microbes through the exposome maybe critical for proper immune education, as transfer of microbes from pets to owners is prevalent and correlates with lower rates of asthma and allergies[37;160]. The prevalence of c-section births has led hospitals to

implement horizontal microbiome transfer of vaginal swabs from mothers to newborns, in many ways setting a long term trajectory by shaping the newborns first exposome[37;125;127;128;161;162]. Extending such a concept could lead to shaping indoor microbiomes in the interest of health, for example to inoculate a competitive iron niche occupying microbe to limit a viral outbreak. A current challenge is packaging and delivery of live probiotics with billions of microbes into a competitive gut ecosystem containing trillions of microbes, and therefore competitive exclusion may be limiting probiotic effectiveness. Sourcing probiotic microbes from human microbiomes instead of common cultures such as Lactobacillus from milk may be a key first step in effective colonization. Still, it may be advantageous to more broadly consider natural ecological patterns of microbial colonization when designing probiotics, such as: priority effects in determining when a probiotic would best colonize an individual, where exposome composition is inadequate for immune education and should be complemented with probiotics, or how characterizing microbial functions can inform probiotic niche specialization that will competitively exclude an opportunistic pathogen.

Community Dynamics and Ecological Stability Shape Microbiomes Throughout Life

While many mechanisms shape community assembly, an established

community is also constantly changing according to ecological forces. **Community dynamics** are simply the variation in a microbiome over time, but the underlying mechanisms of such variation is anything but simple. Changes within microbiomes can occur stochastically, such as with extrinsic influences of dispersal limitations and microbial exposure, or intrinsically such as with random fluctuations in microbial abundance or the niches that are available[53;153;163]. Dynamic change may also be directed by a range of ecological principles already discussed, such as with the progression of microbial communities from low diversity and disparate compositions in newborns, to higher diversity and more interpersonal overlap in older children[125;127;160;164]. Considering time as an aspect of microbiome composition is important because of dynamic fluctuations, but temporal change occurs across many scales[165]. Timescale variation affecting the microbiome may be short such as between day and night or time since last eating, medium such as dietary patterns changing by season or regular menstruation, or long such as expansion of microbiome alpha diversity in children and depletion with old age[165]. Short term dynamics are likely influenced by many of the microbe-microbe and microbe-host ecological principles already discussed, but fecal sampling makes it difficult to get finer than daily resolution of gut microbiome dynamics. This is where model organisms excel by allowing minute to hour

temporal dynamics to be observed, such as in zebrafish gut microbiomes where microbial motility and priority effects help determine dynamic change over short time periods[166]. **Ecological stability** could have a number of interpretations, but is generally defined counter to community dynamics as maintaining microbiome composition over time. Host gut motility creates dynamic microbial turnover through fecal movement, but microbial traits of adherence and motility allow microbes to persist and restore stability[166;167]. This reset of ecological stability may counter dangerous pathogenic overgrowth, and it is speculated that the human appendix has evolved to serve as a microbial reservoir that restocks a stable ecosystem[166-168]. Ecological stability can also define the resiliency of a microbiome to maintain composition while undergoing an extrinsic perturbation, and this aspect of ecological stability has important clinical applications. Lower alpha diversity associates with modern western lifestyles and the onset of many diseases, and is likely a key factor in loss of community stability through many of the ecological principles already discussed[103]. For instance, lower alpha diversity may be linked to community stability through increased invasion by pathogenic microbes due to a lack of competitive exclusion, or loss of functional diversity leading to metabolic networks with ineffective microbial cross-feeding. Ecological resilience is a microbiomes ability maintain a steady state, but ecological stability

can also define a microbiomes reestablishment after strong perturbations like antibiotics[150;169]. Individuals with higher diversity before antibiotics tend to have faster or more pronounced recovery of community composition, and the ability to recover can be accelerated by introducing diverse microbes and slowed by introducing single taxa probiotics[150;169;170]. Ecological resilience could be applied clinically to measure temporal microbiome reestablishment after antibiotics or major medical treatment, where alpha diversity trajectory could be a biomarker of proper recovery. Ecological stability can be measured as intrapersonal beta diversity variation over time, and sudden increases in beta diversity distances could be an early clinical biomarker of health changes.

Another way to think about microbiome variation is overlap across individuals, and **population stability** considers what aspects of community composition are shared by group of people. The core microbiome is a way to assess which individual taxa or functions are maintained across a study population, but population stability considers the maintenance of total community composition. One way to leverage population stability is understanding disease in case control studies, and how the ill-defined term '**dysbiosis**' manifests. In many ways, dysbiosis has been loosely used to describe any microbiome composition which is distinguishable from that of healthy control

individuals, including where a concerted shift in microbiome composition is shared by case individuals. While still hotly debated, one way population stability could define dysbiosis is as a change in microbiome variance between cases and controls. Put more simply, microbiomes of healthy individuals will be more similar (i.e. lower inter-personal beta diversity distance), while microbiomes in dysbiosis will each compositionally breakdown in different ways (i.e. higher inter-personal beta diversity distance)[171]. This sets dysbiosis apart from shared shifts in microbiome composition associated with some diseases, where microbiome variance may remain the same within case and control groups even if microbiomes are distinguishable. Both patterns emerge in a meta-analysis of 28 case-control microbiome studies, with some diseases and studies characterized by a general unstable loss of beneficial taxa, and others by a concerted acquisition of specific pathogens[103]. Coral ecologists have come up with a measure of population stability by comparing the relationship between taxa abundance and the ubiquity of those taxa, known as the **abundance-ubiquity test**[172]. This test is useful to distinguish outlying observations that either appear in high abundance but sporadically across individuals (unstable), or observations that appear consistently across individuals but at a lower than expected abundance (stable)[172]. Applying abundance-ubiquity community-wide as a measure of population

stability may help distinguish disease associations with an indiscriminate loss of stable healthy taxa, apart from disease associations with high abundance blooms of pathogenic microbes. Ultimately our understanding of community stability in the human gut is poor, at least partially because stability depends on so many underlying ecological forces. Other factors may be the limited number of quality datasets that look at human microbiomes temporally, and that model systems may be better suited for understanding stability on short timescales. Still, ecological stability and dynamics can be characterized by diversity measures, and tracking microbiome dynamics temporally may serve as a biomarker that underlying ecological forces are shifting their influence.

Conclusion

Microbiomes assemble and dynamically change throughout the lifetime of host metazoans, yet principles uniting broadly observed microbiome patterns across metazoans have been poorly translated into the clinic. The outlined framework focuses on building from fundamental ecological and evolutionary measures into principles that define patterns in lab and natural populations, and subsequently what such principles can tell us about human health and breakdown in disease. The following work addresses clinically translating ecological and

evolutionary factors affecting microbiomes through three disparate projects: 1) first exploring phylosymbiosis by developing a framework of ecological and evolutionary expectations observed across 24 lab reared animal species, 2) looking at ethnicity as a factor explaining gut microbiome ecology in the United States, and roles that inter-ethnic variation could play in personalized health disparity treatments, and 3) finally working to disentangle the many intrinsic and extrinsic influences on microbiome ecology through a multi-ethnic, dietary controlled microbiome clinical trial. By spanning controlled model organism studies to intervention based human clinical trials, this work provides an example of how ecological and evolutionary principles can be adapted to diverse applications.

CHAPTER II

Phylosymbiosis: Relationships and Functional Effects of Microbial Communities across Host Evolutionary History[1]

Author Contributions

This study was performed by Andrew Brooks (AB), Robert Brucker (RB), Kevin Kohl (KK), Edward van Opstal (EO), and Seth Bordenstein (SB). RB reared and obtained samples from the 24 animal species used throughout primary analyses. Primarily RB and somewhat AB extracted DNA, performed 16S amplification, and sequenced microbiome communities in the 24 main analyses. AB performed the non-functional analytical analyses throughout the paper. EO performed and analyzed *Nasonia* functional microbiome transplants, and KK performed *Peromyscus* functional microbiome transplants. SB was principle investigator and worked with all individuals to plan and develop analyses. Everyone helped write and edit the manuscript.

---

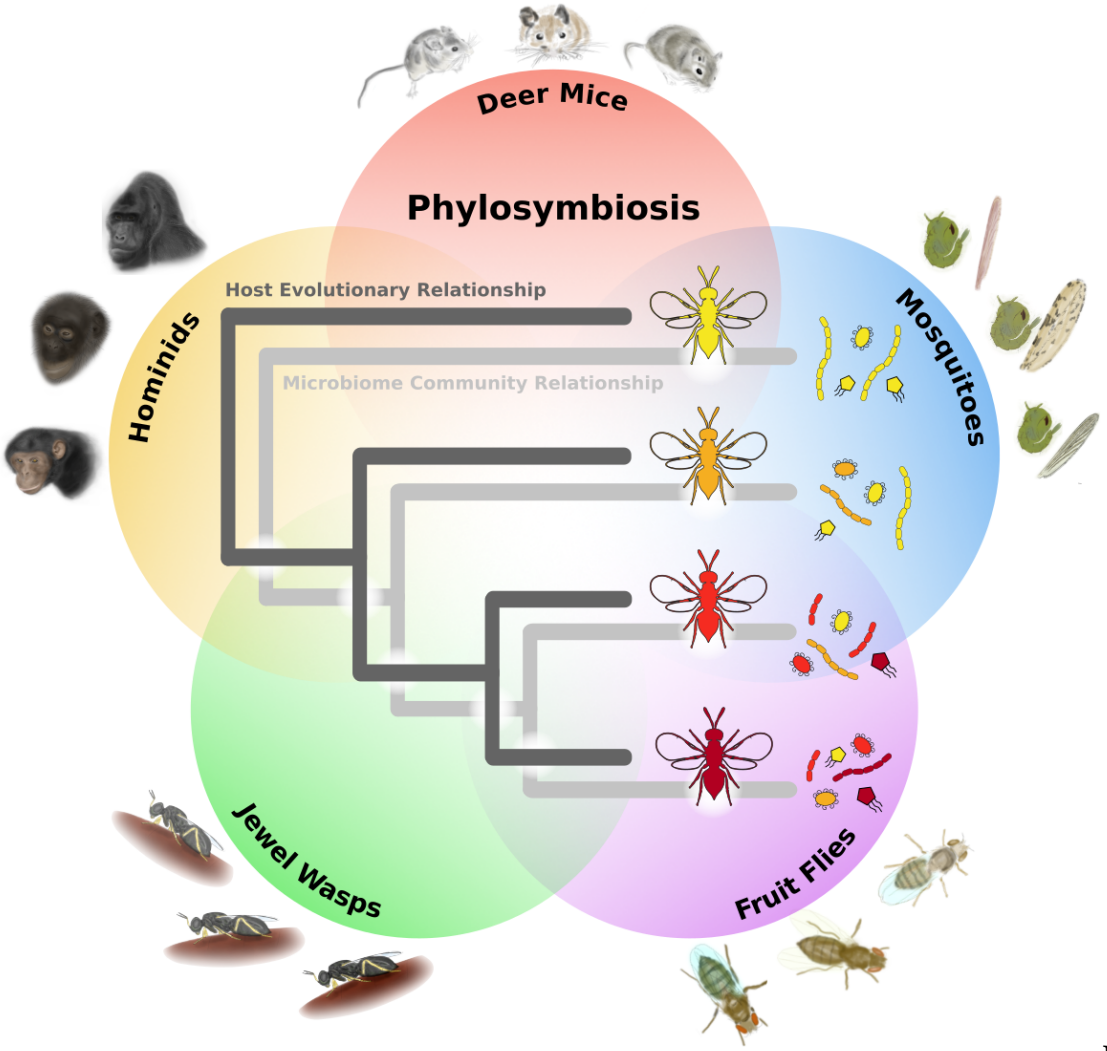
[1] This work is published in *PLOS Biology*: **Brooks AW***, Kohl KD*, Brucker RM*, van Opstal EJ, Bordenstein SR. (2016). Phylosymbiosis: Relationships and Functional Effects of Microbial Communities across Host Evolutionary History. *PLOS Biology*. http://dx.doi.org/10.1371/journal.pbio.2000225. (*Co-first Authors).

Introduction

Abstract

Phylosymbiosis was recently proposed to describe the eco-evolutionary pattern whereby the ecological relatedness of host-associated microbial communities parallels the phylogeny of related host species (Fig 2.1). Here, we test the prevalence of phylosymbiosis and its functional significance under highly controlled conditions by characterizing the microbiota of 24 animal species from four different groups (*Peromyscus* deer mice, *Drosophila* flies, mosquitoes, and *Nasonia* wasps), and we reevaluate the phylosymbiotic relationships of seven species of wild hominids. We demonstrate three key findings. First, intraspecific microbiota variation is consistently less than interspecific microbiota variation, and microbiota-based models predict host species origin with high accuracy across the dataset. Interestingly, the age of host clade divergence positively associates with the degree of microbial community distinguishability between species within the host clades, spanning recent host speciation events (~1 million y ago) to more distantly related host genera (~108 million y ago). Second, topological congruence analyses of each group's complete phylogeny and microbiota dendrogram reveal significant degrees of phylosymbiosis, irrespective of host clade age or taxonomy. Third, consistent with selection on host-

microbiota interactions driving phylosymbiosis, there are survival and performance reductions when interspecific microbiota transplants are conducted between closely related and divergent host species pairs. Overall, these findings indicate that the composition and functional effects of an animal's microbial community can be closely allied with host evolution, even across wide-ranging timescales and diverse animal systems reared under controlled conditions.



Fig

2.1. Graphical abstract for hypothesis of phylosymbiosis: microbiome community similarity will parallel host evolutionary relatedness.

Author Summary

Studies on the assembly and function of host-microbiota symbioses are inherently complicated by the diverse effects of diet, age, sex, host genetics, and endosymbionts. Central to unraveling one effect from the other is an experimental framework that reduces confounders. Using common rearing conditions across four animal groups (deer mice, flies, mosquitoes, and wasps) that span recent host speciation events to more distantly related host genera, this study tests whether microbial community assembly is generally random with respect to host relatedness or "phylosymbiotic," in which the phylogeny of the host group is congruent with ecological relationships of their microbial communities. Across all four animal groups and one external dataset of great apes, we apply several statistics for analyzing congruencies and demonstrate phylosymbiosis to varying degrees in each group. Moreover, consistent with selection on host-microbiota interactions driving phylosymbiosis, transplanting interspecific microbial communities in mice significantly decreased their ability to digest food. Similarly, wasps that received transplants of microbial communities

from different wasp species had lower survival than those given their own microbiota. Overall, this experimental and statistical framework shows how microbial community assembly and functionality across related species can be linked to animal evolution, health, and survival.

Introduction

A large body of literature has documented genetic and environmental influences on the composition of host-associated microbial communities[31;51;75;78;173-178]. Although environmental factors are considered to play a much larger role than host genetics and evolutionary history[10], host influences and their functional consequences are poorly elucidated and thus require systematic study across host-microbiota systems. Several outstanding questions remain regarding the nature of host effects on microbiota assembly. Are host-microbiota associations stochastically assembled, or might there be deterministic assembly mechanisms that predict these associations? How rapidly do microbiota differences form between closely related host species, and are interspecific microbiota differences prone to decay over evolutionary time? Can host-driven assembly of the microbiota be isolated from confounding variables such as diet, age, sex, and endosymbionts? If there are microbiota differences between species,

are they functional in an evolutionarily informed manner, such that mismatches between host and interspecific microbiota lead to reductions in fitness or performance, particularly when interspecific microbiota transplants are conducted between older host species pairs?

If host-associated microbial communities assemble stochastically through environmental acquisition with no host-specific influence, then microbiota compositions across related host species will not differ from expectations based on random community assemblies and dispersal limitations. Therefore, in a common environment, microbiota will form independent of host species (Fig 2.2A), and any interspecific differences in microbiota composition would be arbitrary. In contrast, if hosts influence a sufficient amount of the composition of the microbiota, then under controlled rearing conditions, intraspecific microbial communities will structure more similarly to each other than to interspecific microbial communities (Fig 2.2B). Similarly, if microbial communities are randomly established or are not distinguishable with regard to host evolutionary relationships, then dendrograms illustrating beta diversity distance relationships between microbial communities will not parallel the phylogeny of the host species (Fig 2.2C). However, if microbial communities are distinguishable, then hosts with greater genetic divergence may exhibit more distinguishable microbiota. In

this case, there will be congruence between the host phylogeny and microbiota dendrogram (Fig 2.2D). As this outcome is not likely due to coevolution, cospeciation, or cocladogenesis of the entire microbial community from a last common ancestor, "phylosymbiosis" was proposed as a new term that does not necessarily presume that members of the microbial community are constant, stable, or vertically transmitted from generation to generation[77;78]. Rather, phylosymbiosis refers to an eco-evolutionary pattern in which evolutionary changes in the host associate with ecological changes in the microbiota.

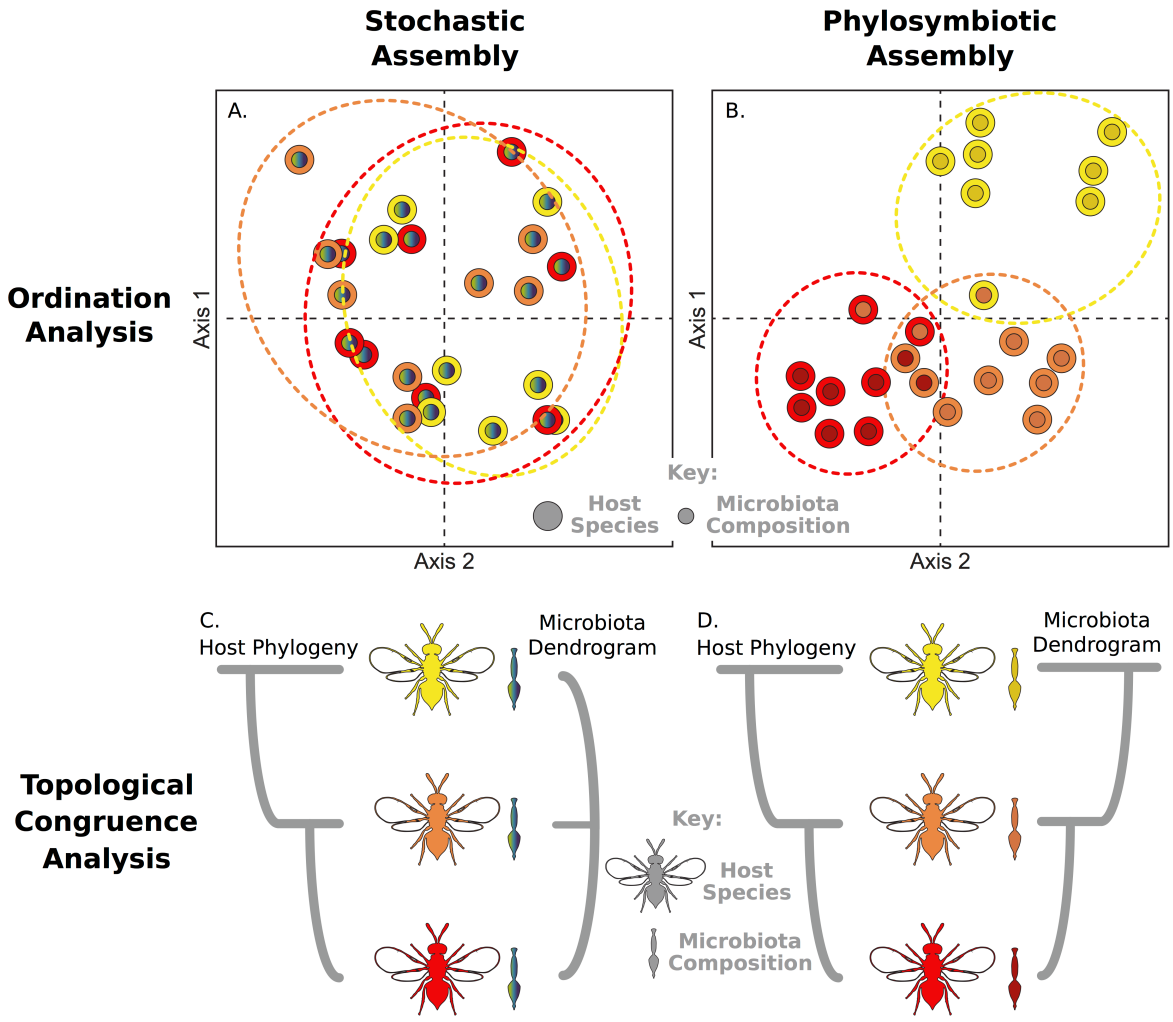


Fig 2.2. Analyses and predictions that can distinguish stochastic host-microbiota assembly from phylosymbiosis under controlled conditions. Two-dimensional ordination plots depict hypothetical microbiota similarity under (A) stochastic versus (B) phylosymbiotic models. Dashed lines represent host-specific clustering. Topological congruence analyses between host phylogeny (evolutionary relatedness) and microbial community dendrogram (ecological relatedness) depict the pattern expected for (C) stochastic versus (D)

phylosymbiotic host-microbiota assembly.

Phylosymbiosis leads to the explicit prediction that as host nuclear genetic differences increase over time, the differences in host-associated microbial communities will also increase. Indeed, phylosymbiosis has been observed in natural populations of sponges[179], ants[178], bats[180], and apes[50;93]. However, other studies on termites[181], flies[182-184], birds[185], and mice[186] have not observed strict patterns of phylosymbiosis or host-specific microbial signatures. In natural population studies, determining the forces driving phylosymbiosis is equivocal, as both environmental and host effects can covary and contribute to microbiota assembly. Importantly, major effects of the environment, age, or sex may overwhelm the ability to detect phylosymbiosis. Indeed, diet is a stronger determinant of whole microbial community structure than genotype in lab-bred mice[187]. Additionally, conjecture about the formation of host-specific communities should be resolved in a wider context, especially their functional significance, as microbiotas may be inconsequential to host biology or uniquely situated for certain host genotypes and fitness. Thus, the prevalence and functional significance of phylosymbiosis is uncertain and requires reductionist approaches to discriminate among the frequently confounded variables of host, environment,

development, sex, and even endosymbiont status.

Here, we quantify phylosymbiosis under laboratory conditions to control for environmental and host rearing variation. Prior investigations of phylosymbiosis have not typically controlled for these confounding variables, with the exception of male *Nasonia* wasps[51;78] and *Hydra*[76;175]. Specifically, we reared 24 species in the laboratory while controlling for sex (virgin females), age, diet, and endosymbionts, thus removing major environmental variables and isolating the contribution of host species on microbiota assembly. The experimental systems, or "host clades," span four species of *Nasonia* parasitic jewel wasps, six species of *Drosophila* fruit flies, eight species of *Anopheles*, *Aedes*, and *Culex* mosquitoes, and six species of *Peromyscus* deer mice. An externally derived dataset with seven members of the hominid lineage[50] provides another mammalian and multigenus clade for reference and facilitates examination of natural populations in which phylosymbiosis was previously documented. Together, the five host clades include 31 distinct taxa and span a range of estimated divergence times from 0.2-108 million y. Last, we test the hypothesis that phylosymbiosis represents a functional association through a series of microbial transplants with autochthonous (intraspecific) and allochthonous (interspecific) microbiota in *Nasonia* and *Peromyscus*. We expect that an experimentally mediated disruption of

phylosymbiosis will have functional costs that may lower host fitness or performance in an evolutionarily informed manner. Our findings demonstrate that a consistent set of controlled experimental and bioinformatic approaches in comparative microbiota studies can isolate host-driven phylosymbiosis.

Results

Host Clade Differentiates Microbial Communities

Phylosymbiosis predicts that host clades will harbor distinguishable microbial communities (e.g., jewel wasps versus fruit flies versus deer mice, etc.) and that more closely related host clades will exhibit more similar microbial communities (e.g., insects versus mammals). Indeed, at a broad scale, we found that host clades harbored relatively distinct microbial communities (Fig 2.3A, ANOSIM, R = 0.961, p < 1e-6). Furthermore, there was significant microbiota differentiation between the mammalian and invertebrate host clades in the principle coordinates analysis (PCoA) (Fig 2.3A, ANOSIM, R = 0.905, p < 1e-6). The PCoA shows insect groups separating along two dimensions of a plane, with the mammals distinguished orthogonally from that plane in a third dimension, suggesting that variance in insect microbial communities is fundamentally different than that in mammals. As is well established, the gut communities of

mammals were dominated by the bacterial classes Clostridia (Firmicutes) (Fig 2.3B, hominid 42%, *Peromyscus* 37%) and Bacteroidia (Bacteroidetes) (Fig 2.3B, hominid 15%, *Peromyscus* 37%), while the insect clades were dominated by Proteobacteria (Fig 2.3B, *Drosophila* 78%, mosquito 69%, *Nasonia* 77%). This same bacterial divide is also seen in the network analysis, with significant clustering of the insect microbial communities around Proteobacteria, and the mammal microbial communities around subsets of shared and unique Firmicutes and Bacteroidetes (G-test, $p < 1e\text{-}6$, Fig 2.3C). Microbial diversity as measured by the Shannon index[188] was approximately 35% higher in mammalian hosts compared to insects, indicating more diverse symbiont communities among the mammalian clades (Fig 2.3D; Nested analysis of variance [ANOVA]: phylum effect [mammals versus insects]: $F_{1,302} = 419.82$, $p < 0.001$; clade effect nested within phylum: $F_{3,298} = 18.46$, $p < 0.001$; species effect nested within clade and phylum: $F_{26,272} = 7.94$, $p < 0.001$).

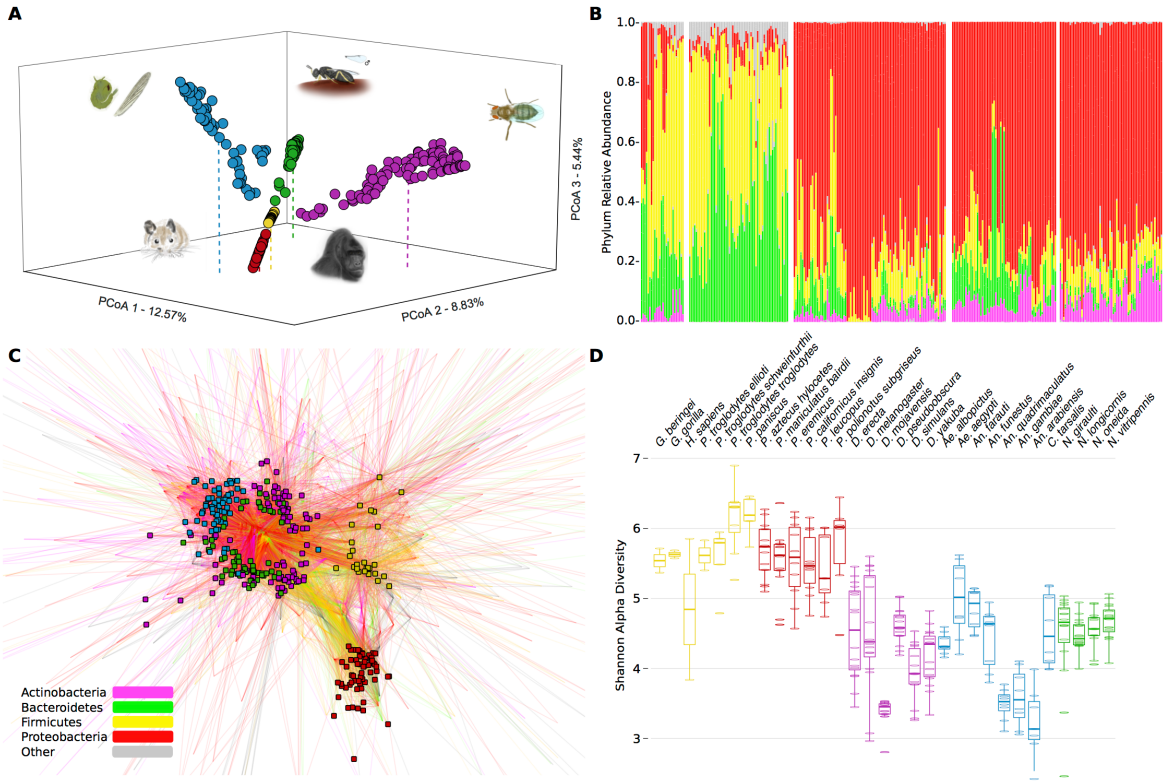


Fig 2.3. Meta-analysis of microbiota variation across five host clades. (A) PCoA analysis of Bray-Curtis ecological similarity in three dimensions based on 99% operational taxonomic unit (OTU) cutoff, with colors depicting clade of origin. (B) Phylum level relative abundance for all samples, with a key provided in C. (C) Network analysis in which small squares depict samples, with their color indicating clade of origin. Lines connect genus-level OTUs to samples and are weighted by occurrence and colored by OTU phylum. (D) Shannon alpha diversity for each host species. Small ellipses depict individual samples, and dark lines indicate the species' median diversity. The lower and upper end of each box represent the 25th and 75th quartiles, respectively. Whiskers denote the 1.5

interquartile range.

We implemented a random forest classifier (RFC) supervised learning algorithm to quantify the degree to which individual microbial communities can be classified into their respective host clade. RFC models show a strong ability to classify microbial communities to their correct host clades based on OTUs (98.5% classification accuracy) (S2.1 Table). Additionally, models distinguish mammals and insect samples with high accuracy (95.9% classification accuracy) (S2.1 Table). Cross-validation prevents overfitting by ensuring that classification accuracy is assessed using only samples excluded from model training. We also used RFC models to identify the most distinguishing bacterial taxonomic level for both interclade distinction and the divide between mammals and insects. Genera provided the strongest ability to predict host clade (99.0% classification accuracy) (S2.1 Table); however, the major groups of insects and mammals were better distinguished by family-level community classification (98.3% classification accuracy) (S2.1 Table). Taken together, these results illustrate that evolutionary relationships of the host clades broadly covary with differences in microbial communities. While differentiation of the five clades could in part be attributable to varied experimental conditions for each animal group (since they were reared

separately), clustering of the vertebrate microbial communities from the insect microbial communities is independent of rearing conditions and suggests a host-assisted structuring of microbial communities.

Intraspecific Microbial Communities Are Distinguishable within Host Clades

Phylosymbiosis predicts that an individual's microbial community will exhibit higher similarity to communities of the same host species than to those from different host species. The degree of similarity can be variable but should correlate with genetic relatedness of the host species. Pairwise comparisons of beta diversity distances between all individuals within each host clade reveal that the average distance between microbial communities within a species is always less than between species (S2.1 Fig). Summarized beta diversity also reveal lower intraspecific versus interspecific distances, with significant differences observed for all clades (Fig 2.4A, Each dataset: Mann-Whitney U, $p < 1e\text{-}6$).

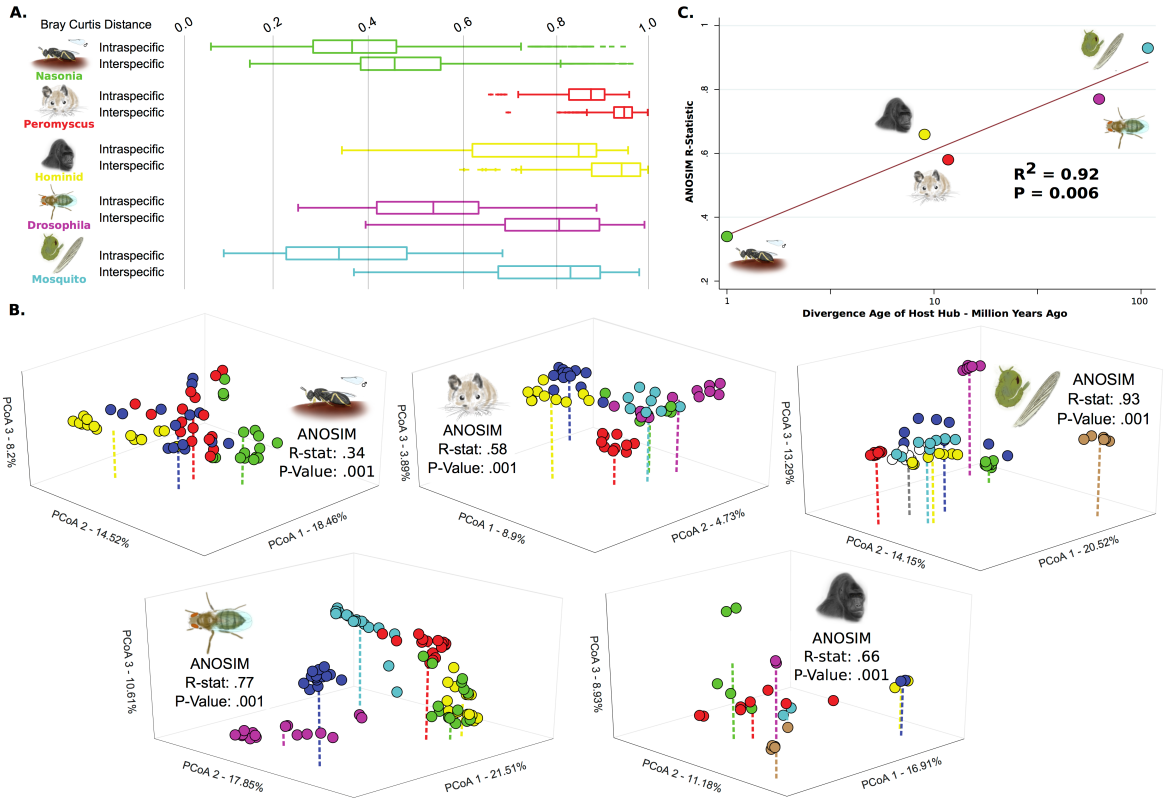


Fig 2.4. Intraspecific versus interspecific microbial community variation within and between host clades. (A) Box-and-whisker plot of intraspecific and interspecific Bray-Curtis distances between samples for each clade. Boxes represent the 25th to 75th quartiles, with the central line depicting the group median and whiskers showing the 1.5 interquartile extent. (B) PCoA of Bray-Curtis distances with first three most distinguishing dimensions shown. Colors represent different species and correspond to the colors in Fig 2.5. (C) Regression analysis measuring the correlation between the evolutionary age of host clade divergence on a log scale and the ANOSIM R-values of intraspecific microbiota distinguishability from part B for each host clade.

We next evaluated intraspecific microbiota clustering through Bray-Curtis beta diversity interrelationships with PCoA and statistically assessed the strength of interspecific microbiota distinguishability with ANOSIM (Fig 2.4B). Visualization of the first three principle components revealed that individual samples clustered around their respective species' centroid position. In all host clades, each host species harbored significantly distinguishable microbial communities (Fig 2.4B, ANOSIM $p < 0.001$ for all host clades). Notably, the ANOSIM R-values of interspecific microbiota distinguishability within a host clade positively correlated with the maximal age of divergence of the species in the host clades (Fig 2.4C, Regression Analysis Log Transformed Clade Age, $R^2 = 0.92$, $p = 0.006$; Untransformed Clade Age, $R^2 = 0.70$, $p = 0.048$). Thus, host clades with higher total divergence times between species had stronger degrees of microbiota distinguishability, while less diverged host clades exhibited less microbiota distinguishability. For example, with an estimated host divergence time of 108 million y[189], mosquitoes showed the greatest distinguishability of their microbiota. Conversely, in *Nasonia* jewel wasps, which only diverged between 200,000 and 1 million y ago[190], the relative strength of clustering was less distinct but still statistically significant. The three intermediate aged clades

showed corresponding intermediate levels of clustering: *Drosophila* had an estimated divergence time of 62.9 million y[191], hominids diverged 9 million y ago[192], and *Peromyscus* diverged 11.7 million y ago[193]. Therefore, the phylosymbiotic prediction that host species will exhibit significant degrees of specific microbiota assembly was supported in these observations, even under highly controlled conditions in the laboratory models. Microbiota specificity was maintained among very closely related and very divergent species, and a connection was observed between the magnitude of host genetic divergence and microbiota similarity.

Supervised Classification: Microbiota Composition Predicts Host Species

As microbiota clustering was supported within species across all five animal clades, it should be possible to model the strength of how well communities of bacteria predict their host species and how specific members of the microbiota affect these predictions. We therefore used RFC models trained on the microbiota of each host clade to evaluate classification accuracy (i.e., the percentage of assigning microbiota to their correct host species) and the expected predicted error (EPE, i.e., the ratio of model accuracy relative to random classification). RFC results indicated that the operational taxonomic units

(OTUs) for *Drosophila* and *Peromyscus* and genus taxonomic levels for hominid, mosquito and *Nasonia* have the highest classification accuracies, with significant EPE observed for all clades (EPE > 2, S2.1 Table). At the genus level, the mosquito and *Drosophila* host clades exhibited the strongest results (mosquito, classification accuracy = 99.8%, EPE = 558.9; *Drosophila*, classification accuracy = 97.2%, EPE = 31.7). Other host clades demonstrated significant but comparatively lower strength models. The reduced predictive power of these models may be due to a number of factors, such as a lower number of host species (*Nasonia*, classification accuracy = 88.7%, EPE = 13.4), uneven sample representation from each species (hominid, classification accuracy = 53.4%, EPE = 2.1), and lower sequencing coverage (*Peromyscus*, classification accuracy = 61.4%, EPE = 2.5).

To determine the most distinguishing genera of the bacterial community, we examined the resulting loss of model classification accuracy when each genus was excluded from RFCs (S2.2 Table). Distinguishability within the *Drosophila*, *Nasonia*, and mosquito clades was driven primarily by genera in Proteobacteria, which represent five (14.0% model accuracy), seven (11.3% model accuracy), and eight (18.2% model accuracy) of the top ten genera, respectively. Three of the ten most distinguishing genera in *Drosophila* females are from the Acetobacteraceae

family (9.5% model accuracy), previously recognized to be "core" microbiota members[194;195]. Three of the twenty most distinguishing genera in *Nasonia* females were closely related symbionts from the Enterobacteriaceae family (genera: *Proteus, Providencia, Morganella*; 3.1% model accuracy), consistently found in our previous studies of *Nasonia* males[51;78]. Eight genera from the phylum Proteobacteria dominate mosquito female distinguishability, primarily three Gammaproteobacteria of the order Pseudomonadales (8.2% model accuracy), and three Betaproteobacteria of the family Comamonadaceae (5.9% model accuracy). Hominid interspecific distinguishability was driven by the phylum Firmicutes, particularly of the order *Clostridiales* that contains three of the most distinguishing genera (1.5% model accuracy). The genus *Allobaculum* conferred nearly double the distinguishing power of any other bacteria in *Peromyscus* (3.8% model accuracy), and it is associated with low-fat diet, obesity, and insulin resistance in mice[196]. As may be expected, genera of the abundant phyla Firmicutes and Bacteroidetes dominated the majority of distinguishability in *Peromyscus* (10.6% model accuracy), but genera from Proteobacteria in the family Helicobacteraceae comprised four of the top eleven genera (4.4% model accuracy). Overall, microbiota composition can be used to predict host species with high accuracy, and genera commonly observed in other studies of these host clades underlie

interspecific distinguishability.

Phylosymbiosis Is Common within Host Clades

The major prediction of phylosymbiosis is that phylogenetic relatedness will correlate with beta diversity relationships of microbial communities among related host species. Microbiota dendrograms were constructed by collapsing individual samples to generate an aggregate microbial community for each species and then by comparing relationships of their beta diversity metrics. The matching cluster and Robinson-Foulds tree metrics were utilized to calculate host phylogenetic and microbiota dendrogram topological similarity, with normalized distances ranging from 0.0 (complete congruence) to 1.0 (complete incongruence)[197]. Matching cluster weights topological congruency of trees, similar to the widely used Robinson-Foulds metric[197;198]. However, matching cluster takes into account sections of subtree congruence and therefore is a more refined evaluation of small topological changes that affect incongruence. Significance of the matching cluster and Robinson-Foulds analyses was determined by the probability of randomized bifurcating dendrogram topologies yielding equivalent or more congruent phylosymbiotic patterns than the microbiota dendrogram. Additionally, using the same methodology, matching

cluster and Robinson-Foulds metrics were evaluated for Bray-Curtis, unweighted UniFrac[56], and weighted UniFrac[56] beta diversity dendrograms at both 99% and 97% clustered OTUs (S2.2 Fig). The cytochrome oxidase I (COI) gene was used to construct the phylogeny for each host clade, which compared well to established phylogenetic or phylogenomic trees for all species included in the study (*Nasonia*[190]; *Drosophila*[191]; hominids[192]; mosquitoes[189]). *Peromyscus* was further resolved with an additional marker (arginine vasopressin receptor 1A [AVPR1A]) to reflect the latest phylo- genetic estimates[199;200].

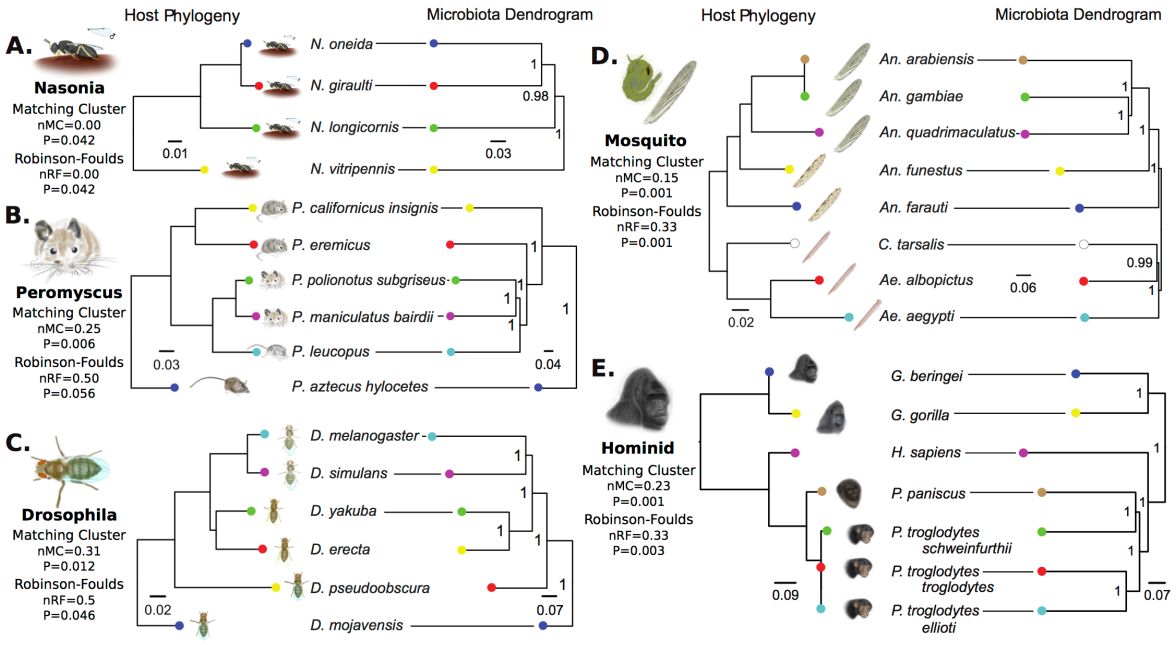


Fig 2.5. Phylosymbiosis between host phylogeny and microbiota dendrogram relationships. Topological congruencies are quantified by the normalized Robinson-Foulds (RF) metric, which takes into account symmetry in

rooted tree shape on a scale from 0 (complete congruence) to 1 (incomplete incongruence). The normalized matching cluster (MC) metric is a refined version of the RF metric that sensitively accounts for incongruences between closely related branches. Horizontal lines connect species whose position is concordant between host phylogeny and microbiota dendrogram based on 99% OTU cutoffs, therefore requiring no topological shift to demonstrate phylosymbiosis.

Nasonia female wasps exhibited an equivalent phylogenetic tree and microbial community dendrogram, representing exact phylosymbiosis (*Nasonia* wasps, Fig 2.5A). These results parallel previous findings in *Nasonia* males[51;78]. Despite congruency, the *Nasonia* clade has limited topological complexity with only four species, therefore resulting in a relatively marginal significance. Mice also show nearly perfect congruence, with the exception of *Peromyscus eremicus* (Fig2.5B). *Drosophila* fruit flies (Fig 2.5C) showed the lowest topological congruency but were still moderately significant. Four of the six species show correct topological relationships, while the microbial community relationships of *Drosophila pseudoobscura* and *D. erecta* are topologically swapped. These results are different from previous findings in *Drosophila* that utilized a different experimental design, set of taxa, and sequencing technology[183]. However, the evidence for

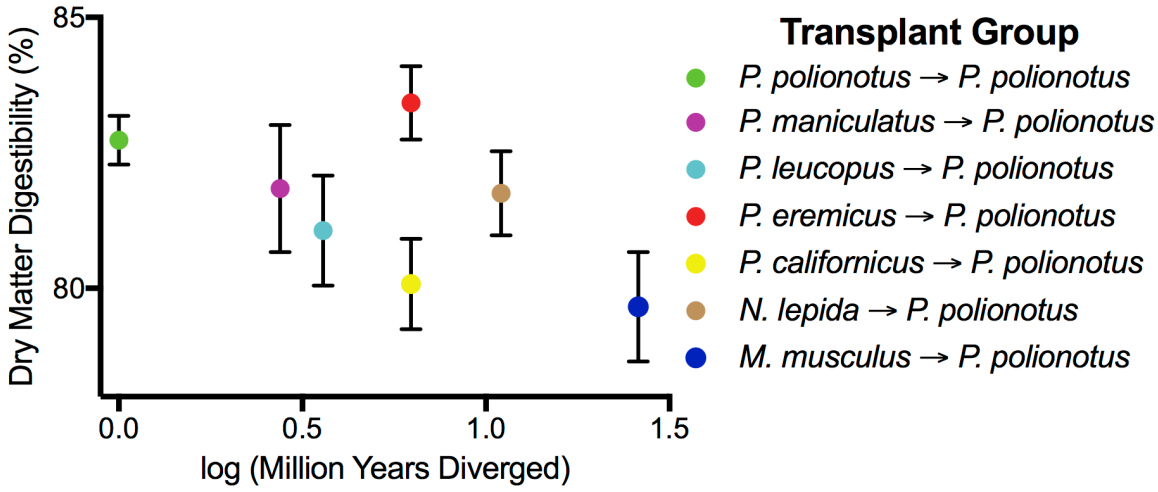
phylosymbiosis is tentative in *Drosophila* as, unlike other clades, there is no significant congruence for either unweighted or weighted UniFrac metrics (S2.2 Fig). Previous studies detected no pattern of phylosymbiosis across *Drosophila* species[183], which could be attributed to *Drosophila's* constant replenishment of microbes from the environment[182;184] or the dominance by the bacterial genus *Acetobacter*, which is important for proper immune and metabolic development[183]. The two additional clades, mosquitoes and hominids, showed significant phylosymbiosis (Fig 2.5D and 2.5E). Specifically, the mosquitoes showed accurate separation of *Culex* and *Aedes* genera from *Anopheles*, and the topological departures from phylosymbiosis appeared in two of the bifurcations between closely related species. The hominid microbial community dendrogram reflects the correct branching of *Gorilla* from *Homo sapiens*, followed by bonobos and chimpanzees, with the exception that one of the chimpanzee subspecies grouped more closely with the bonobo lineage. These results are similar to previous observations that the relationships of the microbial communities parallel those in the host phylogeny[50]. With the exception of *Drosophila*, which yielded variable evidence for host-microbiota congruence, significant degrees of phylosymbiosis were observed across clades with varying tree similarity metrics and microbiota beta diversity analyses.

Phylosymbiosis Represents a Functional Association

Microbiota-host distinguishability and topological congruence does not strictly imply that the phylosymbiotic associations are fitness directed, though it naturally follows that a particular host species may be more ideally suited for an autochthonous versus allochthonous microbiota. We therefore performed a series of microbial transplants to test the prediction that inoculated microbiota from a different species would decrease aspects of host performance or fitness in contrast to inoculated microbiota from the same species. Moreover, if there is selection on host-microbiota interactions such that microbiotas are uniquely or better situated for resident host backgrounds, then transplanted microbiota from a divergent species could drive more pronounced reductions in host functions than transplanted microbiota from a closely related species.

In *Peromyscus,* we followed a previously established protocol[201] to transplant the microbial communities from six rodent donor species into a single recipient species, *P. polionotus*, as well as a control group in which the microbial communities from *P. polionotus* were introduced to intraspecific individuals of *P. polionotus*. Inventories of fecal microbiota from donor and recipient mice revealed that portions of the donor microbiota successfully transferred. The estimated

amount of transplanted OTUs and their relative abundance ranged from 6.5%-26.2% and 11.4%-40.7%, respectively, when analyzed at the 99% OTU cutoff level. Variation in the transfer of foreign microbes was dependent on donor species and its divergence from the recipient species (S2.3 Fig). We then measured dry matter digestibility, or the proportion of food material that is digested by the animal. Consistent with selection on host-microbiota interactions, mice that were inoculated with microbial communities from more distantly related hosts exhibited decreased dry matter digestibility (Fig 2.6). These results were only significant when the group receiving feces from *P. eremicus* donors was removed (Fig 2.6). Notably, the microbiota of *P. eremicus* is not congruent with our predictions of phylosymbiosis (Fig 2.5). Thus, only the taxa showing phylosymbiosis exhibited the functional trend with digestibility. Distantly related donor species (*Neotoma lepida* and *Mus musculus*) did not drive significance, as the correlation remained statistically significant when investigating only *Peromyscus* donors (excluding *P. eremicus*; Fig 2.6).



	R^2	P-value
All data	0.32	0.19
Remove *P. eremicus* → *P. polionotus*	0.70	**0.036**
Peromyscus donors only	0.09	0.61
Peromyscus, excluding *P. eremicus* → *P. polionotus*	0.94	**0.029**

Fig 2.6. Effects of allochthonous and autochthonous microbial communities on the digestive performance of recipient mice. Dry matter digestibility is calculated as (g dry food ingested-g dry feces produced) / g dry food ingested. Divergence times between *P. polionotus* and donor species were determined from previously published phylogenies[199;200]. Points represent mean values ± standard error for each group (n = 5-6 recipients per group).

In the most extreme cases in which mice were inoculated with the microbial communities from *P. californicus* or *M. musculus*, there was approximately a 3% decrease in dry matter digestibility, which is on par with the decrease in

digestibility observed as a result of helminth infections in *Peromyscus*[202]. Animals must consume more food to meet energy demands when faced with decreases in digestibility. Indeed, mice inoculated with microbial communities from *P. californicus* or *M. musculus* exhibited significantly higher food intakes than the control group (S2.4 Fig; Tukey's honest significant difference (HSD) test: $p = 0.001$ for *P. californicus* to *P. polionotus*; $p = 0.044$ for *M. musculus* to *P. polionotus*). The mice inoculated with the microbes from *P. eremicus* performed just as well, if not better, than the control groups in terms of dry matter digestibility (Fig 2.6) but still had slightly higher food intakes (S2.4 Fig).

In *Nasonia*, we used an *in vitro* rearing system to transplant heat-killed microbial communities from three *Nasonia* donor species into larvae of *N. vitripennis* or *N. giraulti*[203]. We then measured the survival of the recipients from first instar larva to adulthood. In both *N. vitripennis* and *N. giraulti* hosts, interspecific microbiota transplantations exhibited significant decreases in survival to adulthood when compared to intraspecific microbial transplantations (Fig 2.7). Specifically, *N. giraulti* with a *N. vitripennis* microbiota yielded a 24.5% average survival decrease in comparison to a *N. giraulti* microbiota (Fig 2.7A, Mann-Whitney U, $p = 0.037$). Interestingly, *N. giraulti* with a microbiota from the more closely related *N. longicornis* exhibited a similar but nonsignificant survival

reduction (23.7%, Fig 2.7A, Mann-Whitney U, $p = 0.086$). *N. vitripennis* with a *N. girault*i or *N. longicornis* microbiota exhibited a 42.6% (Fig 2.7B, Mann-Whitney U, $p < 0.0001$) and 23.3% (Fig 2.7B, Mann-Whitney U, $p = 0.003$) average survival decrease in comparison to a *N. vitripennis* microbiota, respectively (Fig 2.7A, Mann-Whitney U, $p < 0.0001$). Comparisons were also made between noninoculated hosts and those inoculated with interspecific backgrounds (*N. girault*i background: *N. vitripennis* inoculum $p = 0.07$, *N. longicornis* inoculum $p = 0.26$; *N. vitripennis* background: *N. girault*i inoculum $p = 0.001$, *N. longicornis* inoculum $p = 0.15$).

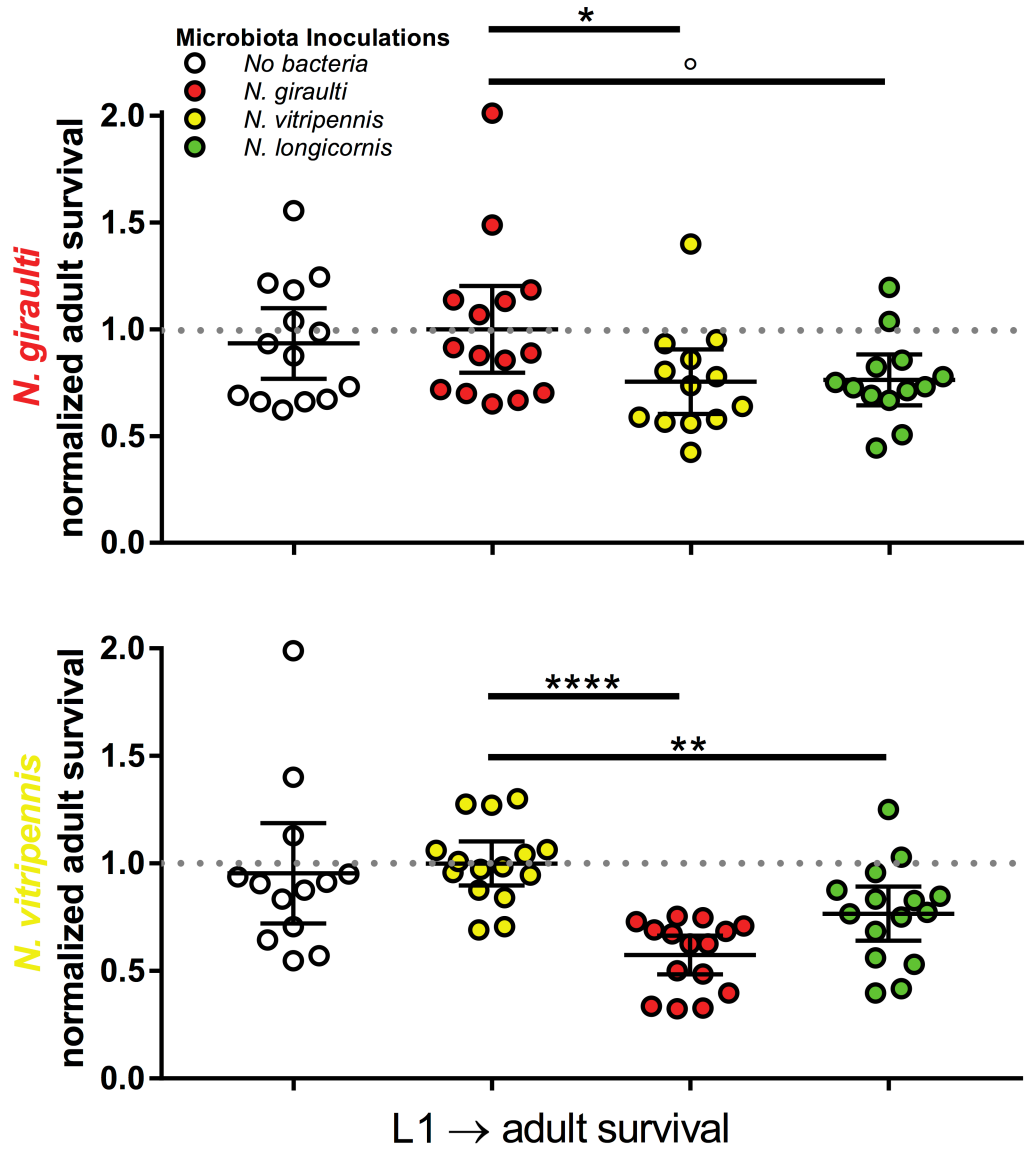


Fig 2.7. Effects of allochthonous and autochthonous microbial communities on the survival of *Nasonia* wasps. (A) Normalized larval-to-adult survival of *N. girault* wasps harboring no, self, or foreign microbiota. (B) Normalized larval-to-adult survival of *N. vitripennis* wasps harboring no, self, or foreign microbiota. Adult survival is calculated as number of adults in a transwell / number of first instar larvae in a transwell. Adult survival was normalized to the average survival

of the autochthonous microbiota transplantation. Circles represent individual transwell samples, and the dashed line represents the average survival of the autochthonous microbiota transplantation normalized to 1; error bars represent 95% confidence intervals. Mann-Whitney U statistics, $p < 0.1$, *$p < 0.05$, ** $p < 0.01$, and **** $p < 0.0001$.

Discussion

Under phylosymbiosis, host-associated microbial communities form, in part, as a result of interactions with the host rather than through purely stochastic processes associated with the environment. Specifically, we predicted that given closely related animals reared in controlled environments, the relationships of the microbiota would be congruent with the evolutionary relationships of the host species. Previous evidence for phylosymbiosis under controlled regimes existed in *Nasonia*[51;78] and *Hydra*[76], and wild populations of sponges[179], ants[178], and apes[50;93] also exhibited this pattern. Here, in a comprehensive analysis of phylosymbiosis in a diverse range of model systems, we report the widespread occurrence of this pattern under strictly controlled conditions as well as a functional basis in the context of host digestive performance in mice and survival in wasps. These results represent the first

evidence for phylosymbiosis in *Peromyscus* deer mice, *Drosophila* flies, a variety of mosquito species spanning three genera, and *Nasonia* wasp females with the inclusion of *N. oneida*. Previous studies in *Nasonia* measured male phylosymbiosis and did not include *N. oneida*[51;78]. By rearing closely related species from the same host clade in a common environment, and by controlling age, developmental stage, endosymbiont status, and sex, the experiments rule out confounding variables that can influence microbiota relationships in comparative analyses. Eliminating these variables is important because they often substantially correlate with inter-specific differences. Thus, our findings demonstrate that a uniform experimental and bioinformatic methodology can excavate host effects on phylosymbiosis from other potentially confounding variables in comparative microbiota studies.

We observed marked differences in microbial diversity and community structure between mammalian and invertebrate host clades. Mammalian communities were more diverse and dominated by Bacteroidetes and Firmicutes, while insect-associated communities were less diverse and primarily dominated by Proteobacteria. These results are consistent with previous microbial inventories conducted in mammals and insects[75;204]. Together, these findings suggest large-scale differences in the host-microbiota interactions between

mammals and insects. These differences across host phyla could be due to a variety of possibilities, including host genetics, diet, age, and rearing environment. To remove confounding variables that structure host-microbiota assemblages and to rigorously test phylosymbiosis, we utilized an experimental design within four host clades that isolated the effects of host evolutionary relationships from other effects (i.e., diet, age, rearing environment, sex, endosymbionts). We found that host species consistently harbored distinguishable microbiota within each host clade. Additionally, we found significant degrees of congruence between the evolutionary relationships of host species and ecological similarities in their microbial communities, which is consistent with the main hypothesis of phylosymbiosis. These results importantly expand previous evidence for this eco-evolutionary pattern and demonstrate that related hosts reared under identical conditions harbor distinguishable microbial assemblages that can be likened to microbial community markers of host evolutionary relationships. It is conceivable that recently diverged species (i.e., those younger than several hundred thousand years) would have less genetic variation and fewer differences in microbiota composition. Furthermore, divergent hosts may have vast differences in physiology that overwhelm the likelihood of observing phylosymbiosis. Surprisingly, we observed phylosymbiosis

to varying degrees in all host clades, and the age of clade divergence positively correlates with the level of intraspecific microbiota distinguishability. Thus, as host species diverge over time, microbial communities become more distinct[77;78], and the limits of detecting phylosymbiosis may occur at extreme scales of incipient or ancient host divergence times.

The mechanisms by which phylosymbiosis is established requires systematic investigation. Perhaps the most apparent regulator of host-microbiota interactions is the host immune system. A previous study of phylosymbiosis in *Hydra* demonstrated that antimicrobial peptides of the innate immune system are strong dictators of community composition, and expression of antimicrobial peptides are necessary for the formation of host-specific microbiota[49;175]. Furthermore, genome-wide association studies in humans[9], mice[176], and *Drosophila*[205] have identified a large immune effect in which host immune genes can explain variation in microbial community structure. Interestingly, host immune genes often exhibit rapid evolution and positive selection compared to genes with other functions[206;207]. While this trend is often explained by the host-pathogen arms race[206], it is also likely due to host evolutionary responses for recruiting and tending a much larger collection of nonpathogenic microbes.

Other host pathways may also underlie the observed species-specific

microbiota signatures. Hosts produce glycans and mucins on the gut lining that may serve as biomolecular regulators of microbial communities[208;209]. For example, knocking out the gene for α1-2 fucosyltransferase inhibits production of fucosylated host glycans on the gut surface and significantly alters microbial community structure[210]. Additional knockout studies have demonstrated the roles of circadian clock genes[211], microRNAs[212], and digestive enzymes[213] in determining microbial community structure. These various physiological systems might also interact with one another and may have even evolved in tandem to regulate microbial community structure.

Alternatively, rather than hosts "controlling" their microbiota, microbes may be active in selecting which host niches to colonize. For example, hosts have been compared to ecological islands, where environmental selection of the microbiota through niche availability may occur[214]. However, given the large number of studies that demonstrate the role of microbes in improving host performance[215], we find it unlikely that hosts would assume a solely passive role in these interactions. An elegant study allowed microbial communities from various environments (soil, termite gut, human gut, mouse gut, etc.) to compete within the mouse gut[216]. This study found that a foreign community of the human gut microbiota exhibited an early competitive advantage and colonized the mouse

gut first. Later, the mouse gut microbiota dominated and outcompeted the human gut microbiota[216]. Thus, community assembly is not a monolithic process of host control but likely a pluralistic combination of host control, microbial control, and microbe-microbe competition. In this context, both population genetic heritability and community heritability measurements of the microbiota will be useful in prescribing the varied genetic influences of a foundational host species on microbiota assembly[217].

The acquisition route of microbes could also influence our understanding of phylosymbiosis. If phylosymbiosis is observed when the microbiota is acquired horizontally from other hosts, the environment, or some combination of the two, then phylosymbiosis is presumably influenced by host-encoded traits such as control of or susceptibility to microbes. However, maternal transmission of microbes is argued to be a common trend in animals[218]. For example, sponges exhibit vertical transmission of a diverse set of microbes in embryos[219]. Transmission of full microbial communities is unlikely in most systems, given that the communities of developing animals tend to exhibit markedly lower diversity and distinct community structure compared to adults[8;78;220]. Thus, it is improbable that phylosymbiotic relationships are explained simply by community drift over host evolutionary divergence. There could be a subset of microbial taxa that are more

likely to be transmitted from mother to offspring that in turn affect what other microbes colonize. For instance, in humans, the family Christensenellaceae is situated as a hub in a co-occurrence network containing several other gut microbes and has a significant population genetic heritability[91]. When *Christensenella minuta* was introduced into the guts of humanized mice, the microbial community structure was significantly altered[91]. This microbe, as well as others, can therefore be likened to a keystone taxa or "microbial hub" that can impact community structure despite low abundance[91;221-223]. Thus, one could hypothesize that phylosymbiotic relationships in some systems may be driven by host transmission of microbial hubs that determine whole community structure through ensuing microbe-microbe interactions. However, further work is needed to test this hypothesis.

The congruent relationships between hosts and associated microbial communities are likely maintained through their positive effects on host performance and fitness but could be neutral or harmful as well. While the importance and specificity of hosts and microbes in bipartite associations has been demonstrated on host performance[224], it is unclear whether such effects commonly occur for hosts and their complex microbial communities. If they exist, disruption of phylosymbiosis via hybridization or microbiota transplants should

lead to reduced fitness or performance. For instance, hybridization experiments demonstrate negative interactions or "hybrid breakdown" between host genetics and the gut microbiota that drives intestinal pathology in house mice[225] and severe larval lethality between *N. vitripennis* and *N. giraulti* wasps[51]. Furthermore, transplant experiments show that all microbes are not equal for the host. An early study demonstrated that germ-free rabbits inoculated with a mouse gut microbiota exhibited impaired gastrointestinal function compared to those given a normal rabbit microbiota[226]. Together, these functional studies and others suggest that interactions between hosts and their microbiota are not random and instead occur at various functional levels.

Here, we add an evolutionary component to these ideas by demonstrating that microbial communities from more evolutionarily distant hosts can be prone to more pronounced reductions in host performance or fitness. Specifically, *Peromyscus* deer mice inoculated with microbial communities from more distantly related species tended to exhibit lower food digestibility. The exception to this trend was the *P. eremicus* to *P. polionotus* group, which did not exhibit any decrease in digestibility. It should be noted that *P. eremicus* also did not follow phylosymbiosis (Fig 2.5B), which may explain the departure from our expected trend in digestibility. For example, deviations from phylosymbiosis could be due

to a microbial community assembly that is inconsequential to host digestibility. Therefore, transferring a nonphylosymbiotic community between host species may not yield performance costs.

An alternative explanation for our results could be that hosts are acclimated to their established microbiota, and the introduction of foreign microbiota either elicits a host immune response or disrupts the established microbiota, thus decreasing digestibility. One technique to distinguish between adaptation and acclimation would be to conduct experiments in germ- free *P. polionotus* recipients. However, the derivation of germ-free mammals is a difficult and expensive process[227] and has not been conducted for *Peromyscus*. Earlier studies utilizing germ-free mammals demonstrate that microbial communities from evolutionarily distant hosts negatively impact gastrointestinal function[226] and immune development[228], thus supporting our hypothesis of functional matching between host and the gut microbiota.

Additionally, among very closely related species, *Nasonia* exposed to interspecific micro- biota have lower fitness than those exposed to intraspecific microbiota. While this experiment utilized heat-killed bacteria to avoid shifts in the microbiota composition during media growth, the protocol is sufficient to test the predictions of phylosymbiosis. First, isolated microbial products can exert

drastic effects on eukaryotic partners. For example, a sulfonolipid purified from bacteria can induce multicellularity in choanoflagellates[229]. Additionally, the insect immune system can respond with strain-level specificity to heat-killed bacteria[230]. Therefore, we hypothesize that each *Nasonia* host species evolved to the products of their own gut microbiota rather than those of gut microbiota from related host species. Together, results from the *Peromyscus* and *Nasonia* functional experiments reveal the importance of host evolutionary relationships when considering interactions between hosts and their gut microbial communities and ultimately the symbiotic processes that can drive adaptation and speciation[231;232]. The molecular mechanisms underlying the functional bases of phylosymbiosis in various systems demand further studies

Overall, we have established phylosymbiosis as a common, though not universal, phenomenon under controlled rearing with functional effects on host performance and survival. It is worth emphasizing again that this term is explicit and different from many other similar terms, such as coevolution, cospeciation, cocladogenesis, or codiversification[233]. While cospeciation of hosts and specific environmentally or socially acquired microbes-e.g., hominids and gut bacterial species[113] or the bobtail squid and *Vibrio* luminescent bacteria[112] could contribute in part to phylosymbiosis, concordant community structuring with the host

phylogeny is not dependent on parallel gene phylogenies but instead on total microbiota compositional divergence. Phylosymbiosis does not assume congruent splitting from an ancestral species because it does not presume that microbial communities are stable or even vertically transmitted from generation to generation[234;235]. Rather, phylosymbiosis predicts that the congruent relationships of host evolution and microbial community similarities could have varied assembly mechanisms in space and time and be newly assembled each generation (though see our discussion of transmission routes above). Moreover, the findings here imply that across wide-ranging evolutionary timescales and animal systems, there is a functional eco-evolutionary basis for phylosymbiosis, at least under controlled conditions.

It may be difficult to detect phylosymbiosis in natural populations because of extensive environmental variation that overwhelms the signal. We suggest that one way to potentially overcome this challenge is to start with laboratory-controlled studies that identify (i) phylosymbiotic communities and (ii) the discriminating microbial taxa between host species. Resultantly, investigations can test whether these microbial signatures exist in natural populations, albeit perhaps in a smaller fraction of the total microbiota that is mainly derived by environmental effects. Another advantage of controlled studies is that the

functional effects, both positive and negative, of a phylosymbiotic community assembly can be carefully measured in the context of host evolutionary history.

Materials and Methods

Ethics Statement

Procedures involving functional microbiota transplants in *Peromyscus* mice were approved by the University of Utah Institutional Animal Care and Use Committee under protocol 12- 12010. Mice obtained from the *Peromyscus* Genetic Stock Center were reared under IACUC approved protocols, and only fecal samples were directly utilized. While our paper contains data for several primate species, this data was conducted by another research group, has been previously published, and is now publicly available. Thus, there was no requirement of approved protocols for the primate species.

Nasonia Husbandry and Sample Collection

Nasonia were reared as previously described[236]. Four strains were used: *Nasonia vitripennis* (strain 13.2), *N. longicornis* (IV7U-1b), *N. giraulti* (RV2x(u)), *N. oneida* (NAS_NONY(u)). To collect individuals for microbiota analysis, virgin females were sorted as pupae into sterile glass vials and collected within the first

24 h of eclosing as adults. Subsequently, they were rinsed with 70% ETOH for 2 min, a 1:10 bleach solution for 2 min, followed by two rinses in sterile water. Individuals were then placed in 1.5 ml tubes and flash frozen in liquid nitrogen. They were then stored at -80 C until DNA extractions. Fifty individuals were collected per strain.

Drosophila Husbandry and Sample Collection

Nine strains of *Drosophila* were obtained from the University of California San Diego *Drosophila* Species Stock Center. Six strains were used in the microbiome analysis because they were *Wolbachia*-free: *Drosophila melanogaster* (Strain Dmel, stock number 14021-0248.25), *D. simulans* (Dsim, 14021-0251.195), *D. yakuba* (Dyak, 14021-0261.01), *D. erecta* (Dere, 14021-0224.01), *D. pseudoobscura* (Dpse, 14011-121.94), and *D. mojavensis* (Dmow, 15081-1352.22). The three strains that tested positive for *Wolbachia* (method described below) were: *D. sechellia* (14021-0248.25), *D. ananassae* (14021-0371.13), and *D. willistoni* (14030-0811.24). All strains were reared on a cornmeal media (*Drosophila* Species Stock Center: http://stockcenter.ucsd.edu/info/food_cornmeal.php) with a sterile Braided Dental Roll (No. 2, Crosstex, Atlanta, Georgia, US) inserted into the surface of the media. All stocks were incubated at 25 C with a 12-h light-dark

cycle and monitored every 24 h. Every 14 d, stock vials were cleared of any emerged adults, and 6 h later, ten virgin females and three males were transferred to new food vials. This conditioning on the same food was done for five generations before setting up media vials for sample collection. For each of the six strains, five virgin females were mated with two males and allowed to oviposit for 24 h; afterwards, the parents were removed and the vials were incubated as per above.

After 12 d, vials were cleared and virgin females were collected every 4-6 h over a 36 h period. All females were rinsed with 70% ETOH for 2 min, a 1:10 bleach solution for 2 min, followed by two rinses in sterile water. Individual adult flies were then placed in 1.5 ml tubes and flash frozen in liquid nitrogen. They were then stored at -80 C until DNA extractions. Approximately 25-30 virgin adult females were collected per strain.

Mosquito Husbandry and Sample Collection

Mosquitoes were acquired from the Malaria Research and Reference Reagent Resource Center as eggs on damp filter paper within 24 h of being laid. Eight strains were used: *Anopheles funestus* (strain name FUMOZ), *An. farauti s.s.* (FAR1), *An. quadrimaculatus* (GORO), *An. arabiensis* (SENN), *An. gambiae* (MALI

NIH), *Aedes aegypti* (COSTA RICA), *Ae. albopictus* (ALBO), and *Culex tarsalis* (YOLO F13). Eggs were floated in 350 ml of sterile water with 1.5 ml of 2% yeast slurry and autoclaved within a sterile and lidded clear plastic container. Containers were enclosed within a larger sterile clear container and placed inside an incubator set at 25° C with a 12-h light-dark cycle and monitored every 24 h. After 48 h, the hatched larvae were sorted out and 100-150 of each species were placed in new sterile water (150 ml) with 30 mg of powdered koi food (Laguna Goldfish & Koi all season pellets). Water level was maintained at 150 ml, and larvae were fed 30 mg of powdered koi food every day for a total of 13 d. All pupae were discarded (frozen and autoclaved) on day 10, and new pupae were collected every 12 h on day 11, 12, and 13. Water samples were also collected and frozen for microbial analysis on day 11.

To collect individuals for microbiota analysis, pupae were sorted according to sex, and all females were rinsed with 70% ETOH for two min, then 1:10 bleach solution for two min, followed by two rinses in sterile water. Individual pupae were then placed in 1.5 ml tubes and flash frozen in liquid nitrogen. They were then stored along with their corresponding water sample at -80 C until DNA extractions. Ten to 25 individuals were collected per strain.

Peromyscus Husbandry and Sample Collection

Fecal samples were collected from the *Peromyscus* Genetic Stock Center at the University of South Carolina. Six stock species of *Peromyscus* were used: *P. maniculatus* (stock BW), *P. polionotus subgriseus* (PO), *P. leucopus* (LL), *P. californicus insignis* (IS), *P. aztecus hylocetes* (AM), and *P. eremicus* (EP). All mice were reared using their standard care practices at the stock center on the same mouse chow diet. Cages were cleaned at regular intervals for all species, and all species were caged within the same facility. Individuals from non-mating cages of females (five to six per cage) were used for collections. Fecal pellets were collected on a single morning from individual mice directly into a sterile tube and placed on dry ice before being stored at -80 C for 24 h. Samples were then shipped overnight on dry ice and again stored at -80 C until DNA extractions. One to three pellets from 15 individuals were collected per strain.

In order to eliminate the introduction of confounding factors and exclude any subjects that had a pinworm infection at the time of sample collection, we conducted a screen to confirm the pinworm status of each mouse. Pinworm status was confirmed by PCR. Primers utilized to amplify the 28S rDNA D1 and D2 domains of multiple pinworm species were developed and confirmed with

positive DNA samples of *Syphacia obvelata* and *Aspiculuris tetraptera* (received from the Feldman Center for Comparative Medicine at the University of Virginia). The C1 primer 5'-ACCCGCTGAATTTAAGCAT-3' and the D1 primer 5'-TCCGTGTTTCAAGACGG-3' were amplified under the following reaction conditions: 94 C for 1 min; 35 cycles of 94 C for 30 s, 55 C for 30 s, 72 C for 30 s; and a final elongation time at 72 C for 2 min. The resultant samples were then visualized on a 1% agarose gel. Of the 84 fecal specimens analyzed, 8 of the samples showed amplification at 750 bp corresponding to the expected amplification size of the pinworm DNA sequence. For confirmation, the 750 bp bands were extracted using a Wizard Gel Extraction Kit (Promega Corporation, Madison, Wisconsin, US) and sequenced (GENEWIZ, Inc, New Jersey, US). Sequence results confirmed the presence of *Aspiculuris tetraptera* infection, and these 8 samples and were excluded from further analysis.

Wolbachia Screens of Stock Insect Lines

The presence or absence of *Wolbachia* was checked using two replicates of three individuals per species. DNA extraction was performed with PureGene DNA Extraction Kit (Qiagen), and fragments of the 16S rDNA gene were PCR amplified using primer set WolbF and WolbR3[237]. Only stock strains that were

Wolbachia negative were used in the experiments.

Insect DNA Extraction

Individual insects (and the mosquitoes' corresponding water samples) were mechanically homogenized with sterile pestles while frozen within their collection tube. The samples were then thawed to room temperature for 30 s and flash frozen again in liquid nitrogen with additional mechanical homogenization. The samples were finally processed using the ZR-Duet DNA/RNA MiniPrep Kit (Zymo Research, Irvine, California, US). Samples were then quantified using the dsDNA BR Assay kit on the Qubit 2.0 Fluorometer (Life Technologies).

DNA Isolation from Mouse Samples

The PowerSoil DNA isolation kit (Mo Bio Laboratories, Carlsbad, California, US), was utilized to extract DNA from 20 mg of mouse fecal material per sample according to manufacturer's protocol after being mechanically homogenized with sterile pestles while frozen within their collection tube. Samples where then quantified using the dsDNA BR Assay kit on the Qubit 2.0 Fluorometer.

PCR, Library Prep, and Sequencing

Total genomic DNA was quantified using dsDNA HS Assay kit on the Qubit. Using two µl of DNA, a 20 µl PCR reaction of 28S general eukaryotic amplification was conducted on each sample, with only 25 cycles. Products were purified using Agencourt AMPure XP, quantified using the dsDNA HS Assay kit on the Qubit, and compared to the amount of 16S amplification from the same DNA volume and PCR reaction volume as previously described[236]. PCR amplification of the bacteria 16S rRNA was performed with the 27F 5'-AGAGTTTGATCCTGGCT- CAG-3' and 338R 5'-GCTGCCTCCCGTAGGAGT-3' "universal" bacterial primers with the NEBNext High-Fidelity 2X PCR Master Mix; duplicate reactions were generated per sample, which were pooled together postamplification. For sequencing runs 1 (*Peromyscus*) and 2 (*Nasonia*, mosquito, and *Drosophila*), 16S PCR products that were made into libraries had their concentrations normalized relative to about 1,000 ng/ml and 2,000 ng/ml of the 28S quantity for library prep respectively.

Using the Encore 384 Multiplex System (NuGEN, San Carlos, California, US), each samples' 16S product was ligated with Illumina NGS adaptors and a unique barcode index (after the reverse adaptor). The samples were then purified

using Agencourt AMPure XP and quantified using the dsDNA HS Assay kit on the Qubit. Samples were subsequently pooled.

Each pooled library was run on the Illumina MiSeq using either the MiSeq Reagent Kit V2 or V3 for paired-end reads. Run 1 was conducted at the University of Georgia Genomics Facility and run 2 was conducted at Vanderbilt Technologies for Advanced Genomics (VANTAGE).

Sequence Quality Control

Sequence quality control and OTU analyses were carried out using QIIME version 1.8.0[238]. Forward and reverse paired-end sequences were joined and filtered if they met the following criteria: they fell below an average Phred quality score of 25, contained homopolymer runs or ambiguous bases in excess of 6 nucleotides, or were shorter than 200 base pairs. Sequences were also removed if there were errors in the primer sequence or if barcodes contained errors and could not be assigned to a sample properly. A total of 5,065,121 reads passed quality control for the meta-analysis, with an average read length of 310 ± 48 nucleotides. *Drosophila*: 648,676 reads, average length 315 ± 23. hominid: 1,292,542 reads, average length 247 ± 38. mosquito: 664,350 reads, average length 328 ± 19. *Nasonia*: 864,969 reads, average length 322 ± 15. *Peromyscus*:

295,752 reads, average length 347 ± 12.

OTU Analysis

Chimeric sequences were evaluated and removed using the UCHIME algorithm[239] for the intersection of de novo and GreenGenes 13_5 non-chimeras[240]. The sequences were then clustered into OTUs at 94%, 97%, and 99% similarity using the USEARCH open-reference method[241]. OTUs were mapped at the respective percent against the GreenGenes 13_5 database and screened for a minimum group size of two counts, with dereplication based on full sequences[240]. Representative sequences were chosen as the most abundant representative in each OTU cluster and aligned using GramAlign[242]. A phylogenetic tree of the representative sequences was built in QIIME[238] with the FastTree method and midpoint rooting[243]. Taxonomy was then assigned to the OTU representatives with the UCLUST method against the GreenGenes 13_5 database[240]. OTU tables were constructed in QIIME[238] and sorted by sample IDs alphabetically.

Sample and OTU Quality Control

OTU tables were screened to remove any OTUs classified as chloroplast, unassigned, and *Wolbachia*. Individual samples were assessed for low sequence

coverage affecting community profiles and diversity as well as for processing errors based on minimum count thresholds assessed against group means. Following rarefaction, counts were subsequently chosen as the highest rarefaction number allowed by the smallest sample's count representation in each respective clade and the meta-analysis. Alpha diversity was measured using Shannon and Chao1 metrics generated with the QIIME alpha_rarefaction script. Plots of alpha diversity at a range of rarefied levels were used to assess and remove samples with low diversity.

Meta-Analysis

The PCoA (Fig 2.3A) components for the meta-analysis were constructed using the QIIME jackknifed_beta_diversity script. The OTU table first underwent rarefaction, followed by the computation of Bray-Curtis beta diversity distances for each rarefied table. PCoA plots of the first three coordinate dimensions were generated using a custom Python script. Individual samples are each depicted as a point and are colored by host clade of origin.

The community profile (Fig 2.3B) for the meta-analysis was generated using a custom Python script and BIOM tools[244]. OTU tables were first converted to relative abundance for each sample, and bacterial taxonomy was collapsed at

the class level. Bacterial classes were sorted alphabetically, and a stacked bar chart representing the relative abundance for each sample was constructed.

The network analysis (Fig 2.3C) was visualized using Cytoscape[245]. OTU tables were first collapsed by bacterial taxonomy at the genus level, and QIIME's make_otu_network script was used to construct connections between each bacterial genus to individual hosts based on relative abundance. Network files were then imported into Cytoscape, where the network was computed using an edge-weighted force directed layout. Nodes were colored by host clade, and connections were colored by key bacterial phylum observed in high abundance (i.e., Actinobacteria, Bacteroidetes, Firmicutes, Proteobacteria) and gray for additional phylum.

Alpha diversity plots (Fig 2.3D) were prepared using the Phyloseq package[246]. OTU tables collapsed by host species were imported into Phyloseq, and the plot_richness function was used to generate box-and-whisker plots of Shannon alpha-diversity. Plots were colored by host clade of origin.

Microbiota Dendrograms

Microbiota dendrograms were constructed using the QIIME jackknifed_beta_diversity script. OTU table counts were first collapsed by host

species of origin to get representative species microbiota profiles. The pipeline script performed 1,000 rarefactions on each table and calculated Bray-Curtis beta diversity distances for each. Bray-Curtis distance matrices were UPGMA clustered to give dendrograms of interspecific relatedness. The role of 97% versus 99% OTU clustering cutoffs and weighted and unweighted UniFrac beta diversity measures (S2.2 Fig) were evaluated for Robinson-Foulds and matching cluster congruence with host phylogeny.

Host Phylogenies

Host phylogenetic trees were constructed using sequences for each host species' cytochrome oxidase gene downloaded from the NCBI. COI was chosen as a highly conserved molecular marker, and it is widely used for interspecific phylogenetic comparison[247]. Sequences were initially aligned using Muscle v3.8.31[248]. Gap positions generated through inserts and dele- tions were removed, and overhanging sequence on 5^0 and 3^0 ends were trimmed. Models of molecular evolution were evaluated using jModelTest v2.1.7[249], and the optimal model was used for final alignment and tree building in RaxML v8.0.0[250]. The *Nasonia* and *Peromyscus* clades were carried out using the same methodology-except for final alignment and tree building in PhyML v3.0[251] and for *Peromyscus* the AVPR1A gene

was concatenated with COI to further resolve the phylogeny. All trees are concordant with well-established phylogenies from literature references noted in the Results section.

Robinson-Foulds and Matching Cluster Congruency Analysis

Quantifying congruence between host phylogeny and microbiota dendrogram relationships (Fig 2.5) was carried out with a custom Python script and the TreeCmp program[252]. The topologies of both trees were constructed, and the normalized Robinson-Foulds score[198] and normalized matching cluster score[253] were calculated as the number of differences between the two topologies divided by the total possible congruency score for the two trees. Next, 100,000 random trees were constructed with the same number of leaf nodes, and each was compared to the host phylogeny. The number of trees which had an equivalent or better score than the actual microbiota dendrogram were used to calculate the significance of observing that topology under stochastic assembly. Normalized results of both statistics have been provided to facilitate comparison. Matching cluster and Robinson-Foulds *p*-values were determined by the probability of 100,000 randomized bifurcating dendrogram topologies yielding equivalent or more congruent phylosymbiotic patterns than the microbiota dendrogram.

Intraspecific Versus Interspecific Beta Diversity Distances

Within each clade, the Bray-Curtis distances calculated by the jackknife_beta_diversity script (Fig 2.4A) were separated by those that compared microbiota within a host species and those that compared between host species. The box-and-whisker plots were constructed in Python. Coloring indicates host clade of origin, and all intraspecific and interspecific distances are rep- resented for each clade. These distances were then compared between the groups using a non- parametric, two-tailed Mann-Whitney U test implemented in SciPy[254;255].

ANOSIM Clustering

To evaluate intraspecific clustering (Fig 2.4B), the ANOSIM test was used to calculate the distinguishability of Bray-Curtis distances based on species of origin. Bray-Curtis distance matrices were generated using the QIIME jackknifed_beta_diversity script on tables of individuals rarefied 1,000 times. The QIIME script compare_categories was used to calculate ANOSIM scores using the Bray-Curtis distance matrix and host species as categories. 1,000 permutations were used to calculate the significance of clustering for each clade. Three-dimensional PCoA plots were generated in Python using components generated

from Bray-Curtis distance matrices in QIIME, and the first three components are shown. Points are colored by host species within each clade, and colors correlate with the species labels in Fig 5 for reference.

Correlation of ANOSIM Clustering and Clade Age

A general linear regression was performed to test the correlation between age of clade origin and the intraspecific clustering measured through ANOSIM R-statistic scores. Cladogenesis Age was Log10 transformed to normalize the distance scale between samples (1, 10, 100 MYA). The regression was carried out in Stata v12.0 to determine the coefficient (R^2) and significance (*p*-value).

Random Forest Analyses

OTU tables were first collapsed at each bacterial taxonomic level (i.e., phylum...genus) using the QIIME script summarize_taxa. Then, both the raw OTU table and each collapsed table underwent ten rarefactions to an even depth using the QIIME script multiple_rarefactions_even_depth. RFC models were constructed with the supervised_learning script for 1,000 rounds of ten-fold Monte Carlo cross validation on each table. At each level, the results were collated and averages were taken for the ten rarefied tables. Host species were used as the

category for RFC model distinguishability, testing the ability to assign samples to their respective host species. The average class error for each clade was subtracted from 100 to get the percent accuracy of the models at each taxonomic level. The same methodology was used for constructing RFC models for the meta-analysis, with the only exception being that host species, host clade, and vertebrate or invertebrate categories were tested for distinguishability.

Microbiota Transplants

Peromyscus. We tested the effects of allochthonous microbial communities on host performance by conducting a series of microbial transplants from various donor rodent species into a single recipient species, the oldfield mouse (*Peromyscus polionotus*). We obtained virgin, female *Peromyscus* species (*P. polionotus*, *P. maniculatus*, *P. leucopus*, *P. eremicus*, *P. californicus*) from the *Peromyscus* stock center. We also obtained three female individuals of *Neotoma lepida* (*Neotoma* is the sister genus of *Peromyscus*) from Dr. M. Denise Dearing (University of Utah). Additionally, we obtained six female individuals of outbred *Mus musculus* from Dr. Wayne Potts (University of Utah). The founding animals of this colony were collected from near Gainesville, Florida, US, and the animals have been randomly bred in captivity for roughly 13 generations and are still highly

outbred[256;257]. All rodent species were maintained on powdered laboratory rodent chow (Formula 8904, Harlan Teklad, Madison, Wisconsin, US) except for woodrats, which were fed powdered rabbit chow (Formula 2031, Harlan Teklad, Madison, Wisconsin, US), given that woodrats are herbivorous. All procedures involving rodents were approved under the University of Utah Institutional Animal Care and Use Committee protocol #12-12010.

To conduct microbial transplants, we followed a protocol that was previously established to transplant the microbiota from *Neotoma lepida* into *Rattus norvegicus*[201]. First, donor feces were collected from three to six individuals of each donor species by placing rodents in wire-bottom metabolic cages overnight and collecting feces the next morning. Feces were then ground with a mortar and pestle and mixed into powdered laboratory chow (Formula 8904, Harlan Teklad, Madison, Wisconsin, US) at a ratio of 15% w/w. Recipient animals (five to six individuals per group) were fed food containing feces of a particular donor species for two nights. Then, recipient animals were fed normal laboratory diets for 6 d, which is a sufficient time for the clearance of transient, ingested microbes[258]. We then measured food intake and dry matter digestibility by placing animals into wire-bottom metabolic cages. Animals were presented with a known amount of powdered rodent chow overnight. The next morning, remaining food was

weighed, and feces were collected, dried overnight, and weighed. Food intake was calculated as g dry food presented-g dry food remaining. Dry matter digestibility was calculated as (g dry food ingested-g dry feces produced) / g dry food ingested.

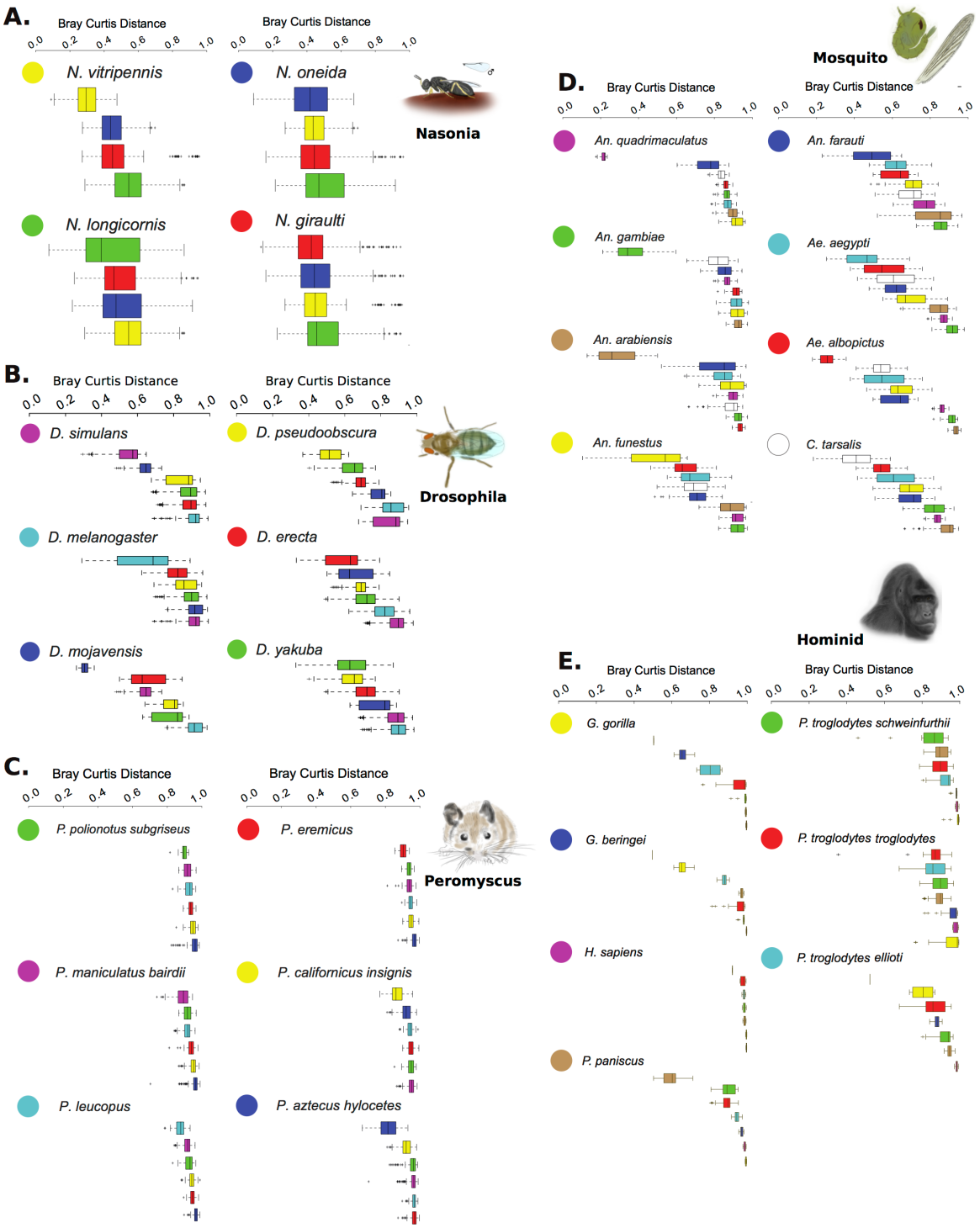
We investigated whether microbial communities from more distantly related hosts affected performance metrics in recipients. We compared food intake using ANOVA and Tukey's HSD test across recipient groups. We also conducted correlations of dry matter digestibility and estimated divergence times based off of previously published phylogenies[200;259]. We performed correlations using both untransformed divergence times and log-transformed divergence times.

Nasonia. We tested the effects of allochthonous microbial communities on host survival by exposing two recipient species (*N. vitripennis* or *N. giraulti*) to a suspension of heat-killed microbes isolated from three donor *Nasonia* species (*N. vitripennis*, *N. giraulti*, and *N. longicornis*). We reared *Nasonia* in an in vitro rearing system[203] and inoculated germ-free larvae in 6 mm diameter transwell inserts with autochthonous microbiota, allochthonous microbiota, and sterile phosphate-buffered saline (PBS) for the first 8 d after embryo hatching. Microbiota were purified from fourth instar larvae of *Nasonia* by filtration through a 5 um filter and centrifugation at 10,000 rpm for 3 min. The pellet was suspended in a sterile PBS

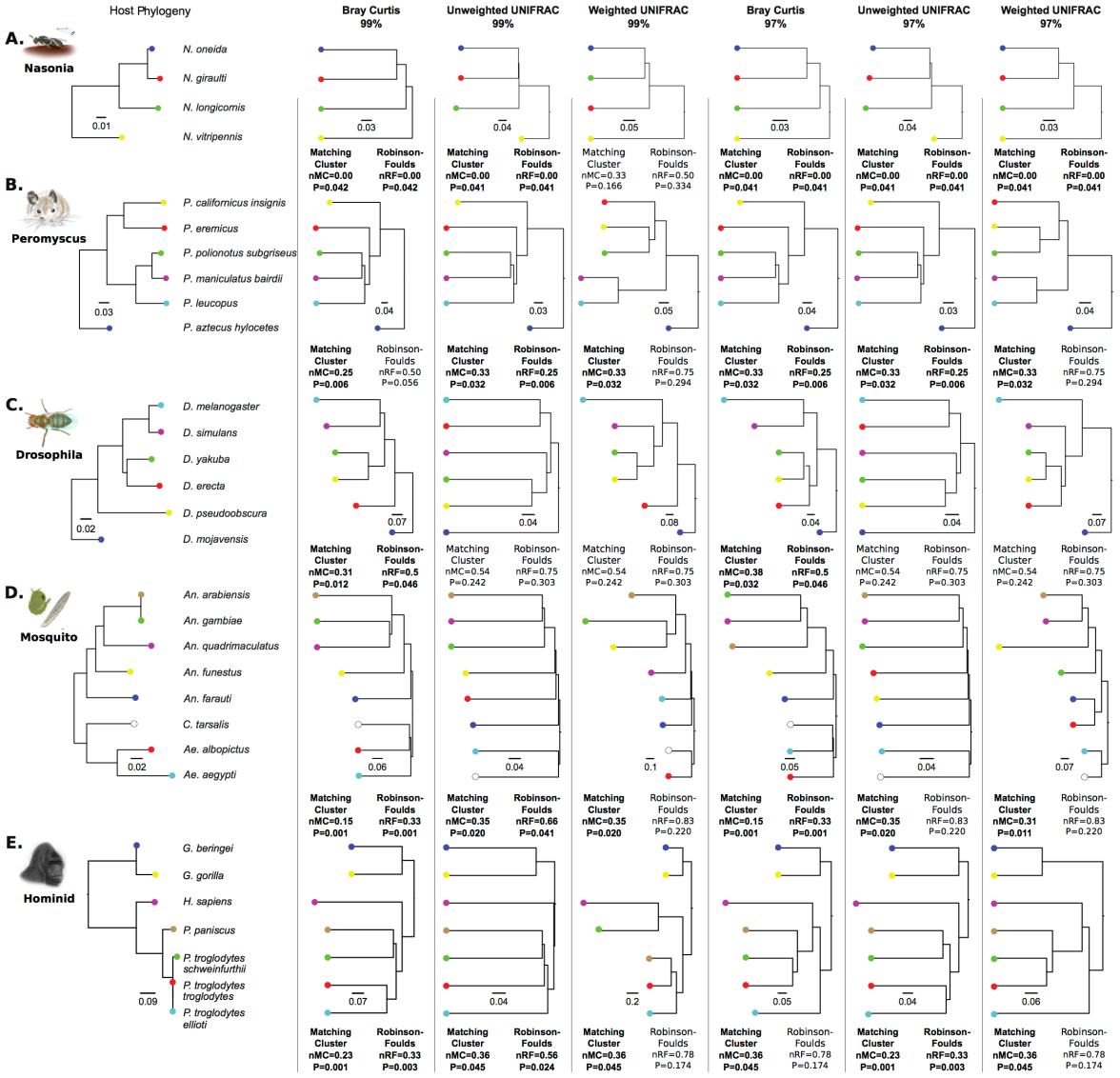
solution at a concentration of 5 x 10^6 CFU of microbiota bacteria (determined by tryptic soy agar plating) per milliliter. 20 uL of this microbiota suspension was added to the transwell inserts for each of the 8 inoculation days. *Nasonia* rearing media was replaced daily just before the inoculations.

Measurements of *Nasonia* survival from first instar larvae to adulthood were determined using transwell insert images taken with an AmScope MT1000 camera. For each transwell, live larval counts were recorded 3 d post-embryo hatching. Adult counts were determined by recording the number of remaining larvae and pupae in each transwell sample 20 d after embryo hatching (5-7 d after first adult eclosion) and subtracting that number from the larval counts previously recorded. Normalized adult survival per transwell sample was calculated as the percent survival of *Nasonia* from 3 d to 20 d after embryo hatching divided by the average percent survival of the autochthonous microbiota treatment group. We compared survival between the autochthonous and allochthonous treatment groups using Mann-Whitney U tests.

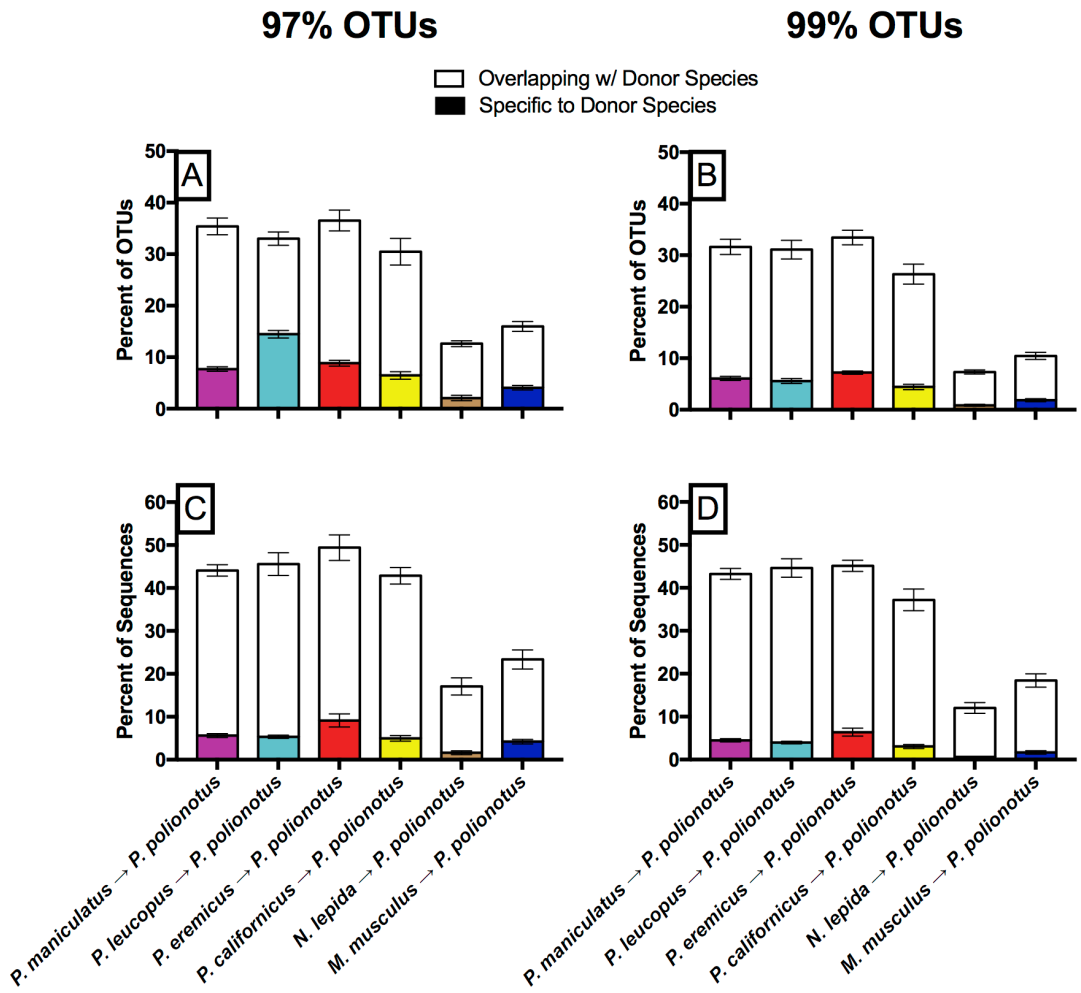
Supporting Information



S2.1 Fig. Comparisons of intraspecific and interspecific Bray-Curtis distances for pairwise combinations of all species. Bray-Curtis beta diversity distances were computed for all pairs of individuals within each clade from 99 percent OTUs. Colored circles denote the named species, and colors within box-and-whisker plots denote to which species it is being compared. Boxes represent the 25^{th} to 75^{th} quartiles with the central line depicting the group median, and whiskers showing the 1.5 interquartile extent.

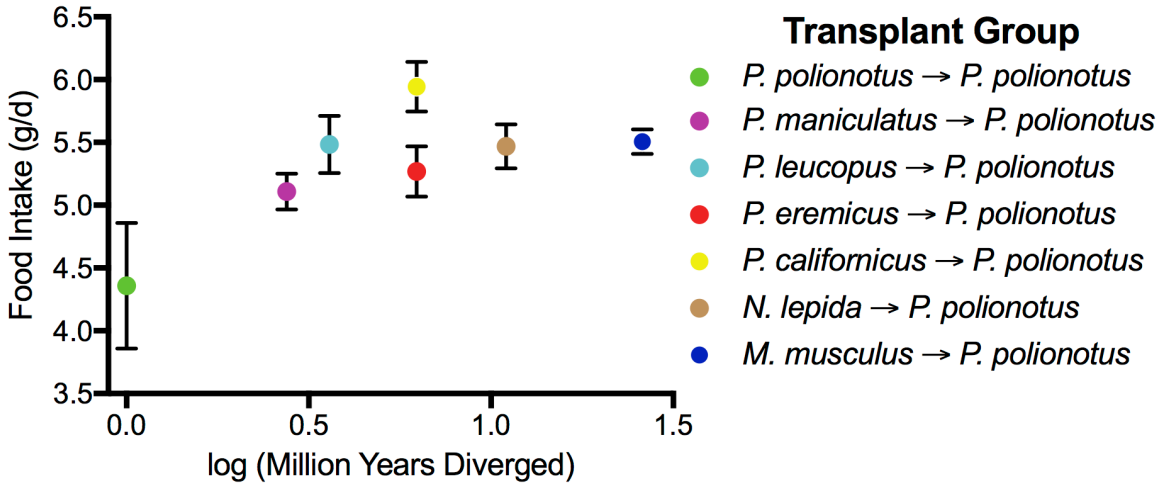


S2.2 Fig. Phylosymbiosis analysis for alternative beta-diversity metrics and OTU clustering cutoffs. The normalized Robinson-Foulds metric and the normalized Matching Cluster metric were used to evaluate the congruence between host phylogenies and microbiota dendrograms for Bray Curtis, Unweighted UniFrac, and Weighted UniFrac beta-diversity metrics at both 97 and 99 percent clustered OTUs.



S2.3 Fig. Fine-resolution overlap between donor and recipient microbial communities. White bars represent shared OTUs between donor and recipients and thus the possible range of transfer. Colored bars represent the portion of shared OTUs that are donor-specific and thus transfer of unique OTUs between donor and recipients. Panels (A) and (B) depict the mean ± s.e.m. percentage of OTUs. Panels (C) and (D) show the mean ± s.e.m abundance of total sequences.

These analyses were conducted with OTU-picking at both 97% and 99% sequence identities.



S2.4 Fig. Effects of allochthonous versus autochthonous microbial communities on the food intake of recipient mice. Divergence times between *P. polionotus* and donor species were determined from previously published phylogenies[200;259]. Points represent mean values ± s. e.m. for each group (n = 5-6 recipients per group).

S2.1 Table. Table of Random Forest accuracy in classifying the microbiota by host species in each host clade, and by host species, clade, and mammal or invertebrate taxonomy in the meta-analysis. Models were generated using OTUs or abundance collapsed by bacterial taxonomy. Red boxes highlight the highest classification accuracy. Ten-fold cross validation assessed the percent

classification accuracy for test sets excluded from model training.

https://doi.org/10.1371/journal.pbio.2000225.s005

S2.2 Table. Table of Random Forest model mean decrease in accuracy when genera are excluded from classification of the microbiota in each host clade. Random Forest models were generated using genera collapsed bacterial taxonomies. Genera are ordered by those that contribute the most accuracy to the model to those that contribute the least accuracy to the model, measured in the form of decrease in model accuracy when a genus is excluded from model construction. Standard deviations of mean decrease in model accuracy are also provided.

https://doi.org/10.1371/journal.pbio.2000225.s006

S2.3 Table. Tables of microbiota taxon in the meta-analysis with varying abundance between host clades or between vertebrates and invertebrates. The meta-analysis OTU table was collapsed at each bacterial taxonomic level (Phylum to Genus), and converted to relative abundance. Kruskal-Wallis tests were performed on microbial taxon within each table, testing for differences in

the mean abundance across host clades or vertebrates and invertebrates. The results were sorted from high to low significance of p-values, which are provided alongside False Discovery Rate and Bonferroni corrected p-values. Mean abundances of each taxon within host clades or vertebrates and invertebrates are provided as a heatmap, with dark blue indicating high abundance, light blue centered at the 5% most abundant values and fading to white for low abundance or non-existent taxon.

https://doi.org/10.1371/journal.pbio.2000225.s005

CHAPTER III

Finer-Scale Phylosymbiosis: Insights from Insect Viromes[2]

Author Contributions

This study was performed by Brittany Leigh (BL), Sarah R. Bordenstein (SRB), Andrew Brooks (AB), Aram Mikaelyan (AM), and Seth Bordenstein (SB). BL reared *Nasonia* and extracted viral particles, sequenced and assembled viral contigs, and assembled virome community profiles. BL, SRB, AB, AM contributed to statistical analysis, with AB performing topological congruency tests. SB was the principle investigator and worked with all study participants to develop analyses. All participants contributed to writing / editing of the manuscript.

Introduction

Abstract

Phylosymbiosis was recently proposed to describe the eco-evolutionary pattern whereby the ecological relatedness (e.g., beta diversity relationships) of

---

[2] This work is published in *mSystems*: Leigh BA, Bordenstein SR, **Brooks AW**, Mikaelyan A, Bordenstein SR. 2018. Finer-scale phylosymbiosis: insights from insect viromes. https://doi.org/10.1128/mSystems.00131-18.

host-associated microbial communities parallels the phylogeny of the host species. Representing the most abundant biological entities on the planet and common members of the animal-associated microbiome, viruses can be influential members of host-associated microbial communities that may recapitulate, reinforce, or ablate phylosymbiosis. Here we sequence the metagenomes of purified viral communities from three different parasitic wasp *Nasonia* species, one cytonuclear introgression line of *Nasonia*, and the flour moth outgroup *Ephestia kuehniella*. Results demonstrate complete phylosymbiosis between the viral metagenome and insect phylogeny. Across all *Nasonia* contigs, 69% of the genes in the viral metagenomes are either new to the databases or uncharacterized, yet over 99% of the contigs have at least one gene with similarity to a known sequence. The core *Nasonia* virome spans 21% of the total contigs, and the majority of that core is likely derived from induced prophages residing in the genomes of common *Nasonia*-associated bacterial genera: *Proteus*, *Providencia*, and *Morganella*. We also assemble the first complete viral particle genomes from *Nasonia*-associated gut bacteria. Taken together, results reveal the first complete evidence for phylosymbiosis in viral metagenomes, new genome sequences of viral particles from *Nasonia*-associated gut bacteria, and a large set of novel or uncharacterized genes in the *Nasonia* virome. This work suggests that

phylosymbiosis at the host-microbiome level will likely extend to the host-virome level in other systems as well.

Importance

Viruses are the most abundant biological entity on the planet and interact with microbial communities with which they associate. The virome of animals is often dominated by bacterial viruses, known as bacteriophages or phages, which can (re)structure bacterial communities potentially vital to the animal host. Beta diversity relationships of animal-associated bacterial communities in laboratory and wild populations frequently parallel animal phylogenetic relationships, a pattern termed phylosymbiosis. However, little is known about whether viral communities also exhibit this eco-evolutionary pattern. Metagenomics of purified viruses from recently diverged species of *Nasonia* parasitoid wasps reared in the lab indicates for the first time that the community relationships of the virome can also exhibit complete phylosymbiosis. Therefore, viruses, particularly bacteriophages here, may also be influenced by animal evolutionary changes either directly or indirectly through the tripartite interactions among hosts, bacteria, and phage communities. Moreover, we report several new bacteriophage genomes from the common gut bacteria in *Nasonia*.

Introduction

Ecological similarity of host-associated microbial communities between species can often mirror phylogenetic similarity of hosts across a wide range of animal taxa[24;50;94;178;179;260]. This eco-evolutionary pattern, termed phylosymbiosis[24;51], can arise from a variety of biotic or abiotic factors. Resultantly, phylosymbiosis does not *a priori* presume stable or long-term, transgenerational associations between microbial communities and their hosts. Phylosymbiosis may change with environments, lifestyles, or multipartite interactions that shift assembly of microbial communities. For example, phages (i.e., bacteriophages; viruses that infect bacteria) can outnumber bacteria in both free-living and host-associated communities[261;262], represent the majority of viruses within animal microbiomes[261;263-266], and may drive or ablate bacterial phylosymbiosis as they prey on bacteria.

A phage can exhibit two main life cycles: lytic and temperate. A lytic phage infects its bacterial host and immediately replicates and lyses the bacterial cell. A temperate phage, however, can integrate into and replicate as part of the bacterial genome until a biotic or abiotic trigger causes it to excise and enter the lytic cycle. In mammalian host-associated phage communities, the temperate life cycle

dominates[263;267-269], presumably due to environmental parameters such as host density[270] and mucosal tissue structure[271]. Phage integration into animal-associated bacterial genomes (i.e., prophage) can alter the phenotype of the host bacterium through lysogenic conversion[272;273], as well as enhance biofilm formation and thereby horizontal gene transfer among co-occurring bacteria[274;275]. The prevalence of temperate phages in host- associated microbiomes suggests that these phages may more intimately evolve with their bacterial hosts and/or shape the composition of the bacterial communities. Additionally, the discovery of intraspecific and interspecific core viromes dominated by phages across animal systems is often reflective of the core bacterial communities described in these same organisms[264;276-278]. Although it has been suggested previously[277], phylosymbiosis at the viral level has yet to be explicitly demonstrated, and evidence for this tripartite association pattern could underpin new ecological and functional interactions between an animal host, its bacterial community, and the viruses infecting both.

Results

Virome Samples and Assemblies

Viral purifications from adults of three species of *Nasonia*, a *Nasonia*

introgression line, and the Mediterranean flour moth *Ephestia kuehniella* were sequenced. Each of the pure *Nasonia* species (*N. vitripennis*, *N. longicornis*, and *N. giraulti*) maintains their natural *Wolbachia* infections from supergroup A. The introgression line IntG has the genome of *N. giraulti* and the cytoplasm of *N. vitripennis*, including the maternally inherited supergroup A *Wolbachia* strain *w*VitA from *N. vitripennis*[279]. *E. kuehniella* harbors a supergroup B *Wolbachia* strain named *w*CauB[280]. Viral particle sequencing and single sample assembly statistics are outlined in S3.1 Table in the supplemental material.

Phylosymbiosis of viral metagenomes

Phylosymbiosis describes a significant host phylogenetic signal on host-associated microbiome communities[24]. Bacterial communities frequently, but not universally, exhibit this relationship under wild and laboratory conditions[24;51]. For viromes, there is no *a priori* reason to expect that phylosymbiosis will occur because inducible proviruses and/or lytic viruses, i.e., the targets of this study, may constitute a small subset of the total viral DNA in bacterial and eukaryotic genomes, and active viral particles have the potential to lyse and shift bacterial communities that may disrupt phylosymbiosis. Here we evaluate if the *Nasonia* viromes form phylosymbiotic community relationships.

The phylogeny of *Nasonia* spp. rooted with the outgroup *E. kuehniella* is based on DNA sequences of the cytochrome oxidase I (COI) gene as previously shown[24;51;281-284]. It resulted in the same branching pattern as the dendrogram generated from Bray-Curtis beta diversity of the viral metagenomes across the host species (Fig 3.1). The matching cluster and Robinson-Foulds tree metrics were utilized to calculate host phylogenetic and virome dendrogram topological congruence, which is highly significant based on both metrics with 100,000 randomly bifurcating trees to simulate stochastic virome assembly[24] (*P* value = 0.00451). Additionally, using the same methodology, matching cluster and Robinson-Foulds metrics were evaluated by the Binary Jaccard beta diversity index, which produced identical results using viral presence and absence within each sample. Taken together, these findings comprise one of the first lines of evidence for phylosymbiosis in host-associated viral communities. We next evaluated the number and types of viruses that comprise these phylosymbiotic communities.

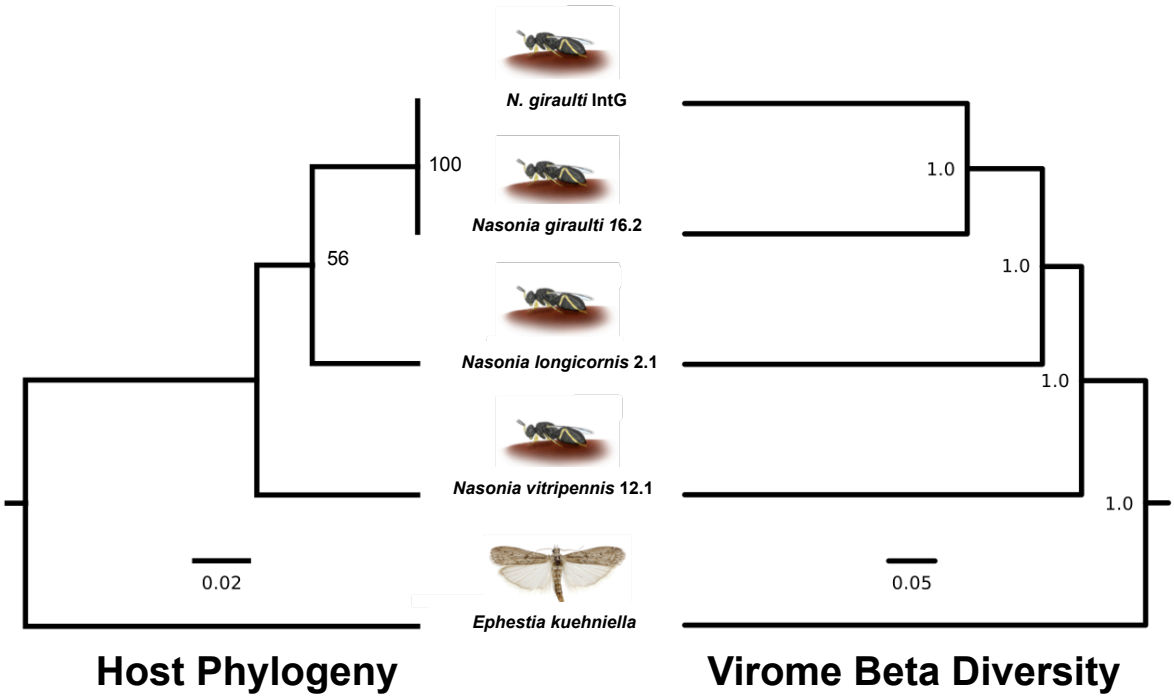


Figure 3.1. Phylosymbiosis occurs between insects and their viral communities. The host phylogeny is constructed with PHYML from 385bp of the cytochrome oxidase I gene, and the UPGMA hierarchical cluster relationships of the viromes are based on Bray-Curtis beta diversity distances. Significance of topological congruence was determined using a previously described method[24] based on the rooted Robinson-Foulds (P-value: 0.00451) and rooted matching cluster (P-value: 0.00451) with a total of 100,000 randomized topologies simulating a null hypothesis of stochastic virome ecological assembly.

Characterizing Host Genetic Effects, the Virome Core, and Toxins

Unlike many environmental viral metagenomes, the majority of the viral

contigs from the insects studied here had at least one gene with BLASTx similarity to either known lytic viruses or genes from their potential respective hosts. An average of 30.9% of the genes identified in each of the samples have a predicted annotation and function (Fig 3.2A). Therefore, to identify groups of proteins independent of the database annotations, unique protein clusters, defined as groups of proteins with significant sequence similarity (>70%), were determined in each of the samples by the protein clustering tool vContact (Fig 3.2B). The protein cluster networks identified *N. giraulti* and IntG as the most diverse viromes, which share a *N. giraulti* genetic background but vary in the origin of their cytotype. This result suggests that host genotype rather than cytotype more strongly impacts diversity of the host-associated viral metagenome, either through interactions with phage directly or through interactions with the bacteria harboring these phages. *N. longicornis* and *N. vitripennis* yielded approximately 50% fewer unique protein groups in their viromes.

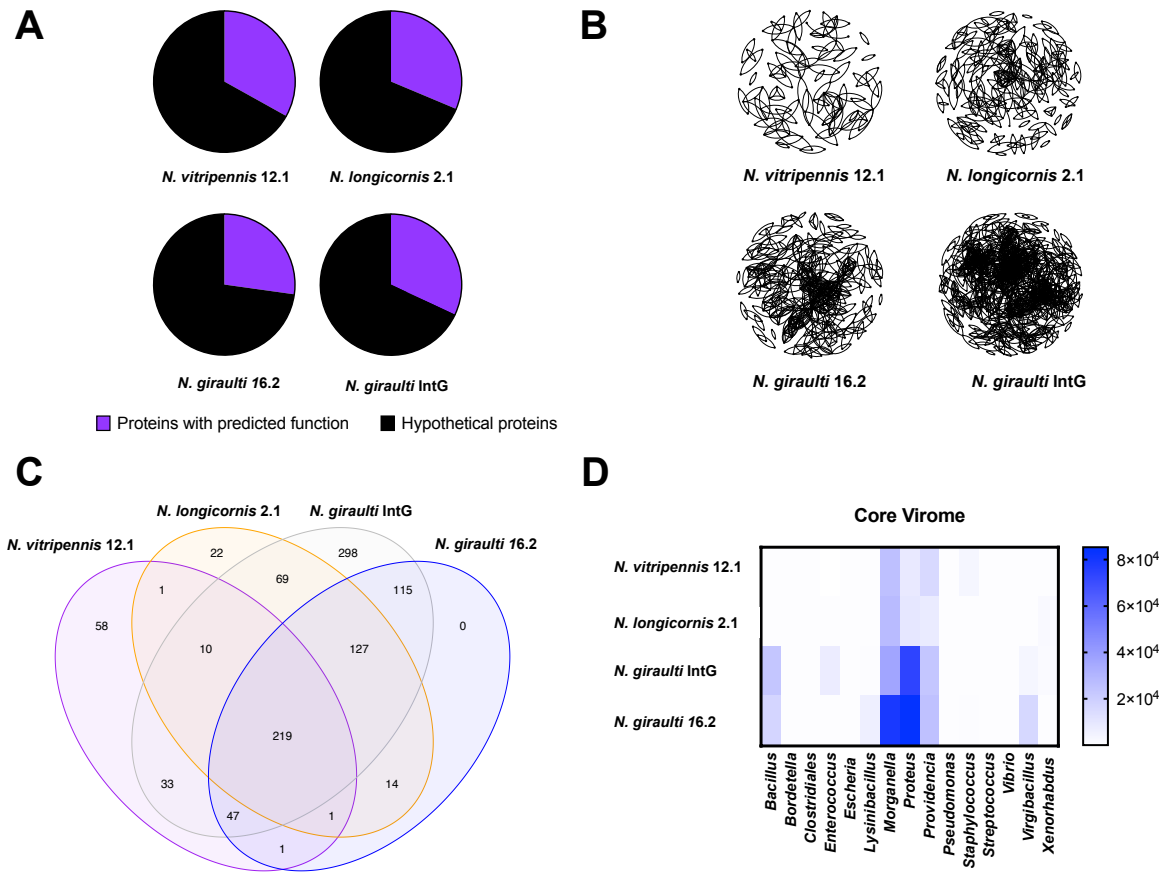


Figure 3.2. *Nasonia* species harbor a modest core virome. (A) Percent viral contigs with at least one functionally annotated gene as determined by Pfam analysis. (B) Viral protein cluster analysis illustrating diversity of viral proteins within each virome. Each dot represents a unique viral protein and connecting lines indicate >70% sequence similarity between two proteins. (C) Venn diagram illustrating the viral contigs unique within and shared between the *Nasonia* species. (D) Taxonomic affiliation of the 219 members of the identified core virome as determined by BLASTx against the nr database. Shading indicates the relative abundance of each member within single viromes and was determined by

read mapping to viral contigs.

To assess identity and diversity of proteins with predicted function in the viromes, contigs from all samples were compared to the protein family (Pfam) database. Each *Nasonia* species virome maintained a small, host-specific set of Pfams ranging from 6.7% to 14.8% of the Pfams (S3.2 Table). Precisely 24.4% of the Pfams (n=173) were shared among all of the *Nasonia* samples, which parallels the 21% of the total contigs described as the core virome below. Across all species, the most abundant Pfam (4.7% of total Pfam predictions) was the helix-turn-helix (HTH) DNA-binding motif (PF01381) followed by the phage integrase Pfam (PF00589, 2.9% of total Pfam predictions).

To further explore the protein content of the viromes and the interactions that could underpin phylosymbiosis between hosts and their viromes, we assessed if domains similar to known toxins or domains that interact with eukaryotic hosts were present in these viruses using the Pfam annotations. Proteins identified as toxins and eukaryotic-interacting domains span immunoglobulin peptidases, virulence genes, lysins, and others (indicated by boldface in S3.2 Table). Domains identified within these groups were found in viral contigs isolated from *N. giraulti* and IntG where 36 and 34 unique identifiable toxin and eukaryotic-interacting

proteins spanned 0.045% and 0.067% of the total contigs, respectively. *N. vitripennis* and *N. longicornis* maintained 17 and 25, which spanned 0.025% and 0.098% of the contigs, respectively. One identified domain is the hemolysin-encoding XhlA (PF10779) detected in *Bacilli* class-associated contigs in all of the samples, which was also the most abundant in the *N. giraulti* and the IntG introgression samples. This family of hemolysins, first observed in the entomopatho- genic *Xenorhabdus nematophila*, notably lyses insect immune cells[285].

Next, core viral contigs shared among all samples were determined by read mapping to the assembled contigs using the iVirus pipeline[286]. Across the *Nasonia* samples, the core was comprised of 219 viral contigs or 21% of total *Nasonia* viral contigs (Fig 3.2C). Of these core viral contigs, the majority (84%) are homologous to members infecting species of the most abundant bacterial genera found within the *Nasonia* gut microbiome: *Morganella*, *Proteus*, and *Providencia* (Fig 3.2D). Additionally, 14 of the core viral contigs are homologous to sequences from the *Bacilli* class, all of which are relatively more abundant in *N. giraulti* and IntG. Two core viral contigs showed amino acid similarity to sequences in the genome of the entomopathogenic *Xenorhabdus innexi*[287];they contain phage structural genes typical of active phage particles. Additionally, the complete genome of *w*VitA phage WO, a prophage of the obligate intracellular bacterium

Wolbachia that infects each of these aforementioned *Nasonia* species[288], was detected only in *N. vitripennis*. The genome of this prophage was described previously[289] and produces viral particles as seen in transmission electron microscopy in *N. vitripennis*[288].

Viral diversity among Nasonia species

The number of reads mapped to each viral contig adjusted for contig size varied among species, highlighting distinct relative abundance differences of *Proteus*, *Providencia*, *Morganella*, and *Bacilli* phages among the *Nasonia* species (Fig 3.3). *Proteus* phages dominate the *N. giraulti* virome at 34.3% of the total contigs, and *Morganella* phages make up the next largest portion at 30.8%. *Morganella* phages dominate the *N. longicornis* virome at 45.9%, and *Providencia* dominates the *N. vitripennis* virome at 41.4%. Phages with similarity to the *Bacillaceae* family outnumber all other groups in introgression line IntG (38.7%), followed by *Proteus* (26.8%). Thus, a different family of phages dominates each individual host genotype as was similarly shown for bacterial communities associated with these wasps[24]. For example, *Providencia* bacteria dominate the *N. vitripennis* microbiome[24], which correlates with the highest abundance of *Providencia* phages in the sequenced virome.

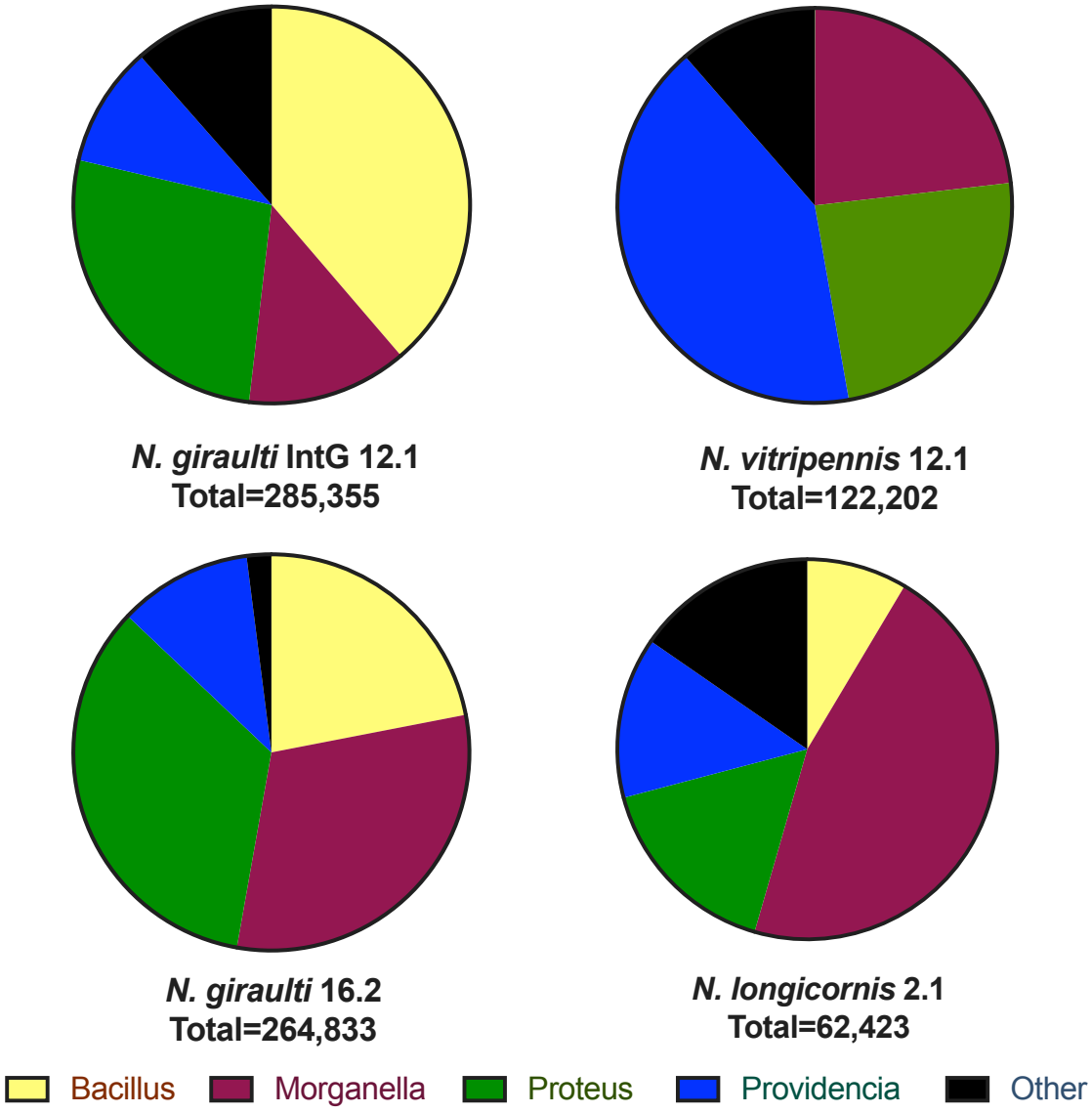
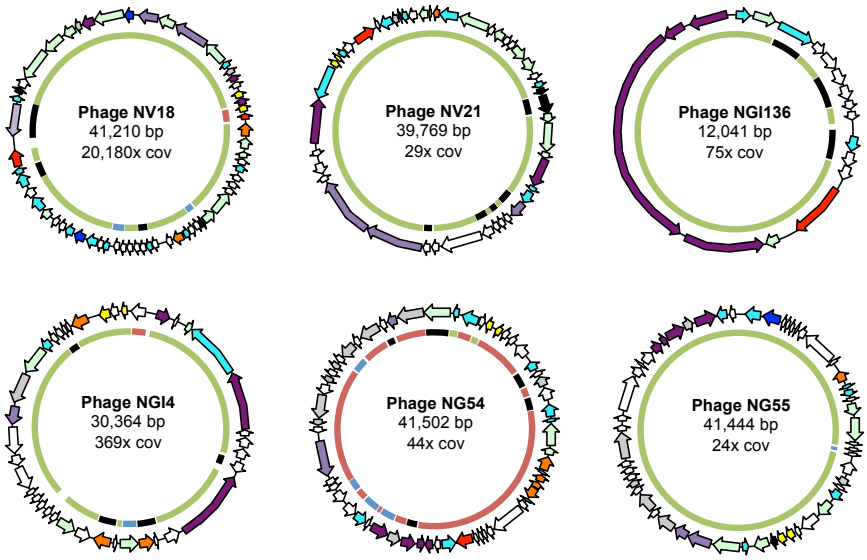


Figure 3.3: Viral communities are distinguishable between *Nasonia* species and dominated by a few taxa. The assigned taxa are bacterial genera that harbor sequences, presumably prophages, homologous to the viral protein sequences. The relative abundance of viral contigs within each species is variable. Taxonomy is determined by highest similarity through tBLASTx against the nr database.

Complete and abundant viral genomes

Six putative circular phage genomes in the core virome with moderate amino acid similarity (>70% homology) to sequences in members of *Proteus* and *Morganella* were identified using the viral classification program VirSorter and annotated using BLASTx against the RefSeq database (Fig 3.4). None of the circular phage genomes were previously reported as prophages in the bacterial genomes from which they were identified, nor have they been previously described as forming lytic phage particles. Genes in five of these circular phage particle genomes have closest matches in the *Proteus* bacterial genus. The other, phage NG54, contains genes most homologous to *Morganella* spp. Thus, these six newly assembled phage genomes, as well as most contigs recovered here, establish the hypothesis that homologous regions in close bacterial relatives of those that colonize *Nasonia* are prophages with the potential to form phage particles.

Circular viruses



Linear viral contigs

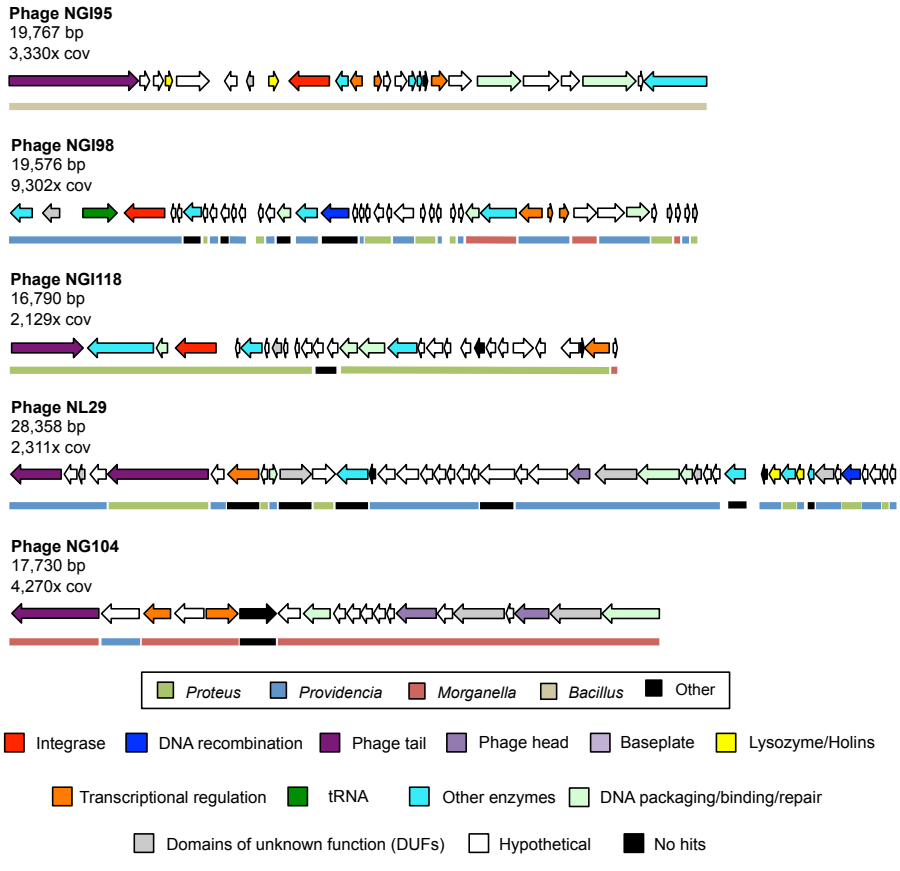


Figure 3.4: Taxonomy and function of circular genomes and most abundant linear viral contigs in *Nasonia*. Six complete and circular viral genomes were part of the core *Nasonia* virome, five of which consisted mostly of open reading frames (ORFs) with similarity to *Proteus* proteins as determined via BLASTx through NCBI against the nr database (denoted by colored line of inner circles). A total of six viral contigs (circular Phage NV18 and five linear contigs above) composed >50% of reads from each of the samples. Each of these dominant viruses were shared among all of the samples with the exception of Phage NGI95 which was only present in *N. giraulti* and *N. giraulti* IntG. Colored arrows indicate predicted gene function, and colored inner circles represent the genus of the closest BLASTx hit of each gene to the nr database.

To determine the most abundant phage variants within the *Nasonia* virome, reads were mapped to each of the viral contigs, and six contigs with total read coverage over 2,000 were identified as the most abundant. These six phage genomes, one of which was circular (phage NV18, Fig 3.4) and five of which were incomplete genomes (contigs), represented 26% of the total reads in *N. longicornis* and over 50% of reads in the other three samples. Five of these six most abundant phages were dominated by ORFs with similarity to *Proteus*, *Providencia*, and

Morganella as well. However, phage NGI95 (Fig 3.4) shared the most similarity with *Bacilli* proteins and was detected only in the introgression line IntG and *N. girault*i. Again, a large number of these genes encode unannotated hypothetical proteins, and of these abundant linear viral contigs, three of the five maintain identifiable integrase genes.

Last, two additional novel circular phage genomes recovered from *N. vitripennis* (phage NV11X) and *N. girault*i (phage NG24X) are composed of *Xenorhabdus* genes, have 94% nucleotide similarity to each other, and maintain predicted phage structural and hypothetical proteins (S3.1 Fig). These two *Xenorhabdus* phages show an average of 64% amino acid identity and complete genome synteny to predicted proteins of *Xenorhabdus innexi* and KK7.4, suggesting that prophages are present within these two bacterial genomes. *Xenorhabdus* bacteria are insect pathogens that suppress the immune system and produce numerous virulence factors such as hemolysin and cytotoxin that result in insect lethality[290-292]. Although hemolysins were found in these viromes, they were associated with *Bacilli* phages and not these *Xenorhabdus* phages, consistent with previous reports that the *Xenorhabdus* bacteria themselves encode these toxins[290-292].

Discussion

Phylosymbiosis between host and bacterial communities is emerging as a trend in microbiome studies of the animal world, across both vertebrate and invertebrate species[23;24;293;294]. While the genetic and biochemical mechanisms underlying phylosymbiosis require more study, animal performance or fitness is often highest when animals contain a homospecific microbiome in comparison to a heterospecific microbiome[24;295]. These findings imply that there are mechanisms by which animals differentially respond to the membership of the microbiome and/or vice versa. Animal-associated viromes, often composed of mostly phages, have generally received much less study than bacterial microbiomes, and there is no *a priori* reason to expect that phylosymbiosis will occur in phage metagenomes because animals are not expected to directly exert influence on membership, nor is the phage community expected to directly determine which animal it occurs in. However, evidence for direct phage protein interactions within insect hosts is found in endosymbionts where a stable association among the phage, bacterium, and animal has been established[296-298]. The bacterial endosymbionts of *Nasonia*, *Wolbachia* and its prophage (WO), represent another potential case as the phage-encoded Cif proteins cause[299] and rescue[300] reproductive parasitism phenotypes in arthropod hosts. Additionally, phage particles can bind animal mucus on

epithelial tissues via immunoglobulin domains found on the surface of some phage capsids, providing a form of immunity against colonizing bacteria[301;302]. The phages in this environment can also be transcytosed across the epithelial membrane and trafficked through the Golgi apparatus via the endomembrane system[303], further highlighting a direct interaction between phages and animals.

While bacteriophages may simply exhibit phylosymbiosis in a passive manner by association with phylosymbiotic bacterial communities, inducible prophages and/or lytic phages that are the subject of study here may only constitute a small subset of the phage DNA in bacterial genomes. Moreover, active phage particles have the potential to lyse and shift bacterial communities that may disrupt phylosymbiosis. Thus, there is no preferred reason to expect the metagenome of the purified community of virus particles will exhibit phylosymbiosis. Similarly to other animal viromes[261;263-265;267;277], the majority of viruses within *Nasonia* species are phages, and they appear to be derived mainly from prophages predicted in the most prevalent bacterial genera in *Nasonia*: *Proteus*, *Providencia*, and *Morganella*. Previous reports in *Hydra* also showed that viromes were host species specific, composed mostly of phages, and partially phylosymbiotic, although congruence of the host and virome topologies was not investigated[277]. Interestingly, wild-caught and lab strains of the same species

(*Hydra vulgaris*) harbor significantly different bacterial communities[304;305] and therefore maintained unique viral communities as well[277].

Here we describe the first report of phylosymbiosis among host-associated viromes in the parasitoid wasp genus *Nasonia*. Members of this genus diverged very recently, between 200,000 and 1 million years ago[284], and controlled rearing of each species leads to distinguishable, phylosymbiotic microbiomes that significantly impact development and survival[24;51]. Indeed, interspecific microbiota transplantation causes 25 to 42% decreases in *Nasonia* survival to adulthood compared to intraspecific microbial transplantations[24]. Moreover, hybrid death in the F2 generation is due to a breakdown in phylosymbiosis whereby inoculations of resident gut bacterial species into germfree hybrids recapitulate hybrid lethality[51].

The results here are consistent with the model that if bacterial communities show phylosymbiosis with animal hosts, so too will their viromes. More simply put, viral phylosymbiosis appears to emerge as a by-product of host-bacterium phylosymbiosis. From a methodological perspective, the result is striking given that the sequencing methods to build the bacterial and viral community dendrograms are fundamentally different: 16S amplicon sequencing versus shotgun viral metagenomics. Machine learning on 16S amplicon data

previously specified that three of the major distinguishing bacterial genera in *Nasonia* are closely related symbionts from the *Enterobacteriaceae* family (genera *Proteus*, *Providencia*, and *Morganella*)[24]. Interestingly, abundant phages of *Proteus*, *Providencia*, and *Morganella* dominate the virome identified within all of the pure *Nasonia* species (Fig 3.2D and Fig 3.3). Nonetheless, distinguishability of the viromes between *Nasonia* species is evident through at least two observations: (i) one of the most abundant viruses, phage NGI95, is solely found in the *Nasonia girault*i genotype and (ii) the majority of the phage particle genetic diversity within *N. girault*i and IntG is represented by a shared group of abundant *Bacillaceae* phages (Fig 3.2B and Fig 3.2C). Similarities between the samples with an *N. girault*i genetic background support the hypothesis that host genotype, rather than cytotype, plays a role in shaping elements of the phage community structure.

Many of the dominant bacteria present within *Nasonia* are related to well-studied human pathogens present in enteric diseases[306-312] in addition to other insects[313-315], and genomes are therefore available[167;316;317]. However, most prophage genomes present within these bacteria have not yet been described, and 69% of the genes remain annotated as encoding hypothetical proteins. Thus, the majority of the viruses found in this study were active, unannotated phages of the most prevalent types of bacteria found in *Nasonia*.

We assembled five complete *Proteus* phages and one *Morganella* phage (Fig 3.3). Four of these phages (phages NGI4, NV18, and NG55 [*Proteus*] and phage NG54 [*Morganella*]) maintained an integrase gene, indicating likely integration into their host's genome as a prophage. One of the circular *Proteus* phages maintaining an integrase, phage NV18, was by far the most prevalent phage in all of the samples with over 20,000-fold read coverage from *N. vitripennis* compared to the 10- to 200-fold coverage of most other viral contigs. This phage genome is composed of mostly hypothetical proteins and proteins with domains of unknown function. Many of these phages show amino acid similarity to sequences within the *Proteus*, *Providencia*, and *Morganella* genera (Fig 3.3). These similarities suggest that the described phages may be able to infect members across these sister genera, integrating and acquiring or leaving behind genes in the process.

The discovery of animal-bacterial-viral phylosymbiosis provides a new insight into the tritrophic relationships between animal evolution, bacterial communities, and their phage communities. We note that phylosymbiosis does not equate to coevolution, codiversification, or cospeciation because these are evolutionary processes that assume divergence from a common ancestor. Phylosymbiosis is an eco-evolutionary pattern whereby ecological similarities in the microbiome, or virome in this case, parallel phylogenetic relationships of the

host. These patterns are not necessarily ones that occur long term, and they can change rapidly in time or space. However, the detection of phylosymbiosis of the virome is consistent with host identity providing either a direct or indirect influence that partitions clustering relationships of viral particle communities in a manner that reflects animal evolution among closely related species. Whether these patterns hold in wild populations will require future study.

The microbiome has now been widely recognized as a key component of many animal functions, and alterations of this bacterial community can result in performance or fitness reductions[318-320]. Prophages are more common than lytic phages in stable host-associated microbial communities[261;263], outnumber bacteria ~3:1, and represent a potential structural force for establishment and maintenance of a microbiome[321-323]. Intimate associations among phages, bacteria, and their animal hosts are complex, and further studies investigating phylosymbiotic phage communities throughout the animal kingdom are necessary to gain a fuller understanding of the role that microbiomes and viromes play in animal functions and evolution.

Materials and Methods

Sample Collection and Sequencing

Nasonia species were reared as previously described[51]. Four strains were used in this study: *Nasonia vitripennis* (strain 12.1), *N. longicornis* (2.1), *N. giraulti* (16.2), and *N. giraulti* (IntG 12.1). Each strain maintains *Wolbachia* infections of the A supergroup. The IntG line was generated by repeatedly backcrossing *N. vitripennis* 12.1 females to uninfected *N. giraulti* RV2R males for nine generations to generate a line that contains *w*VitA-infected cytoplasm of *N. vitripennis* in the genetic background of *N. giraulti*[279]. Each strain was maintained under constant light at 25°C and raised on flesh fly pupae (*Sarcophaga bullata*). The transfected line of the Mediterranean flour moth *Ephestia kuehniella* harboring *Wolbachia* strain *w*CauB was obtained from Takema Fukatsu and Tetsuhiko Sasaki[280]. Moths were maintained at 24°C and 70% humidity on a diet consisting of wheat bran, glycerol, and dried yeast (20:2:1 [wt/wt]).

Whole insects were suspended in sterile SM buffer and homogenized to release the viruses from the animal tissue. Viral particles were PEG precipitated as previously described[289] and filtered through an 0.22-m filter. Viral DNA was extracted using the Qiagen MinElute Virus Spin kit, amplified using the Qiagen

REPLI-g minikit, and sequenced on the Illumina HiSeq 2000 platform with paired-end reads (2x 100 bp).

Bioinformatics

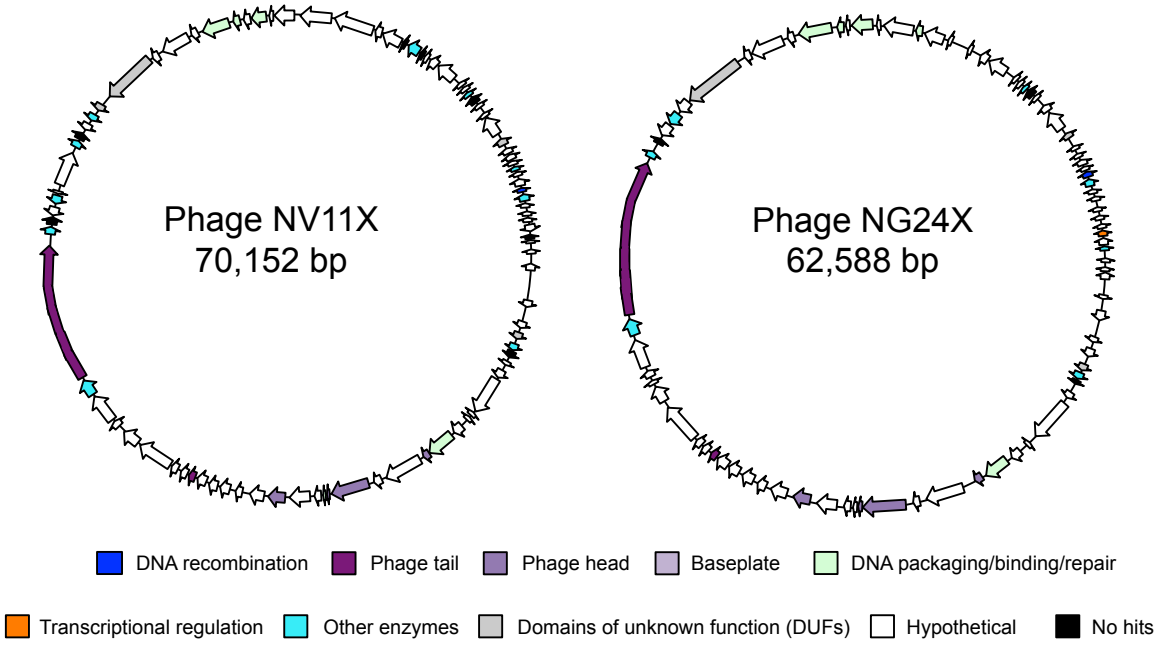
Mate-pair reads from the viromes were analyzed using the iVirus pipeline[286]. First, the sequences were trimmed using Trimmomatic 0.35.0[324] and quality checked using FastQC. *De novo* assembly of mate-pair reads was completed using SPAdes 3.6.0[325] with a k-mer value of 63 and default parameters. Assembly quality was determined by QUAST[326] and is reported in S3.1 Table in the supplemental material. All samples were coassembled with SPAdes 3.6.0 with a k-mer of 63 to generate a single reference file and run through VirSorter[327] in addition to the single assemblies. Viral contigs less than 500 bp and with coverage of less than five were removed from further analysis. Reads were then mapped back to the VirSorter viral contig outputs to estimate the relative abundance of each viral contig for each sample. BowtieBatch[286] was used to run bowtie2 on all samples of the coassembled contigs and produced BAM output files read by Read2RefMapper to generate relative abundance and coverage plots for each viral contig within each metagenome. To consider a contig present within an individual sample, reads from that sample needed to cover 75% of the viral

contig from the coassembled virome. Venn diagrams were generated using the VennDiagram package[328] through R v 3.3.2 software to display the overlap of contigs in different gut compartments using the mapping data from Read2RefMapper[286]. Relative abundance plots were illustrated using GraphPad's Prism v 7.0c. Viral protein cluster diversity was determined using vContact through the iVirus pipeline[286;329] and visualized using Gephi 0.9.2[330].

The VirSorter output for single assemblies was used for taxonomic classification against the NCBI-nr protein database. All taxonomic classifications were determined using the top BLASTx hit, with a threshold score of 50 on BLAST bitscore. Open reading frames (ORFs) were predicted and annotated using Prokka v1.12.0[331]. Additionally, protein families (PFAMs) within the ORFs were identified with InterProScan v5.26.65[332]. jModelTest v2.1.7[333] was performed to determine the optimal model of host gene evolution and, using this model (PHYML with the JC69 substitution model), a phylogenetic tree was constructed for the host species (as previously described in reference 1 [24]) from a nucleotide alignment of the mitochondrial cytochrome oxidase subunit I (COI) genes. Virome similarities as determined by Bray-Curtis beta diversity unweighted pair group method with arithmetic mean (UPGMA) clustering were determined using read coverage counts of viral contigs. These count profiles were rarefied 10 times

to a depth of 16,400 counts for each host virome to normalize for differential sequencing coverage. Bray-Curtis beta diversity and resulting UPGMA clustergrams between host viromes were calculated, and UPGMA trees were averaged to generate a consensus clustergram across the rarefied community profiles. Phylosymbiosis as measured through topological similarity between the host phylogeny and the virome clustergram was evaluated using the rooted Robinson-Foulds and rooted matching cluster methods previously described[24]. Significance was determined by comparing the observed degree of congruence to the congruence obtained across 100,000 randomized tree topologies using a custom script with methods previously described[24].

Supplementary Information



Supplementary Figure 3.1: Complete *Xenorhabdus* phage genomes. Two similar (94% nucleotide similarity) circular *Xenorhabdus* phage genomes were recovered, one from *N. vitripennis* (Phage NV11X) and the other from *N. giraulti* (Phage NG24X).

	***N. vitripennis* 12.1**	***N. longicornis* 2.1**	***N. girault*i IntG**	***N. girault*i 16.2**
# reads (bp)	**55,018,874**	**16,967,060**	**54,305,634**	**51,603,058**
# contigs (>=0 bp)	5,338	10,288	11,154	7,625
# contigs (>=1000 bp)	658	961	1,677	1,060
# contigs (>=5000 bp)	136	198	352	241
# contigs (>=10000 bp)	82	105	191	146
# contigs (>=25000 bp)	34	36	75	83
# contigs (>=50000 bp)	13	6	23	36
Total length (>=0 bp)	6,233,827	7,400,287	14,313,288	10,974,448
Total length (>=1000 bp)	4,574,682	4,746,396	10,687,227	8,488,860
Total length (>=5000 bp)	3,539,370	3,281,465	8,066,250	6,923,197
Total length (>=10000 bp)	3,164,282	2,636,751	6,950,497	6,263,109
Total length (>=25000 bp)	2,380,066	1,550,630	5,146,661	5,246,785
Total length (>=50000 bp)	1,631,890	488,296	3,224,513	3,479,271
# contigs	1,478	2,247	3,730	2,388
GC (%)	40.57	42.08	39.73	39.63
N50	21,830	8,245	16,642	34,618
N75	3,076	1,710	2,686	4,157
L50	42	125	120	65
L75	219	545	629	269
#N's per 100 kbp	0	0	0	0

Supplemental Table 3.1: Assembly statistics.

Supplementary Table 3.2: Pfam assignments in viral metagenomes.

https://msystems.asm.org/content/3/6/e00131-18/figures-only

CHAPTER IV

American Gut: Gut Microbiota Diversity across Ethnicities in the United States[3]

Author Contributions

This study was performed by Andrew Brooks (AB), Sambhawa Priya (SP), Ran Blekhman (RB), and Seth Bordenstein (SB). SP and RB performed the random forest modeling, with input provided by AB and SB. AB performed all other analyses in the manuscript, under the direction of SB. Everyone helped write and edit the manuscript.

Introduction

Abstract

Composed of hundreds of microbial species, the composition of the human gut microbiota can vary with chronic diseases underlying health disparities that disproportionally affect ethnic minorities. However, the influence of ethnicity on

[3] This work is published in *PLOS Biology*: **Brooks AW**, Priya S, Blekhman R, Bordenstein SR. (2018) Gut Microbiota Diversity across Ethnicities in the United States. https://doi.org/10.1371/journal.pbio.2006842

the gut microbiota remains largely unexplored and lacks reproducible generalizations across studies. By distilling associations between ethnicity and differences in two US-based 16S gut microbiota data sets including 1,673 individuals, we report 12 microbial genera and families that reproducibly vary by ethnicity. Interestingly, a majority of these microbial taxa, including the most heritable bacterial family, Christensenellaceae, overlap with genetically associated taxa and form co-occurring clusters linked by similar fermentative and methanogenic metabolic processes. These results demonstrate recurrent associations between specific taxa in the gut microbiota and ethnicity, providing hypotheses for examining specific members of the gut microbiota as mediators of health disparities.

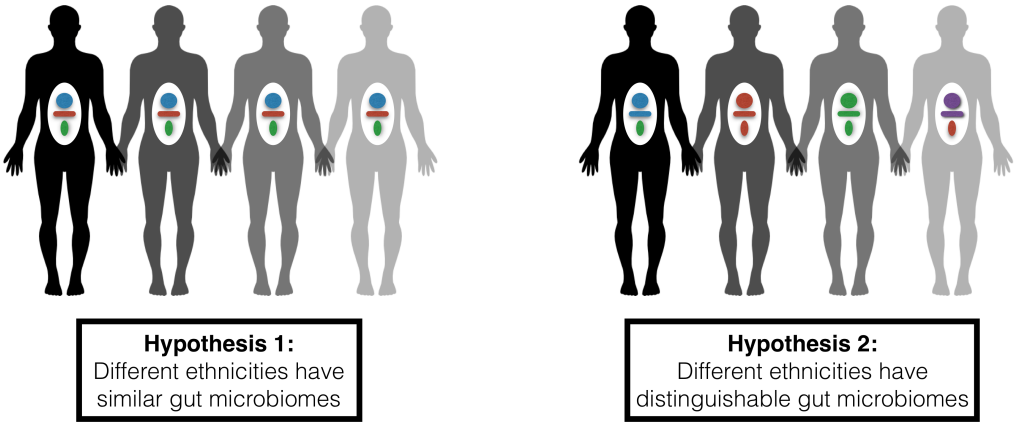


Fig 4.1. Graphical abstract of ethnicity-specific microbiota composition.

Author Summary

Understanding microbiota similarities and differences across ethnicities has the potential to advance approaches aimed at personalized microbial discovery and treatment, particularly those involved in ethnic health disparities. Here, we explore whether or not self-declared ethnicity consistently varies with gut microbiota composition across 1,673 healthy individuals in the United States. We find subtle but significant differences in taxonomic composition between four ethnicities, and we replicate these results across two study populations. Within the gut microbiota of Americans, there are at least 12 microbial taxa, which reproducibly vary in abundance across ethnicities. These taxa tend to correlate in abundance and metabolic functions and overlap with previously identified taxa that are associated with human genetic variation. We discuss the roles these taxa play in digestion and disease and propose hypotheses for how they may relate to ethnic health disparities. This study highlights the need to consider and potentially account for ethnic diversity in microbiota research and therapeutics.

Introduction

The human gut microbiota at fine resolution varies extensively between

individuals [6;7;334], and this variability frequently associates with diet[4;30;31;335], age[4;8;11], sex[4;11;336], body mass index (BMI)[4;7], and diseases presenting as health disparities[337-340]. The overlapping risk factors and burden of many chronic diseases disproportionally affect ethnic minorities in the United States, yet the underlying biological mechanisms mediating these substantial disparities largely remain unexplained. Recent evidence is consistent with the hypothesis that ethnicity associates with variation in microbial abundance, specifically in the oral cavity, gut, and vagina[121;122;341]. To varying degrees, ethnicity can capture many facets of biological variation including social, economic, and cultural variation, as well as aspects of human genetic variation and biogeographical ancestry. Ethnicity also serves as a proxy to characterize health disparity incidence in the United States, and while factors such as genetic admixture create ambiguity of modern ethnic identity, self-declared ethnicity has proven a useful proxy for genetic and socioeconomic variation in population scale analyses, including in the Human Microbiome Project[87;342;343]. Microbiota differences have been documented across populations that differ in ethnicity as well as in geography, lifestyle, and sociocultural structure; however, these global examinations cannot disconnect factors such as intercontinental divides and hunter-gatherer versus western lifestyles from ethnically structured differences[98;99;344]. Despite the importance of

understanding the interconnections between ethnicity, microbiota, and health disparities, there are no reproducible findings about the influence of ethnicity on differences in the gut microbiota and specific microbial taxa in diverse United States populations, even for healthy individuals[4].

Here, we comprehensively examine connections between self-declared ethnicity and gut microbiota differences across more than a thousand individuals sampled by the American Gut Project (AGP, N=1375)[1] and the Human Microbiome Project (HMP, N=298)[4]. Previous studies demonstrated that human genetic diversity in the HMP associates with differences in microbiota composition[9], and genetic population structure within the HMP generally delineates self-declared ethnicity[87]. Ethnicity was not found to have a significant association with microbiota composition in a Middle Eastern population, however factors such as lifestyle and environment that influence microbiota variation across participants was homogenous compared to the ethnic, sociocultural, economic, and dietary diversity found within the United States[35]. While ethnic diversity is generally underrepresented in current microbiota studies, evidence supporting an ethnic influence on microbiota composition among first generation immigrants has been recently demonstrated in a Dutch population[345]. The goal of this examination is to evaluate, for the first time, if there are reproducible

differences in gut microbiota across ethnicities within an overlapping United States population, as ethnicity is one of the key defining factors for health disparity incidence in the United States. Lifestyle, dietary, and genetic factors all vary to different degrees across ethnic groups in the United States, and it will require more even sampling of ethnic diversity and stricter phenotyping of study populations to disentangle which factors underlie ethnic microbiota variation in the AGP and HMP.

Results

Microbiota are Subtly Demarcated by Ethnicity

We first evaluate gut microbiota distinguishability between AGP ethnicities (Fig 4.2A, family taxonomic level, Asians-Pacific Islanders (N=88), Caucasians (N=1237), Hispanics (N=37), and African Americans (N=13)), sexes (female (N=657), male (N=718)), age groups (years grouped by decade), and categorical BMI (underweight (N=70), normal (N=873), overweight (N=318), and obese (N=114)) (Demographic details in S4.1A Table). Age, sex, and BMI were selected as covariates because they are consistent across the AGP and HMP datasets. Additionally, 31 other categorical factors measuring diet, environment, and geography were compared for pairwise differences between two ethnicities

using proportions tests, and very few (10 / 894) tests significantly varied (S4.1 Table additional sheets). Interindividual gut microbiota heterogeneity clearly dominates; however, Analyses of Similarity (ANOSIM) reveal subtle but significant degrees of total microbiota distinguishability for ethnicity, BMI, and sex, but not for age (Fig 4.2B, Ethnicity; Fig 4.2C, BMI; Fig 4.2D, Sex; Fig 4.2E, Age)[57]. Recognizing that subtle microbiota distinguishability between ethnicities may be spurious, we independently replicate the ANOSIM results from HMP African Americans (N=10), Asians (N=34), Caucasians (N=211) and Hispanics (N=43) (S4.2A Table, R=0.065, p=0.044). We again observe no significant distinguishability for BMI, sex, and age in the HMP. Higher rarefaction depths increase microbiota distinguishability in the AGP across various beta diversity metrics and categorical factors (S4.2B Table), and significance increases when individuals from overrepresented ethnicities are subsampled from the average beta diversity distance matrix (S4.2C Table). Supporting the ANOSIM results, Permutational Multivariate Analysis of Variance (PERMANOVA) models with four different beta diversity metrics showed that while all factors had subtle but significant associations with microbiota variation when combined in a single model, effect sizes were highest for ethnicity in 7 out of 8 comparisons across beta diversity metrics and rarefaction depths in the AGP and HMP (S4.2D Table).

We additionally test microbiota distinguishability by measuring the correlation between beta diversity and ethnicity, BMI, sex, and age with an adapted BioEnv test (S4.2E Table)[346]. Similar degrees of microbiota structuring occur when all factors are incorporated (Spearman Rho=0.055, p-values: Ethnicity=0.057, BMI<0.001, Sex<0.001, Age=0.564). Firmicutes and Bacteroidetes dominated the relative phylum abundance, with each representing between 35% and 54% of the total microbiota across ethnicities (S4.1 Fig).

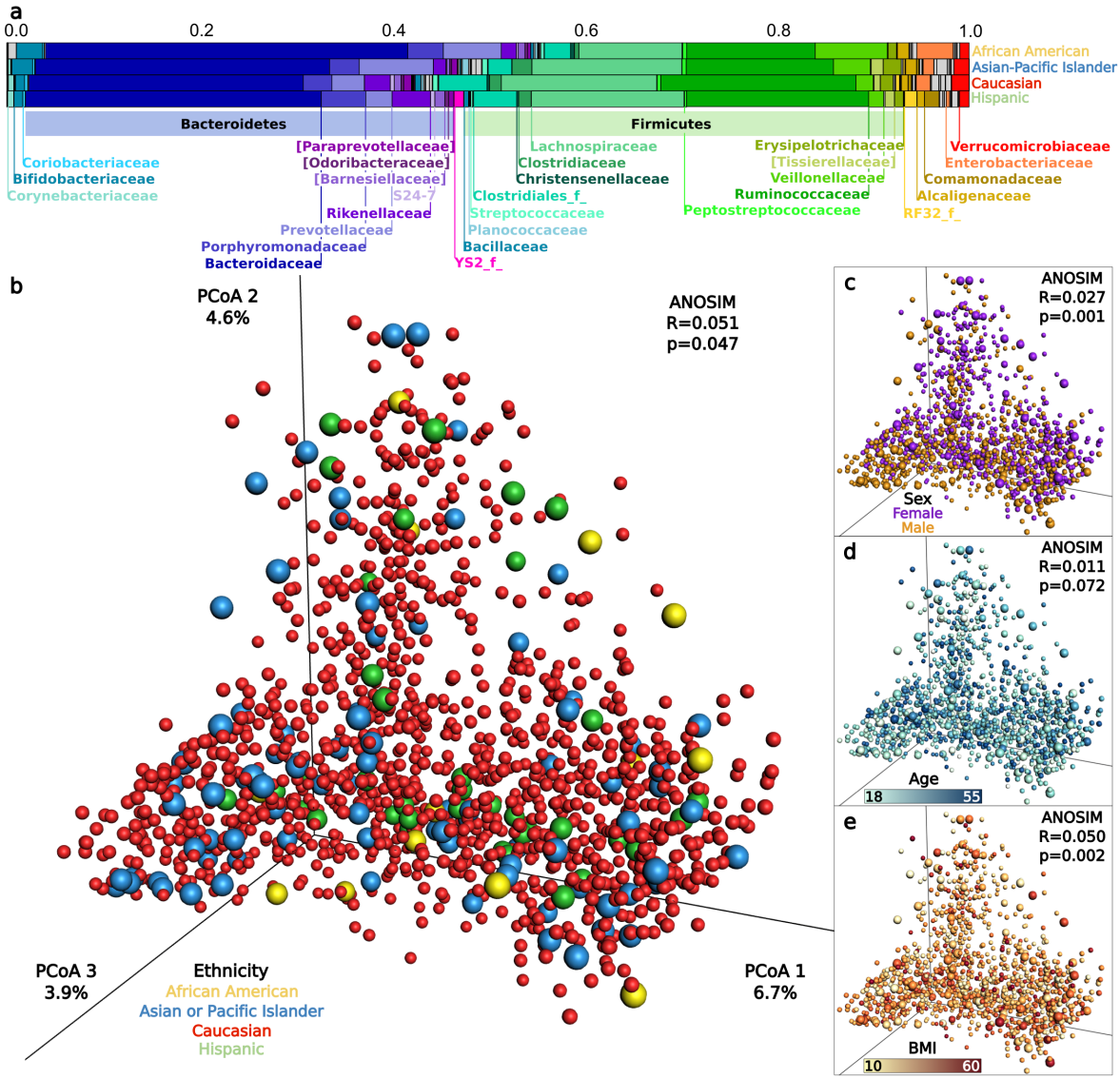


Fig 4.2. Gut microbiota composition and distinguishability by ethnicity, sex, age, and BMI. (A) The average relative abundance of dominant microbial families for each ethnicity. (B-E) Principle coordinates analysis plots of microbiota Bray-Curtis beta diversity and ANOSIM distinguishability for: (B) Ethnicity, (C) Sex, (D) Age, (E) BMI. In B-E, each point represents the microbiota of a single sample, and colors reflect metadata for that sample. Caucasian points are reduced in size

to allow clearer visualization, and *p*-values are not corrected across factors that have different underlying population distributions.

We next test for ethnicity signatures in the gut microbiota by analyzing alpha and beta diversity, abundance and ubiquity distributions, distinguishability, and classification accuracy[60]. Shannon's Alpha Diversity Index[188], which weights both microbial community richness (observed operational taxonomic units [OTUs]) and evenness (Equitability), significantly varies across ethnicities in the AGP data set (Kruskal-Wallis, $p = 2.8e\text{-}8$) with the following ranks: Hispanics > Caucasians > Asian-Pacific Islanders > African Americans (Fig 4.3A). In the HMP, there is a significantly lower Shannon diversity for Asian-Pacific Islanders relative to Caucasians and a trend of lower Shannon diversity for Asian-Pacific Islanders relative to Hispanics; African Americans change position in diversity relative to other ethnicities, potentially as a result of undersampling bias. Five alpha diversity metrics, two rarefaction depths, and separate analyses of Observed OTUs and Equitability generally confirm the results (S4.3A Table).

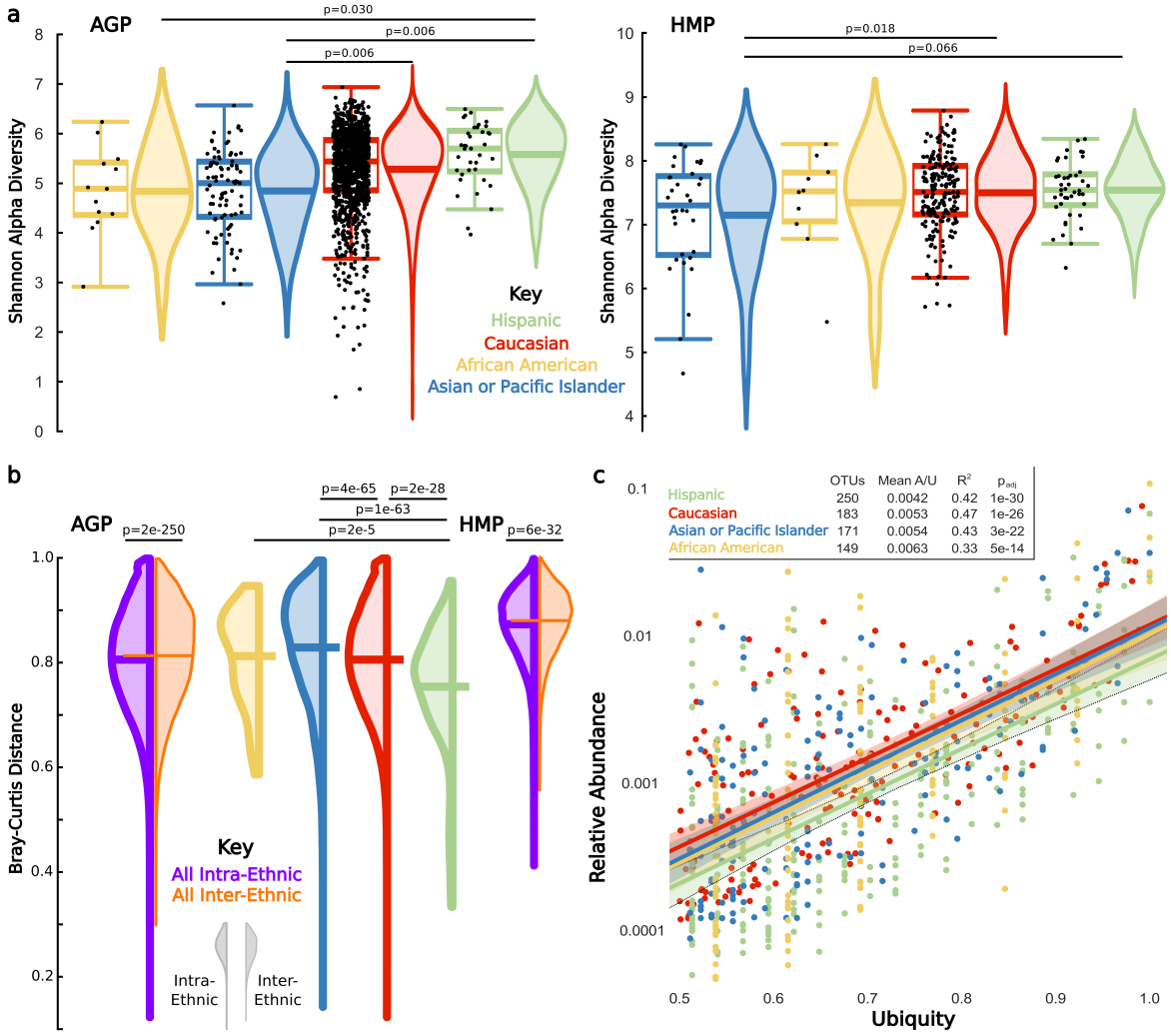


Fig 4.3. Ethnicity associates with diversity and composition of the gut microbiota. (A) Center lines of each boxplot depict the median by which ethnicities were ranked from low (left) to high (right); the lower and upper ends of each box represent the 25th and 75th percentiles, respectively; whiskers denote the 1.5 interquartile range; and black dots represent individual samples. Lines in the middle of violin plots depict the mean, and *p*-values are Bonferroni corrected within each data set. (B) Left extending violin plots represent intraethnic

distances for each ethnicity, and right extending violin plots depict all interethnic distances. Center lines depict the mean beta diversity. Significance bars above violin plots depict Bonferroni corrected pairwise Mann-Whitney U comparisons of the intra-intra- and intra-interethnic distances. (C) Within each ethnicity, OTUs shared by at least 50% of samples. Colored lines represent a robust ordinary least squares regression within OTUs of each ethnicity, shaded regions represent the 95% confidence interval, R^2 denotes the regression correlation, the OTUs column indicates the number of OTUs with >50% ubiquity for that ethnicity, Mean A/U is the average abundance/ubiquity ratio, and the p_{adj} is the regression significance adjusted and Bonferroni corrected for the number of ethnicities.

If ethnicity impacts microbiota composition, pairwise beta diversity distances (ranging from 1/completely dissimilar to 0/identical) will be greater between ethnicities than within ethnicities. While average gut microbiota beta diversities across all individuals are high (Bray-Curtis = 0.808), beta diversities between individuals of the same ethnicity (intraethnic, Bray-Curtis = 0.806) are subtly but significantly lower than those between ethnicities in both the AGP (interethnic, Bray-Curtis = 0.814) and HMP data sets (intraethnic, Bray-Curtis = 0.870 versus interethnic, Bray-Curtis = 0.877) (Fig 4.3B). We confirm AGP

results by subsampling individuals from over-represented ethnicities across beta metrics and rarefaction depths (S4.4A and S4.4B Table). Finally, we repeat analyses across beta metrics and rarefaction depths using only the average distance of each individual to all individuals from the ethnicity to which they are compared (S4.4C and S4.4D Table).

Next, we explore interethnic differences in the number of OTUs shared in at least 50% of individuals within an ethnicity, as the likelihood of detecting a biological signal is improved in more abundant organisms relative to noise that may predominate in lower abundance OTUs. Out of 5,591 OTUs in the total AGP data set, 101 (1.8%) OTUs meet this ubiquity cutoff in all ethnicities, and 293 (5.2%) OTUs meet the cutoff within at least one ethnicity. Hispanics share the most ubiquitous OTUs and have the lowest average abundance/ubiquity (A/U) ratio (Fig 4.3C), indicating stability, whereby stability represents a more consistent appearance of OTUs with lower abundance but higher ubiquity[172]. This result potentially explains their significantly lower intraethnic beta diversity distance and thus higher microbial community overlap relative to the other ethnicities (Fig 4.3B). Comparisons in the AGP between the higher sampled Hispanic, Caucasian, and Asian-Pacific Islander ethnicities also reveal a trend wherein higher intraethnic community overlap (Fig 4.3B) parallels higher

numbers of ubiquitous OTUs (Fig 4.3C), higher Shannon alpha diversity (Fig 4.3A), and higher stability of ubiquitous OTUs as measured by the A/U ratio (Fig 4.3C).

We next assess whether a single ethnicity disproportionately impacts total gut microbiota distinguishability in the AGP by comparing ANOSIM results from the consensus beta diversity distance matrix when each ethnicity is sequentially removed from the analysis (Fig 4.4A and S4.2E Table). Distinguishability remains unchanged when the few African Americans are removed but is lost upon removal of Asian-Pacific Islanders or Caucasians, likely reflecting their higher beta diversity distance from other ethnicities (Fig 4.4A). Notably, removal of Hispanics increases distinguishability among the remaining ethnicities, which may be due to a higher degree of beta diversity overlap observed between Hispanics and other ethnicities (S4.4B Table). Results conform across rarefaction depths and beta diversity metrics (S4.2F Table), and pairwise combinations show strong distinguishability between African Americans and Hispanics (ANOSIM, $R = 0.234$, $p = 0.005$) and Asian-Pacific Islanders and Caucasians (ANOSIM, $R = 0.157$, $p < 0.001$).

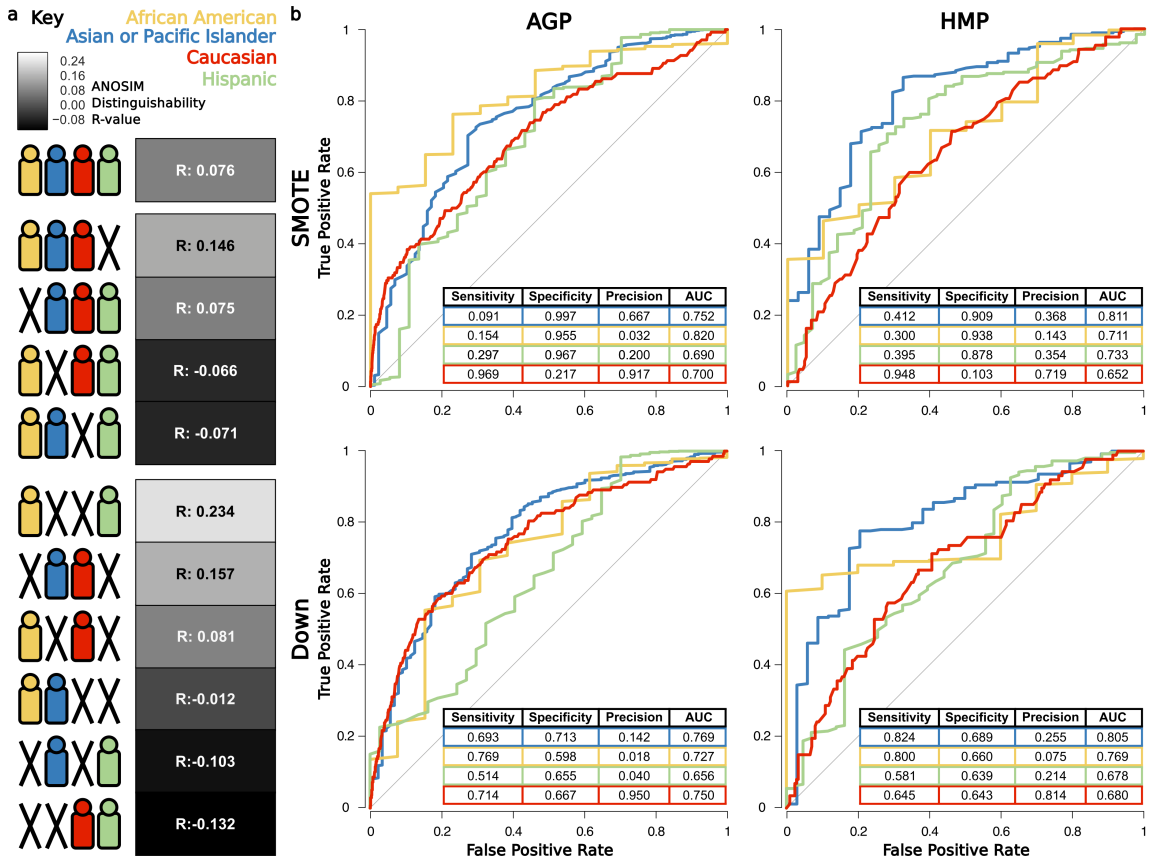


Fig 4.4. Microbiota distinguishability and classification ability across ethnicities. (A) ANOSIM distinguishability between all combinations of ethnicities. Symbols depict specific ethnicities included in the ANOSIM tests, and boxes denote the R-value as a heatmap, in which white indicates increasing and black indicates decreasing distinguishability relative to the R-value with all ethnicities. (B) Average ROC curves (for 10-fold cross-validation) and prediction performance metrics for one-versus-all RF classifiers for each ethnicity, using SMOTE[347] and down subsampling approaches for training.

Finally, to complement evaluation with ecological alpha and beta diversity, we implement a random forest (RF) supervised learning algorithm to classify gut microbiota from genus-level community profiles into their respective ethnicity. We build four one-versus-all binary classifiers to classify samples from each ethnicity compared to the rest and use two different sampling approaches to train the models synthetic minority oversampling technique (SMOTE)[347] and downsampling for overcoming uneven representation of ethnicities in both the data sets (see Materials and methods). Given that the area under the receiver operating characteristic (ROC) curve (or AUC) of a random guessing classifier is 0.5, the models classify each ethnicity fairly well (Fig 4.4B), with average AUCs across sampling techniques and data sets of 0.78 for Asian-Pacific Islanders, 0.76 for African Americans, 0.69 for Hispanics, and 0.70 for Caucasians. Ethnicity distinguishing RF taxa and out-of-bag error percentages appear in (S4.2 Fig).

Recurrent Taxon Associations with Ethnicity

Subtle to moderate ethnicity-associated differences in microbial communities may in part be driven by differential abundance of certain microbial taxa. 16.2% (130/802) of the AGP taxa and 20.6% (45/218) of HMP taxa across all classification levels (i.e., phylum to genus, S4.5 Table) significantly vary in

abundance across ethnicities (Kruskal-Wallis, $p_{FDR} < 0.05$). Between data sets, 19.2% (25/130) of the AGP and 55.6% (25/45) of the HMP varying taxa replicate in the other data set, representing a significantly greater degree of overlap than would be expected by chance (ethnic permutation analysis of overlap, $p < 0.001$ each taxonomic level and all taxonomic levels combined). The highest replication of taxa varying by abundance occurs with 22.0% of families (nine significant in both data sets / 41 significantly varying families in either data set), followed by genus with 13.4% (nine significant in both data sets / 67 significantly varying genera in either data set).

Among 18 reproducible taxa, we categorize 12 as taxonomically distinct (Fig 4.5) and exclude six in which nearly identical abundance profiles between family/genus taxonomy overlap. Comparing relative abundance differences between pairs of ethnicities for these 12 taxa in the AGP reveals 30 significant differences, of which 20 replicate in the HMP ($p < 0.05$, Mann- Whitney U). Intriguingly, all reproducible pairwise differences are a result of decreases in Asian-Pacific Islanders (Fig 4.5). We also test taxon abundance and presence/absence associations with ethnicity separately in the AGP using linear and logistic regression models, respectively, and we repeat the analysis while incorporating categorical sex and continuous age and BMI as covariates (S4.6

Table). Clustering microbial families based on their abundance correlation reveals two co-occurrence clusters: (i) a distinct cluster of six Firmicutes and Tenericutes families in the HMP and (ii) an overlapping but more diverse cluster of 20 families in the AGP (S4.3 Fig). Nine of the 12 taxa found to recurrently vary in abundance across ethnicities are represented in these clusters (Fig 4.5), with four appearing in both clusters and the other five appearing either in or closely correlated with members of both clusters (S4.3 Fig). Furthermore, 90% (18/20) of families in the AGP cluster and 66% (4/6) of taxa in the HMP cluster significantly vary in abundance across ethnicities. We also found overlap for AGP and HMP data sets between taxa significantly varying in abundance across ethnicities (with false discovery rate [FDR] < 0.05) and taxa in RF models with percentage importance greater than 50% for an ethnicity (S4.2B Fig). Taken together, these results establish general overlap of the most significant ethnicity-associated taxa between these methods, reproducibility of microbial abundances that vary between ethnicities across data sets, and patterns of co-occurrence among these taxa, which could suggest they are functionally linked.

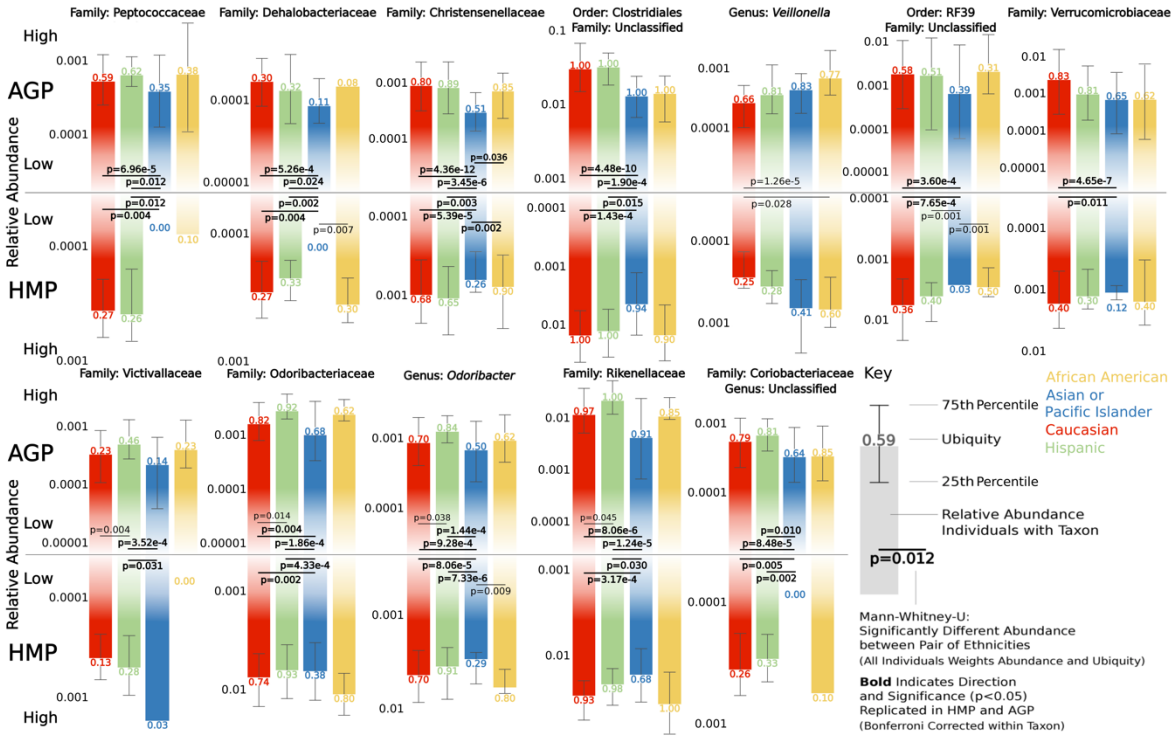


Fig 4.5. Ethnicity-associated taxa match between the HMP and AGP. Bar plots depict the log10 transformed relative abundance for individuals possessing the respective taxon within each ethnicity, ubiquity appears above (AGP) or below (HMP) bars, and the 25th and 75th percentiles are shown with extending whiskers. Mann-Whitney U tests evaluate differences in abundance and ubiquity for all individuals between pairs of ethnicities; for example, the direction of change in Victivallaceae is driven by ubiquity while abundance is higher for those possessing the taxon. Significance values are Bonferroni corrected for the six tests within each taxon and data set, and bold p-values indicate that significance ($p < 0.05$) and direction of change replicate in the AGP and HMP.

Most heritable taxon of bacteria varies by ethnicity

Identified as the most heritable taxon in the human gut[13;91], the family Christensenellaceae exhibits the second strongest significant difference in abundance across ethnicities in both AGP and HMP data sets (S4.5 Table, Family: AGP, Kruskal-Wallis, p_{FDR} = 1.55e-9; HMP, Kruskal-Wallis, p_{FDR} = 0.0019). Additionally, Christensenellaceae is variable by sex and BMI (AGP: Sex, Kruskal-Wallis, p_{FDR} = 1.22e-12; BMI, Kruskal-Wallis, p_{FDR} = 0.0020) and represents some of the strongest pairwise correlations with other taxa in both co-occurrence clusters (S4.3 Fig). There is at least an eight-fold and two-fold reduction in average Christensenellaceae abundance in Asian-Pacific Islanders relative to the other ethnicities in the AGP and HMP, respectively (S4.5 Table), and significance of all pairwise comparisons in both data sets show reduced abundance in Asian-Pacific Islanders (Fig 4.5). Christensenellaceae also occurs among the top 10 most influential taxa for distinguishing Asian-Pacific Islanders from other ethnicities using RF models for both AGP and HMP data sets (S4.2A Fig). Abundance in individuals possessing Christensenellaceae and presence/absence across all individuals significantly associate with ethnicity (S4.6 Table, Abundance, Linear Regression, $p_{Bonferroni}$ = 0.006; Presence/Absence, Logistic Regression,

$p_{Bonferroni}$ = 8.802e-6), but there was only slight correlation between the taxon's relative abundance and BMI (S4.4 Fig). Confirming previous associations with lower BMI[348], we observe that AGP individuals with Christensenellaceae also have a lower BMI (Mean BMI, 23.7 ± 4.3) than individuals without it (Mean BMI, 25.0 ± 5.9; Mann- Whitney U, $p < 0.001$). This pattern is separately reflected in African Americans, Asian-Pacific Islanders, and Caucasians but not Hispanics (Fig 4.6), suggesting that each ethnicity may have different equilibria between the taxon's abundance and body weight.

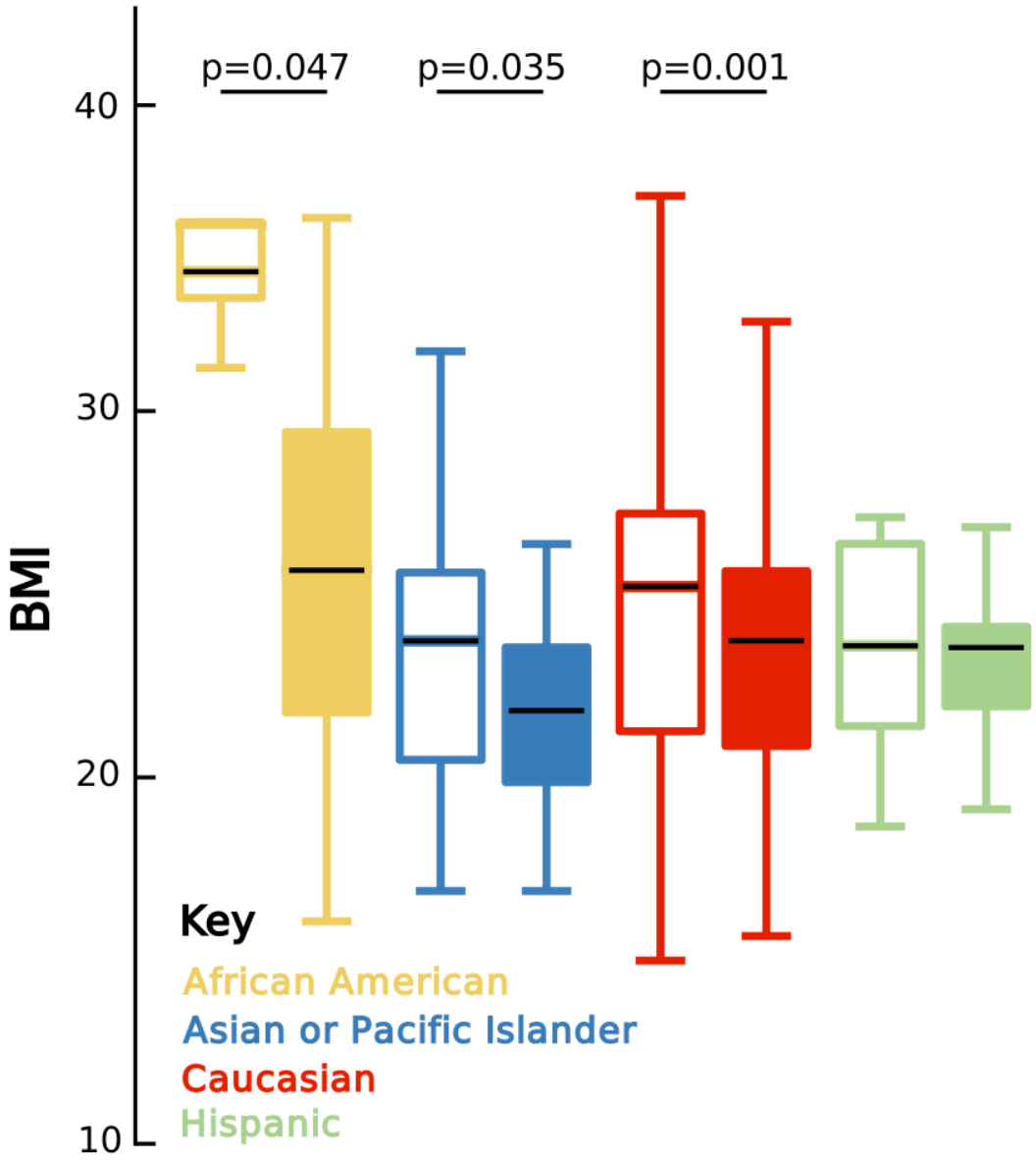


Fig 4.6. Christensenellaceae variably associate with BMI across ethnicities. Boxplots of BMI for individuals without (unfilled boxplots) and with (filled boxplots) Christensenellaceae. Significance was determined using one-tailed Mann-Whitney U tests for lower continuous BMI values. Black lines indicate the mean relative abundance; the lower and upper end of each box represent the 25th

and 75th percentiles, respectively; and whiskers denote the 1.5 interquartile range.

Genetic- and ethnicity-associated taxa overlap

Many factors associate with human ethnicity, including a small subset of population specific genetic variants (estimated approximately 0.5% genome wide) that vary by biogeographical ancestry[199;200]; self-declared ethnicity in the HMP is delineated by population genetic structure[87]. Here, we investigate whether ethnicity-associated taxa overlap with (i) taxa that have a significant population genetic heritability in humans[13;90;91;349] and (ii) taxa linked with human genetic variants in two large Genome-Wide Association Studies (GWAS)-microbiota analyses[13;90]. All recurrent ethnicity-associated taxa except one were heritable in at least one study, with seven replicating in three or more studies (Table 4.1). Likewise, abundance differences in seven recurrent ethnicity-associated taxa demonstrate significant GWAS associations with at least one variant in the human genome. Therefore, we assess whether any genetic variants associated with differences in microbial abundance exhibit significant rates of differentiation (fixation index [F_{ST}]) between 1,000 genome superpopulations[199;200]. Out of 49 variants associated with ethnically varying taxa, 21 have higher F_{ST} values

between at least one pair of populations than that of 95% of other variants on the same chromosome and across the genome; the F_{ST} values of five variants associated with Clostridiaceae abundance rank above the top 99% (S4.7 Table). Since taxa that vary across ethnicities exhibit lower abundance in Asian-Pacific Islanders, it is notable that the F_{ST} values of 18 and 11 variant comparisons for East Asian and South Asian populations, respectively, are above that of the 95% rate of differentiation threshold from African, American, or European populations. Cautiously, the microbiota and 1,000 genomes data sets are not drawn from the same individuals, and disentangling the role of genetic from social and environmental factors will still require more controlled studies.

Recurrent Ethnicity-Associated Taxa	Heritability	Genetic Associations
Family: Peptococcaceae	0.1213[A], 0.2154[C], 0.26[E]	rs143179968[E]
Family: Dehalobacteriaceae	0.6878[B], 0.3087[C]	
Family: Christensenellaceae	0.3819[A], 0.6170[B], 0.4230[C],	
Order: Clostridiales, Family: Unclassified	0.2914[A], 0.4020[B], 0.1330[C]	*40 Genetic Variants[C]
Genus: *Veillonella*	0.1370[A], 0.2168[D]	rs347941[C]
Order: RF39, Family: Unclassified	0.2341[A], 0.6618[B], 0.3074[C]	rs4883972[C]
Family: Verrucomicrobiaceae	0.1257[A], 0.5973[B], 0.1394[C]	
Family: Victivallaceae		
Family Odoribacteriaceae	0.1389[A], 0.1917[D], 0.34[E]	chr7:96414393[E], rs115795847[E]
Genus: *Odoribacter*	0.1916[D]	
Family: Rikenellaceae	0.1299[D], 0.29[E]	rs17098734[C], rs3909540[C], rs147600757[E]
Family: Coriobacteriaceae, Genus: Unclassified	0.1364[A], 0.2822[B], 0.1609[C]	rs9357092[E]

Table 4.1. Most recurrent ethnicity-associated taxa are previously reported heritable and genetically-associated taxa. The table shows population genetic heritability estimates and associated genetic variants for the 12 recurrent ethnically varying taxa. The minimum heritability cutoff was chosen as >0.1, and only exactly

overlapping taxonomies were considered. Studies examined: [A]UKTwins (2014, 'A' measure of additive heritability in ACE model)[91], [B]Yatsunenko (2014, 'A' measure of additive heritability in ACE model)[91], [C]UKTwins (2016, 'A' measure of additive heritability in ACE model)[13], [D]Lim (2016, H2r measure of polygenic heritability in SOLAR[350])[349], [E]Turpin (2016, H2r measure of polygenic heritability in SOLAR[350]). *indicates excessive variants were excluded from table.

Discussion

Many common diseases associate with microbiota composition and ethnicity, raising the central hypothesis that microbiota differences between ethnicities can occasionally serve as a mediator of health disparities. Self-declared ethnicity in the US can capture socioeconomic, cultural, geographic, dietary, and genetic diversity, and a similarly complex array of interindividual and environmental factors influence total microbiota composition. This complexity may result in challenges when attempting to recover consistent trends in total gut microbiota differences between ethnicities. The challenges in turn emphasize the importance of reproducibility, both through confirmation across analytical methods and replication across study populations[87;121;122;341;345;351]. In order to robustly substantiate the ethnicity-microbiota hypothesis, we evaluated recurrent associations between self-declared ethnicity and variation in both total gut

microbiota and specific taxa in healthy individuals. Results provide hypotheses for examining specific members of the gut microbiota as mediators of health disparities.

Our findings from two American data sets demonstrate that (i) ethnicity consistently captures gut microbiota with a slightly stronger effect size than other variables such as BMI, age, and sex; (ii) ethnicity is moderately predictable from total gut microbiota differences; and (iii) 12 taxa recurrently vary in abundance between the ethnicities, of which the majority have been previously shown to be heritable and associated with human genetic variation. Whether shaped through socioeconomic, dietary, healthcare, genetic, or other ethnicity-related factors, reproducibly varying taxa represent sources for novel hypotheses addressing health disparities. For instance, the family Odoribacteriaceae and genus *Odoribacter* are primary butyrate producers in the gut, and they have been negatively associated to severe forms of Crohns disease and Ulcerative Colitis in association with reduced butyrate metabolism[352-354]. Asian-Pacific Islanders possess significantly less Odoribacteriaceae and *Odoribacter* than Hispanics and Caucasians in both data sets, and severity of Ulcerative Colitis upon hospital admission has been shown to be significantly higher in Asian Americans[355]. Considering broader physiological roles, several ethnicity-associated taxa are

primary gut anaerobic fermenters and methanogens[356;357] and associate with lower BMI and blood triglyceride levels[348;358]. Indeed, Christensenellaceae, Odoribacteriaceae, *Odoribacter*, and the class Mollicutes containing RF39 negatively associate with metabolic syndrome and demonstrate significant population genetic heritability in twins[349]. Implications for health outcomes warrant further investigation but could be reflected by positive correlations of Odoribacteriaceae, *Odoribacter*, Coriobacteriaceae, Christensenellaceae, and the dominant Verrucomicrobiaceae lineage *Akkermansia* with old age[359;360]. *Akkermansia* associations with health and ethnicity in Western populations may reflect recently arising dietary and lifestyle effects on community composition, as this mucus-consuming taxon is rarely observed in more traditional cultures globally[99]. Moreover, these findings raise the importance of controlling for ethnicity in studies linking microbiota differences to disease because associations between specific microbes and a disease could be confounded by ethnicity of the study participants.

Based on correlations in individual taxon's abundance, a similar pattern of co-occurrence previously identified as the "Christensenellaceae Consortium" includes 11 of the 12 recurrent ethnically varying taxa[91], and members of this consortium associate with genetic variation in the human formate oxidation gene,

aldehyde dehydrogenase 1 family member 1 (*ALDH1L1*), which is a genetic risk factor for stroke[13;361;362]. Formate metabolism is a key step in the pathway reducing carbon dioxide to methane[363;364], and increased methane associates with increased Rikenellaceae, Christensenellaceae, Odoribacteriaceae, and *Odoribacter*[365]. Products of methanogenic fermentation pathways include short chain fatty acids such as butyrate, which, through reduction of proinflammatory cytokines, is linked to cancer cell apoptosis and reduced risk of colorectal cancer[366;367]. Asian Americans are the only ethnic group where cancer surpasses heart disease as the leading cause of death, and over 70% of Asian Americans were born overseas, which can affect assimilation into Western lifestyles, leading to reduced access to healthcare and screening[366;368-370]. Preliminary results from other groups suggest that the gut microbiome of Southeast Asian immigrants changes after migration to the US[371]. Indeed, as countries in Asia shift toward a more Western lifestyle, the incidence of cancers, particularly gastrointestinal and colorectal cancers, are increasing rapidly, possibly indicating incompatibilities between traditionally harbored microbiota and Western lifestyles[372-375]. Asian Americans have higher rates of type 2 diabetes and pathogenic infections than Caucasians[376], and two metagenomic functions enriched in control versus type 2 diabetes cases appear to be largely conferred by cluster-associated butyrate-

producing and motility-inducing Verrucomicrobiaceae and Clostridia taxa reduced in abundance among AGP and HMP Asian-Pacific Islanders[337]. Both induction of cell motility and butyrate promotion of mucin integrity can protect against pathogenic colonization and associate with microbial community changes[337;367;377]. Levels of cell motility and butyrate are key factors suspected to underlie a range of health disparities including inflammatory bowel disease, arthritis, and type 2 diabetes[337;378-380]. Patterns of ethnically varying taxa across ethnicities could result from many factors including varying diets, environmental exposures, sociocultural influences, human genetic variation, and others. However, regardless of the mechanisms dictating assembly, these results suggest that there is a reproducible, co-occurring group of taxa linked by similar metabolic processes known to promote homeostasis.

The utility of this work is establishing a framework for studying ethnicity-associated taxa and hypotheses of how changes in abundance or presence of these taxa may or may not shape health disparities, many of which also have genetic components. Differing in allele frequency across three population comparisons and associated with the abundance of Clostridiales, the genetic variant rs7587067 has a significantly higher frequency in African (minor allele frequency [MAF] = 0.802) versus East Asian (MAF = 0.190, F_{ST} = 0.54, Chromosome = 98.7%,

Genome-Wide = 98.9%, See Methods), admixed American (MAF = 0.278, F_{ST} = 0.44, Chromosome = 99.0%, Genome-Wide = 99.1%), and European populations (MAF = 0.267, F_{ST} = 0.45, Chromosome = 98.7.3%, Genome-Wide = 98.7%). This intronic variant for the gene *HECW2* is a known expression quantitative trait locus (eQTL) (GTEx, eQTL Effect Size = -0.18, p = 7.4e-5)[381;382], and *HECW2* encodes a ubiquitin ligase linked to enteric gastrointestinal nervous system function through maintenance of endothelial lining of blood vessels[383;384]. Knockout of *HECW2* in mice reduced enteric neuron networks and gut motility, and patients with Hirschsprung's disease have diminished localization of *HECW2* to regions affected by loss of neurons and colon blockage when compared to other regions of their own colon and healthy individuals[385]. Hirschsprung's disease presenting as full colon blockage is rare and has not undergone targeted examination as a health disparity; however, a possible hypothesis is that lower penetrance of the disease in individuals with the risk allele at rs7587067 could lead to subtler effects on gut motility resulting in Clostridiales abundance differences.

Despite the intrigue of connecting the human genome, microbiota, and disease phenotypes, evaluating such hypotheses will require more holistic approaches including incorporating metagenomics and metabolomics to identify

whether enzymes or metabolic functions reproducibly vary across ethnicities, as well as direct functional studies in model systems to understand if correlation is truly driven by causation. Further limitations should also be considered, including recruitment biases for the AGP versus HMP, variation in sample processing and OTU clustering, and uneven sampling, which could only be addressed with downsampling of over-represented ethnicities. Still, despite these confounders, care was taken to demonstrate the reproducibility of results across statistical methods, ecological metrics, rarefaction depths, and study populations. Summarily, this work suggests that abundance differences of specific taxa, rather than whole communities, may represent the most reliable ethnic signatures in the gut microbiota. A reproducible co-occurring subset of these taxa link to a variety of overlapping metabolic processes and health disparities and contain the most reproducibly heritable taxon, Christensenellaceae. Moreover, a majority of the microbial taxa associated with ethnicity are also heritable and genetically associated taxa, suggesting that there is a possible connection between ethnicity and genetic patterns of biogeographical ancestry that may play a role in shaping these taxa. Our results emphasize the importance of sampling ethnically diverse populations of healthy individuals in order to discover and replicate ethnicity signatures in the human gut microbiota, and they highlight a need to account for

ethnic variation as a potential confounding factor in studies linking microbiota differences to disease. Further reinforcement of these results may lead to generalizations about microbiota assembly and even consideration of specific taxa as potential mediators or treatments of health disparities.

Materials and methods

Ethics statement

Access to HMP data was obtained through dbGaP approval granted to SRB and AWB. Institutional Review Board approval was granted with nonhuman subjects determination IRB161231 by Vanderbilt University.

Data acquisition

AGP data was obtained from the project FTP repository located at ftp://ftp.microbio.me/AmericanGut/. AGP data generation and processing prior to analysis can be found at https://github.com/biocore/American-Gut/tree/master/ipynb/primary-processing. All analyses utilized the rounds-1-25 data set, which was released on March 4, 2016. Throughout all analyses, QIIME v1.9.0 was used in an Anaconda environment (https://continuum.io) for all script calls, and custom scripts and notebooks were run in the QIIME 2 Anaconda

environment with python version 3.5.2, and plots were postprocessed using Inkscape (https://inkscape.org/en/)[238]. Ethnicity used in this study was self-declared by AGP study participants as one of four groups: African American, Asian or Pacific Islander (Asian-Pacific Islander), Caucasian, or Hispanic. Sex was self-declared as either male, female, or other. Age was self-declared as a continuous integer of years old, and age categories defined by the AGP by decade (i.e., 20's, 30's, etc.) were used in this study. BMI was self-declared as an integer, and BMI categories defined by AGP of underweight, healthy, overweight, and obese were utilized. A total of 31 categorical metadata factors were assessed for structuring across ethnicities with a two proportion Z test between pairs of ethnicities using a custom python script (S4.1 Table additional sheets). The *p*-values were Bonferroni corrected within each metadata factor for the number of pairwise ethnic comparisons. 97% OTUs generated for each data set are utilized throughout to maintain consistency with other published literature; however, microbial taxonomy of the HMP is reassigned using the Greengenes reference database[386]. Communities characterized with 16S rDNA sequencing of variable region four followed an identical processing pipeline for all samples, which was developed and optimized for the Earth Microbiome Project[387]. HMP 16S rDNA data processed using QIIME for variable regions 3-5 was obtained from

http://hmpdacc.org/HMQCP/. Demographic information for individual HMP participants was obtained through dbGaP restricted access to study phs000228.v2.p1, with dbGaP approval granted to SRB and nonhuman subjects determination IRB161231 granted by Vanderbilt University. Ethnicity and sex were assigned to subjects based on self-declared values, with individuals selecting multiple ethnicities being removed unless they primarily responded as Hispanic, while categorical age and BMI were established from continuous values using the same criteria for assignment as in the AGP. The HMP Amerindian population was removed due to severe under-representation. This filtered HMP table was used for community level analyses (ANOSIM, alpha diversity, beta intra-inter); however, to allow comparison with the AGP data set, community subset analyses (cooccurrence, abundance correlation, etc.) were performed with taxonomic assignments in QIIME using the UCLUST method with the GreenGenes_13_5 reference.

Quality control

AGP quality control was performed in Stata v12 (StataCorp, 2011) using available metadata to remove samples (Raw N = 9,475) with BMI more than 60 (removed −988 total remaining [8,487]) or less than 10 (−68 [8,419]); missing

age (−661 [7,758]), with age greater than 55 years old (−2,777 [4,981]) or less than 18 years old (−582 [4,399]); and blank samples or those not appearing in the mapping file (−482 [3,917]), with unknown ethnicity or declared as other (−131 [3786]), not declared as a fecal origin (−2,002 [1784]), with unknown sex or declared as other (−98 [1686]) or located outside of the US (−209 [1477]). No HMP individuals were missing key metadata or had other reasons for exclusion (−0 [298]). Final community quality control for both the AGP and HMP was performed by filtering OTUs with less than 10 sequences and removing samples with less than 1,000 sequences (AGP, −102 [1375]; HMP, −0 [298]). All analyses used 97% OTUs generated by the AGP or HMP, and unless otherwise noted, results represent Bray-Curtis beta diversity and Shannon alpha diversity at a rarefaction depth of 1,000 counts per sample.

ANOSIM, PERMANOVA, and BioEnv distinguishability

The ANOSIM test was performed with 9,999 repetitions on each rarefied table within a respective rarefaction depth and beta diversity metric (Fig 4.2 and S4.2A-S4.2B Table), with R values and *p*-values averaged across the rarefactions. Consensus beta diversity matrices were calculated as the average distances across the 100 rarefied matrices for each beta diversity metric and depth. Consensus

distance matrices were randomly subsampled 10 times for subset number of individuals from each ethnic group with more than that subset number prior to ANOSIM analysis with 9,999 repetitions, and the results were averaged evaluating the effects of more even representations for each ethnicity (S4.2C Table). Consensus distance matrices had each ethnicity and pair of ethnicities removed prior to ANOSIM analysis with 9,999 repetitions, evaluating the distinguishability conferred by inclusion of each ethnicity (Fig 4.4A, S4.2F Table). Significance was not corrected for the number of tests to allow comparisons between results of different analyses, metrics, and depths. PERMANOVA analyses were run using the R language implementation in the Vegan package[58], with data handled in a custom R script using the Phyloseq package[246]. Categorical variables were used to evaluate the PERMANOVA equation (Beta Diversity Distance Matrix ~ Ethnicity + Age + Sex + BMI) using 999 permutations to evaluate significance, and the R and *p*-values were averaged across 10 rarefactions (S4.2D Table). The BioEnv test, or BEST test, was adapted to allow evaluation of the correlation and significance between beta diversity distance matrices and age, sex, BMI, and ethnicity simultaneously (S4.2E Table)[346]. At each rarefaction depth and beta diversity metric, the consensus distance matrix was evaluated for its correlation with the centered and scaled Euclidian distance matrix of individuals

continuous age and BMI, and categorical ethnicity and sex encoded using patsy (same methodology as original test) (https://patsy.readthedocs.io/en/latest/#). The test was adapted to calculate significance for a variable of interest by comparing how often the degree of correlation with all metadata variables (age, sex, BMI, ethnicity) was higher than the correlation when the variable of interest was randomly shuffled between samples 1,000 times.

Alpha diversity

Alpha diversity metrics (Shannon, Simpson, Equitability, Chao1, Observed OTUs) were computed for each rarefied table (QIIME: alpha_diversity.py), and results were collated and averaged for each sample across the tables (QIIME: collate_alpha.py). Pairwise nonparametric *t* tests using Monte Carlo permutations evaluated alpha diversity differences between the ethnicities with Bonferroni correction for the number of comparisons (Fig 4.3A, S4.3 Table, QIIME: compare_alpha_diversity.py). A Kruskal-Wallis test implemented in python was used to detect significant differences across all ethnicities.

Beta diversity

Each consensus beta diversity distance matrix had distances organized

based on whether they represented individuals of the same ethnic group or were between individuals of different ethnic groups. All values indicate that all pairwise distances between all individuals were used (Fig 4.3B, S4.4A and S4.4B Table), and mean values indicate that for each individual, their average distance to all individuals in the comparison group was used as a single point to assess pseudo-inflation (S4.4C and S4.4D Table). A Kruskal-Wallis test was used to calculate significant differences in intraethnic distances across all ethnicities. Pairwise Mann-Whitney U tests were calculated between each pair of intraethnic distance comparisons, along with intra-versus-interethnic distance comparisons. Significance was Bonferroni corrected within the number of intra-intraethnic and intra-interethnic distance groups compared, with violin plots of intra- and interethnic beta diversity distances generated for each comparison.

Random forest

RF models were implemented using taxa summarized at the genus level, which performed better compared to RF models using OTUs as features, both in terms of classification accuracy and computational time. We first rarefied OTU tables at a sequence depth of 10,000 (using R v3.3.3 package *vegan's* rrarefy() function) and then summarized rarefied OTUs at the genus level (or a higher

characterized level if genus was uncharacterized for an OTU). We filtered for rare taxa by removing taxa present in fewer than half of the number of samples in rarest ethnicity (i.e., fewer than 10/2 = 5 samples in HMP and 13/2 = 6 [rounded down] in AGP), retaining 85 distinct taxa in the HMP data set and 322 distinct taxa in the AGP data set at the genus level. The resulting taxa were normalized to relative abundance and arcsin-sqrt transformed before being used as features for the RF models. We initially built a multiclass RF model, but since the RF model is highly sensitive to the uneven representation of classes, all samples were identified as the majority class, i.e., Caucasian. In order to even out the class imbalance, we considered some sampling approaches, but most existing techniques for improving classification performance on imbalanced data sets are designed for binary class imbalanced data sets and are not effective on data sets with multiple under-represented classes. Hence, we adopted the binary classification approach and built four one-versus-all binary RF classifiers to classify samples from each ethnicity compared to the rest. 10-fold cross-validation (using R package *caret* [388]) was performed using ROC as the metric for selecting the optimal model. The performance metrics and ROC curves were averaged across the 10 folds (Fig 4.4B). Without any sampling during training the classifiers, most samples were identified as the majority class, i.e., Caucasian, by

all four one-versus-all RF classifiers. In order to overcome this imbalance in class representation, we applied two sampling techniques inside cross-validation: i) downsampling and ii) SMOTE[347]. In the downsampling approach, the majority class is downsampled by random removal of instances from the majority class. In the SMOTE approach, the majority class is downsampled, and synthetic samples from the minority class are generated based on the k-nearest neighbors technique[347]. Note, the sampling was performed inside cross-validation on training set, while the test was performed on unbalanced held-out test set in each fold. In comparison to a no-sampling approach, which classified most samples as the majority class, i.e., Caucasians, our sampling-based approach leads to improved sensitivity for classification of minority classes on unbalanced test sets. Nevertheless, the most accurate prediction remains for the inclusion in the majority class. The ROC curves and performance metrics table in Fig 4.4B show the sensitivity-specificity tradeoff and classification performance for one-versus-all classifier for each ethnicity for both the sampling techniques applied on both of the data sets. For both of the data sets, downsampling shows higher sensitivity and lower specificity and precision for minority classes (i.e., African Americans, Asian-Pacific Islanders, and Hispanics) compared to SMOTE. However, for the majority class (i.e., Caucasian), downsampling lowers the sensitivity and

increases the specificity and precision compared to SMOTE. The sensitivity-specificity tradeoff, denoted by the AUC, is reduced for Hispanics in both the data sets. The most important taxa with >50% importance for predicting an ethnicity using RF model with SMOTE sampling approach are shown in S4.2A Fig. Among the 10 most important taxa for each ethnicity, there are nine taxa that overlap between the AGP and HMP data sets (highlighted by the blue rectangular box); however, which ethnicity, they best distinguish varies between the two data sets. Within each data set we highlighted taxa that are distinguishing in RF models and have distinguishing differential abundance in S4.2B Fig, reporting both the FDR corrected significance for Kruskal-Wallis tests of differential abundance, and the percent importance for the most distinguished ethnicity of each in RF models. We also report out-of-bag errors for the final RF classifier that was built using the optimal model parameters obtained from cross-validation approach corresponding to each ethnicity and sampling procedure for both AGP and HMP data sets in S4.2C Fig.

Taxon associations

Taxon differential abundance across categorical metadata groups was performed in QIIME (QIIME: group_significance.py, S4.5 Table) to examine

whether observation counts (i.e., OTUs and microbial taxon) are significantly different between groups within a metadata category (i.e., ethnicity, sex, BMI, and age). The OTU table prior to final community quality control was collapsed at each taxonomic level (i.e., Phylum-Genus; QIIME: collapse_taxonomy.py), with counts representing the relative abundance of each microbial taxon. Differences in the mean abundance of taxa between ethnicities were calculated using Kruskal-Wallis nonparametric statistical tests. *p*-values are provided alongside false discovery rate and Bonferroni corrected *p*-values, and taxon were ranked from most to least significant. Results were collated into excel tables by taxonomic level and metadata category being examined, with significant (FDR and Bonferroni *p*-value < 0.05) highlighted in orange, and taxa that were false discovery rate significant in both data sets were colored red. The Fisher's exact test for the overlap of number of significant taxa between data sets was run at the online portal (http://vassarstats.net/tab2x2.html), with the expected overlap calculated as 5% of the number of significant taxa at all levels within the respective data set, and the observed 25 taxa that overlapped in our analysis. The permutation analysis was performed by comparing the number of significant taxa (S4.5 Table, $p_{FDR} < 0.05$) overlapping between the AGP and HMP to the number overlapping when the Kruskal-Wallis test was performed 1,000 times with

ethnicity randomly permuted. In 1/ 1,000 runs, there was one significant taxon overlapping at the family level and one in 3/1,000 permutations at the genus level, with no significant taxa overlapping in any repetitions at higher taxonomic levels. The 12 families and genera that were significantly different were evaluated to not be taxonomically distinct if their abundances across ethnicities at each level represented at least 82%-100% (nearly all >95%) of the overlapping taxonomic level, and the genera was used if classified and family level used if genera was unclassified (g__). Average relative abundances on a log10 scale among individuals possessing the taxon were extracted for each taxon within each ethnicity, and the abundance for 12 families and genera were made into bar chart figures (Fig 4.5). The external whisker (AGP above, HMP below) depicts the 75th percentile of abundance, and the internal whisker depicts the 25th percentile. Pairwise Mann-Whitney U tests were performed between each pair of ethnicities using microbial abundances among all individuals and were Bonferroni corrected for the six comparisons within each taxon and data set. Bonferroni significant *p*-values are shown in the figure and shown in bold if significance and direction of change replicate in both data sets. Ubiquity shown above or below each bar was calculated as the number of individuals in which that taxon was detected within the respective ethnicity. Additional confirmation of ethnically varying abundance

was also performed at each taxonomic level (S4.6 Table), at which the correlation of continuous age and BMI along with categorically coded sex and ethnicity were simultaneously measured against the log10 transformed relative abundance of each taxon among individuals possessing it using linear regression (S4.6 Table, Abundance) and against the presence or absence of the taxon in all individuals with logistic regression (S4.6 Table, Presence Absence). Significance is presented for the models each with ethnicity alone and with all metadata factors included (age, sex, BMI), alongside Bonferroni corrected *p*-values and individual effects of each metadata factor.

Co-occurrence analysis

Bacterial taxonomy was collapsed at the family level, Spearman correlation was calculated between each pair of families using SciPy[389], cluster maps were generated using seaborn (S4.3 Fig), and ethnic associations were drawn from S4.5 Table. Correlations were masked where Bonferroni corrected Spearman *p*-values were >0.05, and clusters were identified as the most prominent (strongest correlations) and abundance enriched. Enrichment of ethnic association was evaluated by measuring the Mann-Whitney U of cluster families' ethnic associations (*p*-values, S4.5 Table) compared to the ethnic associations of

noncluster taxa. Cluster-associated families were identified as having at least three significant correlations with families within the cluster.

Christensenellaceae analysis

The abundance of the family Christensenellaceae was input as relative abundance across all individuals from the family level taxonomic table. Individuals were subset based on the presence/absence of Christensenellaceae, and BMIs were compared using a one-tailed Mann-Whitney U test, then each was further subset by ethnicity and BMI compared using one-tailed Mann-Whitney U tests and boxplots within each ethnicity (Fig 4.6).

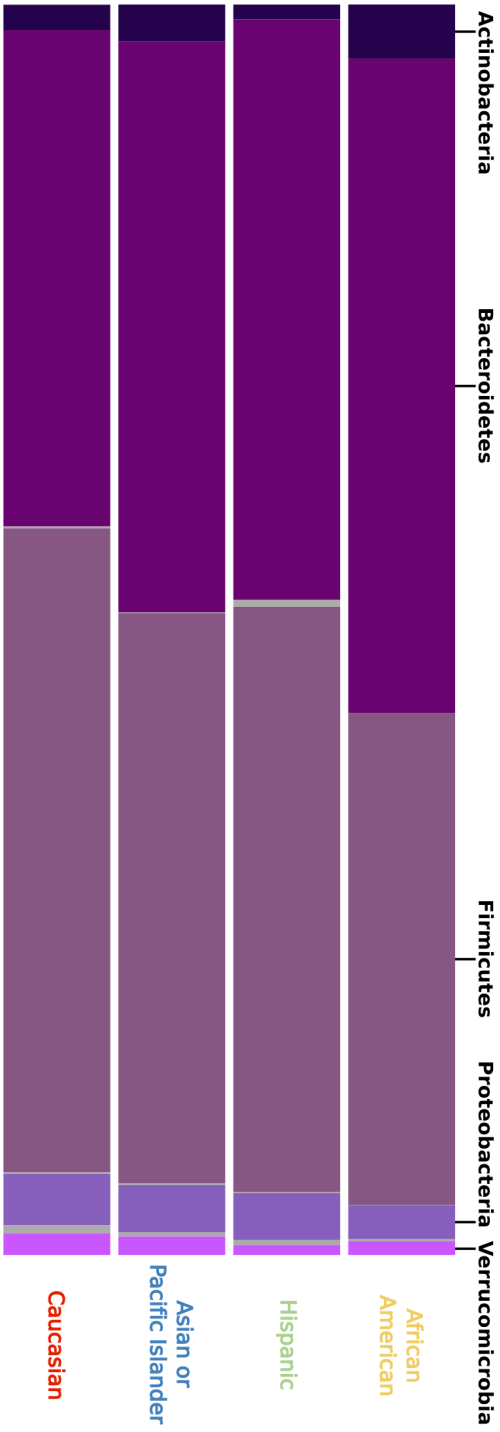
Genetically associated, heritable, and correlated taxa analysis

Genetically associated taxa from population heritability studies[13;90;91;349] with a minimum heritability (A in ACE models or H2r) >0.1 and from GWAS studies[13;90] were examined for exact taxonomic overlap with our 12 ethnically-associated taxa. The 42 genetic variants associated with Unclassified Clostridiales are rs16845116, rs586749, rs7527642, rs10221827, rs5754822, rs4968435, rs17170765, rs1760889, rs6933411, rs2830259, rs7318523, rs17763551, rs2248020, rs1278911, rs185902, rs2505338, rs6999713, rs5997791, rs7236263,

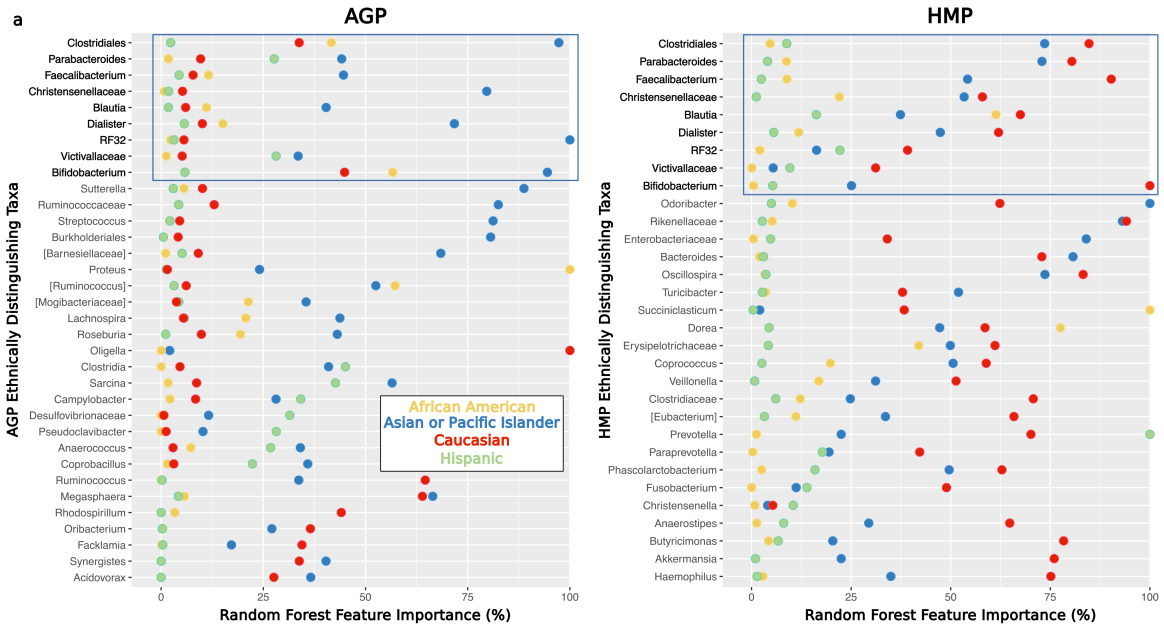
rs10484857, rs9938742, rs1125819, rs4699323, rs641527, rs7302174, rs2007084, rs2293702, rs9350764, rs2170226, rs2273623, rs9321334, rs6542797, rs9397927, rs2269706, rs4717021, rs7499858, rs10148020, rs7524581, rs11733214, and rs7587067 from[13]. These 40 variants along with variants in Table 4.1 except for chr7:96414393 (total = 49) were then assessed in 1,000 Genomes individuals for significant differentiation across superpopulations[390]. The 1,000 Genomes VCF files were downloaded (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/), and variants with a minor allele frequency less than 0.01 were removed, with F_{ST} calculated between each pair of superpopulations using vcftools[391]. The East Asian versus South Asian F_{ST} rates were not used in the analysis. A custom script was used to examine the F_{ST} for each of the 49 variants and was compared to the F_{ST} of all variants on the same chromosome and all variants genome-wide for that pair of populations, with percentile calculated and the number of variants with a higher F_{ST} divided by the total number of variants. The eQTL value and significance for rs7587067 were drawn from the GTEx database[382].

Supporting information

Supporting Figures



S4.1 Fig. The average relative abundance of dominant microbial phyla for each ethnicity.



b

AGP Taxonomic Lineage	Differential Abundance FDR	RF % Importance for Most Distinguished Ethnicity
Oligella	6.81E-19	100% for Caucasian
Bifidobacterium	1.63E-5	94.49% for Asian-Pacific Islander
Ruminococcaceae	3.92E-6	82.47% for Asian-Pacific Islander
Christensenellaceae	7.35E-9	79.64% for Asian-Pacific Islander
Dialister	0.0358	71.73% for Asian-Pacific Islander
Barnesiellaceae	0.0106	68.41% for Asian-Pacific Islander
Megasphaera	1.57E-5	66.45% for Asian-Pacific Islander

HMP Taxonomic Lineage	Differential Abundance FDR	RF % Importance for Most Distinguished Ethnicity
Odoribacter	0.0006	100% for Asian-Pacific Islander
Succiniclasticum	1.31E-7	100% for African American
Rikenellaceae	0.0042	94.09% for Caucasian
Bacteroides	0.0102	80.69% for Asian-Pacific Islander
Dorea	0.0362	77.50% for African American
Erysipelotrichaceae	0.0362	61.10% for Caucasian
Christensenellaceae	0.0027	57.95% for Caucasian
Veillonella	0.0432	51.32% for Caucasian

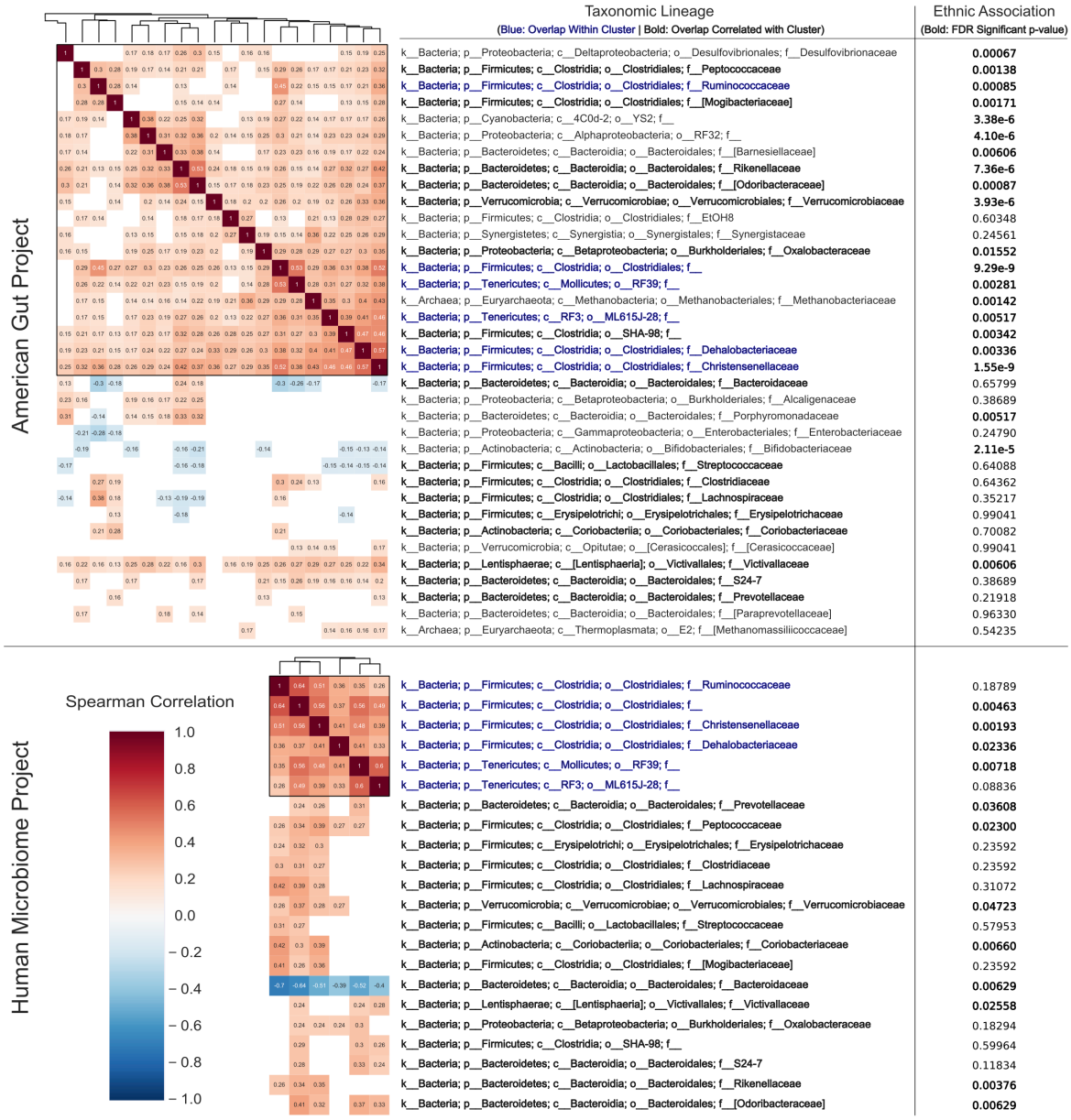
c

Sampling Method	African American	Asian or Pacific Islander	Caucasian	Hispanic
SMOTE	8.79%	11.69%	11.70%	8.49%
Down	42.31%	31.82%	29.35%	60.81%

Sampling Method	African American	Asian or Pacific Islander	Caucasian	Hispanic
SMOTE	10.00%	7.56%	10.67%	12.29%
Down	40.00%	33.82%	38.51%	43.02%

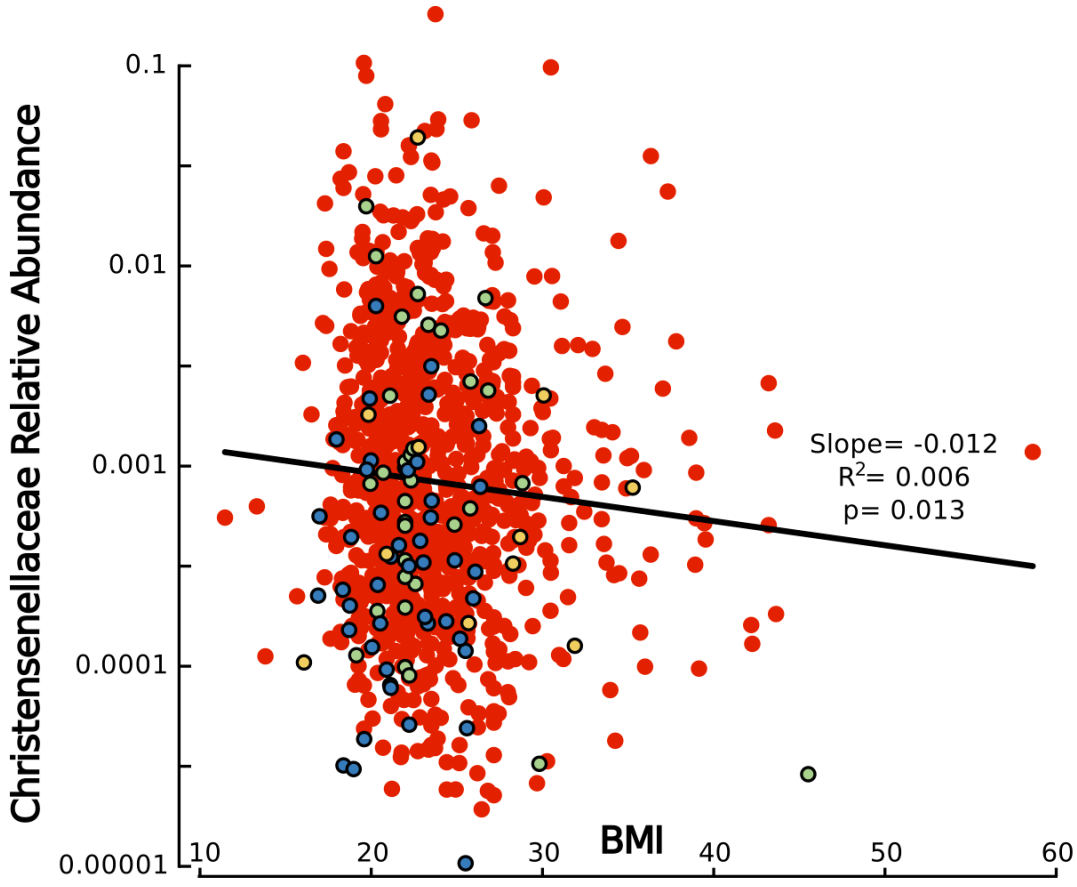
S4.2 Fig. Summary of RF distinguishing taxa and out-of-bag error for each ethnicity. (A) Importance of taxa for predicting each ethnicity using RF models with SMOTE sampling approach are shown as percentage contributions, highlighted by color for each ethnicity. Among the 10 most important taxa for each ethnicity, nine overlap between the AGP and HMP data sets (highlighted by the blue rectangular box); however, which ethnicity they best distinguish varies between the two data sets. (B) Taxa that are distinguishing in RF models and have distinguishing differential abundance in S4.5 Table. The FDR corrected

significance for Kruskal-Wallis tests of differential abundance and the percent importance for the most distinguished ethnicity of each in RF models are shown. (C) Out-of-bag error percentages for the final RF classifier that was built using the optimal model parameters obtained from cross-validation approach corresponding to each ethnicity and sampling procedure for both AGP and HMP datasets.



S4.3 Fig. Abundance correlation of microbial families. Spearman correlation cluster maps of bacterial abundance for families in the AGP and HMP. Numbers within boxes depict the spearman correlation value with heatmap coloration from blue negative correlation (−1), white no correlation (0), to red positive correlation (1). Positions have been masked based on Bonferroni significance

<0.05 for the total cluster map of all microbial families. Taxa within boxes were identified as a highly correlated cluster, and taxa outside the boxes share multiple correlations with those within the cluster. Blue taxonomic names indicate overlap of taxa within boxes of both the AGP and HMP, while black indicate multiple correlations with the clusters in both data sets. The ethnic association column depicts FDR corrected *p*-values from Kruskal-Wallis tests in S4.5 Table, which are bolded if <0.05.



S4.4 Fig. Correlation of BMI with Christensenellaceae abundance. The

relationship for each individual between log10 transformed Christensenellaceae abundance on the y-axis and BMI on the x-axis, with statistics slope, R^2, and *p* fit with a linear regression. Coloration of each point indicates ethnicity: yellow, African American; blue, Asian-Pacific Islander; green, Hispanic; red, Caucasian.

Supporting Tables

https://doi.org/10.1371/journal.pbio.2006842.s005

S4.1 Table. Demographic information for the AGP. Breakdown of age and BMI by sex and ethnicity. Heatmaps were constructed within each statistic and category (bounded by black box). The means for all sex and ethnic groups were used as the center (white), with higher values indicated in red and lower in blue. HMP data is not shown because of data access restrictions on participant metadata, available through dbGaP application. Additional sheets depict proportions tests of ethnic structuring for 31 metadata factors, each on their own sheet.

https://doi.org/10.1371/journal.pbio.2006842.s006

S4.2 Table. Microbiota distinguishability by ethnicity, age, sex, and BMI. (A) AGP and HMP ANOSIM distinguishability by ethnicity, age, sex, and BMI at a

rarefaction depth of 1,000 and across four ecological metrics (more details in table). (B) AGP ANOSIM distinguishability by ethnicity, age, sex, and BMI at rarefaction depths of 1,000 and 10,000. (C) ANOSIM results for consensus distance matrix while subsampling the maximum number of individuals from each ethnic group. (D) BioEnv results of correlation between ethnicity, age, sex, and BMI together with outcome as multivariate beta diversity distance matrices (Distance Matrix = Ethnicityx1 + Categorical Agex2 + Categorical BMIx3 + Sexx4 + B). (E) ANOSIM results for consensus distance matrix when each ethnicity and group of ethnicities are sequentially removed from the analysis.

https://doi.org/10.1371/journal.pbio.2006842.s007

S4.3 Table. Alpha diversity by ethnicity, age, sex, and BMI. Alpha diversity for ethnicity, age, sex, and BMI across varying rarefaction depths and beta diversity metrics in the AGP (Fig 4.5A and Fig 4.5C-4.5E) and for ethnicity in the HMP (Fig 4.5B). Results are based on nonparametric permutation-based *t* tests, and *p*-values are Bonferroni corrected within each factor of interest, depth, and metric.

https://doi.org/10.1371/journal.pbio.2006842.s008

S4.4 Table. Comparison of beta diversity distances for within and between ethnicities. All values depicted are Mann-Whitney U *p*-values. (A) All distances between pairs of individuals within each ethnicity were compared between ethnicities across rarefaction depths 1,000 and 10,000, four beta diversity metrics, and with subsampling over-represented ethnicities. (B) All distances between pairs of individuals within and between each ethnicity were compared between ethnicities. (C) Mean distances between pairs of individuals within each ethnicity were compared between ethnicities. (D) Mean distances between pairs of individuals within and between each ethnicity were compared between ethnicities.

https://doi.org/10.1371/journal.pbio.2006842.s009

S4.5 Table. Taxa that are differentially abundant by ethnicity, sex, BMI, and age in the AGP and HMP. Kruskal-Wallis results for differential taxa abundance across metadata groupings, including FDR and Bonferroni corrected *p*-values, and taxa abundance averages within each group. Metadata factors and taxonomic levels are separated by excel tabs.

https://doi.org/10.1371/journal.pbio.2006842.s010

S4.6 Table. Taxa that are correlated with ethnicity, sex, BMI, and age in the AGP. Results of linear (Abundance) and logistic (Presence Absence) regression results for differential taxa abundance across metadata factors separated by taxonomic level. Columns in order indicate the taxon name, the number of individuals with nonzero abundance; then the *p*-value for ethnicity alone, the *p*-value Bonferroni corrected, the f-test statistic, and R^2; then the same values for the regression with ethnicity, age, sex, and BMI together; then the abundances in each ethnic group; and finally the *p*-values for each factor broken down.

https://doi.org/10.1371/journal.pbio.2006842.s011

S4.7 Table. Genetic variants with taxa associations and detailed 1,000 Genomes population differentiation rates (FST). Variants in red indicate the variant has at least one F_{ST} above the 95th percentile for high differentiation between at least one pair of populations. Columns I-BU represent the values for calculating variant F_{ST} and percentiles. The first two spaces indicate the two superpopulations being compared. F_{ST} indicates the rate of differentiation for that variant between that pair of populations. Higher indicates the number of variants genome- wide with a higher F_{ST}, and total indicates the total genome-wide variants examined. The columns with chromosome indicate the number of

variants with higher F_{ST} and total variants on the same chromosome as the variant of interest. Percent indicates the number of variants with a higher F_{ST} divided by the total number of variants.

CHAPTER V

Vanderbilt Microbiome Initiative[4]

Author Contributions

Timothy Olszewski, Katie A. Friese, and Dr. Heidi J. Silver at the Vanderbilt Nutrition Center recruited and screened study subjects. Timothy Olszewski obtained subject consent, supervised visits, and gathered samples at the Nutrition Center. Andrew Brooks stored and aliquoted fecal and oral samples and then performed DNA extractions and Illumina library preparation prior to sequencing. Andrew Brooks worked with Karen Beeri at Vantage to perform metagenomic sequencing. Holly M. Smith and Jane F. Ferguson stored blood and urine samples, submitted samples to Metabolon for metabolomics profile generation, and extracted DNA from blood samples for human genotyping which was performed at Vantage. James C. Poland and John A. McLean performed fecal metabolomics profiling with funding provided by the Vanderbilt Institute for Infection, Immunology, and Inflammation to Andrew Brooks. William Beavers and Eric P. Skaar performed the metallomics profiling. Angela M. Eeds, Akos

[4] Analysis of the VMI clinical trial is currently underway, and this chapter will be updated once significant results have been generated. A multi'omics publication describing results from the first 18 participants will be prepared.

Ledeczi, Hamid Zare, Andrew Brooks, and Seth R. Bordenstein worked on study questionnaires and the survey application. Seth R. Bordenstein was the principle investigator and supervised all aspects of the project.

Introduction

'Omics describes techniques measuring many components of a complex biological system, and rapid technological advances in data generation and processing now allow researchers to look at multiple 'omics systems throughout the human body simultaneously. Combinatorial multi'omic studies allow deeper insights about interconnections between different physiological systems, particularly when measuring characteristics of human hosts, the trillions of microorganisms that call them home, and the multitude of abiotic molecules on which both depend. However, the novelty of multi'omic studies leads to challenges in the cost of data generation, limited tools for inter-dataset comparison, and researchers lacking experience in the caveats specific to each type of 'omics data. Yet a more holistic view of human health can emerge with multi'omics, particularly if cause and effect can be attributed across systems. Multi'omics are particularly promising for incorporating the composition and genetic capacities of complex human-associated microbiomes with host systems

like metabolism and immunity. Additionally, each 'omics profile serves as a complex snapshot in time, but temporal multi'omics sampling may be key to understanding how different system affect one another. Finally, diet is frequently the dominant explanatory variable of microbiome composition, making interpersonal dietary diversity a key confounding variable in human studies. The Vanderbilt Microbiome Initiative (VMI) seeks to comprehensively address these points by leveraging: multi'omics measures of systems throughout the human body, replication of those measures temporally, and performing this temporal sampling as individuals progress through a week-long controlled dietary intervention.

Sample Location	**'Omics**	**Time Points**	**Profiling**
Oral (Saliva)	Metagenomics	2	Oral microbiome composition and functional capacity
Fecal	Metagenomics	3-8	Gut microbiome composition and functional capacity
Fecal	Metabolomics	3-8	Gut metabolites
Fecal	Metallomics	3-8	Gut trace mineral abundance
Fecal	Viromics	3-8	Gut viral and phage genomes
Blood	Metabolomics	2	Bloodstream circulating metabolites

Blood	Genomics	2	Human SNPs genome-wide
Urine	Metabolomics	2	Urine excreted metabolites

Table 5.1: 'Omics datasets sampled in the VMI clinical trial (N=18). Yellow shades depict samples that were used in multiple 'omics methods. Orange shades depict 'omics methods that were performed at multiple body sites.

As in human genetics research, representative diversity has been lacking in the early days of microbiome research, which is predominantly represented by white, ancestrally-European individuals. A focus of the VMI is to capture multiethnic diversity across a range of 'omics systems. Further, interpersonal diversity is to be weighed across 'omics as individuals shift western diets to a shared vegetarian diet for four days, before returning to their own unique western diet. As many factors can affect the microbiome, some confounding influences were restricted in selection criteria, including: sexual diversity (all female), age diversity (18-40 years), dietary diversity (all western) and body mass index diversity (BMI – normal). Ideally, a study population should represent as much of humanities' diversity as possible to improve extensibility of results to a wider sampling of people. Unfortunately, study size and diversity was limited to account

for costs of participant diets and multi'omics, challenges in broader ethnic recruitment, and maintaining statistical power to detect multi'omic and inter-ethnic variation. Other factors were controlled for in an attempt to address two questions: 1) whether intra'omics distinguishability between ethnicities will increase or decrease on a shared diet, and 2) whether multi'omics profiles will compositionally converge on a shared diet, and diverge upon returning to unique western diets. Currently, all of the datasets (except for the fecal viromics) have been generated and prepared for analysis.

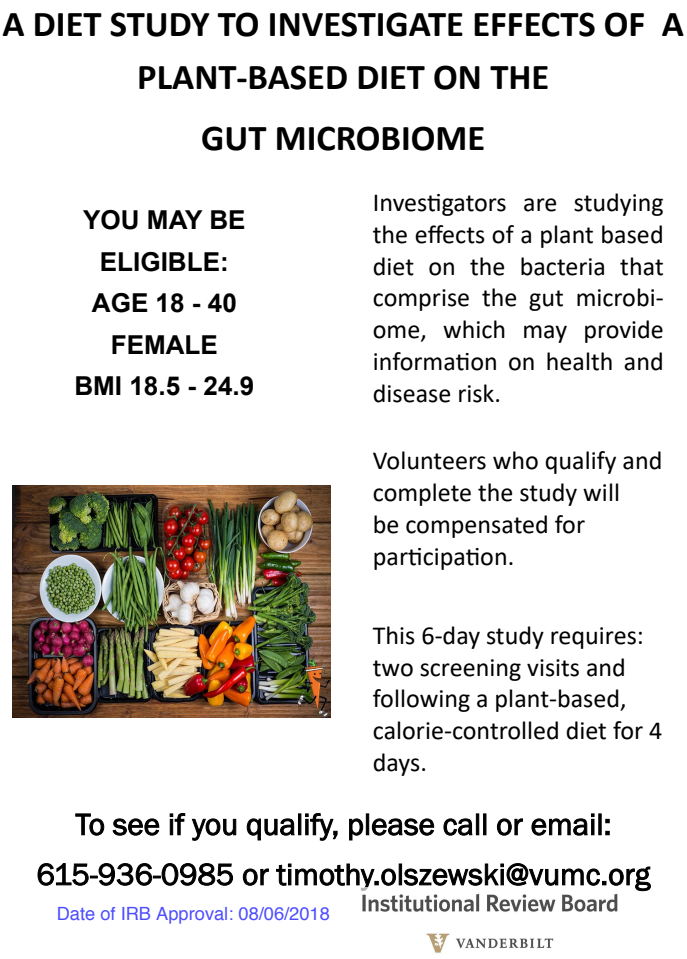


Figure 5.1: Recruitment poster for VMI study. Participants were recruited through the Vanderbilt Nutrition Center, with N=18 study subjects completing sampling in 2018.

Materials and Methods

Inclusion and Exclusion Criteria

Inclusion criteria: Single self-declared ethnicity (Black, White) for participant

and both parents. Female. Age 18-40 years. BMI 18.5-24.9 kg/m^2. Stable weight over past three months.

Exclusion criteria: No medication or dietary supplementation over past three months. No history of chronic disease or current illness / infection / inflammatory state. No tobacco use. No history of drug or alcohol abuse (> 1-2 drinks per week). No current pregnancy or lactation validated by blood test. No dietary restrictions / food allergies / food intolerances. No vegetarian or vegan diet.

Participant Recruitment and Visits

Participants were recruited from the greater area of Nashville, Tennessee, USA. All participants provided written consent forms approved by the Vanderbilt Institutional Review Board (IRB#: 171170). Recruitment was carried out on local college campuses around the Nashville area, therefore participants are biased toward those who attend or work at a local college and are not necessarily representative of the larger Nashville population. Initial visits involved consent and medical background screening, measurements of vitals (blood pressure, pulse, respiration, temperature, height, weight), and collection of the first round of oral saliva, urine, and blood samples (see following sampling sections for

specific protocols). At the initial visit to the Vanderbilt Nutrition Center, four days of vegetarian food were provided (three meals and one snack per day) alongside fecal sampling kits, gloves, and FecesCatchers. Participants were then asked to provide two days of fecal samples while on their normal western diet, consume the provided diet for four days while collecting fecal samples, then return to their normal western diet for two days with fecal samples collected. At the end of this period participants returned to the Vanderbilt Nutrition Center, where the final set of oral saliva, urine, and blood samples were collected. Questionnaires were filled out to collate personal metadata using one initial pre-survey, daily-surveys with each fecal sample collected, and a post-survey for participant feedback. All personal or identifiable information was stored in Vanderbilt's secure RedCap clinical trial system (https://redcap.vanderbilt.edu/), and survey questions were approved by the Vanderbilt Institutional Review Board (IRB#: 171170).

Oral Sampling

Study participants self-collected saliva samples using the OMNIgene Oral Kit (DNA Genotek). All samples were collected in the morning at the participants time of visit to the Vanderbilt Nutrition Clinic for pre- and post-diet time points.

Participants were asked to avoid tooth brushing, flossing, and use of mouthwash for 12 hours prior to sampling. Participants were asked to avoid eating, drinking, or chewing gum for 30 minutes prior to sampling. At the time of sampling, participants were asked to wash their hands and rinse their mouth with fresh water. One minute after expelling water rinse, participants spit fresh saliva into OMNIgene Oral collection funnels to the specified fill line and closed the lid to introduce stabilizing solution. Finally, collection tubes were sealed and shaken for 10 or more seconds to homogenize the sample among the stabilization solution, before being submitted to the research team for storage in negative 80-degree Celsius freezers.

Fecal Sampling

Study participants self-collected fecal samples using Zymo DNA/RNA Shield Fecal Collection Tubes. Participants were instructed to collect samples from the first bowel movement of the day. Participants were asked to wash their hands and wear gloves, then place FecesCatchers across the toilet to catch fecal samples. After depositing the stool on the FecesCatcher, participants were instructed to collect a small (~1 gram) sample using the scoop in the DNA/RNA Shield Fecal Collection Tube, then flush the FecesCatcher and remaining stool

down the toilet. The collection scoop was reconnected back into the collection tube and vigorously shaken for 30 seconds to thoroughly homogenize the sample with DNA/RNA Shield solution. Samples were stored at room temperature until the post-study participant visit to the Vanderbilt Nutrition Center, where all samples were returned to researchers and stored in negative 80-degree Celsius freezers.

Dual DNA/RNA Extraction

Sterility Protocol: All of the following steps of aliquoting, extraction and metagenomics library preparation were performed in a SterilGARD III Advance - class II biological safety cabinet. Prior to every use the interior of the cabinet was thoroughly cleaned with 70% ethanol, and left for at least 15 minutes under UV exposure. At no point were sample tubes opened outside of the biosafety hood. An Eppendorf 24 sample centrifuge (ID 5424) was thoroughly cleaned with 70% ethanol and left in the biosafety hood throughout all extractions to minimize movement in and out of the hood.

Sample Homogenization and Aliquoting: Stool samples were thawed in collection tubes within the biosafety hood, thoroughly shaken, and aliquoted into 1.5mL microcentrifuge tubes in the following amounts for downstream 'Omics: 1mL was

set aside for metallomics, 1mL for fecal metabolomics, 400uL for fecal metagenomics, 400uL for viromics. Oral samples were homogenized and 400uL were stored for oral metagenomics.

Metagenomic Bead Beating: Samples totaling 300uL were aliquoted into Zymo Research BashingBead 2mL Lysis Tube with 0.1 and 0.5mm beads provided in the ZymoBIOMICS DNA/RNA Miniprep Kit (Cat. No. R2002) immediately prior to bead beating. Tubes were then secured into a Biospec Products Mini-Beadbeater-96 (Cat. No. 1001, Mini-Beadbeater-96, 115 volt). Bead beating was performed at maximum speed for 3 minutes, left to sit for 2 minutes in bead beater to prevent sample overheating, and then bead beat again at maximum speed for 3 minutes.

Metallomics Profiling

200 μL of each sample or buffer (some samples did not have 200 μL, so less was added as described in the sample key) was transferred to preweighed metal-free tubes (VWR, Radnor, PA). Tubes were weighed again to get the weight of each sample. Samples were acid digested in 2 mL Optima grade 70 % nitric acid (ThermoFisher, Waltham, MA) and 500 μL 30 % hydrogen peroxide (Sigma, St. Louis, MO) for 24 h at 60 °C. After digestion, 10 mL UltraPure (Invitrogen, Carlsbad, CA) water was added to each sample. Elemental quantification on acid-

digested liquid samples was performed using an Agilent 7700 inductively coupled plasma mass spectrometer (Agilent, Santa Clara, CA). The following settings were fixed for the analysis Cell Entrance = -40 V, Cell Exit = -60 V, Plate Bias = -60 V, OctP Bias = -18 V, and collison cell Helium Flow = 4.5 mL/min. Optimal voltages for Extract 2, Omega Bias, Omega Lens, OctP RF, and Deflect were determined empirically before each sample set was analyzed. Element calibration curves were generated using ARISTAR ICP Standard Mix (VWR) diluted from 10 ppm to 1 ppb in 10-fold intervals. Samples were introduced by peristaltic pump with 0.5 mm internal diameter tubing through a MicroMist borosilicate glass nebulizer (Agilent). Samples were initially up taken at 0.5 rps for 30 s followed by 30 s at 0.1 rps to stabilize the signal. Samples were analyzed in Spectrum mode at 0.1 rps collecting three points across each peak and performing three replicates of 100 sweeps for each element analyzed. Data were acquired and analyzed using the Agilent Mass Hunter Workstation Software version A.01.02.

CHAPTER VI

Conclusion

Summary

The body of research presented here addresses a diverse set of hypotheses about how animal and human hosts shape their associated microbiomes. This breadth of topics reflects the newly appreciated importance of host-associated microbiomes, and exemplifies the many important questions that still need to be addressed in such a fledgling field. Microbiome research as a field of study may be new, but it draws foundational principles developed in a variety of scientific disciplines including ecology, evolution, microbiology, genetics, biochemistry, mathematics, medicine, and many others. Recent technological advances across a range of 'omics approaches have unveiled an inner complexity of life within our bodies, and lay the groundwork to ask many fundamental questions across such diverse disciplines. A primary aim is translating microbiome research into clinical settings, with goals of improving human health, building equity across health outcomes, and eliminating diseases. Still, such a young field lacks uniform standards, methodologies, and frameworks that help bridge basic science principles into clinical settings. There is a long road toward proper

standardization, but fortunately researchers can compare human studies with patterns observed in more easily controlled animal model and lab settings. Characterizing clinical implications of basic science principles on human microbiome assembly will require combining basic reductionist and clinical results. If properly investigated, clinically translating such principles could provide researchers with new biomarkers and toolkits to address public health and human disease. The three primary projects in this body of research span a basic science approach to connecting host evolution and microbiome ecology, a big data examination of how microbiome ecology relates to ethnicity and health disparities, and a temporal clinical trial to holistically examine how dietary intervention affects meta'omics throughout the human body. While disparate questions were addressed in each project, the results presented here reveal novel insights about how animals and human shape their complex communities of associated microorganisms.

Future Directions

The Extent of Phylosymbiosis

Phylosymbiosis was proposed as an alternative hypothesis to stochastic microbiome assembly, where host genetic variation shapes species-specific microbiomes that reflect the host evolutionary relationships. This understanding was proposed[51;77] and then methodologically framed in lab settings[24;49], but a key question is the extent to which phylosymbiosis can be detected in natural animal populations. Initial studies were performed on model organisms in labs because it had been demonstrated that diet and physiology confounded detection of an evolutionary signal, making strict controls necessary[25;75]. The extent to which those factors and many more can be controlled in natural settings is limited, and across mammals it appears that diet obfuscates phylogenetic signals in the microbiome across species that diverged more than 100mya[23]. Still, in natural and lab settings the observation of phylosymbiosis has utility. For one, phylosymbiosis connects host evolution and microbial ecology, and its observation could indicate of host filtering mechanisms and possibly even hologenome level selection, but only if other factors can be accounted for in study design. It was also discussed how the underlying hypotheses of phylosymbiosis

could be extended to understand roles of human evolution in modern microbiome assembly, particularly in framing divergence of human microbiomes in western and modernized societies. Modern lifestyles have affected human microbiomes, as observed through higher community divergence than would be expected relative to inter- and intra-specific comparisons with ancestral and more traditionally shaped microbiomes[28;50;94;97]. As the multitude of intrinsic and extrinsic influences on the microbiome are quantified, attributing microbiome variation to each confounding effect could help uncover evolutionary signals. Under controlled conditions phylosymbiosis could offer utility as a null hypothesis, against which alternative evolutionary principles could be tested for microbiome associations. In this light, it will be interesting to explore under which conditions phylosymbiotic signals disappear. In simulations for example, it was shown that modeling phylosymbiosis could manifest a signal simply through host filtering related to a single trait, like changing pH in the gut environment that correlates with phylogeny[392]. Among natural populations there are likely many more factors with influential roles on a phylosymbiotic signal, but this reductionist modeling highlights how simulations leveraging ecological principles can explain real world patterns. This is because phylosymbiosis provides useful hypotheses that are extensible across circumstances and diverse metazoan. Like other ecological and

evolutionary principles, it will likely take many years to disentangle the factors that shape phylosymbiosis, but that process would itself yield many interesting discoveries. Ultimately, a wide variety of other questions could be asked pertaining to phylosymbiosis, but for clinical applications the most interesting will be what aspects of phylosymbiotic hypotheses can be applied to modern human microbiomes.

How could Ethnicity-Associated Microbiomes Contribute to Personalized Therapies?[5]

Recently, studies have explored the role that ethnicity plays in gut microbiome assembly, uncovering subtle but reproducible differences in microbiome composition that could result from ethnic individuality in factors of diet, lifestyle, socioeconomic conditions, cultural practices, and genetic ancestry[35;393;394]. Variation within each factor has been linked to microbiome composition, and thus each could underlie ethnicity-associated microbiomes[1;4;7;11;13;25;30-32;36;37;94;105;335;344;395-399]. Unfortunately, in studies across global populations factors like geography and lifestyle covary with ethnicity,

[5] This work is published in *Future Microbiology*: **Brooks AW**. (2019). How could ethnicity-associated microbiomes contribute to personalized therapies? *Future Microbiology*.

making individual effects indistinguishable[27;98;99;344;400;401]. However, three recent studies have eliminated geographic variation and reduced social and cultural variation by examining multiethnic populations within the Netherlands, United States, and Israel[35;393;394]. Microbiome variation was observed at the community and individual taxon levels across ethnicities in the Netherlands and United States[393;394], but environmental effects dominated in the Israeli population where authors note homogeneous lifestyles across the six ethnic groups[35]. This highlights the important nuance that no two ethnicities are the same, and that ethnicity-associated microbiome differences will be subjective to the ethnic groups being compared and the larger context in which they reside. Given the subtlety with which ethnicity-associated microbiome variation has been observed, and the subjectivity of such variation to the individual, ethnicity, and larger national context, what utility could ethnicity-associated microbiomes contribute to personalized therapies?

Notably, ethnicity demarcates risk for many diseases with disproportionate burdens in one or more ethnic groups, and these health disparities cost hundreds of billions of dollars annually in the United States alone[162;342;355;366;402-404]. Ethnic health disparities present a key target for personalized medicine, where health interventions will be tailored to each individual, their health circumstance, and

ethnicity as one of many potential criteria. Many health disparities also associate with microbiome composition, establishing a tripartite relationship between ethnicity, the gut microbiome, and health outcomes[43;103;339;340;380;405-408]. These associations inform the hypothesis that ethnicity-associated microbiome composition can be used to link the influence of factors like ethnicity-associated dietary or cultural patterns with health disparity etiology. While individuals, organizations, and governments are working to address public policy and societal factors underlying health disparities, medical researchers also have the capacity to leverage ethnicity-by-microbiome and disease-by-microbiome associations to build a foundation for ethnic-specific therapies. An obvious but worthwhile approach would be to target ethnicity-associated microbiome compositions directly with tools like probiotics, fecal microbiome transplantation, or even phage therapies. However, our understanding of ethnicity-associated microbiomes is in its infancy, and any such interventions will require years of further research and clinical trials before widespread implementation is possible. To spur much needed inclusion of diversity in microbiome research and build a foundational understanding of the ethnicity-microbiome-disparity intersection, two potential approaches of how ethnicity-associated microbiome could be leveraged today will be considered.

Ethnicity-associated microbiomes: a proxy for factors explaining microbiome assembly

Ethnicity captures a complex array of dietary, socioeconomic, cultural, lifestyle, and genetic factors to varying degrees, making it difficult to disentangle which, if any, of these factors play a role in microbiome composition. However, when microbiome composition covaries with ethnicity and one of these factors, it could serve as an indication that the factor itself may explain ethnicity-associated composition. The question is whether such patterns are informative in disentangling the role of each factor in ethnicity-associated microbiome assembly without validation in large, comprehensively phenotyped, multiethnic studies. Many studies are examining how factors like diet or genetics shape factor-associated microbiomes, but as with much of microbiome research these are in smaller and generally ethnically homogeneous populations[4;7;11;13;335;409;410]. Still, through reductionist means these studies are isolating microbiome composition shaped by associated factors;factors which may also vary across ethnicities and contribute to ethnicity-associated microbiomes. The problem is that identifying which factors like diet or lifestyle vary between ethnicities and lead to meaningful biological variation in the microbiome is difficult. For researchers though,

identifying factor-associated microbiome variation that overlaps with ethnicity-associated microbiome variation could serve as an indicator that a factor varies ethnically and is playing a biologically meaningful role in microbiome assembly. Critically, using data available today would mean comparing microbiome patterns across different study populations, and the caveat must be acknowledged that most factor-associated microbiome variation identified to date is in Caucasian majority populations. With potential pitfalls, what makes this approach worthwhile?

Ethnicity-associated microbiome variation appears to be subtle[393;394], requiring large, multiethnic study populations to detect its signal. Fortunately, large microbiome projects and biobanks are now emerging that facilitate ethnicity-microbiome studies, and ethnicity is in general consistently and reliably declared in individual's electronic health records[67;342;343;411]. On the other hand, factor-associated microbiome studies require careful controls to eliminate confounding variables, favoring smaller and ethnically homogeneous populations. Factors like diet and lifestyle have a complex array of underlying variables that could be measured, few of which are regularly assessed during medical visits and seldom appear in electronic health records. This is notable because measuring all complex factors that could influence the microbiome in large multiethnic

populations is prohibitively invasive and costly, while carrying out factor-associated microbiome studies between each pair of ethnicities in their subjective context would quickly become exhaustive. By comparing microbiome variation in smaller factor-associated studies with microbiome variation observed across ethnicities in large studies, the amount of overlapping variation could be used as a proxy ranking for which factors more likely contribute to ethnicity-associated microbiome assembly. Caveats of this data-driven approach include cross population comparisons, ethnicity-associated microbiome composition potentially overlapping variation of multiple confounding factors, and the inability to directly prove causal effects on microbiome assembly. Still, such an approach could be a powerful tool to generate hypotheses about which factors play roles in ethnicity-associated microbiome formation. Further, prescribing changes to factors like diet or lifestyle identified in this approach could serve as therapies to manipulate ethnicity-associated microbiomes, especially if composition is linked to health disparities, as discussed below.

Ethnicity-associated microbiome composition in health disparity etiology

Health disparity inequality manifests as different disease risks across ethnicities, and therefore ethnicity-associated microbiome composition could

associate with disparity risk[342;343;355]. Many health disparities have been correlated with microbiome composition as well, but it has not been established if disease-associated microbiome variation overlaps with ethnicity-associated microbiome variation[43;103;339;340;380;405-408]. Just as overlapping factor-associated and ethnicity-associated microbiome variation could serve as proxies for factors contributing to microbiome assembly, overlapping disease-associated and ethnicity-associated microbiome variation could serve as a proxy for microbiome variation more likely to be linked with disparity etiology. Certainly, identifying overlap would not establish whether microbiome variation is causal, consequential, or only tangentially associated to disparity etiology. With this said, overlap could be used to rank which disparities have the potential to be mediated by ethnicity-associated microbiome variation.

The true strengths of this approach lie in more nuanced examination. Microbiome variation associated with age and ethnicity had little overlap in younger and healthier individuals in the United States[394]. However, it is unknown how ethnicity-associated microbiomes change as individuals age, which will be important to investigate, as most disparities have onset at later ages. If ethnicity-associated variation changes with age, does overlap with disparity-associated microbiome composition exist lifelong, increase with age, or manifest around the

normal age of disparity onset? Ethnicity-associated and disparity-associated microbiome variation could also overlap across multiple diseases, indicating that while disease phenotypes may be distinct, the underlying biological etiology may be shared. It will be particularly interesting to explore whether groups like autoimmune or metabolic disparities that share increasing risk within an ethnicity will also overlap in the same sets of ethnicity-associated microbiome composition. Combined with the first approach, overlapping factor-associated, ethnicity-associated, and disparity-associated microbiome composition could provide unique insights into the mechanisms worth investigating as drivers of disparities. As available data grows the variety of questions that could be asked grows as well, and many insights about the potential for personalized therapies could be gained in the near term by creatively connecting ethnicity-associated microbiome assembly to disease etiology.

Conclusion

Growing awareness that ethnicity-associated microbiome composition exists will hopefully lead researchers to increase ethnic diversity in microbiome study recruitment, particularly for studies of diseases that present as health disparities. Ultimately, multiethnic studies targeting a single factor or disparity will be necessary to attribute causality in the ethnicity-microbiome-disparity intersection, but the insights that could be gained by leveraging big data in approaches like these may help inform which targets are worth pursuing. Ethnicity is a complex concept subjective to the individual and environmental context, with each ethnicity varying differently for a range of factors like diet, lifestyle, culture, and genetics. Exhaustively investigating every combination of ethnicities, factors, and disparities for their microbiome relationship is untenable, and so creative and targeted approaches will be necessary to accelerate personalized therapies related to ethnicity-associated microbiomes. The details of each of these approaches are less important than fostering the mindset that existing ethnicity-associated microbiome composition can be utilized now, and that barriers in cost and recruitment time for multiethnic clinical trials should not prevent investigations of how the microbiome could mediate health disparities. As a culture, addressing injustices in access to healthcare and fresh food, socioeconomic mobility, and

many other factors underlying disparities is an ideal but protracted avenue to pursue equality in health outcomes. Right now, practical approaches leveraging ethnicity-associated microbiomes for clues about contributing factors, underlying etiologies, and lifestyle interventions for health disparities are worth pursuing if they can ameliorate even some inequality in disparity risk.

Closing Remarks

"I am large, I contain multitudes." – Walt Whitman (*Leaves of Grass*)

The multitude of viewpoints, beliefs, and emotions that make each of us unique in Walt Whitman's eyes are analogous to the multitudes of microorganisms, enzymes, metabolites, and minerals that make each of us biologically distinct. In search of health equality our biological complexity must be characterized across the breadth of human diversity to develop treatments that work for all. Even broader is the diversity of metazoans with which we share this world, and it would be foolish to discount the insights we can make by marrying basic, model organism, and clinical studies across humans and animals. Microbiome research as a new scientific field is still like the wild west, it begs for structure and standardization, but also allows creative minds to investigate without preconceived notions and assumptions. As our understanding develops alongside novel tools like probiotics and FMT, perturbing the microbiome will become more accessible, effective, and hopefully provide a leap forward in our ability to shape human health.

BIBLIOGRAPHY

1. McDonald, D., Birmingham, A., and Knight, R. (2015). Context and the human microbiome. Microbiome 3, 52.
2. Integrative, H.M.P.R.N.C. (2014). The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host Microbe 16, 276-289.
3. Human Microbiome Project, C. (2012). A framework for human microbiome research. Nature 486, 215-221.
4. Human Microbiome Project, C. (2012). Structure, function and diversity of the healthy human microbiome. Nature 486, 207-214.
5. Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J.R., Prifti, E., Nielsen, T., et al. (2014). An integrated catalog of reference genes in the human gut microbiome. Nature Biotechnology 32, 834-841.
6. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464, 59-65.
7. Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., et al. (2009). A core gut microbiome in obese and lean twins. Nature 457, 480-484.
8. Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., et al. (2012). Human gut microbiome viewed across age and geography. Nature 486, 222-227.
9. Blekhman, R., Goodrich, J.K., Huang, K., Sun, Q., Bukowski, R., Bell, J.T., Spector, T.D., Keinan, A., Ley, R.E., Gevers, D., et al. (2015). Host genetic variation impacts microbiome composition across human body sites. Genome Biol 16, 191.
10. Davenport, E.R. (2016). Elucidating the role of the host genome in shaping microbiome composition. Gut Microbes 7, 178-184.
11. Davenport, E.R., Cusanovich, D.A., Michelini, K., Barreiro, L.B., Ober, C., and Gilad, Y. (2015). Genome-Wide Association Studies of the Human Gut Microbiota. PLoS One 10, e0140301.
12. Davenport, E.R., Sanders, J.G., Song, S.J., Amato, K.R., Clark, A.G., and Knight, R. (2017). The human microbiome in evolution. BMC Biol 15, 127.
13. Goodrich, J.K., Davenport, E.R., Beaumont, M., Jackson, M.A., Knight, R., Ober, C., Spector, T.D., Bell, J.T., Clark, A.G., and Ley, R.E. (2016). Genetic Determinants of the Gut Microbiome in UK Twins. Cell Host Microbe 19, 731-743.
14. Borenstein, E. (2012). Computational systems biology and in silico modeling of the human microbiome. Brief Bioinform 13, 769-780.
15. Noecker, C., Eng, A., Srinivasan, S., Theriot, C.M., Young, V.B., Jansson, J.K., Fredricks, D.N., and Borenstein, E. (2016). Metabolic Model-Based Integration of Microbiome Taxonomic and Metabolomic Profiles Elucidates Mechanistic Links between Ecological and Metabolic Variation. mSystems 1.
16. Belkaid, Y., and Harrison, O.J. (2017). Homeostatic Immunity and the Microbiota. Immunity 46, 562-576.
17. Levy, M., Blacher, E., and Elinav, E. (2016). Microbiome, metabolites and host immunity. Curr Opin Microbiol 35, 8-15.
18. Proal, A.D., Albert, P.J., and Marshall, T.G. (2013). The human microbiome and autoimmunity. Current Opinion in Rheumatology 25, 234-240.
19. Rooks, M.G., and Garrett, W.S. (2016). Gut microbiota, metabolites and host immunity. Nat Rev Immunol 16, 341-352.
20. Thaiss, C.A., Zmora, N., Levy, M., and Elinav, E. (2016). The microbiome and innate immunity. Nature 535, 65-74.

21. Amato, K.R. (2016). Incorporating the gut microbiota into models of human and non-human primate ecology and evolution. Am J Phys Anthropol 159, S196-215.
22. Bordenstein, S.R., and Theis, K.R. (2015). Host Biology in Light of the Microbiome: Ten Principles of Holobionts and Hologenomes. PLoS Biol 13, e1002226.
23. Groussin, M., Mazel, F., Sanders, J.G., Smillie, C.S., Lavergne, S., Thuiller, W., and Alm, E.J. (2017). Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. Nat Commun 8, 14319.
24. Brooks, A.W., Kohl, K.D., Brucker, R.M., van Opstal, E.J., and Bordenstein, S.R. (2016). Phylosymbiosis: Relationships and Functional Effects of Microbial Communities across Host Evolutionary History. PLoS Biol 14, e2000225.
25. Ley, R.E., Lozupone, C.A., Hamady, M., Knight, R., and Gordon, J.I. (2008). Worlds within worlds: evolution of the vertebrate gut microbiota. Nat Rev Microbiol 6, 776-788.
26. Gupta, V.K., Paul, S., and Dutta, C. (2017). Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity. Front Microbiol 8, 1162.
27. Jha, A.R., Davenport, E.R., Gautam, Y., Bhandari, D., Tandukar, S., Ng, K.M., Fragiadakis, G.K., Holmes, S., Gautam, G.P., Leach, J., et al. (2018). Gut microbiome transition across a lifestyle gradient in Himalaya. PLoS Biol 16, e2005396.
28. Schnorr, S.L., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., Turroni, S., Biagi, E., Peano, C., Severgnini, M., et al. (2014). Gut microbiome of the Hadza hunter-gatherers. Nat Commun 5, 3654.
29. Candela, M., Biagi, E., Brigidi, P., O'Toole, P.W., and De Vos, W.M. (2014). Maintenance of a healthy trajectory of the intestinal microbiome during aging: a dietary approach. Mech Ageing Dev 136-137, 70-75.
30. David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E., Ling, A.V., Devlin, A.S., Varma, Y., Fischbach, M.A., et al. (2013). Diet rapidly and reproducibly alters the human gut microbiome. Nature 505, 559-563.
31. Muegge, B.D., Kuczynski, J., Knights, D., Clemente, J.C., González, A., Fontana, L., Henrissat, B., Knight, R., and Gordon, J.I. (2011). Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science 332, 970-974.
32. Singh, R.K., Chang, H.W., Yan, D., Lee, K.M., Ucmak, D., Wong, K., Abrouk, M., Farahnik, B., Nakamura, M., Zhu, T.H., et al. (2017). Influence of diet on the gut microbiome and implications for human health. J Transl Med 15, 73.
33. Voreades, N., Kozil, A., and Weir, T.L. (2014). Diet and the development of the human intestinal microbiome. Front Microbiol 5, 494.
34. Zmora, N., Suez, J., and Elinav, E. (2019). You are what you eat: diet, health and the gut microbiota. Nat Rev Gastroenterol Hepatol 16, 35-56.
35. Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., Costea, P.I., Godneva, A., Kalka, I.N., Bar, N., et al. (2018). Environment dominates over host genetics in shaping human gut microbiota. Nature 555, 210-215.
36. Spor, A., Koren, O., and Ley, R. (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. Nat Rev Microbiol 9, 279-290.
37. Levin, A.M., Sitarik, A.R., Havstad, S.L., Fujimura, K.E., Wegienka, G., Cassidy-Bushrow, A.E., Kim, H., Zoratti, E.M., Lukacs, N.W., Boushey, H.A., et al. (2016). Joint effects of pregnancy, sociocultural, and environmental factors on early life gut microbiome structure and diversity. Sci Rep 6, 31775.
38. Theis, K.R., Venkataraman, A., Dycus, J.A., Koonter, K.D., Schmitt-Matzen, E.N., Wagner, A.P., Holekamp, K.E., and Schmidt, T.M. (2013). Symbiotic bacteria appear to mediate hyena social odors. Proc Natl Acad Sci U S A 110, 19832-19837.
39. Ursell, L.K., Metcalf, J.L., Parfrey, L.W., and Knight, R. (2012). Defining the human microbiome. Nutr Rev 70 Suppl 1, S38-44.
40. Voigt, A.Y., Costea, P.I., Kultima, J.R., Li, S.S., Zeller, G., Sunagawa, S., and Bork, P. (2015). Temporal and technical variability of human gut metagenomes. Genome Biol 16, 73.

41. Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A.B., Brady, A., Creasy, H.H., McCracken, C., Giglio, M.G., et al. (2017). Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 550, 61-66.
42. Backhed, F., Fraser, C.M., Ringel, Y., Sanders, M.E., Sartor, R.B., Sherman, P.M., Versalovic, J., Young, V., and Finlay, B.B. (2012). Defining a healthy human gut microbiome: current concepts, future directions, and clinical applications. Cell Host Microbe 12, 611-622.
43. Wu, G.D., Bushmanc, F.D., and Lewis, J.D. (2013). Diet, the human gut microbiota, and IBD. Anaerobe 24, 117-120.
44. Belizario, J.E., and Napolitano, M. (2015). Human microbiomes and their roles in dysbiosis, common diseases, and novel therapeutic approaches. Front Microbiol 6, 1050.
45. Lloyd-Price, J., Abu-Ali, G., and Huttenhower, C. (2016). The healthy human microbiome. Genome Med 8, 51.
46. Muszer, M., Noszczynska, M., Kasperkiewicz, K., and Skurnik, M. (2015). Human Microbiome: When a Friend Becomes an Enemy. Arch Immunol Ther Exp (Warsz) 63, 287-298.
47. Kostic, A.D., Gevers, D., Siljander, H., Vatanen, T., Hyotylainen, T., Hamalainen, A.M., Peet, A., Tillmann, V., Poho, P., Mattila, I., et al. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. Cell Host Microbe 17, 260-273.
48. Wu, X., Chen, H., and Xu, H. (2015). The genomic landscape of human immune-mediated diseases. J Hum Genet 60, 675-681.
49. Franzenburg, S., Walter, J., Kunzel, S., Wang, J., Baines, J.F., Bosch, T.C., and Fraune, S. (2013). Distinct antimicrobial peptide expression determines host species-specific bacterial associations. Proc Natl Acad Sci U S A 110, E3730-3738.
50. Ochman, H., Worobey, M., Kuo, C.H., Ndjango, J.B., Peeters, M., Hahn, B.H., and Hugenholtz, P. (2010). Evolutionary relationships of wild hominids recapitulated by gut microbial communities. PLoS Biol 8, e1000546.
51. Brucker, R.M., and Bordenstein, S.R. (2013). The hologenomic basis of speciation: gut bacteria cause hybrid lethality in the genus Nasonia. Science 341, 667-669.
52. Pollock, F.J., McMinds, R., Smith, S., Bourne, D.G., Willis, B.L., Medina, M., Thurber, R.V., and Zaneveld, J.R. (2018). Coral-associated bacteria demonstrate phylosymbiosis and cophylogeny. Nat Commun 9, 4921.
53. McCafferty, J., Muhlbauer, M., Gharaibeh, R.Z., Arthur, J.C., Perez-Chanona, E., Sha, W., Jobin, C., and Fodor, A.A. (2013). Stochastic changes over time and not founder effects drive cage effects in microbial community assembly in a mouse model. ISME J 7, 2116-2125.
54. McClelland, J., and Koslicki, D. (2018). EMDUniFrac: exact linear time computation of the UniFrac metric and identification of differentially abundant organisms. Journal of Mathematical Biology 77, 935-949.
55. Lozupone, C., Hamady, M., and Knight, R. (2006). UniFrac--an online tool for comparing microbial community diversity in a phylogenetic context. BMC Bioinformatics 7, 371.
56. Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71, 8228-8235.
57. Clarke, K.R. (1993). Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology 18, 117-143.
58. Anderson, M.J. (2001). A new method for non-parametric multivariate analysis of variance. Australian Journal of Ecology 26, 32-46.
59. Mantel, N. (1967). The Detection of Disease Clustering and a Generalized Regression Approach. Cancer Research 27, 209-220.
60. Knights, D., Costello, E.K., and Knight, R. (2011). Supervised classification of human microbiota. FEMS Microbiol Rev 35, 343-359.
61. Anderson, M.J., and Walsh, D.C. (2013). PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? Ecological Monographs 83(4), 557-574.
62. Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53, 325-338.

63. Kenkel, N.C., and Orlóci, L. (1986). Applying Metric and Nonmetric Multidimensional Scaling to Ecological Studies: Some New
Results. Ecology 67, 919-928.
64. Hardoon, D.R., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical Correlation Analysis: An Overview with Application to Learning Methods. Neural Computation 16, 2639-2664.
65. Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., Huttenhower, C., and Ley, R.E. (2013). A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. PLoS Comput Biol 9, e1002863.
66. Cornwell, W., and Nakagawa, S. (2017). Phylogenetic comparative methods. Curr Biol 27, R333-R336.
67. Liu, X.Q., Paterson, A.D., John, E.M., and Knight, J.A. (2006). The role of self-defined race/ethnicity in population structure control. Ann Hum Genet 70, 496-505.
68. Morton, N.E., Yasuda, N., Miki, C., and Lee, S. (1967). Population Structure of the ABO Blood Groups in Switzerland. American Journal of Human Genetics.
69. Relethford, J.H., and Lees, F.C. (1982). The Use of Quantitative Traits in the Study of Human Population Structure. Annual Review of Genetics 3, 53-74.
70. Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Sato, H., Hori, M., Nakamura, Y., et al. (2002). Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 32, 650-654.
71. Ikegawa, S. (2012). A short history of the genome-wide association study: where we were and where we are going. Genomics Inform 10, 220-225.
72. Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet 9, 255-266.
73. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature 461, 747-753.
74. Polderman, T.J., Benyamin, B., de Leeuw, C.A., Sullivan, P.F., van Bochoven, A., Visscher, P.M., and Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nat Genet 47, 702-709.
75. Ley, R.E., Hamady, M., Lozupone, C., Turnbaugh, P.J., Ramey, R.R., Bircher, J.S., Schlegel, M.L., Tucker, T.A., Schrenzel, M.D., Knight, R., et al. (2008). Evolution of mammals and their gut microbes. Science 320, 1647-1651.
76. Fraune, S., and Bosch, T.C.G. (2007). Long-term maitenance of species-specific bacterial microbiota in the basal metazoan *Hydra*. Proc Natl Acad Sci 104, 13146-13151.
77. Brucker, R.M., and Bordenstein, S.R. (2012). Speciation by symbiosis. Trends Ecol Evol 27, 443-451.
78. Brucker, R.M., and Bordenstein, S.R. (2012). The roles of host evolutionary relationships (genus: Nasonia) and development in structuring microbial communities. Evolution 66, 349-362.
79. van Veelen, H.P.J., Falcao Salles, J., and Tieleman, B.I. (2017). Multi-level comparisons of cloacal, skin, feather and nest-associated microbiota suggest considerable influence of horizontal acquisition on the microbiota assembly of sympatric woodlarks and skylarks. Microbiome 5, 156.
80. Ross, A.A., Muller, K.M., Weese, J.S., and Neufeld, J.D. (2018). Comprehensive skin microbiome analysis reveals the uniqueness of human skin and evidence for phylosymbiosis within the class Mammalia. Proc Natl Acad Sci U S A 115, E5786-E5795.
81. Chiarello, M., Auguet, J.C., Bettarel, Y., Bouvier, C., Claverie, T., Graham, N.A.J., Rieuvilleneuve, F., Sucre, E., Bouvier, T., and Villeger, S. (2018). Skin microbiome of coral reef fish is highly variable and driven by host phylogeny and diet. Microbiome 6, 147.
82. Bletz, M.C., Archer, H., Harris, R.N., McKenzie, V.J., Rabemananjara, F.C.E., Rakotoarison, A., and Vences, M. (2017). Host Ecology Rather Than Host Phylogeny Drives Amphibian Skin Microbial Community Structure in the Biodiversity Hotspot of Madagascar. Front Microbiol 8, 1530.
83. Bost, A., Martinson, V.G., Franzenburg, S., Adair, K.L., Albasi, A., Wells, M.T., and Douglas, A.E. (2018). Functional variation in the gut microbiome of wild Drosophila populations. Mol Ecol 27, 2834-2845.

84. Cook, G.C., and Al-Torki, M.T. (1975). High Intestinal Lactase Concentrations in Adult Arabs in SaudiArabia. British Medical Journal 3, 135-136.
85. Gerbault, P., Liebert, A., Itan, Y., Powell, A., Currat, M., Burger, J., Swallow, D.M., and Thomas, M.G. (2011). Evolution of lactase persistence: an example of human niche construction. Philos Trans R Soc Lond B Biol Sci 366, 863-877.
86. Lynch, J., Tang, K., Priya, S., Sands, J., Sands, M., Tang, E., Mukherjee, S., Knights, D., and Blekhman, R. (2017). HOMINID: A framework for identifying associations between host genetic variation and microbiome composition. GigaScience.
87. Kolde, R., Franzosa, E.A., Rahnavard, G., Hall, A.B., Vlamakis, H., Stevens, C., Daly, M.J., Xavier, R.J., and Huttenhower, C. (2018). Host genetic variation and its microbiome interactions within the Human Microbiome Project. Genome Med 10, 6.
88. Weissbrod, O., Rothschild, D., Barkan, E., and Segal, E. (2018). Host genetics and microbiome associations through the lens of genome wide association studies. Curr Opin Microbiol 44, 9-19.
89. Bonder, M.J., Kurilshikov, A., Tigchelaar, E.F., Mujagic, Z., Imhann, F., Vila, A.V., Deelen, P., Vatanen, T., Schirmer, M., Smeekens, S.P., et al. (2016). The effect of host genetics on the gut microbiome. Nature Genetics 48, 1407-1412.
90. Turpin, W., Espin-Garcia, O., Xu, W., Silverberg, M.S., Kevans, D., Smith, M.I., Guttman, D.S., Griffiths, A., Panaccione, R., Otley, A., et al. (2016). Association of host genome with intestinal microbial composition in a large healthy cohort. Nature Genetics 48, 1413-1417.
91. Goodrich, J.K., Waters, J.L., Poole, A.C., Sutter, J.L., Koren, O., Blekhman, R., Beaumont, M., Van Treuren, W., Knight, R., Bell, J.T., et al. (2014). Human genetics shape the gut microbiome. Cell 159, 789-799.
92. Davenport, E.R. (2016). Elucidating the role of the host genome in shaping microbiome composition. Gut Microbes 7, 178-184.
93. Moeller, A.H., Li, Y., Ngole, E.M., Ahuka-Mundeke, S., Lonsdorf, E.V., Pusey, A.E., Peeters, M., Hahn, B.H., and Ochman, H. (2014). Rapid changes in the gut microbiome during human evolution. Proc Natl Acad Sci 111, 16431-16435.
94. Moeller, A.H., Li, Y., Mpoudi Ngole, E., Ahuka-Mundeke, S., Lonsdorf, E.V., Pusey, A.E., Peeters, M., Hahn, B.H., and Ochman, H. (2014). Rapid changes in the gut microbiome during human evolution. Proc Natl Acad Sci U S A 111, 16431-16435.
95. Clayton, J.B., Vangay, P., Huang, H., Ward, T., Hillmann, B.M., Al-Ghalith, G.A., Travis, D.A., Long, H.T., Tuan, B.V., Minh, V.V., et al. (2016). Captivity humanizes the primate microbiome. Proceedings of the National Academy of Sciences 113, 10376-10381.
96. Weyrich, L.S., Duchene, S., Soubrier, J., Arriola, L., Llamas, B., Breen, J., Morris, A.G., Alt, K.W., Caramelli, D., Dresely, V., et al. (2017). Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. Nature 544, 357-361.
97. Christina J Adler, Keith Dobney, Laura S Weyrich, John Kaidonis, Alan W Walker, Wolfgang Haak, Corey J A Bradshaw, Grant Townsend, Arkadiusz Sołtysiak, Kurt W Alt, et al. (2013). Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. Nature Genetics 45, 450-456.
98. Clemente JC, P.E., Blaser MJ, Sandhu K, Gao K, Wang B, Magda M, Hidalgo G, et al. (2015). The microbiome of uncontacted Amerindians. Science Advances 3.
99. Smits, S.A., Leach, J., Sonnenburg, E.D., Gonzalez, C.G., Lichtman, J.S., Reid, G., Knight, R., Manjurano, A., Changalucha, J., Elias, J.E., et al. (2017). Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science 357, 802-806.
100. Carlotta De Filippo, Duccio Cavalieria, Monica Di Paolab, Matteo Ramazzottic, Jean Baptiste Poulletd, Sebastien Massartd, Silvia Collinib, Giuseppe Pieraccinie, and Lionettib, P. (2010). Impact of diet in shaping gut microbiota revealed
by a comparative study in children from Europe
and rural Africa. PNAS 107, 14691-14696.

101. Vangay, P., Johnson, A.J., Ward, T.L., Al-Ghalith, G.A., Shields-Cutler, R.R., Hillmann, B.M., Lucas, S.K., Beura, L.K., Thompson, E.A., Till, L.M., et al. (2018). US Immigration Westernizes the Human Gut Microbiome. Cell 175, 962-972 e910.
102. Gorvitovskaia, A., Holmes, S.P., and Huse, S.M. (2016). Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. Microbiome 4, 15.
103. Duvallet, C., Gibbons, S.M., Gurry, T., Irizarry, R.A., and Alm, E.J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun 8, 1784.
104. Hsu, T., Joice, R., Vallarino, J., Abu-Ali, G., Hartmann, E.M., Shafquat, A., DuLong, C., Baranowski, C., Gevers, D., Green, J.L., et al. (2016). Urban Transit System Microbial Communities Differ by Surface Type and Interaction with Humans and the Environment. mSystems 1.
105. Pehrsson, E.C., Tsukayama, P., Patel, S., Mejia-Bautista, M., Sosa-Soto, G., Navarrete, K.M., Calderon, M., Cabrera, L., Hoyos-Arango, W., Bertoli, M.T., et al. (2016). Interconnected microbiomes and resistomes in low-income human habitats. Nature 533, 212-216.
106. Aminov, R.I. (2010). A brief history of the antibiotic era: lessons learned and challenges for the future. Front Microbiol 1, 134.
107. Vincent, C., Stephens, D.A., Loo, V.G., Edens, T.J., Behr, M.A., Dewar, K., and Manges, A.R. (2013). Reductions in intestinal Clostridiales precede the development of nosocomial Clostridium difficile infection . Microbiome 1.
108. Atarashi, K., Suda, W., Luo, C., Kawaguchi, T., Motoo, I., Narushima, S., and al., e. (2017). Ectopic colonization of oral bacteria in the intestine drives TH1 cell induction and inflammation. Science 358, 359-365.
109. Russell, C.W., Bouvaine, S., Newell, P.D., and Douglas, A.E. (2013). Shared metabolic pathways in a coevolved insect-bacterial symbiosis. Appl Environ Microbiol 79, 6117-6123.
110. Lai, C.Y., Baumann, L., and Baumann, P. (1994). Amplification of trpEG: Adaptation of Buchnera aphidicola to an endosymbiotic association with aphids
. Proc Natl Acad Sci 91, 3819-3823.
111. McFall-Ngai, M., Heath-Heckman, E.A., Gillette, A.A., Peyer, S.M., and Harvie, E.A. (2012). The secret languages of coevolved symbioses: insights from the Euprymna scolopes-Vibrio fischeri symbiosis. Semin Immunol 24, 3-8.
112. Nishiguchi, M.K. (2002). Host-symbiont recognition in the environmentally transmitted sepiolid squid-Vibrio mutualism. Microb Ecol 44, 10-18.
113. Moeller, A.H., Caro-Quintero, A., Mjungu, D., Georgiev, A.V., Lonsdorf, E.V., Muller, M.N., Pusey, A.E., Peeters, M., Hahn, B.H., and Ochman, H. (2016). Cospeciation of gut microbiota with hominids. Science 353, 380-382.
114. Hosokawa, T., Koga, R., Kikuchi, Y., Meng, X.Y., and Fukatsu, T. (2010). Wolbachia as a bacteriocyte-associated nutritional mutualist. Proc Natl Acad Sci U S A 107, 769-774.
115. Weiss, B.L., Wang, J., and Aksoy, S. (2011). Tsetse immune system maturation requires the presence of obligate symbionts in larvae. PLoS Biol 9, e1000619.
116. Sloan, D.B., and Moran, N.A. (2012). Genome reduction and co-evolution between the primary and secondary bacterial symbionts of psyllids. Mol Biol Evol 29, 3781-3792.
117. Sabree, Z.L., Huang, C.Y., Okusu, A., Moran, N.A., and Normark, B.B. (2013). The nutrient supplying capabilities of Uzinura, an endosymbiont of armoured scale insects. Environ Microbiol 15, 1988-1999.
118. Bennett, G.M., and Moran, N.A. (2013). Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a Phloem-feeding insect. Genome Biol Evol 5, 1675-1688.
119. Sloan, D.B., and Moran, N.A. (2013). The evolution of genomic instability in the obligate endosymbionts of whiteflies. Genome Biol Evol 5, 783-793.
120. Vogel, K.J., and Moran, N.A. (2013). Functional and evolutionary analysis of the genome of an obligate fungal symbiont. Genome Biol Evol 5, 891-904.
121. Ravela, J., Gajera, P., Abdob, Z.G., Schneiderc, M., Koeniga, S.S.K., McCullea, S.L., Peraltae, L., and Forneyc, L.J. (2011). Vaginal microbiome of reproductive-age women. Procedings of the National Academy of Sciences 108, 4680-4687.

122. Fettweis, J.M., Brooks, J.P., Serrano, M.G., Sheth, N.U., Girerd, P.H., Edwards, D.J., Strauss, J.F., 3rd, Vaginal Microbiome, C., Jefferson, K.K., and Buck, G.A. (2014). Differences in vaginal microbiome in African American women versus women of European ancestry. Microbiology 160, 2272-2282.
123. France, M.T., Mendes-Soares, H., and Forney, L.J. (2016). Genomic Comparisons of Lactobacillus crispatus and Lactobacillus iners Reveal Potential Ecological Drivers of Community Composition in the Vagina. Appl Environ Microbiol 82, 7063-7073.
124. Si, J., You, H.J., Yu, J., Sung, J., and Ko, G. (2016). Prevotella as a Hub for Vaginal Microbiota under the Influence of Host Genetics and Their Association with Obesity. Cell Host Microbe.
125. Yassour, M., Jason, E., Hogstrom, L.J., Arthur, T.D., Tripathi, S., Siljander, H., Selvenius, J., Oikarinen, S., Hyoty, H., Virtanen, S.M., et al. (2018). Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. Cell Host Microbe 24, 146-154 e144.
126. Loo, E.X.L., Sim, J.Z.T., Loy, S.L., Goh, A., Chan, Y.H., Tan, K.H., Yap, F., Gluckman, P.D., Godfrey, K.M., Van Bever, H., et al. (2017). Associations between caesarean delivery and allergic outcomes: Results from the GUSTO study. Ann Allergy Asthma Immunol 118, 636-638.
127. Mitselou, N., Hallberg, J., Stephansson, O., Almqvist, C., Melen, E., and Ludvigsson, J.F. (2018). Cesarean delivery, preterm birth, and risk of food allergy: Nationwide Swedish cohort study of more than 1 million children. J Allergy Clin Immunol 142, 1510-1514 e1512.
128. Ferretti, P., Pasolli, E., Tett, A., Asnicar, F., Gorfer, V., Fedi, S., Armanini, F., Truong, D.T., Manara, S., Zolfo, M., et al. (2018). Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. Cell Host Microbe 24, 133-145 e135.
129. Moossavi, S., Sepehri, S., Robertson, B., Bode, L., Goruk, S., Field, C.J., Lix, L.M., de Souza, R.J., Becker, A.B., Mandhane, P.J., et al. (2019). Composition and Variation of the Human Milk Microbiota Are Influenced by Maternal and Early-Life Factors. Cell Host Microbe 25, 324-335 e324.
130. Boehm, G., and Stahl, B. (2007). Oligosaccharides from Milk. American Society for Nutrition.
131. Thurl, S., Munzert, M., Boehm, G., Matthews, C., and Stahl, B. (2017). Systematic review of the concentrations of oligosaccharides in human milk. Nutr Rev 75, 920-933.
132. Ayechu-Muruzabal, V., van Stigt, A.H., Mank, M., Willemsen, L.E.M., Stahl, B., Garssen, J., and Van't Land, B. (2018). Diversity of Human Milk Oligosaccharides and Effects on Early Life Immune Development. Front Pediatr 6, 239.
133. Ojala, T., Kankainen, M., Castro, J., Cerca, N., Edelman, S., Westerlund-Wikström, B., Paulin, L., Holm, L., and Auvinen, P. (2014). Comparative genomics of Lactobacillus crispatus suggests novel mechanisms for the competitive exclusion of Gardnerella vaginalis. BMC Genomics 15.
134. Bai, Y., Tao, J., Zhou, J., Fan, Q., Liu, M., Hu, Y., Xu, Y., Zhang, L., Yuan, J., Li, W., et al. (2018). Fucosylated Human Milk Oligosaccharides and N-Glycans in the Milk of Chinese Mothers Regulate the Gut Microbiome of Their Breast-Fed Infants during Different Lactation Stages. mSystems 3.
135. Lerner, A., Matthias, T., and Aminov, R. (2017). Potential Effects of Horizontal Gene Exchange in the Human Gut. Front Immunol 8, 1630.
136. Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., et al. (2009). Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet 5, e1000344.
137. Shafquat, A., Joice, R., Simmons, S.L., and Huttenhower, C. (2014). Functional and phylogenetic assembly of microbial communities in the human microbiome. Trends Microbiol 22, 261-266.
138. Franzosa, E.A., Morgan, X.C., Segata, N., Waldron, L., Reyes, J., Earl, A.M., and al., e. (2014). Relating the metatranscriptome and metagenome of the human gut. Proc Natl Acad Sci 10, 2329-2338.
139. Rowland, I., Gibson, G., Heinken, A., Scott, K., Swann, J., Thiele, I., and Tuohy, K. (2018). Gut microbiota functions: metabolism of nutrients and other food components. Eur J Nutr 57, 1-24.
140. Flint, H.J., Duncan, S.H., Scott, K.P., and Louis, P. (2015). Links between diet, gut microbiota composition and gut metabolism. Proc Nutr Soc 74, 13-22.
141. Reichardt, N., Duncan, S.H., Young, P., Belenguer, A., McWilliam Leitch, C., Scott, K.P., Flint, H.J., and Louis, P. (2014). Phylogenetic distribution of three pathways for propionate production within the human gut microbiota. ISME J 8, 1323-1335.

142. Levy, R., Carr, R., Kreimer, A., Freilich, S., and Borenstein, E. (2015). NetCooperate: a network-based tool for inferring host-microbe and microbe-microbe cooperation. BMC Bioinformatics 16, 164.
143. Gielda, L.M., and DiRita, V.J. (2012). Zinc competition among the intestinal microbiota. MBio 3, e00171-00112.
144. Lopez, C.A., and Skaar, E.P. (2018). The Impact of Dietary Transition Metals on Host-Bacterial Interactions. Cell Host Microbe 23, 737-748.
145. Drakesmith, H., and Prentice, A. (2008). Viral infection and iron metabolism. Nat Rev Microbiol 6, 541-552.
146. Andreesen, J., Schaupp, A., Neurauter, C., Brown, A., and Ljungdahl, L. (1973). Fermentation of Glucose, Fructose, and Xylose by Clostridium thermoaceticum: Effect of Metals on Growth Yields, Enzymes, and the Synthesis of Acetate from CO2. Journal of Bateriology 114, 743-751.
147. Shinichi, N., Satoshi, N., lKiyotaka, Y., Naomi, T., and Shoki, N. (1982). Carbohydrate Fermentation byClostridium difficile. Microbiology and Immunology 26, 107-111.
148. Kaiko, G.E., Ryu, S.H., Koues, O.I., Collins, P.L., Solnica-Krezel, L., Pearce, E.J., Pearce, E.L., Oltz, E.M., and Stappenbeck, T.S. (2016). The Colonic Crypt Protects Stem Cells from Microbiota-Derived Metabolites. Cell 165, 1708-1720.
149. Raskov, H., Kragh, K.N., Bjarnsholt, T., Alamili, M., and Gogenur, I. (2018). Bacterial biofilm formation inside colonic crypts may accelerate colorectal carcinogenesis. Clin Transl Med 7, 30.
150. Suez, J., Zmora, N., Zilberman-Schapira, G., Mor, U., Dori-Bachash, M., Bashiardes, S., Zur, M., Regev-Lehavi, D., Ben-Zeev Brik, R., Federici, S., et al. (2018). Post-Antibiotic Gut Mucosal Microbiome Reconstitution Is Impaired by Probiotics and Improved by Autologous FMT. Cell 174, 1406-1423 e1416.
151. Zmora, N., Zilberman-Schapira, G., Suez, J., Mor, U., Dori-Bachash, M., Bashiardes, S., Kotler, E., Zur, M., Regev-Lehavi, D., Brik, R.B., et al. (2018). Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is Associated with Unique Host and Microbiome Features. Cell 174, 1388-1405 e1321.
152. Cobey, S., and Lipsitch, M. (2013). Pathogen diversity and hidden regimes of apparent competition. Am Nat 181, 12-24.
153. Obadia, B., Guvener, Z.T., Zhang, V., Ceja-Navarro, J.A., Brodie, E.L., Ja, W.W., and Ludington, W.B. (2017). Probabilistic Invasion Underlies Natural Gut Microbiome Stability. Curr Biol 27, 1999-2006 e1998.
154. Putignani, L., and Dallapiccola, B. (2016). Foodomics as part of the host-microbiota-exposome interplay. J Proteomics 147, 3-20.
155. Jiang, C., Wang, X., Li, X., Inlora, J., Wang, T., Liu, Q., and Snyder, M. (2018). Dynamic Human Environmental Exposome Revealed by Longitudinal Personal Monitoring. Cell 175, 277-291 e231.
156. Moeller, A.H., Suzuki, T.A., Lin, D., Lacey, E.A., Wasser, S.K., and Nachman, M.W. (2017). Dispersal limitation promotes the diversification of the mammalian gut microbiota. Proc Natl Acad Sci U S A.
157. Afshinnekoo, E., Meydan, C., Chowdhury, S., Jaroudi, D., Boyer, C., Bernstein, N., Maritz, J.M., Reeves, D., Gandara, J., Chhangawala, S., et al. (2015). Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. Cell Syst 1, 72-87.
158. Stamper, C.E., Hoisington, A.J., Gomez, O.M., Halweg-Edwards, A.L., Smith, D.G., Bates, K.L., Kinney, K.A., Postolache, T.T., Brenner, L.A., Rook, G.A., et al. (2016). The Microbiome of the Built Environment and Human Behavior: Implications for Emotional Health and Well-Being in Postmodern Western Societies. Int Rev Neurobiol 131, 289-323.
159. Meadow, J.F., Altrichter, A.E., and Green, J.L. (2014). Mobile phones carry the personal microbiome of their owners. PeerJ 2, e447.
160. O'Connor, G.T., Lynch, S.V., Bloomberg, G.R., Kattan, M., Wood, R.A., Gergen, P.J., Jaffee, K.F., Calatroni, A., Bacharier, L.B., Beigelman, A., et al. (2018). Early-life home environment and risk of asthma among inner-city children. J Allergy Clin Immunol 141, 1468-1475.
161. Costello, E.K., Carlisle, E.M., Bik, E.M., Morowitz, M.J., and Relman, D.A. (2013). Microbiome assembly across multiple body sites in low-birthweight infants. MBio 4, e00782-00713.
162. Manuck, T.A. (2017). Racial and ethnic differences in preterm birth: A complex, multifactorial problem. Semin Perinatol.

163. Sala, C., Vitali, S., Giampieri, E., do Valle, I.F., Remondini, D., Garagnani, P., Bersanelli, M., Mosca, E., Milanesi, L., and Castellani, G. (2016). Stochastic neutral modelling of the Gut Microbiota's relative species abundance from next generation sequencing data. BMC Bioinformatics 17 Suppl 2, 16.
164. Neu, J., and Rushing, J. (2011). Cesarean versus vaginal delivery: long-term infant outcomes and the hygiene hypothesis. Clin Perinatol 38, 321-331.
165. Uhr, G.T., Dohnalova, L., and Thaiss, C.A. (2019). The Dimension of Time in Host-Microbiome Interactions. mSystems 4.
166. Stephens, W.Z., Wiles, T.J., Martinez, E.S., Jemielita, M., Burns, A.R., Parthasarathy, R., Bohannan, B.J., and Guillemin, K. (2015). Identification of Population Bottlenecks and Colonization Factors during Assembly of Bacterial Communities within the Zebrafish Intestine. MBio 6, e01163-01115.
167. Pearson, M.M., Sebaihia, M., Churcher, C., Quail, M.A., Seshasayee, A.S., Luscombe, N.M., Abdellah, Z., Arrosmith, C., Atkin, B., Chillingworth, T., et al. (2008). Complete genome sequence of uropathogenic Proteus mirabilis, a master of both adherence and motility. J Bacteriol 190, 4027-4037.
168. Smith, H.F., Parker, W., Kotzé, S.H., and Laurin, M. (2017). Morphological evolution of the mammalian cecum and cecal appendix. Comptes Rendus Palevol 16, 39-57.
169. Relman, D.A. (2012). The human microbiome: ecosystem resilience and health. Nutr Rev 70 Suppl 1, S2-9.
170. Moss, E.L., Falconer, S.B., Tkachenko, E., Wang, M., Systrom, H., Mahabamunuge, J., Relman, D.A., Hohmann, E.L., and Bhatt, A.S. (2017). Long-term taxonomic and functional divergence from donor bacterial strains following fecal microbiota transplantation in immunocompromised patients. PLoS One 12, e0182585.
171. Zaneveld, J.R., McMinds, R., and Vega Thurber, R. (2017). Stress and stability: applying the Anna Karenina principle to animal microbiomes. Nat Microbiol 2, 17121.
172. Hester, E.R., Barott, K.L., Nulton, J., Vermeij, M.J., and Rohwer, F.L. (2016). Stable and sporadic symbiotic communities of coral and algal holobionts. ISME J 10, 1157-1169.
173. Burns, A.R., Stephens, W.Z., Stagaman, K., Wong, S., Rawls, J.F., Guillemin, K., and Bohannan, J.M. (2015). Contribution of neutral processes to the assembly of gut microbial communities in the zebrafish over host development. ISME Journal 10, 655-664.
174. David, L.A., Materna, A.C., Friedman, J., Campos-Baptista, M.I., Blackburn, M.C., Perrotta, A., Erdman, S.E., and Alm, E.J. (2014). Host lifestyle affects human microbiota on daily timescales. Genome Biol 15, R89.
175. Franzenburg, S., Walter, J., Künzel, S., Wang, J., Baines, J.F., Bosch, T.C., and Fraune, S. (2013). Distinct antimicrobial peptide expression determines host species-specific bacterial associations. Proc Natl Acad Sci 110, E3730-3738.
176. Org, E., Parks, B.W., Joo, J.W.L., Emert, B., Schwartzman, W., Kang, E.Y., Mehrabian, M., Pan, C., Knight, R., Gunsalus, R., et al. (2015). Genetic and environmental control of host-gut microbiota interactions. Genome Res 25, 1558-1569.
177. Rawls, J.F., Mahowald, M.A., Ley, R.E., and Gordon, J.I. (2006). Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. Cell 127, 423-433.
178. Sanders, J.G., Powell, S., Kronauer, D.J., Vasconcelos, H.L., Frederickson, M.E., and Pierce, N.E. (2014). Stability and phylogenetic correlation in gut microbiota: lessons from ants and apes. Mol Ecol 23, 1268-1283.
179. Easson, C.G., and Thacker, R.W. (2014). Phylogenetic signal in the community structure of host-specific microbiomes of tropical marine sponges. Front Microbiol 5, 532.
180. Phillips, C.D., Phelan, G., Dowd, S.E., McDonough, M.M., Ferguson, A.W., Hanson, J.D., Siles, L., Ordóñez-Garza, N., San Francisco, M., and Baker, R.J. (2012). Microbiome analysis among bats describes influences of host phylogeny, life history, physiology, and geography. Mol Ecol 21, 2617-2627.
181. Dietrich, C., Köhler, T., and Brune, A. (2014). The cockroach origin of the termite gut microbiota: patterns in bacterial community structure reflect major evolutionary events. Appl Environ Microbiol 80, 2261-2269.
182. Chandler, J.A., Lang, J.M., Bhatnagar, S., Eisen, J.A., and Kopp, A. (2011). Bacterial communities of diverse Drosophila species: ecological context of a host-microbe model system. PLoS Genet 7, e1002272.
183. Wong, A.C., Chaston, J.M., and Douglas, A.E. (2013). The inconstant gut microbiota of Drosophila species revealed by 16S rRNA gene analysis. ISME J 7, 1922-1932.

184. Staubach, F., Baines, J.F., Kunzel, S., Bik, E.M., and Petrov, D.A. (2013). Host species and environmental effects on bacterial communities associated with Drosophila in the laboratory and in the natural environment. PLoS One 8, e70749.
185. Hird, S.M., Sánchez, C., Carstens, B.C., and Brumfield, R.T. (2015). Comparative gut microbiota of 59 Neotropical bird species. Front Microbiol 6, 1403.
186. Baxter, N.T., Wan, J.J., Schubert, A.M., Jenior, M.L., Myers, P., and Schloss, P.D. (2015). Intra- and interindividual variations mask interspecies variations in the microbiota of sympatric *Peromyscus* populations. Appl Environ Microbiol 81, 396-404.
187. Carmody, R.N., Gerber, G.K., Luevano, J.M., Gatti, D.M., Somes, L., Svenson, K.L., and Turnbaugh, P.J. (2015). Diet dominates host genotype in shaping the murine gut microbiota. Cell Host & Microbe 17, 72-84.
188. Shannon, C.E. (1948). A mathematical theory of communication. Bell Syst Tech J 27, 379-423.
189. Neafsey, D.E., Waterhouse, R.M., Abai, M.R., Aganezov, S.S., Alekseyev, M.A., Allen, J.E., Amon, J., Arcà, B., Arensburger, P., Artemov, G., et al. (2015). Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. Science 347, 1258522.
190. Werren, J.H., Richards, S., Desjardins, C.A., Niehuis, O., Gadau, J., Colbourne, J.K., and al., e. (2010). Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. Science 327, 343-348.
191. Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., and al., e. (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450, 203-218.
192. Steiper, M.E., and Young, N.M. (2006). Primate molecular divergence dates. Mol Phylogenet Evol 41, 384-394.
193. Weber, J.N., and Hoekstra, H.E. (2009). The evolution of burrowing behaviour in deer mice (genus *Peromyscus*). Anim Behav 77, 603-609.
194. Wong, A.C.N., Chaston, J.M., and Douglas, A.E. (2013). The inconstant gut microbiota of *Drosophila* species revealed by 16S rRNA gene analysis. ISME Journal 7, 1922-1932.
195. Shin, S.C., Kim, S.-H., You, H., Kim, B., Kim, A.C., Lee, K.-A., Yoon, J.-H., Ryu, J.-H., and Lee, W.-J. (2011). *Drosophila* microbiome modulates host developmental and metabolic homeostasis via insulin signaling. Science 334, 670-674.
196. Everard, A., Lazarevic, V., Gaia, N., Johansson, M., Stahlman, M., Backhed, F., Delzenne, N.M., Schrenzel, J., Francois, P., and Cani, P.D. (2014). Microbiome of prebiotic-treated mice reveals novel targets involved in host response during obesity. ISME J 8, 2116-2130.
197. Bogdanowicz, D., and Giaro, K. (2013). On a matching distance between rooted phylogenetic trees. International Journal of Applied Mathematics and Computer Science 23.
198. Robinson, D.F., and Foulds, L.R. (1981). Comparison of phylogenetic trees. Math Biosci 53, 131-147.
199. Platt, R.N., 2nd, Amman, B.R., Keith, M.S., Thompson, C.W., and Bradley, R.D. (2015). What Is Peromyscus? Evidence from nuclear and mitochondrial DNA sequences suggests the need for a new classification. J Mammal 96, 708-719.
200. Stepphan, S.J., Adkins, R.M., and Anderson, J. (2004). Phylogeny and divergence-date estimates of rapid radiations in muroid rodents based on multiple nuclear genes. Syst Biol 53, 533-553.
201. Kohl, K.D., Stengel, A., and Dearing, M.D. (2016). Inoculation of tannin-degrading bacteria into novel hosts increases performance on tannin-rich diets. Environ Microbiol Online Ahead of Print.
202. Munger, J.C., and Karasov, W.H. (1989). Sublethal parasites and host energy budgets: tapeworm infection in white-footed mice. Ecology 70, 904-921.
203. Shropshire, J.D., Opstal, E.J., and Bordenstein, S.R. (2016). An optimized approach to germ-free rearing in the jewel wasp *Nasonia*. PeerJ 4, e2316.
204. Yun, J.H., Roh, S.W., Whon, T.W., Jung, M.J., Kim, M.S., Park, D.S., Yoon, C., Nam, Y.D., Kim, Y.J., Choi, J.H., et al. (2014). Insect gut bacterial diversity determined by environmental habitat, diet, developmental stage, and phylogeny of host. Appl Environ Microbiol 80, 5254-5264.
205. Chaston, J.M., Dobson, A.J., Newell, P.D., and Douglas, A.E. (2016). Host genetic control of the microbiota mediates the *Drosophila* nutritional phenotypes. Appl Environ Microbiol 82, 671-679.

206. Obbard, D.J., Welch, J.J., Kim, K.-W., and Jiggins, F.M. (2009). Quantifying adaptive evolution in the *Drosophila* immune system. PLoS Genet 5, e1000698.
207. Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J., et al. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3, e170.
208. Hooper, L.V., and Gordon, J.I. (2001). Glycans as legislators of host-microbial interactions: spanning the spectrum from symbiosis to pathogenicity. Glycobiol 11, 1-10.
209. McLoughlin, K., Schluter, J., Rakoff-Nahoum, S., Smith, A.L., and Foster, K.R. (2016). Host Selection of Microbiota via Differential Adhesion. Cell Host Microbe 19, 550-559.
210. Kashyap, P.C., Macobal, A., Ursell, L.K., Smits, S.A., Sonnenburg, E.D., Costello, E.K., Higginbottom, S.K., Domino, S.E., Holmes, S.P., Relman, D.A., et al. (2013). Genetically dictated change in host mucus carbohydrate landscape exerts a diet-dependent effect on the gut microbiota. Proc Natl Acad Sci 110, 17059-17064.
211. Liang, X., Bushman, F.D., and FitzGerald, G.A. (2015). Rhythmicity of the intestinal microbiota is regulated by gender and the host circadian clock. Proc Natl Acad Sci 112, 10479-10484.
212. Liu, S., Pires da Cunha, A., Rezende, R.M., Cialic, R., Wei, Z., Bry, L., Comstock, L.E., Gandhi, R., and Weiner, H.L. (2016). The host shapes the gut microbiota via fecal microRNA. Cell Host & Microbe 19, 32-43.
213. Malo, M.S., Alam, S.N., Mostafa, G., Zeller, S.J., Johnson, P.V., Mohammad, N., Chen, K.T., Moss, A.K., Ramasamy, S., Faruqui, A., et al. (2010). Intestinal alkaline phosphatase preserves the normal homeostasis of gut microbiota. Gut 59, 1476-1484.
214. Costello, E.K., Stagaman, K., Dethlefsen, L., Bohannan, B.J.M., and Relman, D.A. (2012). The application of ecological theory toward an understanding of the human microbiome. Science 336.
215. McFall-Ngai, M., Hadfield, M.G., Bosch, T.C., Carey, H.V., Domazet-Loso, T., Douglas, A.E., Dubilier, N., Eberl, G., Fukami, T., Gilbert, S.F., et al. (2013). Animals in a bacterial world, a new imperative for the life sciences. Proc Natl Acad Sci U S A 110, 3229-3236.
216. Seedorf, H., Griffin, N.W., Ridaura, V.K., Reyes, A., Cheng, J., Rey, F.E., Smith, M.I., Simon, G.M., Scheffrahn, R.H., Woebken, D., et al. (2014). Bacteria from diverse habitats colonize and compete in the mouse gut. Cell 159, 253-266.
217. van Opstal, E.J., and Bordenstein, S.R. (2015). Rethinking heritability of the microbiome. Science 349, 1172-1173.
218. Funkhouser, L.J., and Bordenstein, S.R. (2013). Mom knows best: the universality of maternal microbial transmission. PLoS Biol 11, e1001631.
219. Sharp, K.H., Eam, B., Faulkner, D.J., and Haygood, M.G. (2007). Vertical transmission of diverse microbes in the tropical sponge *Corticium* sp. Appl Environ Microbiol 73, 622-629.
220. Pantoja-Feliciano, I.G., Clemente, J.C., Costello, E.K., Perez, M.E., Blaser, M.J., Knight, R., and Dominguez-Bello, M.G. (2013). Biphasic assembly of the murine intestinal microbiota during development. ISME Journal 7, 1112-1115.
221. <1080748_HB QIAamp DNA Microbiome Kit (2).pdf>.
222. Agler, M.T., Ruhe, J., Kroll, S., Morhenn, C., Kim, S.T., Weigel, D., and Kemen, E.M. (2016). Microbial hub taxa link host and abiotic factors to plant microbiome variation. PLoS Biol 14, e1002352.
223. Fisher, C.K., and Mehta, P. (2014). Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. PLoS One 9, e102451.
224. Murfin, K.E., Lee, M.-M., Klassen, J.L., McDonald, B.R., Larget, B., Forst, S., Stock, S.P., Currie, C.R., and Goodrich-Blair, H. (2015). *Xenorhabdus bovienii* strain diversity impacts coevolution and symbiotic maintenance with *Steinernema* spp. nematode hosts. mBio 6, e00076-00015.
225. Wang, J., Kalyan, S., Steck, N., Turner, L.M., Harr, B., Künzel, S., Vallier, M., Häsler, R., Franke, A., Oberg, H.-H., et al. (2015). Analysis of intestinal microbiota in hybrid house mice reveals evolutionary divergence in a vertebrate hologenome. Nat Comm 6, 6440.

226. Boot, R., Koopman, J.P., Kruijt, B.C., Lammers, R.M., Kennis, H.M., Lankhorst, A., Mullink, J.W.M.A., Stadhouders, A.M., De Boer, H., Welling, G.W., et al. (1985). The 'normalization' of germ-free rabbits with host-specific caecal microflora. Lab Anim 19, 344-352.
227. Wostmann, B.S. (1996). Germfree and gnotobiotic animal models: background and applications.(Boca Raton, FL: CRC Press).
228. Chung, H., Pamp, S.J., Hill, J.A., Surana, N.K., Edelman, S.M., Troy, E.B., Reading, N.C., Villablanca, E.J., Wang, S., Mora, J.R., et al. (2012). Gut immune maturation depends on colonization with a host-specific microbiota. Cell 149, 1578-1593.
229. Alegado, R.A., Brown, L.W., Cao, S., Dermenjian, R.K., Zuzow, R., Fairclough, S.R., Clardy, J., and King, N. (2012). A bacterial sulfonolipid triggers multicellular development in the closest living relatives of animals. eLife 1, e00013.
230. Roth, O., Sadd, B.M., Schmid-Hempel, P., and Kurtz, J. (2009). Strain-specific priming of resistance in the red flour beetle, *Tribolium castaneum*. Proc R Soc B 276, 145-151.
231. Bordenstein, S.R., and Theis, K.R. (2015). Host biology in light of the microbiome: ten principles of holobionts and hologenomes. PLoS Biol 13, e1002226.
232. Shapira, M. (2016). Gut microbiotas and host evolution: scaling up symbiosis. Trends Ecol Evol Online Ahead of Print.
233. Theis, K.R., Dheilly, N.M., Klassen, J.L., Brucker, R.M., Baines, J.F., Bosch, T.C.G., Cryan, J.F., Gilbert, S.F., Goodnight, C.J., Lloyd, E.A., et al. (2016). Getting the hologenome concept right: An eco-evolutionary framework for hosts and their microbiomes. mSystems 1, e00028-00016.
234. Brucker, R.M., and Bordenstein, S.R. (2012). The roles of host evolutionary relationships (genus: *Nasonia*) and development in structuring microbial communities. Evol 66, 349-362.
235. Brucker, R.M., and Bordenstein, S.R. (2012). Speciation by symbiosis. Trends Ecol Evol 27, 443-451.
236. Brucker, R.M., and Bordenstein, S.R. (2013). The hologenomic basis of speciation: gut bacteria cause hybrid lethality in the genus *Nasonia*. Science 341, 667-669.
237. Lo, N., Casiraghi, M., Salati, E., Bazzocchi, C., and Bandi, C. (2002). How many *Wolbachia* supergroups exist? Mol Biol Evol 19, 341-346.
238. Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Gonzalez Pena, A., Goodrich, J.K., Gordon, J.I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat Method 7, 335-336.
239. Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27, 2194-2200.
240. McDonald, D., Price, M.N., Goodrich, J.K., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME Journal 6, 610-618.
241. Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460-2461.
242. Russell, D.J., Otu, H.H., and Sayood, K. (2008). Grammar-based distance in progressive multiple sequence alignment. BMC Bioinformatics 9, 306.
243. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 5, e9490.
244. McDonald, D., Clemente, J.C., Kuczynski, J., Rideout, J.R., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., et al. (2010). The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. GigaScience 1, 7.
245. Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., et al. (2007). Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2, 2366-2382.
246. McMurdie, P.J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8, e61217.
247. Patwardhan, A., Ray, S., and Roy, A. (2014). Molecular markers in phylogenetic studies - a review. J Phylogen Evol Biol 2, 131.

248. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acid Res 32, 1792-1797.
249. Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. Nat Method 9, 772-772.
250. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312-1313.
251. Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59, 307-321.
252. Bogdanowicz, D., Giaro, K., and Wróbel, B. (2012). TreeCmp: comparison of trees in polynomial time. Evol Bioinform Online 8, 475.
253. Bogdanowicz, D., and Giaro, K. (2013). On a matching distance between rooted phylogenetic trees. Int J Appl Math Comp Sci 23, 669-684.
254. Millman, K.J., and Aivazis, M. (2011). Python for scientists and engineers. Comp Sci Eng 13, 9-12.
255. Oliphant, T.E. (2007). Python for scientific computing. Comp Sci Eng 9, 10-20.
256. Meagher, S., Penn, D., and Potts, W. (2000). Male-male competition magnifies inbreeding depression in wild house mice. Proc Natl Acad Sci 97, 3324-3329.
257. Gaukler, S.M., Ruff, J.S., Galland, T., Underwood, T.K., Kandaris, K.A., Liu, N.M., Morrison, L.C., Veranth, J.M., and Potts, W.K. (2016). Quantification of cerivastatin toxicity supports organismal performance assays as an effective tool during pharmaceutical safety assessment. Evol Appl Online Ahead of Print.
258. Zhang, C., Derrien, M., Levenez, F., Brazeilles, R., Ballal, S., Kim, J., Degivry, M.-C., Quéré, G., Garault, P., van Hylckama Vlieg, J.E.T., et al. (2016). Ecological robustness of the gut microbiota in response to the ingestion of transient food-borne microbes. ISME Journal 10, 2235-2245.
259. Platt, R.N., Amman, B.R., Keith, M.S., Thompson, C.W., and Bradley, R.D. (2015). What Is *Peromyscus*? Evidence from nuclear and mitochondrial DNA sequences suggests the need for a new classification. J Mamm 96, 708-719.
260. Phillips, C.D., Phelan, G., Dowd, S.E., McDonough, M.M., Ferguson, A.W., Delton Hanson, J., Siles, L., Ordonez-Garza, N., San Francisco, M., and Baker, R.J. (2012). Microbiome analysis among bats describes influences of host phylogeny, life history, physiology and geography. Mol Ecol 21, 2617-2627.
261. Reyes, A., Haynes, M., Hanson, N., Angly, F.E., Heath, A.C., Rohwer, F., and Gordon, J.I. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466, 334-338.
262. Suttle, C.A. (2005). Viruses in the sea. Nature 437, 356-361.
263. Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S.A., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2011). The human gut virome: inter-individual variation and dynamic response to diet. Genome Res 21, 1616-1625.
264. Leigh, B.A., Djurhuus, A., Breitbart, M., and Dishaw, L.J. (2018). The gut virome of the protochordate model organism, Ciona intestinalis subtype A. Virus Res 244, 137-146.
265. Fawaz, M., Vijayakumar, P., Mishra, A., Gandhale, P.N., Dutta, R., Kamble, N.M., Sudhakar, S.B., Roychoudhary, P., Kumar, H., Kulkarni, D.D., et al. (2016). Duck gut viral metagenome analysis captures snapshot of viral diversity. Gut Pathog 8, 30.
266. Sachsenroder, J., Twardziok, S.O., Scheuch, M., and Johne, R. (2014). The general composition of the faecal virome of pigs depends on age, but not on feeding with a probiotic bacterium. PLoS One 9, e88888.
267. Kim, M.S., and Bae, J.W. (2018). Lysogeny is prevalent and widely distributed in the murine gut microbiota. ISME J.
268. Minot, S., Bryson, A., Chehoud, C., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2013). Rapid evolution of the human gut virome. Proc Natl Acad Sci U S A 110, 12450-12455.
269. Reyes, A., Wu, M., McNulty, N.P., Rohwer, F.L., and Gordon, J.I. (2013). Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. Proc Natl Acad Sci U S A 110, 20236-20241.
270. Knowles, B., Silveira, C.B., Bailey, B.A., Barott, K., Cantu, V.A., Cobian-Guemes, A.G., Coutinho, F.H., Dinsdale, E.A., Felts, B., Furby, K.A., et al. (2016). Lytic to temperate switching of viral communities. Nature 531, 466-470.

271. Silveira, C.B., and Rohwer, F.L. (2016). Piggyback-the-Winner in host-associated microbial communities. NPJ Biofilms Microbiomes 2, 16010.
272. Gama, J.A., Reis, A.M., Domingues, I., Mendes-Soares, H., Matos, A.M., and Dionisio, F. (2013). Temperate bacterial viruses as double-edged swords in bacterial warfare. PLoS One 8, e59043.
273. Brussow, H., Canchaya, C., and Hardt, W.D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev 68, 560-602, table of contents.
274. Madsen, J.S., Burmolle, M., Hansen, L.H., and Sorensen, S.J. (2012). The interconnection between biofilm formation and horizontal gene transfer. FEMS Immunol Med Microbiol 65, 183-195.
275. Molin, S., and Tolker-Nielsen, T. (2003). Gene transfer occurs with enhanced efficiency in biofilms and induces enhanced stabilisation of the biofilm structure. Curr Opin Biotechnol 14, 255-261.
276. Manrique, P., Bolduc, B., Walk, S.T., van der Oost, J., de Vos, W.M., and Young, M.J. (2016). Healthy human gut phageome. Proc Natl Acad Sci U S A 113, 10400-10405.
277. Grasis, J.A., Lachnit, T., Anton-Erxleben, F., Lim, Y.W., Schmieder, R., Fraune, S., Franzenburg, S., Insua, S., Machado, G., Haynes, M., et al. (2014). Species-specific viromes in the ancestral holobiont Hydra. PLoS One 9, e109952.
278. Soffer, N., Zaneveld, J., and Vega Thurber, R. (2015). Phage-bacteria network analysis and its implication for the understanding of coral disease. Environ Microbiol 17, 1203-1218.
279. Chafee, M.E., Zecher, C.N., Gourley, M.L., Schmidt, V.T., Chen, J.H., Bordenstein, S.R., Clark, M.E., and Bordenstein, S.R. (2011). Decoupling of host-symbiont-phage coadaptations following transfer between insect species. Genetics 187, 203-215.
280. Fujii, Y., Kubo, T., Ishikawa, H., and Sasaki, T. (2004). Isolation and characterization of the bacteriophage WO from Wolbachia, an arthropod endosymbiont. Biochem Biophys Res Commun 317, 1183-1188.
281. Campbell, B.C., Steffen-Campbell, J.D., and Werren, J.H. (1993). Phylogeny of the Nasonia species complex (Hymenoptera: Pteromalidae) inferred from an internal transcribed spacer (ITS2) and 28S rDNA sequences. Insect Mol Biol 2, 225-237.
282. Bordenstein, S.R., O'Hara, F.P., and Werren, J.H. (2001). Wolbachia-induced incompatibility precedes other hybrid incompatibilities in Nasonia. Nature 409, 707-710.
283. Bordenstein, S.R., and Werren, J.H. (2007). Bidirectional incompatibility among divergent Wolbachia and incompatibility level differences among closely related Wolbachia in Nasonia. Heredity (Edinb) 99, 278-287.
284. Werren, J.H., Richards, S., Desjardins, C.A., Niehuis, O., Gadau, J., Colbourne, J.K., Nasonia Genome Working, G., Werren, J.H., Richards, S., Desjardins, C.A., et al. (2010). Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. Science 327, 343-348.
285. Cowles, K.N., and Goodrich-Blair, H. (2005). Expression and activity of a Xenorhabdus nematophila haemolysin required for full virulence towards Manduca sexta insects. Cell Microbiol 7, 209-219.
286. Bolduc, B., Youens-Clark, K., Roux, S., Hurwitz, B.L., and Sullivan, M.B. (2017). iVirus: facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. ISME J 11, 7-14.
287. Lengyel, K., Lang, E., Fodor, A., Szallas, E., Schumann, P., and Stackebrandt, E. (2005). Description of four novel species of Xenorhabdus, family Enterobacteriaceae: Xenorhabdus budapestensis sp. nov., Xenorhabdus ehlersii sp. nov., Xenorhabdus innexi sp. nov., and Xenorhabdus szentirmaii sp. nov. Syst Appl Microbiol 28, 115-122.
288. Bordenstein, S.R., Marshall, M.L., Fry, A.J., Kim, U., and Wernegreen, J.J. (2006). The tripartite associations between bacteriophage, Wolbachia, and arthropods. PLoS Pathog 2, e43.
289. Bordenstein, S.R., and Bordenstein, S.R. (2016). Eukaryotic association module in phage WO genomes from Wolbachia. Nat Commun 7, 13155.
290. Chaston, J.M., Suen, G., Tucker, S.L., Andersen, A.W., Bhasin, A., Bode, E., Bode, H.B., Brachmann, A.O., Cowles, C.E., Cowles, K.N., et al. (2011). The entomopathogenic bacterial endosymbionts Xenorhabdus and Photorhabdus: convergent lifestyles from divergent genomes. PLoS One 6, e27909.

291. Goodrich-Blair, H., and Clarke, D.J. (2007). Mutualism and pathogenesis in Xenorhabdus and Photorhabdus: two roads to the same destination. Mol Microbiol 64, 260-268.
292. Sicard, M., Ferdy, J.B., Pages, S., Le Brun, N., Godelle, B., Boemare, N., and Moulia, C. (2004). When mutualists are pathogens: an experimental study of the symbioses between Steinernema (entomopathogenic nematodes) and Xenorhabdus (bacteria). J Evol Biol 17, 985-993.
293. Kohl, K.D., Varner, J., Wilkening, J.L., and Dearing, M.D. (2018). Gut microbial communities of American pikas (Ochotona princeps): Evidence for phylosymbiosis and adaptations to novel diets. J Anim Ecol 87, 323-330.
294. Kohl, K.D., Dearing, M.D., and Bordenstein, S.R. (2017). Microbial communities exhibit host species distinguishability and phylosymbiosis along the length of the gastrointestinal tract. Mol Ecol.
295. Shultz, L.D., Ishikawa, F., and Greiner, D.L. (2007). Humanized mice in translational biomedical research. Nat Rev Immunol 7, 118-130.
296. Degnan, P.H., and Moran, N.A. (2008). Diverse phage-encoded toxins in a protective insect endosymbiont. Appl Environ Microbiol 74, 6782-6791.
297. Toh, H., Weiss, B.L., Perkin, S.A., Yamashita, A., Oshima, K., Hattori, M., and Aksoy, S. (2006). Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of Sodalis glossinidius in the tsetse host. Genome Res 16, 149-156.
298. Moran, N.A., Degnan, P.H., Santos, S.R., Dunbar, H.E., and Ochman, H. (2005). The players in a mutualistic symbiosis: insects, bacteria, viruses, and virulence genes. Proc Natl Acad Sci U S A 102, 16919-16926.
299. LePage, D.P., Metcalf, J.A., Bordenstein, S.R., On, J., Perlmutter, J.I., Shropshire, J.D., Layton, E.M., Funkhouser-Jones, L.J., Beckmann, J.F., and Bordenstein, S.R. (2017). Prophage WO genes recapitulate and enhance Wolbachia-induced cytoplasmic incompatibility. Nature 543, 243-247.
300. Shropshire, J.D., On, J., Layton, E.M., Zhou, H., and Bordenstein, S.R. (2018). One prophage WO gene rescues cytoplasmic incompatibility in Drosophila melanogaster. Proc Natl Acad Sci U S A 115, 4987-4991.
301. Barr, J.J., Auro, R., Furlan, M., Whiteson, K.L., Erb, M.L., Pogliano, J., Stotland, A., Wolkowicz, R., Cutting, A.S., Doran, K.S., et al. (2013). Bacteriophage adhering to mucus provide a non-host-derived immunity. Proc Natl Acad Sci U S A 110, 10771-10776.
302. Barr, J.J., Youle, M., and Rohwer, F. (2013). Innate and acquired bacteriophage-mediated immunity. Bacteriophage 3, e25857.
303. Nguyen, S., Baker, K., Padman, B.S., Patwa, R., Dunstan, R.A., Weston, T.A., Schlosser, K., Bailey, B., Lithgow, T., Lazarou, M., et al. (2017). Bacteriophage Transcytosis Provides a Mechanism To Cross Epithelial Cell Layers. MBio 8.
304. Fraune, S., and Bosch, T.C. (2007). Long-term maintenance of species-specific bacterial microbiota in the basal metazoan Hydra. Proc Natl Acad Sci U S A 104, 13146-13151.
305. Franzenburg, S., Fraune, S., Altrock, P.M., Kunzel, S., Baines, J.F., Traulsen, A., and Bosch, T.C. (2013). Bacterial colonization of Hydra hatchlings follows a robust temporal pattern. ISME J 7, 781-790.
306. Muller, H.E. (1986). Occurrence and pathogenic role of Morganella-Proteus-Providencia group bacteria in human feces. J Clin Microbiol 23, 404-405.
307. McDermott, C., and Mylotte, J.M. (1984). Morganella morganii: epidemiology of bacteremic disease. Infect Control 5, 131-137.
308. Hola, V., Peroutkova, T., and Ruzicka, F. (2012). Virulence factors in Proteus bacteria from biofilm communities of catheter-associated urinary tract infections. FEMS Immunol Med Microbiol 65, 343-349.
309. Albert, M.J., Alam, K., Ansaruzzaman, M., Islam, M.M., Rahman, A.S., Haider, K., Bhuiyan, N.A., Nahar, S., Ryan, N., Montanaro, J., et al. (1992). Pathogenesis of Providencia alcalifaciens-induced diarrhea. Infect Immun 60, 5017-5024.
310. O'Hara, C.M., Brenner, F.W., and Miller, J.M. (2000). Classification, identification, and clinical significance of Proteus, Providencia, and Morganella. Clin Microbiol Rev 13, 534-546.
311. Armbruster, C.E., Smith, S.N., Johnson, A.O., DeOrnellas, V., Eaton, K.A., Yep, A., Mody, L., Wu, W., and Mobley, H.L. (2017). The Pathogenic Potential of Proteus mirabilis Is Enhanced by Other Uropathogens during Polymicrobial Urinary Tract Infection. Infect Immun 85.

312. Armbruster, C.E., Smith, S.N., Yep, A., and Mobley, H.L. (2014). Increased incidence of urolithiasis and bacteremia during Proteus mirabilis and Providencia stuartii coinfection due to synergistic induction of urease activity. J Infect Dis 209, 1524-1532.
313. Galac, M.R., and Lazzaro, B.P. (2011). Comparative pathology of bacteria in the genus Providencia to a natural host, Drosophila melanogaster. Microbes Infect 13, 673-683.
314. Vicente, C.S., Nascimento, F.X., Espada, M., Barbosa, P., Hasegawa, K., Mota, M., and Oliveira, S. (2013). Characterization of bacterial communities associated with the pine sawyer beetle Monochamus galloprovincialis, the insect vector of the pinewood nematode Bursaphelenchus xylophilus. FEMS Microbiol Lett 347, 130-139.
315. Wei, T., Miyanaga, K., and Tanji, Y. (2014). Persistence of antibiotic-resistant and -sensitive Proteus mirabilis strains in the digestive tract of the housefly (Musca domestica) and green bottle flies (Calliphoridae). Appl Microbiol Biotechnol 98, 8357-8366.
316. Clifford, R.J., Hang, J., Riley, M.C., Onmus-Leone, F., Kuschner, R.A., Lesho, E.P., and Waterman, P.E. (2012). Complete genome sequence of Providencia stuartii clinical isolate MRSN 2154. J Bacteriol 194, 3736-3737.
317. Olaitan, A.O., Diene, S.M., Gupta, S.K., Adler, A., Assous, M.V., and Rolain, J.M. (2014). Genome analysis of NDM-1 producing Morganella morganii clinical isolate. Expert Rev Anti Infect Ther 12, 1297-1305.
318. Cho, I., and Blaser, M.J. (2012). The human microbiome: at the interface of health and disease. Nat Rev Genet 13, 260-270.
319. Greenblum, S., Turnbaugh, P.J., and Borenstein, E. (2012). Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. Proc Natl Acad Sci U S A 109, 594-599.
320. Morgan, X.C., Tickle, T.L., Sokol, H., Gevers, D., Devaney, K.L., Ward, D.V., Reyes, J.A., Shah, S.A., LeLeiko, N., Snapper, S.B., et al. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol 13, R79.
321. Backhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., et al. (2015). Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe 17, 852.
322. Breitbart, M., Haynes, M., Kelley, S., Angly, F., Edwards, R.A., Felts, B., Mahaffy, J.M., Mueller, J., Nulton, J., Rayhawk, S., et al. (2008). Viral diversity and dynamics in an infant gut. Res Microbiol 159, 367-373.
323. Lim, E.S., Zhou, Y., Zhao, G., Bauer, I.K., Droit, L., Ndao, I.M., Warner, B.B., Tarr, P.I., Wang, D., and Holtz, L.R. (2015). Early life dynamics of the human gut virome and bacterial microbiome in infants. Nat Med 21, 1228-1234.
324. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120.
325. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19, 455-477.
326. Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072-1075.
327. Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015). VirSorter: mining viral signal from microbial genomic data. PeerJ 3, e985.
328. Chen, H., and Boutros, P.C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics 12, 35.
329. Bolduc, B., Jang, H.B., Doulcier, G., You, Z.Q., Roux, S., and Sullivan, M.B. (2017). vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. PeerJ 5, e3243.
330. Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. Icwsm 8, 361-362.
331. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068-2069.

332. Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). InterProScan: protein domains identifier. Nucleic Acids Res 33, W116-120.
333. Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9, 772.
334. Huse, S.M., Ye, Y., Zhou, Y., and Fodor, A.A. (2012). A core human microbiome as viewed through 16S rRNA sequence clusters. PLoS One 7, e34242.
335. Wu, G.D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.Y., Keilbaugh, S.A., Bewtra, M., Knights, D., Walters, W.A., Knight, R., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. Science 334, 105-108.
336. Fierera, N., Hamadyc, M., Lauberb, C.L., and Knight, R. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. Procedings of the National Academy of Sciences 105.
337. Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490, 55-60.
338. Frank, D.N., Allison, A.L., Feldman, R.A., Boedeker, E.C., Harpaz, N., and Pace, N.R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Procedings of the National Academy of Sciences, 13780–13785.
339. Walters, W.A., Xu, Z., and Knight, R. (2014). Meta-analyses of human gut microbes associated with obesity and IBD. FEBS Lett 588, 4223-4233.
340. Zackular, J.P., Baxter, N.T., Iverson, K.D., Sadler, W.D., Petrosino, J.F., Chen, G.Y., and Schloss, P.D. (2013). The gut microbiome modulates colon tumorigenesis. MBio 4, e00692-00613.
341. Mason, M.R., Nagaraja, H.N., Camerlengo, T., Joshi, V., and Kumar, P.S. (2013). Deep sequencing identifies ethnicity-specific bacterial signatures in the oral microbiome. PLoS One 8, e77287.
342. Williams, D.R., Priest, N., and Anderson, N.B. (2016). Understanding associations among race, socioeconomic status, and health: Patterns and prospects. Health Psychol 35, 407-411.
343. Mersha, T.B., and Abebe, T. (2015). Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. Human Genomics 9, 1.
344. Rampelli, S., Schnorr, S.L., Consolandi, C., Turroni, S., Severgnini, M., Peano, C., Brigidi, P., Crittenden, A.N., Henry, A.G., and Candela, M. (2015). Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. Curr Biol 25, 1682-1693.
345. Deschasaux, M., Bouter, K.E., Prodan, A., Levin, E., Groen, A.K., Herrema, H., Tremaroli, V., Bakker, G.J., Attaye, I., Pinto-Sietsma, S.J., et al. (2018). Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. Nat Med.
346. Clarke, K.R., and Ainsworth, M. (1993). A method of linking multivariate community structure to environmental variables. Marine Ecology 92, 205-219
.
347. N.V., C., K.W., B., L.O., H., and W.P., K. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16, 321-357.
348. Fu, J., Bonder, M.J., Cenit, M.C., Tigchelaar, E.F., Maatman, A., Dekens, J.A., Brandsma, E., Marczynska, J., Imhann, F., Weersma, R.K., et al. (2015). The Gut Microbiome Contributes to a Substantial Proportion of the Variation in Blood Lipids. Circ Res 117, 817-824.
349. Lim, M.Y., You, H.J., Yoon, H.S., Kwon, B., Lee, J.Y., Lee, S., Song, Y.M., Lee, K., Sung, J., and Ko, G. (2016). The effect of heritability and host genetics on the gut microbiota and metabolic syndrome. Gut.
350. Almasy, L., and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62, 1198-1211.
351. Rothschild, D., Weissbrod, O., Barkan, E., Korem, T., Zeevi, D., Costea, P.I., Godneva, A., Kalka, I., Elinav, E., and Segal, E. (2017). Environmental factors dominate over host genetics in shaping human gut microbiota composition. BioRxiv.
352. Morgan, X.C., Tickle, T.L., Sokol, H., Gevers, D., and Huttenhower, C. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biology 7.

353. Lewis, J.D., Chen, E.Z., Baldassano, R.N., Otley, A.R., Griffiths, A.M., Lee, D., Bittinger, K., Bailey, A., Friedman, E.S., Hoffmann, C., et al. (2015). Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. Cell Host Microbe 18, 489-500.
354. Goker, M., Gronow, S., Zeytun, A., Nolan, M., Lucas, S., Lapidus, A., Hammon, N., Deshpande, S., Cheng, J.F., Pitluck, S., et al. (2011). Complete genome sequence of Odoribacter splanchnicus type strain (1651/6). Stand Genomic Sci 4, 200-209.
355. Castaneda, G., Liu, B., Torres, S., Bhuket, T., and Wong, R.J. (2017). Race/Ethnicity-Specific Disparities in the Severity of Disease at Presentation in Adults with Ulcerative Colitis: A Cross-Sectional Study. Dig Dis Sci.
356. Boucias, D.G., Cai, Y., Sun, Y., Lietze, V.U., Sen, R., Raychoudhury, R., and Scharf, M.E. (2013). The hindgut lumen prokaryotic microbiota of the termite Reticulitermes flavipes and its responses to dietary lignocellulose composition. Mol Ecol 22, 1836-1853.
357. LATHAM, M.J., and WOLIN, M.J. (1977). Fermentation of Cellulose by Ruminococcus flavefaciens in the Presence and Absence ofMethanobacterium ruminantium. Appl Environ Microbiol 34, 297-301.
358. Falony, G., and Raes, J. (2016). Population-level analysis of gut microbiome variation. Science 352, 560-564.
359. Biagi, E., Franceschi, C., Rampelli, S., Severgnini, M., Ostan, R., Turroni, S., Consolandi, C., Quercia, S., Scurti, M., Monti, D., et al. (2016). Gut Microbiota and Extreme Longevity. Curr Biol 26, 1480-1485.
360. Thevaranjan, N., Puchta, A., Schulz, C., Naidoo, A., Szamosi, J.C., Verschoor, C.P., Loukov, D., Schenck, L.P., Jury, J., Foley, K.P., et al. (2017). Age-Associated Microbial Dysbiosis Promotes Intestinal Permeability, Systemic Inflammation, and Macrophage Dysfunction. Cell Host Microbe 21, 455-466 e454.
361. Xie, W., Wood, A.R., Lyssenko, V., and al., e. (2013). Genetic Variants Associated With Glycine Metabolism and Their Role in Insulin Sensitivity and Type 2 Diabetes. Diabetes 62.
362. Williams, S.R., Yang, Q., Chen, F., Liu, X., Keene, K.L., Jacques, P., Chen, W.M., Weinstein, G., Hsu, F.C., Beiser, A., et al. (2014). Genome-wide meta-analysis of homocysteine and methionine metabolism identifies five one carbon metabolism loci and a novel association of ALDH1L1 with ischemic stroke. PLoS Genet 10, e1004214.
363. Petersen, L.M., Bautista, E.J., Nguyen, H., Hanson, B.M., Chen, L., Lek, S.H., Sodergren, E., and Weinstock, G.M. (2017). Community characteristics of the gut microbiomes of competitive cyclists. Microbiome 5, 98.
364. Nakamura, N., Lin, H.C., McSweeney, C.S., Mackie, R.I., and Gaskins, H.R. (2010). Mechanisms of microbial hydrogen disposal in the human colon and implications for health and disease. Annu Rev Food Sci Technol 1, 363-395.
365. Parthasarathy, G., Chen, J., Chen, X., Chia, N., O'Connor, H.M., Wolf, P.G., Gaskins, H.R., and Bharucha, A.E. (2016). Relationship Between Microbiota of the Colonic Mucosa vs Feces and Symptoms, Colonic Transit, and Methane Production in Female Patients With Chronic Constipation. Gastroenterology 150, 367-379 e361.
366. Jackson, C.S., Oman, M., Patel, A.M., and Vega, K.J. (2016). Health disparities in colorectal cancer among racial and ethnic minorities in the United States. J Gastrointest Oncol 7, S32-43.
367. Lopetuso, L.R., Scaldaferri, F., Petito, V., and Gasbarrini, A. (2013). Commensal Clostridia: leading players in the maintenance of gut homeostasis. Gut Pathogens.
368. Sy, D.F. The Center for Asian Health Engages Communities in Research to Reduce Asian American Health Disparities. US Department of Health & Human Services, National Institute on Minority Health and Health Disparities.
369. Hwang, H. (2013). Colorectal Cancer Screening among Asian Americans. Asian Pacific Journal of Cancer Prevention 14, 4025-4032.
370. Oh, K.M., Kreps, G.L., and Jun, J. (2013). Colorectal Cancer Screening Knowledge, Beliefs, and Practices of Korean Americans. American Journal of Health Behavior 37, 381-394.
371. Lab, K. (2018). Immigrant Microbiome Project. In. (www.knightslab.org/immigrant-microbiome-project, University of Minnesota.

372. Sankaranarayanan, R., Ramadas, K., and Qiao, Y.-l. (2014). Managing the changing burden of cancer in Asia. BMC Medicine 12.
373. Pourhoseingholi, M.A. (2012). Increased burden of colorectal cancer in Asia. World J Gastrointest Oncol 4, 68-70.
374. Pourhoseingholi, M.A., Vahedi, M., and Baghestani, A.R. (2015). Burden of gastrointestinal cancer in Asia; an overview. Gastroenterology and Hepatology.
375. Pourhoseingholi, M.A. (2014). Epidemiology and burden of colorectal cancer in Asia-Pacific region: what shall we do now? Translational Gastrointestinal Cancer 3, 169-173.
376. Report, C.H.D.a.I. (2013).
377. Cao, H., Liu, X., An, Y., Zhou, G., Liu, Y., Xu, M., Dong, W., Wang, S., Yan, F., Jiang, K., et al. (2017). Dysbiosis contributes to chronic constipation development via regulation of serotonin transporter in the intestine. Sci Rep 7, 10322.
378. Mosca, A., Leclerc, M., and Hugot, J.P. (2016). Gut Microbiota Diversity and Human Diseases: Should We Reintroduce Key Predators in Our Ecosystem? Front Microbiol 7, 455.
379. Zhang, X., Zhang, D., Jia, H., Feng, Q., Wang, D., Liang, D., Wu, X., Li, J., Tang, L., Li, Y., et al. (2015). The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nat Med 21, 895-905.
380. Singh, V.P., Proctor, S.D., and Willing, B.P. (2016). Koch's postulates, microbial dysbiosis and inflammatory bowel disease. Clin Microbiol Infect 22, 594-599.
381. Sherry ST, W.M., Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Research 29.
382. Consortium, G.T. (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet 45, 580-585.
383. Qiu, X., Wei, R., Li, Y., Zhu, Q., Xiong, C., Chen, Y., Zhang, Y., Lu, K., He, F., and Zhang, L. (2016). NEDL2 regulates enteric nervous system and kidney development in its Nedd8 ligase activity-dependent manner. Oncotarget 7.
384. Wei, R., Qiu, X., Wang, S., Li, Y., Wang, Y., Lu, K., Fu, Y., Xing, G., He, F., and Zhang, L. (2015). NEDL2 is an essential regulator of enteric neural development and GDNF/Ret signaling. Cell Signal 27, 578-586.
385. O'Donnell, A.M., Coyle, D., and Puri, P. (2016). Decreased expression of NEDL2 in Hirschsprung's disease. J Pediatr Surg 51, 1839-1842.
386. DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72, 5069-5072.
387. Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S.M., Betley, J., Fraser, L., Bauer, M., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6, 1621-1624.
388. Kuhn, M. (2017). A short introduction to the caret package.
389. Jones, E., Oliphant, T., and Peterson, P. (2001). Open Source Scientific Tools for Python.
390. Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature 526, 68-74.
391. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics 27, 2156-2158.
392. Mazel, F., Davis, K.M., Loudon, A., Kwong, W.K., Groussin, M., and Parfrey, L.W. (2018). Is Host Filtering the Main Driver of Phylosymbiosis across the Tree of Life? mSystems 3.
393. Deschasaux, M., Bouter, K.E., Prodan, A., Levin, E., Groen, A.K., Herrema, H., Tremaroli, V., Bakker, G.J., Attaye, I., Pinto-Sietsma, S.J., et al. (2018). Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. Nat Med 24, 1526-1531.
394. Brooks, A.W., Priya, S., Blekhman, R., and Bordenstein, S.R. (2018). Gut microbiota diversity across ethnicities in the United States. PLoS Biol 16, e2006842.

395. Griffin, N.W., Ahern, P.P., Cheng, J., Heath, A.C., Ilkayeva, O., Newgard, C.B., Fontana, L., and Gordon, J.I. (2017). Prior Dietary Practices and Connections to a Human Gut Microbial Metacommunity Alter Responses to Diet Interventions. Cell Host Microbe 21, 84-96.
396. Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J.M., Kennedy, S., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. Nature 500, 541-546.
397. Turnbaugh, P.J. (2017). Microbes and Diet-Induced Obesity: Fast, Cheap, and Out of Control. Cell Host Microbe 21, 278-281.
398. Turnbaugh, P.J., Backhed, F., Fulton, L., and Gordon, J.I. (2008). Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. Cell Host Microbe 3, 213-223.
399. Walter, J., and Ley, R. (2011). The human gut microbiome: ecology and recent evolutionary changes. Annu Rev Microbiol 65, 411-429.
400. Pajau Vangay, Abigail J. Johnson, Tonya L. Ward, ..., Purna C. Kashyap, Kathleen A. Culhane-Pera, and Knights, D. (2018). US Immigration Westernizes the Human Gut Microbiome. Cell 175, 962-972.
401. He, Y., Wu, W., Zheng, H.M., Li, P., McDonald, D., Sheng, H.F., Chen, M.X., Chen, Z.H., Ji, G.Y., Zheng, Z.D., et al. (2018). Regional variation limits applications of healthy gut microbiome reference ranges and disease models. Nat Med 24, 1532-1535.
402. Findley, K., Williams, D.R., Grice, E.A., and Bonham, V.L. (2016). Health Disparities and the Microbiome. Trends Microbiol 24, 847-850.
403. Fortenberry, J.D. (2013). The uses of race and ethnicity in human microbiome research. Trends Microbiol 21, 165-166.
404. Williams, D.R., Mohammed, S.A., Leavell, J., and Collins, C. (2010). Race, socioeconomic status, and health: complexities, ongoing challenges, and research opportunities. Ann N Y Acad Sci 1186, 69-101.
405. Jackson, M.A., Verdi, S., Maxan, M.E., Shin, C.M., Zierer, J., Bowyer, R.C.E., Martin, T., Williams, F.M.K., Menni, C., Bell, J.T., et al. (2018). Gut microbiota associations with common diseases and prescription medications in a population-based cohort. Nat Commun 9, 2655.
406. Zackular, J.P., Rogers, M.A., Ruffin, M.T.t., and Schloss, P.D. (2014). The human gut microbiome as a screening tool for colorectal cancer. Cancer Prev Res (Phila) 7, 1112-1121.
407. Gregory, J.C., Buffa, J.A., Org, E., Wang, Z., Levison, B.S., Zhu, W., Wagner, M.A., Bennett, B.J., Li, L., DiDonato, J.A., et al. (2015). Transmission of atherosclerosis susceptibility with gut microbial transplantation. J Biol Chem 290, 5647-5660.
408. Huttenhower, C., Kostic, A.D., and Xavier, R.J. (2014). Inflammatory bowel disease as a model for translating the microbiome. Immunity 40, 843-854.
409. David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E., Ling, A.V., Devlin, A.S., Varma, Y., Fischbach, M.A., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. Nature 505, 559-563.
410. Bonder, M.J., Kurilshikov, A., Tigchelaar, E.F., Mujagic, Z., Imhann, F., Vila, A.V., Deelen, P., Vatanen, T., Schirmer, M., Smeekens, S.P., et al. (2016). The effect of host genetics on the gut microbiome. Nat Genet 48, 1407-1412.
411. Samuels, D.C., Cardena, M.M.S.G., Ribeiro-dos-Santos, Â., Santos, S., Mansur, A.J., Pereira, A.C., and Fridman, C. (2013). Assessment of the Relationship between Self-Declared Ethnicity, Mitochondrial Haplogroups and Genomic Ancestry in Brazilian Individuals. PLoS ONE 8, e62005.