From Metagenomics to Pangenomics: Characterization of Dairy Worker Microbiomes and Development of Novel Statistical Methodology

Pauline Trinh

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:
Amy D. Willis, Chair
Peter M. Rabinowitz, Chair
Karen Levy

Program Authorized to Offer Degree:
Environmental and Occupational Health Sciences

University of Washington

Abstract

From Metagenomics to Pangenomics: Characterization of Dairy Worker Microbiomes and Development of Novel Statistical Methodology

Pauline Trinh

Co-Chairs of the Supervisory Committee:
Amy D. Willis
Department of Biostatistics

Peter M. Rabinowitz
Department of Environmental and Occupational Health Sciences

The complex interplay between routine antibiotic use and zoonotic pathogen presence makes livestock farming environments unique nexuses for the potential emergence of zoonotic diseases and/or antibiotic resistant bacteria and their resistance genes. Livestock can further facilitate transmission and emergence by serving as intermediary or amplifying hosts in which pathogens and antibiotic resistant bacteria and their genes can evolve and spill over into humans. As such, we were interested in understanding differences in the dairy worker microbiota that may arise due to exposure to livestock farming environments to evaluate potential risks of these environments in facilitating global dissemination of zoonotic disease, antibiotic resistant bacteria and antibiotic resistance genes.

We used culture independent methods that go beyond the traditional single pathogen approach to conduct comparisons between the gut microbiota and resistome of dairy workers and of community controls as well as to interrogate functional differences in commensal genomes recovered from both groups. To enable our study of functional differences, we first addressed methodological limitations with novel statistical methods for pangenomics. We developed `happi`, a statistical method for modeling gene presence that accounts for differential genome quality factors (e.g., mean coverage). We evaluated `happi`'s performance

using simulated and shotgun sequencing data and found that `happi` is accurate and robust even in scenarios when genome quality is correlated with the main covariate of interest. `happi` can furthermore be broadly applied to functional comparisons of genomes of other microorganisms beyond bacteria, and used in functional comparisons of metagenomes to adjust for differential quality (e.g., sequencing depths) of metagenomes.

Using `happi` to facilitate our functional comparisons, we conducted a metagenomics and pangenomics investigation of the effects of occupational exposure to dairy farm environments on metagenome differences in taxonomy, diversity and gene presence (i.e., co-abundant gene groups (CAGs), antibiotic resistance genes (ARGs), and virulence factors) and on functional differences of gut commensal bacteria genomes in dairy workers and community controls. A major strength of our study was the multi-level interrogation of dairy worker and community control microbiomes. Our cross-sectional study examining differences in microbial genes and genomes from dairy workers and community controls observed several patterns for further investigation including greater abundance of tetracycline resistance genes and higher occurrence of cephamycin resistance genes in dairy workers' metagenomes; evidence of commensal organism association with plasmid-mediated tetracycline resistance genes found in both dairy workers and community controls; and lower average gene and genome diversity in dairy workers' metagenomes compared to community controls. These findings point towards possible avenues for future research to better understand the impact of exposure to zoonotic pathogens, antibiotic resistant organisms, and ARGs on the microbiome and resistome of livestock workers and others with close animal contact.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I have felt so fortunate to have had the opportunity to pursue my doctoral degree and wish to acknowledge my endless gratitude to the many groups who have made this work possible.

First and foremost, I am profoundly thankful to my entire Doctoral Supervisory Committee: Drs. Amy Willis, Peter Rabinowitz, Karen Levy, Marilyn Roberts, and Amanda Phipps. I feel fortunate to be supported by each of my committee members and to be able to bring together their expert knowledge to improve the rigor of my work and science.

I wanted to give my particular thanks to my co-chairs Drs. Peter Rabinowitz and Amy Willis. Thank you so much Peter for giving me the creative space to craft this interdisciplinary training and for sharing your expertise in One Health. Your support, mentorship, accessibility, enthusiasm, and vision have really helped me throughout my doctoral training and I've been able to make many great connections during my time with COHR.

And I wanted to thank you, Amy, for constantly challenging me to be a better scientist and for always supporting and believing in my ability to develop statistical methods. I have learned so many things from you beyond statistical methods development like strategies for productivity and time management, navigating interpersonal work relationships, critically evaluating the academic institution, thinking through ways to be anti-racist in our work as researchers, and so much more! You have made a lasting impact on my life and there are honestly not enough words for me to express how truly grateful I am for you.

I am incredibly thankful to the members of the StatDivLab, Sarah Teichman, David Clausen, Maria Valdez Cabrera, and Jess Kunke who have all been exceedingly generous and helpful in bolstering my statistical training. I would like to give special additional thanks

to David for his insightful conversations and statistical contributions to the development of `happi`. I could not have asked for a better collaborator and co-author to work with in developing `happi`. I would also be remiss to not mention my gratitude to all the pets of the StatDivLab (PJ, Nutmeg/Catmeg, Shippo) for providing so much more joy to my life each week at coffee time.

I am also incredibly grateful to the COHR lab and DEOHS staff members past and present. Vickie Ramirez, Gemina Garland-Lewis, Pablo Hernandez, Jose Carmona, Pat Janssen, Helen Lee, Mary Saucier, Lindsay Pysson, Grace Wong, Trina Sterry, Brian High, John Yocum–thank you all for bearing with me these past five years as I've fumbled my way through administrative logistics and tried not to be too annoying with my questions.

I would also like to acknowledge and thank all the funding sources that have supported my doctoral training throughout the past five years. These funding sources include the Pacific Northwest Agricultural Safety and Health Center (PNASH), the NIOSH Education & Research Center (ERC), the National Institute of General Medical Sciences (NIGMS), and Biostatistics, Epidemiologic, And Bioinformatic Training in Environmental Health (BEBTEH). In particular, I want to thank Dr. Chris Simpson who heads up the ERC and Dr. Lianne Sheppard who directs BEBTEH. Having financial support to conduct my research and studies has been crucial to my success.

I would additionally like to express my deep gratitude to my academic peers. Taylor Reiter, Evan Bolyen, Bryan Martin, Megumi Matsushita, Sarah Philo, Nancy Carmona, Renee Codsi, Jessica Porter, Lauren Frisbie, and Alicia Hendrix. Your mental and emotional support, jokes, and work sessions throughout these past five years have made the process infinitely more enjoyable.

Finally, I want to thank my family and closest friends. My parents never had the opportunity to go to college and worked tirelessly to make sure my siblings and I had all the opportunities that they never had. Without their support and the support of my immediate family Peter, Paul and Diane, I would not have been able to pursue my doctoral education. To my dearest friends Fendy, Marian, Haylea, K.Lai, John – where would I be without you all keeping me sane. And last but not least, I want to express my deepest gratitude to my partner, Ricardo (and our dog Oliver), for making sure I eat, go out for walks, take care of my mental and physical health, and for being the real MVPs of the household in the final year of my PhD program.

To everyone, thank you all from the bottom of my heart.

Chapter 1

INTRODUCTION

Zoonotic disease emergence and transmission of antibiotic-resistant bacteria or genes are serious biological hazards facing people with animal contact [60]. Modern farming practices and intensification have been linked to increased risks of emergence and amplification of zoonotic diseases and antimicrobial resistance [80, 108, 171]. Antibiotics are routinely administered in modern conventional livestock farming as therapeutics and/or prophylaxis, exerting a selective pressure for antibiotic resistance development in bacteria [46, 129, 94]. Livestock can further facilitate the emergence and transmission of these biological hazards by serving as intermediate or amplifier hosts in which pathogens and other antibiotic resistant organisms can evolve and spill over into humans [80]. The complex interplay between routine antibiotic use and zoonotic pathogen presence has raised concerns about livestock farming environments becoming unique nexuses for the potential emergence of zoonotic diseases and/or antibiotic resistant bacteria and their resistance genes [111].

Previous studies investigating the risk of zoonotic infections and antibiotic resistance acquisition in livestock workers have utilized culture-dependent techniques to examine the transmission and evolution of single pathogens and their antibiotic resistance genes [57, 159, 58, 130, 90]. However, a single pathogen view to understanding transmission dynamics of zoonotic diseases, antibiotic resistant bacteria, and antibiotic resistance genes is limited as bacterial communities of both pathogens and commensals have complex interactions and are capable of acquiring or transferring genes through horizontal gene transfer [133, 69]. Advancements in culture-independent methods such a next-generation sequencing (i.e., amplicon and shotgun metagenomic sequencing) have now allowed for a more comprehensive view of not only pathogens, but also of the structure and function of entire microbial communities. Studies using next-generation sequencing data have revealed the functional importance of commensal organisms in immune homeostasis [6, 77, 5], disease development [55, 200, 122, 27], and even in resisting pathogen invasion and colonization [1, 86, 16, 82]. However, commensal organisms may also serve as reservoirs for transmission of antibiotic resistance genes to pathogens [171, 53] and under certain circumstances (e.g., depletion of microflora or acquisition of pathoadaptive functions) transition and become pathogens [143, 31, 150]. It is therefore important to understand effects of exposure to livestock farming environments on differences that may arise in the commensal microbiota to evaluate potential risks of these environments in facilitating global dissemination of zoonotic disease, antibiotic resistant bacteria and antibiotic resistance genes.

To study the effects of exposure to biological hazards (e.g., zoonotic pathogens, antibiotic resistant bacteria and their resistance genes) in livestock farming environments on the gut microbiota and resistome, I selected stool samples for shotgun metagenomics sequencing from dairy workers and community controls that had been collected as part of the Healthy Dairy Worker (HDW) study. I conducted shotgun metagenomics sequencing of stool samples from this cohort to facilitate detection of antimicrobial resistance genes and virulence factors and to assess functional differences between genomes of organisms recovered from the gastrointestinal tract of dairy workers and community controls.

To enable this study of functions, I first needed to address methodological limitations for functional comparisons in assembled genomes. Metagenome-assembled genomes (MAGs) obtained from shotgun metagenomics provide detailed insight into the functional and genetic differences of culturable and non-culturable microbial genomes [146]. However, MAGs are frequently incomplete or contain errors due to factors such as low sequencing depth, host contamination, choice of assembly, method of binning, and library preparation methods [22]. Failure to account for factors influencing the quality and completeness of MAGs has major implications on both the qualitative and quantitative conclusions of an analysis. To address these challenges, I developed a hypothesis testing approach to modeling gene presence that accounts for differential quality of genomes. This approach can also be extended to modeling gene presence in metagenomes while accounting for differential quality of metagenomes. Chapter 2 presents the hierarchical model for modeling gene presence and illustrates the performance of the method on shotgun sequencing and simulated data.

Chapter 3 details the metagenomics investigation of differences between dairy worker and community control metagenomes. I examined differences in the taxonomic composition, carriage of antibiotic resistance and virulence factor genes, co-abundant gene groups, and gene richness between dairy workers and community controls. I applied the method developed in Chapter 2 to test for differential gene presence between dairy worker and community control metagenomes while accounting for sequencing depth differences. I additionally conducted taxonomic annotation of reconstructed genomic context around tetracycline and cephamycin resistance genes to evaluate potential differences in taxonomic affiliations of these genes between groups.

The work in Chapter 4 complements the work from the previous chapter by focusing on metagenome-assembled genomes of five commensal species recovered from dairy workers and community controls. I applied the methodology developed in Chapter 2 to conduct a pangenomics investigation of functional differences in commensal genomes from dairy workers compared to community controls. I furthermore performed a phylogenomics analysis of the commensal bacteria genomes to see whether there was evidence of phylogenetic organization by group. I conclude Chapter 4 with a comprehensive discussion of results and limitations from both Chapters 3 and 4. Finally, in Chapter 5 I provide concluding remarks of this dissertation research and discuss future directions and on-going complementary work.

Chapter 2

HAPPI: A HIERARCHICAL APPROACH TO PANGENOMICS INFERENCE

2.1 Background

Members of the same bacterial species can display a wide variety of different phenotypes, and intra-species variation in pathogenicity, virulence, drug resistance, environmental range, and stress response has been observed across the tree of life [131, 153, 165, 76, 185]. Variation in phenotypes can in part be explained by genotypic variation, which is also considerable because mechanisms of genetic recombination in bacteria facilitate large genetic variation even within narrow organismal groups. For example, of 7,385 gene clusters observed in a study of 31 genomes in the genus *Prochlorococcus*, only 766 gene clusters were detected in all genomes [32]. We refer to the set of genes shared by all members of a clade as the *core genome* and we refer to the set of genes not shared by all members as the *accessory genome* [178]. Together, these sets of genes comprise a clade's *pangenome*: the entire collection of genes present in one or more organisms within the clade. In this paper, we describe a novel tool for pangenome analysis. Our tool is a statistical method to model the association between gene presence and covariates (predictors). Our method offers interpretable parameter estimates, a fast algorithm for estimation, and a flexible hypothesis testing procedure.

While culture-based studies have historically been used to study the gene content of bacteria, it has become increasingly common to employ shotgun metagenomics to study bacterial genomes and communities. Shotgun metagenomic sequencing involves untargeted sequencing of all DNA in an environment, enabling the study of genomes in their environmental context. Short reads from shotgun sequencing can be assembled into contigs and binned into metagenome-assembled genomes (MAGs), which represent a partial reconstruction of an individual bacterial genome. Despite major advances in methods for binning MAGs, MAGs can contain two types of errors. First, there can be genes that are truly present in the genome the MAG represents, but are unobserved in a MAG. Common reasons for this error include inadequate sequencing depth, high diversity in the metagenomes under study, and the inherent limitations of short read sequencing for reconstructing repetitive regions [37, 202, 154, 146, 164]. A second type of error in MAGs is erroneously observed genes: genes that are included in a MAG that are not truly present in the originating genome. This phenomenon is often referred to as contamination. The use of automated binning tools in the absence of manual inspection and refinement can lead to elevated rates of contamination. For example, the identification of contaminating contigs from manual refinement of MAGs produced by a massive unsupervised genome reconstruction effort removed 30 putative functions from a single contaminated genome[22, 139].

To address the challenges that contaminating and unobserved genes create for detecting enriched genes, our proposed method incorporates information about each genome's quality. Under our proposed model, a gene may be unobserved in a genome either because the gene is not present in the source genome, or because it could not be recovered from the obtained sequencing data. If, for example, the coverage of short reads across the genome was high and most of the expected core genes were observed, then the lack of detection of a given gene is more likely attributable to its true absence. The user can select which variables they believe to be the most informative for genome quality in their dataset. We develop estimators of the parameters of our model, discuss interpretation of model parameters, propose a hypothesis testing approach, and illustrate the performance of our model on shotgun sequencing and simulated data.

2.2 Results

2.2.1 A Hierarchical Model for Gene Presence

We present a hierarchical model for the association between bacterial gene presence and covariates of interest (e.g., host treatment status, environment of origin, relevant confounders, etc.). We consider observations on n genomes, which could be either metagenome-assembled genomes, isolate genomes, reference genomes, or any combination. Let Y_i be an indicator variable for the gene of interest being *observed* in genome i, $Y_i = 1$ if the gene is observed in genome i and $Y_i = 0$ otherwise. However, we are not interested in whether the gene is *observed* in each genome – we are interested in whether it is *present* in each genome. To this end, we define λ_i to be a latent (unobserved) random variable that indicates if the gene is truly present in genome i ($\lambda_i = 1$ if present).

We propose a logistic model to connect gene presence to covariate vector $X_i \in \mathbb{R}^p$:

$$\log\left(\frac{Pr(\lambda_i = 1|X_i)}{Pr(\lambda_i = 0|X_i)}\right) = X_i^T\beta, \tag{2.1}$$

where the λ_i's are conditionally independent given X_i and follow a Bernoulli distribution. Therefore, when comparing groups of genomes that differ by one unit in $X_{\cdot k}$ but are alike with respect to $X_{\cdot 1}, X_{\cdot 2}, \ldots, X_{\cdot, k-1}, X_{\cdot, k+1}, \ldots, X_{\cdot p}$, β_k gives the difference in the log-odds that the gene will be present between these two groups of genomes. To connect λ_i to Y_i we propose the following model

$$Pr(Y_i = 1|\lambda_i = \ell, M_i) = \begin{cases} f(M_i) & \ell = 1 \\ \varepsilon & \ell = 0, \end{cases} \tag{2.2}$$

where Y_i are conditionally independent Bernoulli distribution random variables; ε is the probability that a gene is observed in a genome in which it is absent (e.g., due to contamination or crosstalk); $M_i \in \mathbb{R}^q$ is a vector of genome quality covariates; and $f(\cdot) \colon \mathbb{R}^q \to \mathbb{R}$ is a flexible function to connect quality variables to the probability of detecting a present gene. Relevant quality variables are context-dependent and could include coverage of the gene from metagenomic read recruitment, completion (percentage of single copy core genes observed in the genome), redundancy (percentage of single copy core genes observed more than once in the genome), and an indicator for the genome originating from an isolated bacterial population.

2.2.2 Parameter Estimation

The latent variable structure of our model makes the Expectation-Maximization Algorithm [33] an appealing choice for estimating unknown parameters $\theta = (\beta, f)$. Because we do not observe $\{\lambda_i\}_{i=1}^n$, ε and f are not, in general, jointly identifiable. Therefore, we treat ε as a hyperparameter that can be fixed by the user or leveraged for sensitivity analyses. To improve stability of parameter estimates, we impose a Firth-type penalty on β. The complete data penalized log-likelihood is linear in λ_i, which allows us to simplify the expected complete data penalized log-likelihood at step t of an EM iteration as

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\lambda}|\mathbf{Y},\theta^{(t-1)}}\left[l(\beta,\tilde{f},\tilde{\varepsilon}|\mathbf{Y},\boldsymbol{\lambda})\right] = &\sum_{i=1}^{n}\Bigg(p_i^{(t)}\left[Y_i\tilde{f}(M_i)-\log(1+\exp(\tilde{f}(M_i))\right]\\
&+(1-p_i^{(t)})[Y_i\tilde{\varepsilon}-\log(1+\exp(\tilde{\varepsilon}))]\\
&+\left[p_i^{(t)}X_i^T\beta-\log(1+\exp(X_i^T\beta))\right]\Bigg)\\
&+\frac{1}{2}\log\left|\sum_{i=1}^{n}X_iX_i^T\text{expit}(X_i^T\beta)(1-\text{expit}(X_i^T\beta)\right|,
\end{aligned}\tag{2.3}$$

where $\tilde{\varepsilon} = \text{logit}(\epsilon)$, $\tilde{f}(x) = \text{logit}(f(x))$ for all x, and $p_i^{(t)} = \mathbb{E}[\lambda_i|Y_i,\theta^{(t-1)}]$ can be simplified as

$$p_i^{(t)} = \frac{Pr\left(Y_i|\lambda_i=1,\theta^{(t-1)}\right)Pr\left(\lambda_i=1|\theta^{(t-1)}\right)}{Pr\left(Y_i|\theta^{(t-1)}\right)},\tag{2.4}$$

where the terms in the numerator are given in (2.1) and (2.2), and the denominator is given by

$$\begin{aligned}
Pr\left(Y_i|\theta^{(t-1)}\right) = &Pr(Y_i|\lambda_i=1,\theta^{(t-1)})Pr(\lambda_i=1|\theta^{(t-1)})\\
&+Pr(Y_i|\lambda_i=0,\theta^{(t-1)})Pr(\lambda_i=0|\theta^{(t-1)}).
\end{aligned}\tag{2.5}$$

We maximize the expected complete data penalized log-likelihood separately for β and f. Owing to the form of the expected complete data penalized log-likelihood, efficient algorithms exist to perform each of these maximizations. Optimizing (2.3) with respect to β is equivalent to fitting a binomial generalized linear model with logit link function for outcomes $p_i^{(t)}$ via Firth-penalized maximum likelihood, and we find Newton's method to be stable and fast for this purpose.

Optimizing for f depends on the class of functions in which f falls. We investigated two flexible non-parametric options for f: $f \in \mathcal{F}$, where $\mathcal{F}$ is the class of bounded non-decreasing functions that map from $\mathbb{R}$ to $\mathbb{R}$, and $f \in \mathcal{I}$ where $\mathcal{I}$ is the class of linear combinations of k I-spline basis functions and a constant function where all basis functions have nonnegative coefficients. Both $f \in \mathcal{F}$ and $f \in \mathcal{I}$ result in a monotone estimate for f. To obtain the EM update for $f \in \mathcal{F}$, we use the primal active set algorithm of `isotone` [30] with custom loss function given by the first term in (2.3) plus a penalty term $-cosh\left(\left(\frac{m}{a}\right)^2\right)$ to prevent $\left|\tilde{f}\right|$ from growing without bound. We found that setting $a = 50$ gives a sensible tradeoff between algorithm convergence and numerical stability. To obtain the EM update for $f \in \mathcal{I}$, we fit a logistic regression on $p_i^{(t)}$ with predictors consisting of an I-spline basis with all non-intercept coefficients constrained to be nonnegative. We use the I-spline basis functions implemented in `splines2` [189]. In an analysis where we used short-read subsampling to approximate an empirical f, we found that $f \in \mathcal{I}$ outperformed $f \in \mathcal{F}$ (see Section 2.5.2), and for that reason we consider $f \in \mathcal{I}$ throughout the remainder of this chapter. We run the estimation algorithm for $t_{\max}$ steps or until the relative increase in the log-likelihood is below threshold Δ for 5 consecutive steps.

2.2.3 Hypothesis Testing

To enable inference on the odds that a gene will be present in groups of genomes that differ in their covariate attributes, we construct a hypothesis test for null hypotheses of the form $\mathbf{A}\beta = c$ for $\mathbf{A} \in \mathbb{R}^{h \times p}$ and $c \in \mathbb{R}^h$ where rank$(\mathbf{A}) = h$. This allows testing of null hypotheses including $\beta_k = 0$ (the odds that the gene will be present are equal when comparing groups of genomes that differ in $X_{\cdot k}$ but are alike with respect to $X_{\cdot 1}, X_{\cdot 2}, \ldots, X_{\cdot, k-1}, X_{\cdot, k+1}, \ldots, X_{\cdot, p}$). We propose to use a likelihood ratio test for $\mathbf{A}\beta = c$, rejecting H_0 at level α if $Q_{LRT} = 2[\mathcal{L}(\hat{\theta}) - \mathcal{L}(\hat{\theta}_0)]$ exceeds the upper $100\alpha\%$ quantile of a χ^2_h distribution, where $\hat{\theta}$ is the maximum likelihood estimate of θ; $\hat{\theta}_0$ is the maximum likelihood estimate of θ under the null hypothesis; and $\mathcal{L}$ is the log-likelihood function.

2.2.4 Data Analysis: Saccharibacteria MAGs

We consider a publicly-available dataset of $n = 43$ non-redundant Saccharibacteria (TM7) MAGs recovered from supragingival plaque ($n = 27$) and tongue dorsum ($n = 16$) samples of seven individuals from [163] (see Section 2.5 for more information). The wide variation in mean coverage across the MAGs (1.07 – 26.35×) makes this an appealing dataset on which to illustrate our quality variable-adjusting pangenomics method.

We consider methods that allow us to test the null hypothesis that the probability (equivalently, odds) that a gene is present in Saccharibacteria genomes are equal for tongue and plaque-associated genomes. The alternative hypothesis is that the probabilities differ. We compare our proposed method (`happi`: a Hierarchical Approach to Pangenomics Inference) with three competitors: a logistic regression model for Y_i with a likelihood ratio test (GLM-LRT); a logistic regression model for Y_i with a Rao test (GLM-Rao); and Fisher's exact test (Fisher). Note that these latter three methods test hypotheses about the odds that a gene is observed, while our proposed approach tests hypotheses about the odds that a gene is present, but we believe that results can be reasonably compared between these methods. We consider a single quality variable M_i for our analysis with `happi`: mean coverage across genome i. Our primary comparison is with GLM-Rao, which is the method currently implemented for pangenomics hypothesis testing in anvi'o [163]. We also note that the results from GLM-Rao and GLM-LRT are highly correlated, especially for larger p-values.

Different methods identified different differentially present genes. Out of 713 COG functions tested, `happi` identified 171 differentially present genes when controlling false discovery rate at the 5% level; GLM-LRT identified 219 genes; GLM-Rao identified 175 genes; and

Fisher identified 146 genes. Our proposed method calculated lower p-values for 20%, 35% and 85% of genera compared to GLM-LRT, GLM-Rao, and Fisher's test. We show results from 6 specific model estimates in Figure 2.1: 3 genes for which happi produced greater p-values than GLM-Rao (upper panels), and 3 genes for which it produced smaller p-values than GLM-Rao (lower panels). In all instances where happi produced greater p-values than GLM-Rao, non-detections generally occurred in genomes with low mean coverage. GLM-Rao does not account for coverage information, and so unlike happi, it can conflate gene absence with non-detections due to quality. We believe that statements about significance should be moderated when detection patterns can be attributable to quality variables, and therefore that it is reasonable that p-values are larger in these three cases. In contrast, happi produced smaller p-values than GLM-Rao in instances when non-detections occurred for greater coverage MAGs, or broadly across the range of MAG coverage (lower panels). In these instances, differences in detection are less likely to be attributable to quality factors, and it is reasonable that the significance of findings can be strengthened by including data on quality variables.

We additionally performed a sensitivity analysis using different values of the hyperparameter $\varepsilon \in \{0.01, 0.05, 0.1\}$ compared to $\varepsilon = 0$ to assess the robustness of our results to varying degrees of contamination in our genomes. We interpret $\varepsilon > 0$ as the non-zero probability of observing a gene given that it is truly absent. As shown in Figure 2.4, increasing values of ε resulted in larger p-values with average percent increases in p-values of 8%, 65%, and 209% for $\varepsilon = 0.01, 0.05, \text{and } 0.1$, respectively. We examined the proportion of p-values that had percent changes greater than 5% and found that 32%, 50%, and 49% of p-values from our $\varepsilon = 0.01, 0.05, \text{and } 0.1$ results had percent changes greater than 5%. The larger p-values produced by happi with increasing magnitudes of ε are sensible as we would expect differences in gene presence between groups to be less pronounced when accounting for a non-zero probability of erroneously observing a gene. Furthermore, we can use the results from our sensitivity analyses to moderate findings for gene presence results that were less robust to changes in ε and similarly confirm the robustness of our results for genes that were

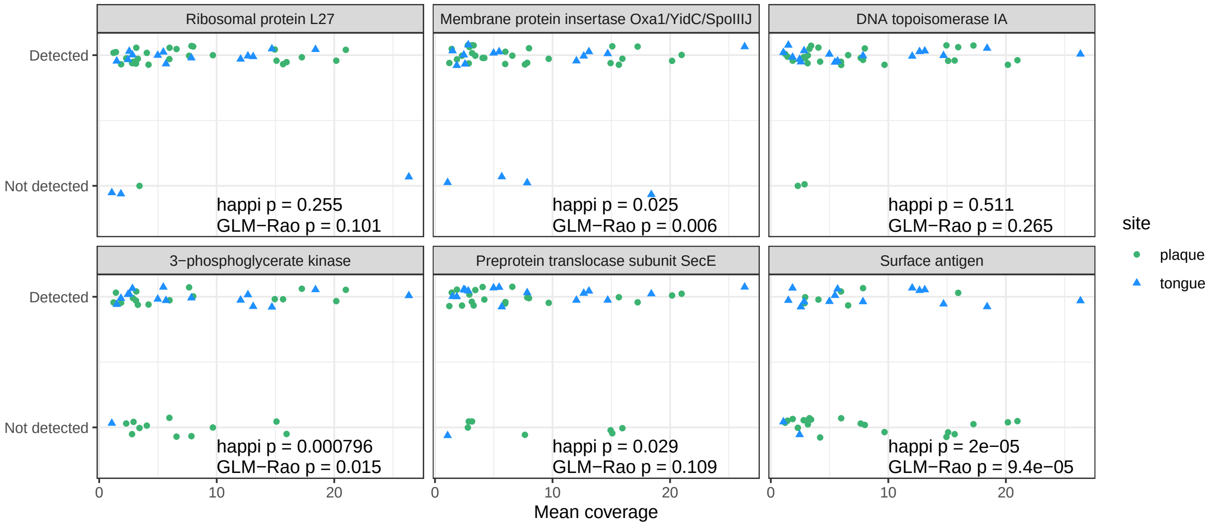


Figure 2.1: We test the null hypothesis that the probability that a gene is present are equal for tongue and plaque-associated Saccharibacteria genomes. The top 3 panels show genes for which our proposed method resulted in greater p-values than existing methods, and the lower 3 panels shows genes for which our proposed method resulted in smaller p-values than existing methods. Our method reduced p-values when differences in detection cannot be attributed to genome quality factors (here, coverage), and increased p-values in situations when non-detection may be conflated with lower quality genomes. Points have been jittered vertically to separate observations.

minimally affected by a change in ε.

2.2.5 Simulation Study

Finally, we investigate the performance of our approach by evaluating its Type 1 error rate and power. To generate data that most realistically reflects the relationship between coverage and gene detection in shotgun metagenomics studies, we construct $f(\cdot)$ for use in this simulation by subsampling short-reads from host-associated *E. coli* genomes ([4]; see Section 2.5.2 and Figure 2.3). We consider $q = 1$ and $q = 2$, and let $M_i = 10 + 30\frac{i-1}{n-1}$,

$X_{i1} = 1$, $X_{i2} = \mathcal{N}(\frac{i-1}{n-1}, \sigma = \sigma_x)$ and $\epsilon = 0$. σ_x is a parameter that controls the degree of correlation between M_i and X_{i2}, with larger values resulting in less correlation between quality variables and the predictor of interest. We simulate data according to the model described in (2.1) and (2.2), with $\beta = (0, 0)^T$ for Type 1 error simulations and $\beta = (0, \beta_1)^T$ with $\beta_1 \neq 0$ for power simulations. Note that because X_{i1} is continuous, a Fisher's exact test cannot be applied in this setting.

The results of Type 1 error rate simulations are shown in Figure 2.2 (left panels). We only show results for GLM-Rao because GLM-LRT and GLM-Rao produced highly similar p-values (mean squared difference 1.3×10^{-5}, correlation = 0.99996, $n_{sim} = 3000$). Notably, the logistic regression methods are anti-conservative, and do not control Type 1 error rates at nominal levels. For example, for a 5%-level test, Type 1 error rates for GLM-LRT range from 8.6% ($n = 30$ and $\sigma_x = 0.5$; 95% CI: 6.1–11.1%) to 31.6% ($n = 100$ and $\sigma_x = 0.25$; 95% CI: 27.5–35.7%). Stated differently, under H_0, GLM-LRT will return p-values that are usually too small, leading to more frequent incorrect conclusions of an association. In contrast, `happi` does control the Type 1 error rate, behaving near-exactly. We estimate that `happi`'s Type 1 error rates for a 5% test when $n = 30$ and $\sigma_x = 0.5$ is 5.2% (95% CI: 3.3–7.2%), and when $n = 100$ and $\sigma_x = 0.25$, `happi`'s empirical Type 1 error rate is 6.0% (95% CI: 3.9–8.1%). Greater correlation between the quality variable (coverage) and the covariate of interest leads to greater anti-conservativeness for logistic regression methods, which incorrectly attribute differences in gene presence to the covariate of interest. However, `happi` appears to control Type 1 error across the range of σ_x investigated here.

We show the power of `happi` to correctly reject a null hypothesis at the 5% level in Figure 2.2 (right panels). We do not evaluate power for GLM-Rao and GLM-LRT because they have uncontrolled Type 1 error rates, making them invalid tests. We observe that the power of `happi` to reject a false null hypothesis increases with the effect size and sample size, but decreases with greater correlation between M_i and X_{i1}. Stated differently, `happi` has low power to detect true associations between gene presence and covariates of interest when covariates are correlated with genome quality, though this can be remedied with larger

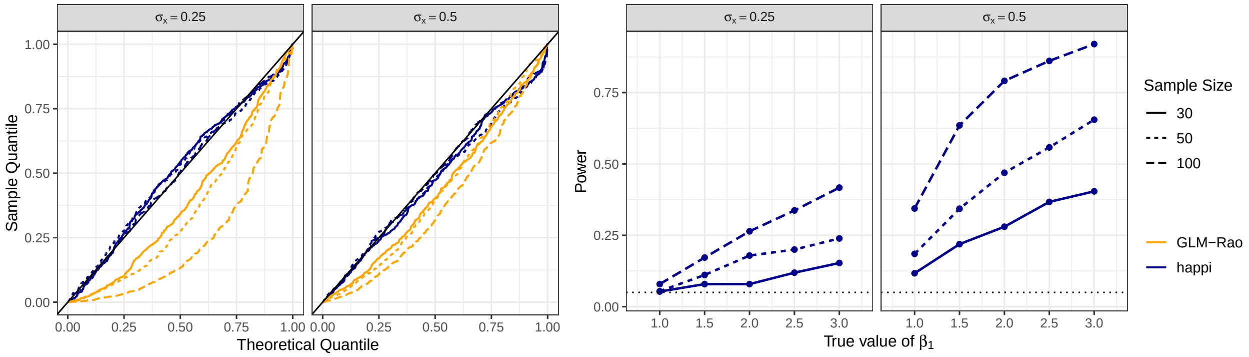


Figure 2.2: Simulations can be useful for evaluating the Type 1 and Type 2 error rates of methods for testing statistical hypotheses. (left) We find that logistic regression methods do not control Type 1 error, while `happi` does control Type 1 error at nominal levels. (right) We evaluate the power of `happi` to reject a false null hypothesis, finding that larger samples have greater power. In situations with greater correlation between quality variables and the covariate of interest, `happi` exhibits comparatively lower power.

sample sizes.

Taken together, these results show that `happi` is robust to potential correlation between covariates of interest and genome quality. This is not the case for logistic regression-based methods, which cannot distinguish between differential gene presence due to genome quality and differential gene presence due to associations with covariates. No method will perform well under the alternative with small sample sizes and high correlation (see Figure 2.2, third panel), but `happi` has some power for large sample sizes and large effect sizes in this setting, and controls Type 1 error at nominal levels regardless of the sample size.

2.3 Discussion

Many tools exist to study associations between microbial genome variation and microbial or host phenotypes [14, 39, 24, 99, 162]. Studies investigating the association between microbial genomes and phenotypes are often referred to as microbial genome-wide association studies (mGWAS) [155, 141]. Most mGWAS tools have been developed for the analysis of pure microbial isolates, and do not account for differential genome quality in genomes analyzed collectively. mGWAS tools may be better-suited when the hypothesized causal direction is that the presence of genetic features gives rise to a phenotypic characteristic, and not the reverse. In this paper, we propose and validate a novel method (happi) to understand how non-microbial variation (e.g., environmental variation) is associated with microbial genome variation. The implied direction of modeling is reversed in our model compared to mGWAS models: our response variable is gene presence rather than phenotype. This allows interrogation of questions about factors influencing selection pressures on genomes, rather than questions about the impact of the microbiome on phenotypic outcomes.

We view the main advantage of happi as its use of data about genome quality factors. To support the increasing use of shotgun metagenomic data to recover fragmented microbial genomes, researchers need methods that are capable of analyzing incomplete and imperfect genomes. While we are not aware of methods for modeling gene enrichment in MAGs, we offer comparisons to commonly used methods for analyzing near-complete genomes, such as Fisher's exact test (used by PanPhlAn3 [157, 9] and Scoary [14]) and logistic regression (used by anvi'o [43, 163]; see also [13]). In situations where differences in gene detection can be attributed to differences in genome quality, happi correctly infers that gene enrichment is ambiguous, and correspondingly identifies associations as less significant compared to competitor methods. However, in situations where genome quality cannot explain gene detection patterns, happi has greater precision than other methods and produces smaller p-values. We show via simulation that the advantages of happi are most pronounced when there is correlation between covariates and quality variables.

Results generated from happi are easily interpretable with reasonable run times on a modern laptop without parallelization, averaging 1.04 seconds per gene over 713 genes in $n = 43$ samples with $t_{\max} = 1000$ and $\Delta = 0.01$ on a 2.6 GHz i7 processor with 16 GB RAM. Since genes are treated independently, this analysis can be trivially parallelized, and furthermore, accuracy in estimation can be traded off for reduced runtime by reducing $t_{\max}$

or increasing Δ.

We suggest several avenues for further research. The first is to study the impact of experimental design on the statistical power of our proposed hypothesis testing procedure. Researchers often have to decide how to allocate budget across number of samples (including replicates and control data) and sequencing depth per sample. While existing guidelines for sequencing depth have focused on taxonomy estimation, MAG reconstruction, and gene detection [154, 202, 146, 63, 66, 166], our proposed modeling approach enables the principled study of the design of shotgun sequencing experiments to maximize power to detect differences in gene presence across sample groups.

Our latent variable model has possible utility for modeling the presence of amplicon sequence variants, and could offer a method for studying patterns of sequence variant presence when shotgun sequencing is infeasible or not preferred. For example, if a sequence variant is observed W_i times in sample i, then it would be reasonable to model $Y_i = \mathbf{1}_{\{W_i>0\}}$. This would permit inference on the equality of the probability that the sequence variant is absent in a sample across sample groups. Notably, by choosing an $\epsilon > 0$ (e.g., via the use of negative control samples), `happi` can adjust for the impact of index switching in studies that leverage multiplexing [97, 75]. We leave the application of `happi` to modeling the presence of amplicon sequence variants to future research.

Collectively, we have shown that `happi` is accurate and robust, even when genome quality is correlated with gene presence predictors. As the recovery of metagenome-assembled genomes becomes increasingly common, statistical tools that account for errors in recovered genomes become increasingly necessary. By leveraging genome quality metrics, `happi` provides sensible and interpretable results in an analysis of metagenome-assembled genome data, improves statistical inference under simulation, and can run efficiently on a local machine. Finally, by distributing open-source software in `R` implementing our proposed estimation and inference methods, we hope that `happi` can be used widely in a variety of genomics research settings. `happi` is available as an open-source `R` package via `https://github.com/statdivlab/happi` under a BSD-3-Clause license.

2.4 Availability of data and materials

happi is available as an open-source R package at https://github.com/statdivlab/happi. The data supporting the conclusions of this article along with code for reproducing our results are made available at https://github.com/statdivlab/happi_supplementary.

2.5 Methods

2.5.1 Methods: Saccharibacteria MAGs

The Saccharibacteria MAGs used in Data Analysis: Saccharibacteria MAGs, were taken from publicly available data [163]. Specifically, data on genome quality metrics (i.e. mean coverage) of these Saccharibacteria MAGs were retrieved from supplementary materials https://doi.org/10.6084/m9.figshare.11634321 and information on the presence or absence of COG functions in each MAG was extracted from the Saccharibacteria pangenome contigs databases and profiles located at https://doi.org/10.6084/m9.figshare.12217811. Functional annotation of the genes was performed using NCBI's Clusters of Orthologous Groups (COG) database [175]. Further details on sampling, assembly, binning, and refinement can be found in [163]. In our data analysis, we specified $t_{\max} = 1000$, $\Delta = 0.01$ and $\epsilon = 0$. We set $\epsilon = 0$ because these MAGs had undergone careful manual refinement to remove contamination from other genomes. We suggest the use of $\epsilon > 0$ when binning is performed automatically and without additional manual refinement.

2.5.2 Methods: simulation studies

Subsampling study of E. coli isolate DRR102664

To investigate the probability of detecting a gene that it is truly present ($Pr(Y_i = 1|\lambda_i = 1, M_i = m)$), we conducted a subsampling simulation study of an *E. coli* isolate genome taken from [4]. We selected *E. coli* isolate DRR102664 to perform our subsampling simulation and the ***eaeA*** gene (K12790) as our target gene of interest. In enteropathogenic

Escherichia coli, the ***eaeA*** gene produces a 94-kDa outer membrane protein called intimin which has been shown to be necessary to produce the attaching-and-effacing lesion. For our subsampling study, we subsampled paired sequences 50 times from the DRR102664 genome at approximate coverages $m = (2\times, 3\times, ..., 24\times, 25\times)$. Coverages were estimated using the calculation $\frac{\text{read count}\times\text{read length}}{\text{genome length}}$. We annotated and identified the ***eaeA*** gene in each set of subsampled sequences and calculated the empirical probability of detection as the fraction of samples of coverage m that detected ***eaeA***. The results of our subsampling investigation of the impact of coverage on $Pr(Y = 1|\lambda = 1)$ are shown in Figure 2.3.

Evaluating estimators for f

Many different choices of functions f could be used to connect the probability of detecting a present gene to quality variables M_i. We evaluated two options under simulation: $f(M_i) \in \mathcal{F}$ for $\mathcal{F}$ the class of bounded non-decreasing functions and $f(M_i) \in \mathcal{I}$ for $\mathcal{I}$ the class of bounded non-decreasing functions. As in Simulation Study, we set $M_i = 10 + 30\frac{i-1}{n-1}$, $X_{i1} = 1$, $X_{i2} = \mathcal{N}(\frac{i-1}{n-1}, \sigma = \sigma_x)$, $\beta_0 = 0$ and $\epsilon = 0$. The true $f(\cdot)$ in this simulation is a generalized additive model with binomial link function [197] fit to the observations shown in Figure 2.3. This was done to select a true detection curve that well-reflects empirical probabilities of detecting a gene at a given coverage, such as gene ***eaeA*** in *E. coli* isolate genome DRR102664. We evaluated all estimators via mean squared error and median squared error for estimating β_1. We investigated all combinations of $n \in \{30, 50, 100\}$, $\beta_1 \in \{0.5, 1, 2\}$ and $\sigma_x \in \{0.25, 0.5\}$, and performed 250 draws for each combination. For 17 out of 18 combinations of n, β_1 and σ_x, we found that $f \in \mathcal{I}$ outperformed $f \in \mathcal{F}$ with respect to median squared error, with an average reduction in median squared error of 54%. For 18 out of 18 combinations, $f \in \mathcal{I}$ outperformed $f \in \mathcal{F}$ with respect to mean squared error, with an average reduction of 51%. For this reason, we chose to set $f \in \mathcal{I}$ as the default option `happi`, and used this class of functions for both our data analyses and error rate simulations.

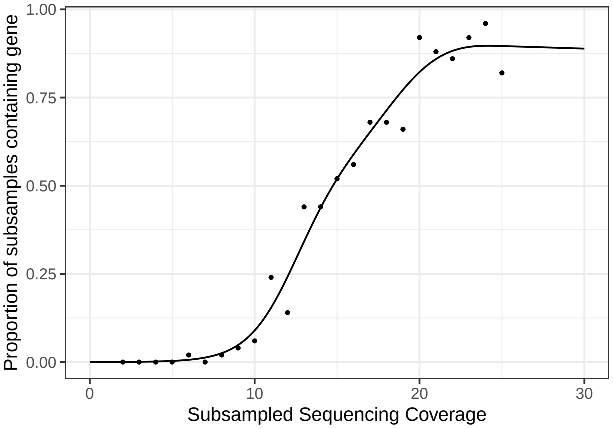


Figure 2.3: We subsampled reads from a publicly available *E. coli* isolate genome to understand the impact of coverage on the probability of detecting a gene, finding that the probability of detection increases with coverage. We use a nonparametric smoother to interpolate this curve and use it as the true function f in our simulations.

Type 1 error and power simulations

For the Type 1 error rate and power simulations shown in Section 2.2.5, we performed 500 simulations for each combination of σ_x, β_1 and n. We set a minimum of 16 EM iterations, $t_{\max} = 50$ and $\Delta = 0.1$ for both the null and alternative models.

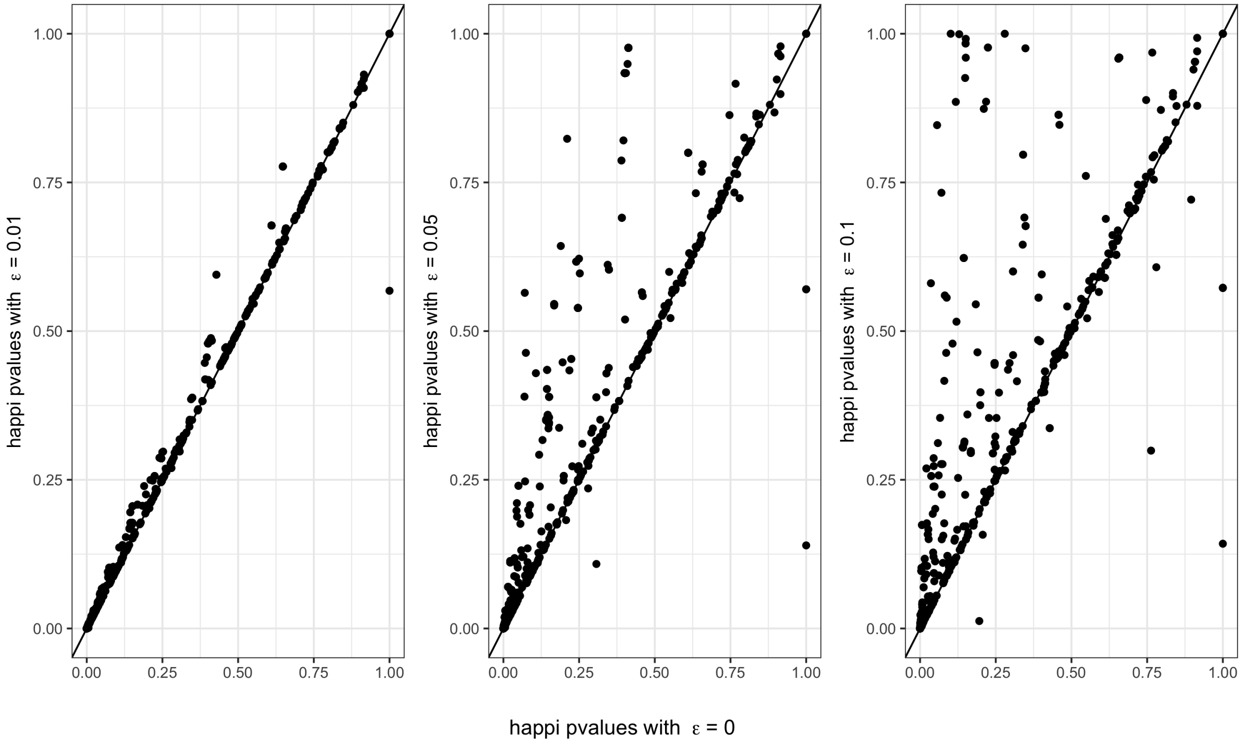


Figure 2.4: We evaluate the impact of increasing values of the hyperparameter $\varepsilon = 0.01, 0.05, \text{and } 0.1$ to assess the robustness of our results when fixing $\varepsilon = 0$. By leveraging ε, we can assess the impact that different probabilities of "contamination" would have on our results. We find that increasing ε results in larger p-values on average compared to when we assume there is a 0% probability of falsely detecting a gene. Differences in gene presence between groups are generally moderated when we account for the non-zero probability of erroneously observing a gene.

Chapter 3

INVESTIGATION OF DIFFERENCES IN DAIRY WORKER AND COMMUNITY CONTROL METAGENOMES

3.1 Background

Advances in next generation sequencing have facilitated further understanding of not only pathogens, but entire microbial communities of culturable and unculturable microorganisms. Increased knowledge of the composition, genetics, and functional capacity of the gut microbiome has revealed the profound impact that microorganisms of the human gut microbiome have on immune homeostasis [6, 77, 5], disease development [55, 200, 122, 27], and even resistance against pathogen invasion [1, 86, 16, 82]. The human gut microbiota is influenced both by host genetics [199, 144] and environmental factors such as diet [103, 28], geography [201], and medications [12, 106]. Recent research suggests, however, that environmental factors outweigh host genetics in shaping the gut microbiome [152, 54]. Consequently, environments that are rich in antibiotic resistant organisms, antibiotic residues, and/or ARGs are of great public health concern, as these environments may serve as hotspots for antibiotic resistance emergence and propagation. Antibiotics are administered frequently for therapeutic and prophylactic purposes in conventional livestock farming, making these antibiotic-rich environments potential nexuses for transmission between livestock and workers of not only antibiotic resistant organisms and their ARGs, but also of zoonotic pathogens that could be resistant to certain antibiotics [111]. Modern farming practices and intensification have previously been linked to the emergence and amplification of zoonotic diseases and antimicrobial resistance (AMR) [80, 108]. Livestock workers on animal farms where antibiotics are commonly administered are therefore an ideal cohort for studying the impact of environmental exposure to antibiotic resistant organisms, ARGs and zoonotic pathogens on the potential for zoonotic infections as well as commensal bacteria adaptations and roles in antibiotic resistance propagation.

Lower rates of asthma and allergic diseases have been observed among persons with early life exposure to farm environments [84, 195, 34] and differences in the gut microbiota between such groups have been proposed as an explanation for these lower rates. In particular, increased microbial diversity in microbiomes of individuals with early life exposure to certain non-occupational farming environments has been associated with a protective “farm effect” against the development of asthma and allergic disease [84, 195, 34]. Studies focusing on the effect of occupational farm exposure on the worker microbiome have demonstrated similar patterns of increased diversity in the microbiomes of workers, in addition to evidence of microbial sharing between the workers and their environments [84, 92, 117]. However, since these studies utilized 16S rRNA gene sequencing, they had limited resolution in studying functional differences in the farm-exposed gut microbiome.

Shotgun metagenomic studies, which involves untargeted sequencing of all genetic content in a sample, can circumvent some of the challenges with amplicon sequencing by providing higher taxonomic and functional resolution of microorganisms; insight into the metabolic, virulence or resistance potential of microbial communities; recovery of whole genome sequences; and simultaneous study of all microorganisms in a sample (including archaea, viruses, bacteria, and eukaryotes) [146, 81]. Several metagenomic studies have looked at the effect of occupational exposure to animal agriculture on ARG carriage and through characterization of the “resistome”, found higher prevalence of ARGs as well as evidence of transmission of ARGs from animal farming environments to workers [171, 184, 36]. While these studies demonstrated the potential impacts of exposure to these ARG rich environments on carriage of ARGs in the gut metagenome, they either focused primarily on understanding the presence of ARGs in total community DNA without assigning ARGs to particular species of commensal bacteria, or they used cultured isolates of a single species of generally commensal bacteria (e.g., *Escherichia coli*) to understand species-level antibiotic resistance transmission [171, 184, 36]. Furthermore, these studies did not simultaneously examine virulence factor genes which encode for functions that can cause disease as well as assist an organism to persist and disseminate [125]. While virulence factors have historically been associated with pathogens [125] they have also been identified on commensal or non-pathogenic genomes [67, 132]. Many of the same strategies used by pathogens to persist and adapt to their environments are similarly used by commensal organisms [125, 65], and transmission of these factors can occur between pathogens and commensals through mobile genetic elements [88, 134]. Previous studies of the effect of occupational exposure to livestock farming environments on the metagenome have thus provided limited understanding of the roles of both culturable and unculturable commensal bacteria in ARG and virulence factor transmission and propagation.

To interrogate the effect of exposure to livestock farming environments on the commensal gut microbiome of workers, we conducted pangenomic and metagenomic comparisons of dairy worker and community control gut microbiota. We studied differences in the taxonomic compositions and community structures of gut metagenomes between dairy workers and community controls, and investigated differences in the carriage of virulence factor genes and ARGs in dairy worker and community control metagenomes. We additionally evaluated potential taxonomic affiliations of genes conferring resistance to beta-lactams (cephamycin and cephalosporins) and tetracyclines, and assessed whether differences in taxonomic context existed based on group association. We continued our metagenomic investigation by analyzing co-abundant gene groups that may be associated with working on a dairy farm and examined differences in genetic diversity. Finally, we investigated whether occupational exposure to livestock on dairy farms resulted in functional differences in commensal bacteria genomes of the gut microbiome and whether there was an increase in carriage of antibiotic resistance genes and/or virulence factors in these bacteria. Our study presents insights into differences that may arise in the human gut microbiome and resistome as a result of occupational exposure to livestock farming environments.

3.2 Results

3.2.1 Study description

Recruitment and sample collection of dairy workers and community controls was conducted under the Healthy Dairy Worker study. The Healthy Dairy Worker study recruited dairy workers from three conventional large (> 5000 animals) farms in Yakima Valley where every dairy worker was approached by study staff for study enrollment. Community controls were enrolled from Yakima Valley neighborhoods using snowball sampling where research participants assisted research staff in identifying other potential participants. Details on the eligibility criteria used by the Healthy Dairy Worker study in enrolling its cohort of dairy workers and community controls is detailed in section 3.4.

We studied 16 gut metagenomes from ten dairy workers and six community controls derived from stool samples obtained at study enrollment by the Healthy Dairy Worker (HDW) study. We selected the 10 dairy worker samples through random sampling of study subjects that met our exclusion criteria of no antibiotic use within three months of sample collection and baseline enrollment. Additionally, dairy worker samples were selected from dairy workers working on one farm. All ten dairy workers identified as white Hispanic or Latino males. We note that numbers of women working on the participating dairy farms of the Healthy Dairy Worker study was low and therefore recruitment of women into the study was low. Selection of the 6 community control samples was done using random sampling of individuals who identified as white Hispanic or Latino males and had no antibiotic use within three months of sample collection and baseline enrollment.

The mean age of dairy workers was lower (38.40 SD 8.99) compared to community controls (49.50 SD 10.75), however these differences were not statistically significant at the 5% level using an independent t-test ($p = 0.06$). We additionally observed similar proportions of community controls who were current smokers (67%) compared to dairy workers (70%). Furthermore, all community controls reported similar occupations as field workers in non-animal agriculture at the time of sample collection and study enrollment. Study enrollment and baseline sample collection began in 2018 for these 16 participants. In January 2017, the Food and Drug Administration completed implementation of the Guidance for Industry (GFI) no. 213 which banned the use of antibiotics for growth promotion purposes and transitioned medically important antibiotics used in drinking water and feed from over-the-counter status to Veterinary Feed Directive (VFD) or prescription status [52, 51]. We note that the collection timeline of the samples used in this metagenomics study occurred at least one year after the full implementation of the FDA's GFI no. 213 policy.

3.2.2 Taxonomic profiling of dairy worker and community control metagenomes

The 16 metagenomic samples were composed of nine distinct phyla: Firmicutes, Bacteroidetes, Actinobacteria, Verrucomicrobia, Proteobacteria, Euryarchaeota, Spirochaetes, unclassified Eukaryota and Synergistetes. Of these phyla, Firmicutes, Bacteroidetes, and Actinobacteria were the three most abundant phyla found across all samples (Figure 3.1, left). The large representation of Firmicutes, Bacteroidetes, and Actinobacteria reflected similar community compositions observed in healthy subjects from the Human Microbiome Project [72]. We also note that while the majority of the phyla identified are from the domain Bacteria, we observed organisms from the domains Archaea (Euryarchaeota) and Eukaryota as well. We detected Euryarchaeota organisms in five dairy worker and six community control samples and unclassified Eukaryota organisms in low abundances in only two dairy worker samples from our study. To examine phylum-level differences between dairy workers and community controls, we tested for differential abundance of phyla between groups and found no significant differences at the 5% significance level in phyla abundances.

At the species-level, we identified 272 different species across the 16 metagenomes. The most prevalent bacteria species observed were *Prevotella copri*, *Faecalibacterium prausnitzii*, *Eubacterium rectale*, *Ruminococcus bromii*, and *Bacteroides vulgatus* (Figure 3.1, right). These five species have been previously shown to be highly abundant organisms found in healthy human gut microflora [105, 93, 8, 191, 177]. Differential abundance testing revealed no statistically significant differences in the abundances of these five organisms between groups at the 5% significance level. However, we did find a single organism (*Clostridium* sp. CAG 167) that was significantly more abundant in community control metagenomes at the 5% significance level ($q = 0.01$). Examination of abundance patterns in species with the largest magnitude test statistics (Figure A.1) demonstrated mixed patterns of abundances, with higher abundances of *Bifidobacterium catenulatum* ($q = 0.21$) and *Blautia wexlerae* ($q = 0.21$) observed in dairy workers and higher abundances of *Clostridium* sp. CAG 167 ($q = 0.01$) and *Ruminococcus callidus* ($q = 0.21$) in community controls.

We further interrogated differences in the community structures of dairy worker and community control metagenomes by examining differences in $\alpha-$ and $\beta-$ diversities. A comparison of the species-level $\alpha-$diversity using Shannon Diversity Index (SDI) showed no significant difference in the $\alpha-$diversity of dairy worker metagenomes compared to community control metagenomes ($p = 0.68$). Similarly, a comparison of differences in the community composition ($\beta-$diversity) of dairy worker and community control metagenomes using the Bray-Curtis dissimilarity metric showed no evidence of differences in community composition between groups (Figure A.2).

3.2.3 Identification of virulence factor genes

We identified 45 different virulence factor genes across only five samples corresponding to three community controls and two dairy workers (Supplementary Table 4). We found that samples with the highest number of identified Virulence Factor Database (VFDB) genes were also those with higher sequencing depth (Figure 3.2C). Using `happi`, the method described in Chapter 2 that adjusts for differential quality of metagenomes or genomes, we tested for differential enrichment of virulence factor genes between dairy worker and community control metagenomes while accounting for differences in metagenome sequencing depth. No virulence factor genes were significantly enriched between dairy worker and community control metagenomes at the 5% significance level (Supplementary Table 6).

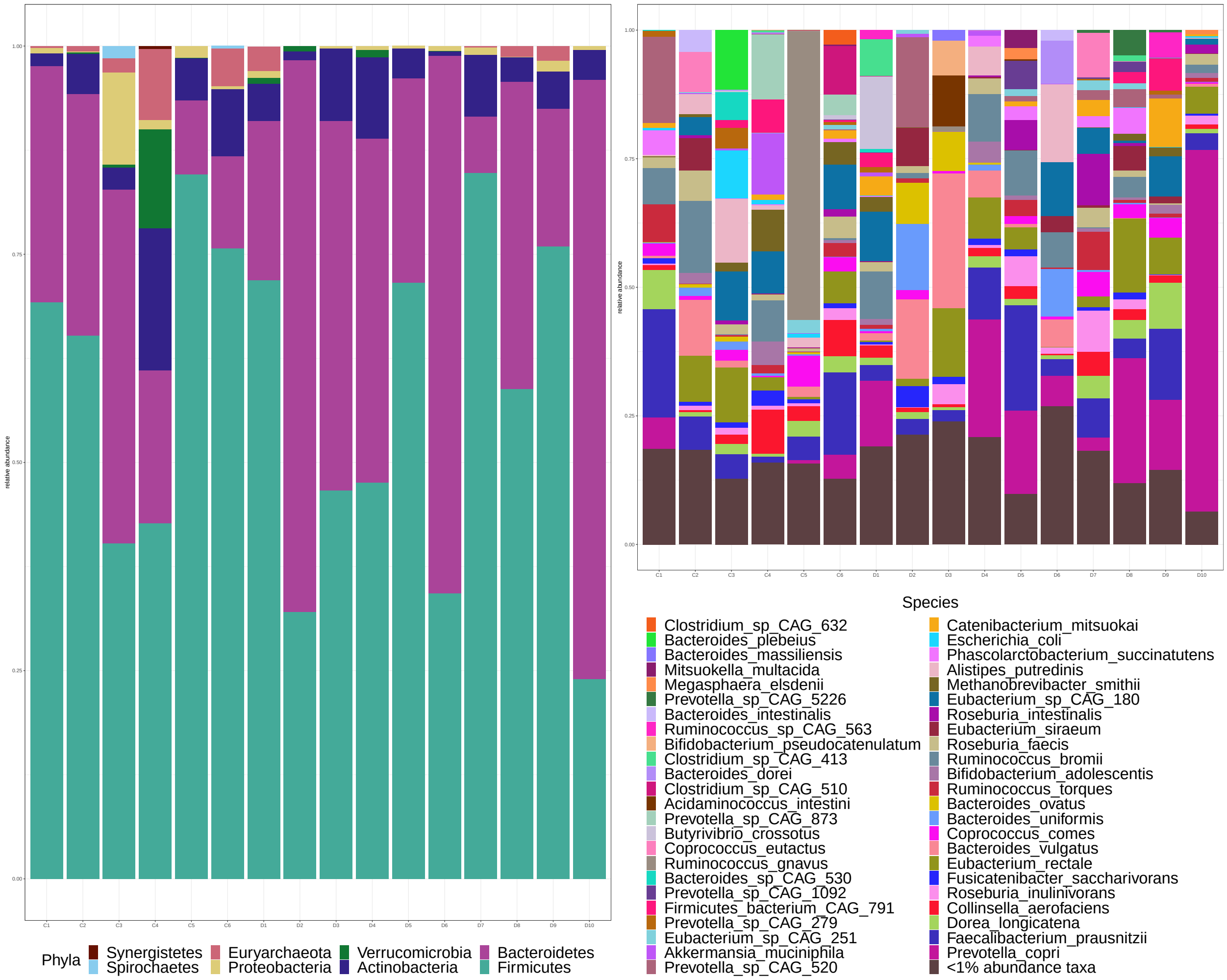


Figure 3.1: Stacked barplots of relative abundances show the most abundant phyla (left) and species (right) within each metagenome. At the phylum-level (left), Firmicutes spp. , Bacteroidetes spp. , and Actinobacteria spp. are the most abundant phyla across all samples. At the species-level (right), the five most abundant and prevalent species across community control and dairy worker metagenomes were *F. prausnitzii*, *E. rectale*, *P. copri*, and Eubacterium sp. CAG-180. Species with relative abundances less than 1% were grouped together. There was insufficient evidence to suggest major differences in the taxonomic composition of dairy worker metagenomes as compared to community controls.

3.2.4 Identification and taxonomic associations of antimicrobial resistance genes (ARGs)

Mass screening of the 16 metagenomes using the Comprehensive Antibiotic Resistance Database (CARD), identified 85 unique ARGs conferring resistance to at least 19 different antibiotic classes (Figure A.3, Supplementary Table 3). On average, a higher number of ARGs were identified in community control metagenomes compared to dairy worker metagenomes (Figure 3.2A). However, differences in the number of ARGs identified may be due, in part, to differences in sequencing depth, as metagenome samples with the highest number of ARGs identified also had higher numbers of sequenced reads (Figure 3.2A). We therefore used `happi` to test for differences in the enrichment of ARGs between dairy worker and community control metagenomes, while accounting for differences in sequencing depth. No ARGs were differentially enriched at the false discovery level of 0.05, but the following ARGs had the largest magnitude test statistics: *sat4* (happi LRT $\chi^2 = 0.01, q = 0.19$), *tet*(W) (happi LRT $\chi^2 = 0.07, q = 0.19$), and *dfrF* (happi LRT $\chi^2 = 0.17, q = 0.19$) (Supplementary Table 5).

We further focused our analyses to ARGs conferring resistance to antibiotic classes considered critically important to human medicine by the World Health Organization (WHO) [158]. Across our study metagenomes, we identified 37 different ARGs conferring resistance to eight antibiotic classes described in the WHO's list of Critically Important Antimicrobials (CIA): aminoglycosides, fluoroquinolones, macrolides, tetracyclines, cephalosporins, cephamycins, glycopeptides, and sulfonamides (Figure 3.3). The most frequently occurring types of antibiotic resistance genes found across the 16 metagenomes included those resistant to tetracyclines ($n = 15$), aminoglycosides ($n = 14$), cephamycins ($n = 13$), and macrolides ($n = 12$) (Figure 3.3). Genes for tetracycline resistance appeared to dominate the resistomes of both dairy workers and community controls with 11 different tetracycline resistance genes identified in 15 of our study metagenomes. We compared relative abundances of genes aggregated by antibiotic class between both groups and found that dairy workers' metagenomes had higher mean relative abundances of tetracycline ($p = 0.68$) and macrolide ($p = 0.82$)

resistance genes than community controls' metagenomes (Figure 3.3, Table A.1); however, these differences were not statistically significant at the 5% level. Similarly, the lower mean relative abundances of aminoglycosides ($p = 0.42$) and cephamycin ($p = 0.62$) resistance genes in dairy workers' metagenomes compared to community controls' metagenomes were also not statistically significant at the 5% level.

To understand whether there were differences in taxonomic affiliation of ARGs between groups, we assessed the taxonomic context of tetracycline and beta-lactam resistance genes. We identified six different genes (*cblA*-1, *cfx*A2, *cfx*A3, *cfx*A4, *cfx*A5, *cfx*A6) that encode for beta-lactamases and confer resistance to beta-lactam antibiotics. Additional details on the presence of each beta-lactam resistance gene in each of our study metagenomes are found in Supplementary Table 3. These six beta-lactam genes have typically been identified on the chromosomes of Bacteroidetes spp. [2]. Taxonomic annotation of the genomic context of these genes in dairy worker and community control metagenomes confirmed their association with organisms from the phylum Bacteroidetes such as *Prevotella copri*, *Bacteroides fragilis*, and *Bacteroides uniformis*. Additionally, we observed no differences in taxonomic affiliation of these beta-lactam genes between dairy workers and community controls (Figure A.5-A.9, Supplementary Table 9).

We identified 9 unique genes (*tet*(40), *tet*(B), *tet*(G), *tet*(W/N/W), *tet*(32), *tet*(M), *tet*(O), *tet*(Q), and *tet*(W)) that encode for efflux pumps or ribosomal protection proteins providing resistance to tetracycline antibiotics. These genes have been associated with plasmids [2], which are small, extrachromosomal DNA molecules that facilitate genetic sharing between and within species [151]. Taxonomic annotation of the assembly graphs for these tetracycline resistance genes demonstrated affiliation of these genes with a variety of both commensal (e.g., *Lawsonia intracellularis*, *Ligilactobacillus animalis*, *Trueperella pyogenes*, *Schaalia turicensis*, and *Faecalibacterium prausnitzii*) and pathogenic (e.g., Campylobacter spp., Clostridium spp.) bacteria. Full annotations of these ARGs to affiliated bacterial organisms can be found in Supplementary Table 9. Finally, while these tetracyline resistance genes were affiliated with a wide range of commensal and pathogenic bacteria, we found no differences in the taxonomic context of tetracycline resistance genes identified in community controls compared to dairy workers.

3.2.5 Co-abundant gene groups

We *de novo* assembled $157,451 - 525,755$ (median: $379,402$) protein coding genes across the 16 study metagenomes. Co-abundance clustering of these genes yielded 58 unique co-abundant gene groups (CAGs) [113, 114]. CAGs are useful for reducing the high-dimensionality of gene-level metagenomics by grouping genes with correlated levels of relative abundances across samples. These groupings can correspond to biologically meaningful elements such as core genomes of species, transposons, or genomic islands [113]. At the false discovery level of 0.05, we found no significantly differentially abundant CAG groups between dairy worker and community control metagenomes. We compared the distribution of association estimates from abundance testing of the 58 CAGs between groups with the $-log_{10}$ unadjusted p-values and found that the majority of CAG groups were less abundant in dairy worker metagenomes. The CAG group with the largest magnitude test statistic ($q = 0.19$, Supplementary Table 7) was less abundant in dairy workers than community controls and consisted of 5,941 genes. Taxonomic assignment of genes from this CAG group associated these genes with commonly represented commensal organisms of the gut microbiome such as *Coprococcus catus*, *Anaerobutyricum hallii*, *Mogibacterium sp.* BX12, and *Romboutsia timonensis.* Genes found within this CAG group are shown in Supplementary Table 8 and encoded for a wide variety of proteins including those related to metabolism, adhesion, replication, and restriction enzymes.

Next, we tested for differences in gene richness (the number of unique genes) between dairy worker and community control metagenomes using `geneshot`'s integration of `breakaway` to estimate the gene richness of each sample [193]. Our results showed significantly lower gene richness in dairy worker metagenomes as compared to community control metagenomes (estimate:-2.0×10^5 SE: 6.8×10^4, $p = 0.003$). To contextualize this finding, we also estimated the genome diversity in each metagenome using single-copy core genes [44]. We found that on average, dairy worker metagenomes had lower genome diversity than community control metagenomes (t-test $p = 0.02$) (Fig. 3.2B). In addition, we observed a positive correlation between genome diversity and gene richness (Figure A.4). This congruence is unsurprising as we would expect species diversity and gene richness to be positively correlated.

3.3 Discussion

In Chapter 4, we recover metagenome-assembled genomes from dairy worker and community control metagenomes, and perform comparative genomics on these recovered microbial genomes. Therefore, to comprehensively discuss our study of dairy worker and community control gut metagenomes, we defer discussion of results from this chapter to Chapter 4.4.

3.4 Methods

3.4.1 Study participant selection

We obtained shotgun metagenomic sequencing data from 16 samples selected from the Healthy Dairy Worker study. The Healthy Dairy Worker study is a prospective cohort study that focuses on the effects of working on a dairy farm on the fecal and nasal microbiome, and immune and respiratory function. The study began collection in May 2017 of fecal and nasal samples, as well as health history data for each participant at baseline enrollment, 3, 6, 12, and 24 months. Recruitment was performed on a rolling basis. Dairy workers were recruited from three conventional large ($> 5,000$ animals) farms in the Yakima Valley of Eastern Washington state and community controls were recruited from surrounding communities. Recruitment of both community controls and dairy workers was done through snowball sampling where research participants assisted in identifying other potential participants. Eligibility to be a participant as a dairy worker required subjects to have been working on a dairy farm for at least 6 months. Eligibility as a community control required participants to have no prior work experience on a dairy farm in the previous five years, to have not lived on a dairy farm, and to have no current household member who worked on a dairy farm in the previous five years. Participants were consented by bilingual study staff and received an incentive payment for enrollment and subsequent sampling. Sample collection and study activities were approved by the University of Washington Institutional Review Board under STUDY00000042.

To conduct the current cross-sectional metagenomics study, we obtained shotgun sequencing data of 16 baseline fecal samples taken from the Healthy Dairy Worker study cohort. These samples came from 10 dairy workers and 6 community controls. The 16 samples were selected at random from dairy workers and community controls who all identified as white Hispanic or Latino male and had no antibiotic use within 3 months of baseline sampling. Additionally, we constrained the selection of the dairy worker samples to one farm. The unbalanced sampling of each group was designed to over-sample dairy workers, as community control samples could be supplemented with additional healthy subjects' metagenomics data from publicly available data (i.e., The Human Microbiome Project).

3.4.2 Sampling, shotgun metagenomic library preparation and sequencing

Stool samples were self-collected by participants using a stool specimen collection kit with instructions on how to collect and package specimens in prelabelled fecal collection tubes and biospecimen bags. Participants were instructed to store stool samples in their refrigerators and to return their stool samples within 24 hours of collection to study staff. Samples were stored in a -20 degree Celsius freezer by field staff at a partner study site for 1-6 months before before being packaged with dry ice and transported to the University of Washington for extraction and storage in a -20 degree Celsius freezer. DNA extraction was performed using the MoBio DNeasy PowerLyzer PowerSoil Kit (Qiagen) following manufacturer's protocols, and quantification of the resulting DNA was conducted using the Quant-iT PicoGreen dsDNA Assay Kit (ThermoFisher/Invitrogen). Extracted DNA samples were packaged on dry ice and transported to the Fred Hutchinson Cancer Research Center for sequencing.

Sequencing libraries were prepared from 250pg gDNA with a quarter reaction workflow using the Nextera XT Library Prep Kit (Illumina, San Diego, CA) and 12 cycles of indexing PCR. Indexed libraries were pooled by volume and post-amplification cleanup was performed with 0.8X Agencourt AMPure XP beads (Beckman Coulter, Indianapolis, IN). The library pool size distribution was validated using the Agilent High Sensitivity D5000 ScreenTape run on an Agilent 4200 TapeStation (Agilent Technologies, Inc., Santa Clara, CA). Additional library QC and cluster optimization was performed using Life Technologies-Invitrogen Qubit® 2.0 Fluorometer (Life Technologies-Invitrogen, Carlsbad, CA, USA). The resulting libraries were sequenced on the Illumina HiSeq 2500 to generate paired-end 150nt sequences for each fragment. Image analysis and base calling were performed with Illumina Real Time Analysis software v1.18.66.3, followed by demultiplexing of dual-indexed reads and generation of FASTQ files with bcl2fastq Conversion Software v1.8.4 [74].

3.4.3 Profiling taxonomic composition

We performed profiling of the microbial composition of the metagenomic short reads using `MetaPhlAn3` v3.0.14 [10]. `MetaPhlAn3` estimates relative abundances by mapping reads to a reference database of clade-specific marker genes from ChocoPhlAn v30 (published in January 2019) [161, 10]. `MetaPhlAn3` performs this read mapping against marker genes using bowtie2 v2.3.5.1 [95, 96]. Default parameters were used when running `MetaPhlAn3` with an additional flag "-t rel_ab_w_read_stats" for outputting relative abundances with estimated number of reads mapping to each clade.

3.4.4 Metagenomic assembly and processing of contigs

We conducted *de novo* assembly and processing of contigs using anvi'o v6.2 [44]. Anvi'o integrates a suite of bioinformatics tools for the processing, analyzing, and visualization of metagenomics, pangenomics, and phylogenomics studies. We used the anvi'o Snakemake [91] metagenomics workflow obtained from "anvi-run-workflow" with "–workflow metagenomics" to conduct our metagenomic assembly and processing of contigs. `Illumina-utils` [45] was used to apply Minoche et al. [112] published guidelines for quality filtering of reads. A median of 41.9 M (IQR: 10 M) reads per sample passed quality filtering. `MEGAHIT` v1.2.9

[100] was used to perform individual assembly of each metagenome. Further processing of the individual assemblies included generating a contigs database using anvi'o v6.2 [44], identifying open read frames using `Prodigal` v2.6.3 [73], predicting gene-level taxonomy using `Centrifuge` [87], functional annotation of genes using NCBI's Clusters of Orthologous Groups (COGs) [175] and Pfams [42], searching for sequences using DIAMOND v0.9.14 [15], identifying single copy core genes (SCGs) using HMMER v3.3 [40] and built-in anvi'o HMM profiles for bacteria and archaea, recruiting reads using `bowtie2` v2.3.5.1 [95], and generating BAM files with `samtools` v1.10 [101]. Prediction of the approximate number of genomes in a metagenomic assembly using SCGs was done using the anvi'o script "anvi-display-contigs-stats".

3.4.5 Metagenome annotation of virulence factors and antibiotic resistance genes

We used `ABRicate` v1.0.1 [160] to perform a mass screening of our *de novo* assembled gene calls for antibiotic resistance genes and virulence factor genes. `ABRicate` uses the Basic Local Alignment Search Tool (BLAST) [3] to annotate genes from a user-specified reference database. We used the the Virulence Factor Database v6.0 [21] and the Comprehensive Antibiotic Resistance Database (CARD) v4.0 [110] as reference databases in our search. Genes were considered present in a given metagenome if they met conservative minimum thresholds of 90% identity and 100% coverage.

Gene abundances were calculated within a metagenome by taking the mean coverage of a target ARG or VF gene divided by the sum of all mean coverages of all protein coding genes identified in a given metagenome. ARG relative abundances were further aggregated by their antibiotic classes by summing the relative abundances of genes within each antibiotic class for each metagenome. AMR genes that conferred resistance to multiple antibiotic classes were defined as "multi-drug" resistance if they were associated with resistance to at least 2 antibiotic classes. We focused our analyses to antibiotic classes that were identified by the World Health Organization (WHO) as Critically Important Antibiotics (CIA) [158].

3.4.6 Reconstruction of genomic context of ARGs

We used our results from ABRicate to extract ARG target sequences from each metagenomic assembly. These sequences were extracted using samtools [101] and were used as "query" sequences in our genomic context reconstruction analyses. ARG query sequences were used to produce query neighborhoods that reassociated unassembled or unbinned (as is the case when using bins for query sequences) reads that are graph-adjacent to the query sequence. To prepare our metagenomic short reads for genomic context reconstruction, we removed adapters and quality trimmed the reads using fastp [23] before removing human host reads using bbduk [17] and the masked human k-mer data [18]. Using our quality trimmed and filtered short reads and our query sequences of interest, we constructed the genomic context of each query sequence using MetaCherchant [128]. MetaCherchant uses a de Bruijn graph assembly approach to build genomic context of query sequences. We used the "environment-finder" tool in parallel and set k-mer length to 31, minimum coverage to 5, and max radius to 1000. Taxonomic annotation of sequences corresponding to graph nodes was done using kraken2 v2.1.2 [196]. A minimum gene coverage of 15 was used in producing and visualizing our final assembly graphs as genes with coverages less than 15 produced fractured assembly graphs. Assembly graph visualizations were created using bandage [192]. Segements of the graph that correspond to the resistance gene sequence are colored in blue and the query neighborhoods are shown in grey. The grey segments represent nodes of the compact de Bruijn graph that are at least k = 31 in size.

Identification of resistance genes located on plasmids or microbial chromosomes was conducted using the Resistance Gene Identifier (RGI) v5 [2]. The RGI integrates with the CARD database to predict AMR genes and their mutations in complete chromosome sequences, predicted genomic islands, complete plasmid sequences, and whole genome shotgun assemblies taken from National Center for Biotechnology Information (NCBI) databases. This is accomplished through prediction of open reading frames using Prodigal [73], alignment to CARD reference sequences using either BLAST [3] or DIAMOND [15], and the use of either protein homolog or protein variant models. The results from RGI's exhaustive search are maintained and updated for each gene catalog on the CARD database.

3.4.7 Co-Abundant Gene Groups

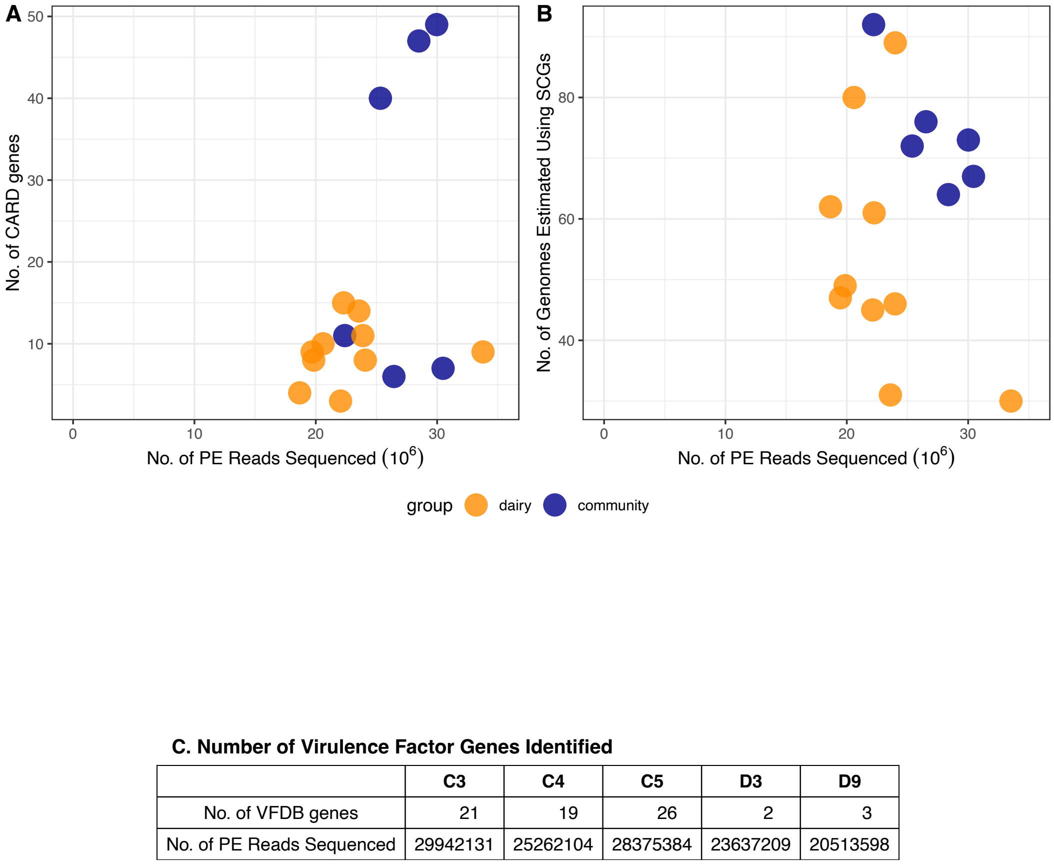
We used geneshot [115, 113] to perform gene-level metagenomic analyses of our study samples and managed our workflow using Nextflow [35]. Our geneshot workflow consisted of (1) removal of human genome contamination, (2) *de novo* assembly of reads using MEGAHIT [100], (3) identification of open reading frames using Prodigal [73], (4) deduplication of genes into a non redundant reference set of genes using MMSeqs2 [168], (5) estimation of gene abundance using DIAMOND [15] to align reads against the gene catalog followed by (6) resolving reads that align to multiple genes into a single unique alignment using FAMLI [59], (7) clustering of genes based on co-abundances across samples to form CAG groupings, (8) taxonomic assignment through alignment with RefSeq [102], (9) functional annotation using eggNOG mapper [71, 19], (10) gene richness estimation using breakaway[193] and testing for differences in gene richness between groups using the betta() function [194], and (11) differential abundance testing of CAG groups by covariates of interest using corncob [109].

3.4.8 Statistical analyses

Differences in demographic characteristics between groups were evaluated using independent t-tests for continuous variables and chi-square tests for independence for categorical variables. To test for differential abundance of species between dairy workers and community controls, we first estimated the number of reads mapping to each clade generated from MetaPhlAn3. Species that were not present in any sample were filtered out of our analyses and a psuedo-count of 1 was used for species with 0 read counts. We centered log-ratio transformed the species-level counts and performed independent t-tests with a false discovery rate correction using q-value estimation through the qvalue v2.26.0 [170] package.

We calculated α-diversity using the plug-in Shannon Diversity Index (SDI) on relative abundances and compared the SDI of dairy worker and community control metagenomes using an independent t-test. Differential abundance testing of antibiotic resistance genes in dairy workers compared to community controls was performed using independent t-tests on centered-log ratio transformed relative abundances. Relative abundances of ARGs were calculated using the mean coverage for a given target gene divided by the sum of all mean coverages of all the protein coding genes identified in a given metagenome. We applied our CLR transformation on our relative abundance data after filtering out genes that had zero abundance.

Testing for differential enrichment of virulence factor genes and ARGs between dairy workers and community controls was performed using `happi` v1.0, a hierarchical approach to modeling gene presence that adjusts for factors that impact the probability of detecting a gene. We used `happi` to account for differences in sequencing depth across our metagenomes when conducting gene enrichment testing and performed a false discovery rate correction for multiple comparisons using the `qvalue` v2.26.0 `R` package for q-value estimation. All statistical analyses were conducted using `R` v4.1.2.



C. Number of Virulence Factor Genes Identified

	C3	C4	C5	D3	D9
No. of VFDB genes	21	19	26	2	3
No. of PE Reads Sequenced	29942131	25262104	28375384	23637209	20513598

Figure 3.2: For each metagenome, we compare the sequencing depth with the number of identified CARD genes, VFDB genes, and estimated number of genomes using single-copy core genes. Sequencing depths for our study metagenomes ranged from $18,687,775 - 33,675,630$ paired end reads per sample. Samples with deeper sequencing had higher numbers of identified genes from the CARD and VFDB databases and higher numbers of estimated genomes. Within the community control group, three samples had the highest number of identified CARD genes out of all samples studied, whereas the remaining three community control samples within the community control group appeared to be indistinguishable from dairy workers in the number of identified CARD genes. This raises questions as to whether the three community control samples with the highest number of identified CARD genes are outliers or if antibiotic use was not accurately reported by these individuals.

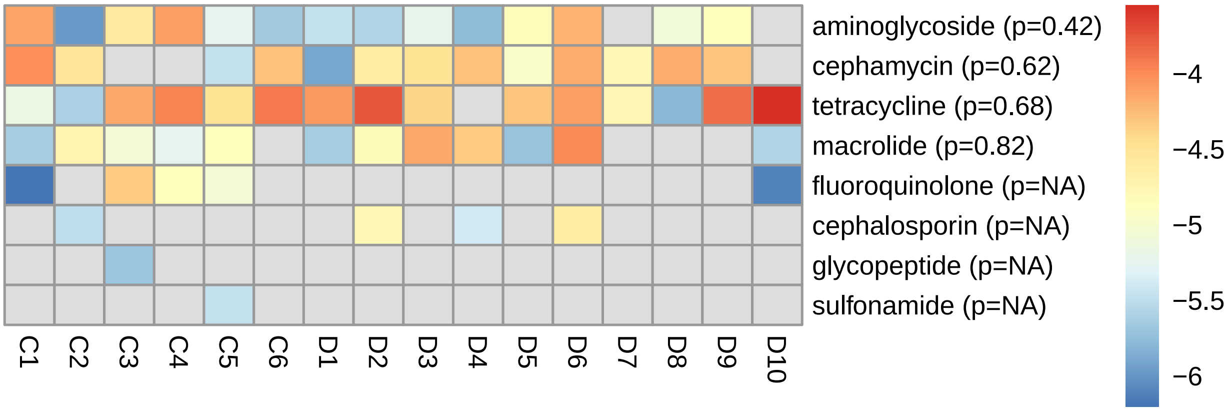


Figure 3.3: We identified ARGs from eight antibiotic classes (rows) listed as critically important to human medicine by the WHO. log_{10} transformed relative abundances of antibiotic resistance genes grouped by these antibiotic classes are colored from lower (blue) to higher (red) relative abundances in each metagenome. Grey squares denote undetected antibiotic resistance genes. Visual inspection displays patterns of increased abundance of tetracycline resistance genes and macrolide resistance genes in dairy worker metagenomes. While the mean relative abundance of cephamycin resistance genes was lower in dairy workers compared to controls, cephamycin resistance genes had a higher occurrence in dairy workers as these genes were identified in 90% of dairy worker samples compared to 67% of community control samples.

Chapter 4

PANGENOMIC INSIGHTS INTO COMMENSAL BACTERIA FROM DAIRY WORKERS AND COMMUNITY CONTROLS

4.1 Introduction

Shotgun metagenomic sequencing can facilitate interrogation of the entire genomic content of unculturable and culturable microorganisms in a sample. The work in Chapter 3 utilized a metagenomics approach to examine differences in taxonomic composition, co-abundant gene group abundances, gene richness, and carriage of antibiotic resistance and virulence factor genes between dairy workers and community controls. In this chapter we conduct a pangenomics investigation to understand functional differences in commensal genomes recovered from dairy workers compared to community controls.

Recent studies have demonstrated the capacity of commensal bacteria to promote resistance to invasive pathogen colonization through direct (competition of nutrients and niche) and indirect (enhancement of host immune defenses) mechanisms [1, 86, 16, 82]. Other studies have demonstrated the association between commensal organisms and the development of obesity, diabetes, and irritable bowel diseases [55, 200, 122, 27]. While commensal organisms are not conventional pathogens, there is concern that they can serve as reservoirs for antibiotic resistant genes (ARGs) that can be transferred to other commensal species and pathogens through mobile genetic elements or bacteriophages [53]. Furthermore, under selective pressure from the host immune system, environmental changes, or antibiotic use, commensal organisms (e.g., *Escherichia coli*) can adopt similar strategies and functions as pathogens (e.g., attachment mechanisms, motility functions) to colonize and persist in their niche environments [65] and can even transition into pathogens by acquiring transposable insertions that encode for pathoadaptive functions [143]. Pathoadaptive functions may include antibiotic resistance genes as well as virulence factors that allow bacteria to build on existing adaptations to either be successful at invading a new niche or persisting in their current niche [143, 31]. The complex roles that commensal organisms play in human health and in the propagation of antibiotic resistance warrants further investigation, especially as interest in developing therapeutics (e.g., probiotics) that manipulate microbiota using commensal organisms grows [105, 140, 70].

We conduct pangenome comparisons to understand functional differences between commensal organisms found in dairy workers compared to community controls. In particular, we are interested in understanding whether exposure to livestock farming environments exerts any selection or differentiation of functions in commensal species found in the gut microbiota of dairy workers compared to community controls. Pangenome analyses further our understanding of gene loss or gain dynamics, gene duplication, and interactions of the genome with mobile elements that are shaped by selection and drift [205]. While metagenomics facilitates characterization of taxonomic composition and functional profiles of entire microbial communities, pangenomics provides insights into the genomic heterogeneity and diversification of a given species [205]. These approaches are thus complementary in our investigation of the effects of exposure to livestock farming environments on the worker gut microbiome. Investigating whether functional differences exist between commensal genomes from dairy workers and community controls provides insight into potential diversification or adaptations of gut commensal bacteria in response to exposure to these unique environments.

4.2 Background

Through quality control and automatic binning procedures detailed in Section 4.5, we recovered a total of 535 draft genomes across the 16 metagenomic samples. The maximum number of draft MAGs identified in a single sample was 50 (from a community control) and the minimum number of draft MAGs identified in a single sample was 13 (from a dairy worker). To conduct our pangenomic comparisons between dairy workers and community controls, we taxonomically annotated our MAGs and identified the most prevalent commensal species recovered across our 16 samples. The top five most prevalent commensal species found in our MAG recovery effort were *Faecalibacterium prausnitzii* (21 genomes recovered; identified in 13 samples), *Ruminoccocus bromii* (13 genomes recovered; identified in 11 samples), *Ruminococcaceae UBA 1417. sp002305575* (9 genomes recovered; identified in 9 samples), *Eubacterium CAG-180 sp000432435* (9 genomes recovered; identified in 9 samples), and *Fusicatenibacter saccharivorans* (8 genomes recovered; identified in 8 samples). Testing for differential abundance of these species in dairy worker and community control microbiomes was conducted in Chapter 3.2.2 where we found no statistically significant differences at the 5% significance level in abundances of these species between groups. We constrain the focus of this chapter to pangenome and phylogenomic analyses to explore differences in functions and evolutionary relationships of these five commensal species between dairy workers and community controls.

4.2.1 Background: Faecalibacterium prausnitzii

Faecalibacterium prausnitzii is a gram-positive, non-motile, non-spore-forming bacterium and one of the main butyrate-producing organisms in the gut microbiomes of humans and animals [47, 104, 156]. There has been increasing interest in *F. prausnitzii* for their role in promoting gut health due to the anti-inflammatory properties that have been attributed to this species [104, 47, 105]. Many studies have also associated the depletion of *F. prausnitzii* in the gut microbiome to onset and progression of inflammatory bowel diseases [116, 167, 56, 149, 107]. Recent interest in using *F. prausnitzii* in probiotics to promote gut health has raised concerns regarding potential propagation of antibiotic resistance genes as evidence of phenotypic antibiotic resistance has been shown in this species [105, 50]. Specifically, phenotypic resistance to antibiotics such as ampicillin, gentamycin, kanamycin and streptomycin have been detected [105].

4.2.2 Background: Ruminococcus bromii

Ruminococcus bromii is an anaerobic, gram-positive, spore-forming, non-motile cocci bacteria that has been proposed as a keystone species of the human colonic microbiome due to its importance in metabolizing dietary resistant starch [204, 26, 120]. *R. bromii* is one of the most abundant bacteria species in the human colonic microbiota [204, 26, 120] and has been observed in both humans and ruminants [89, 120, 127, 181]. *R. bromii* is considered a specialist in metabolizing dietary resistant starches, creating energy products (i.e., glucose) that stimulate the growth of other bacterial organisms [204, 203, 148]. Similar to *F. prausnitzii*, *R. bromii* has been of interest for use in probiotics due to its positive associations with human health [93]. To our knowledge, phenotypic antibiotic resistance has not yet been shown in *R. bromii*.

4.2.3 Background: Fusicatenibacter saccharivorans

Fusicatenibacter saccharivorans is an anaerobic, gram-positive, non-motile, non-spore-forming, spindle-shaped bacteria that was first isolated from human feces in 2013 [173]. While there has been limited research conducted on *F. saccharivorans*, there have been a few studies interested in the relationship between *F. saccharivorans* and intestinal inflammation [174, 145, 61]. These studies focused on individuals with inflammatory bowel diseases (i.e. Crohn's disease, ulcerative colitis) and found that *F. saccharivorans* may have anti-inflammatory effects in the intestine [174, 145, 61]. To our knowledge, no studies have shown evidence of phenotypic antibiotic resistance in *F. saccharivorans*.

4.2.4 Background: Eubacterium sp. CAG-180

Eubacterium sp. CAG-180 is a bacteria identified through a large-scale metagenome-assembled genomes recovery effort across ≈ 400 human stool samples [123]. While there is limited research on *Eubacterium* sp. CAG-180, *Eubacterium* spp. are generally considered beneficial commensal bacteria and contain species known to support gut homeostasis and suppress gut inflammation [119]. To our knowledge, phenotypic antibiotic resistance has not yet been observed in this species.

4.2.5 Background: Ruminococcaceae UBA 1417 sp002305575

Ruminococcaceae (also known as Oscillospiraceae) UBA 1417 sp002305575 is an uncultured species of bacteria identified in a massive metagenome-assembled genomes reconstruction effort from over 1,500 publicly available metagenomes [138]. The Ruminococcaceae family of bacteria include commensal organisms from the genera *Faecalibacterium*, *Ruminococcus*, and *Subdoligranulum* that have been associated with positive metabolic functions such as production of short chain fatty acids and butyrate [62, 11]. Very little is known about *Ruminococcaceae* UBA 1417 sp002305575 especially as this species has only been identified through shotgun metagenomics [138].

4.3 Results

Faecalibacterium prausnitzii pangenome

In our study of 16 metagenomes, we recovered 21 *F. prausnitzii* MAGs that met our thresholds for quality (Figure 4.1). These 21 *F. prausnitzii* MAGs came from 6 community controls and 7 dairy workers. After manual refinement of these MAGs, the completion and redundancy of the 21 *F. prausnitzii* MAGs ranged from $45 - 78\%$ and $4 - 9.8\%$, respectively. In total, $6,076$ gene clusters with $40,928$ genes were identified across these 21 genomes. Genome lengths of completed and published *F. prausnitzii* genomes have ranged from 2.68–3.32 million base pairs [48, 7]. In comparison, the median genome length of our 21 MAGs (1.97 million base pairs IQR 399,709) was smaller than reported sizes of completed *F. prausnitzii* genomes.

AMR and virulence factor gene annotation of these 21 *F. prausnitzii* MAGs identified a single MAG recovered from a dairy worker sample that contained two antibiotic resistance genes: *sat4* and *aph(3')-IIIa* (Figure 4.1). Meanwhile, annotation of Clusters of Orthologous Genes (COG) functions identified the presence of 1,590 functions across the 21 genomes. Testing for differential enrichment of functions between the 10 *F. prausnitzii* MAGs from community controls and the 11 *F. prausnitzii* MAGs from dairy workers showed no differentially enriched COG functions when controlling for the false discovery rate at 5%. While no functions were significantly differentially enriched, we observed notable enrichment patterns of COG functions with the highest magnitude test statistics. 14 of the 23 highest ranked COG functions were more enriched in the community controls' MAGs and consisted of functions in COG categories C (Energy production and conversion), DH (cell cycle control, cell division, chromosome partitioning), E (amino acid transport and metabolism), F (nucleotide transport and metabolism), J (translation, ribosomal structure, and biogenesis), P (inorganic ion transport and metabolism), R (general function prediction only), T (signal transduction mechanisms), and V (defense mechanisms). Nine of the 23 highest ranked COG functions associated with dairy workers' MAGs were functions in COG categories C (Energy production and conversion), G/GE (carbohydrate transport and metabolism), J (translation, ribosomal structure, and biogenesis), K (transcription), L (replication, recombination and repair), T (signal transduction mechanisms). Full details of COG function enrichment and categorization can be found in Supplementary Table 10. Interestingly, amongst the top 15 highest ranked functions, we identified five COG functions (COG2929, COG2274, COG0610, COG0627, COG1518) from category V (Defense mechanisms), that were all more enriched in *F. prausnitzii* MAGs from community controls. This is particularly interesting given that category V defense mechanisms encode for important functional traits for microbial persistence, abundance, and adaptation [68, 79, 49].

4.3.1 Ruminoccocus bromii pangenome

We recovered 13 *R. bromii* MAGs that met our thresholds for quality (Figure 4.2). These 13 *R. bromii* MAGs came from five community controls and six dairy workers. After manual refinement, the completion and redundancy of the 13 *R. bromii* MAGs ranged from 49% to 95% and 0% to 2.8%, respectively. 4,846 gene clusters with 25,718 genes were identified across these 13 MAGs. Reported genome sizes of *R. bromii* have ranged from 2.15-2.4 million base pairs in length [120]. By comparison, we observed a similar median genome length of 2.1 million base pairs (IQR 119,770) for our 13 *R. bromii* MAGs.

AMR and virulence factor gene annotation of these 13 *R. bromii* MAGs found no presence of ARGs or virulence factor genes on these MAGs. Annotation of COG functions, however, identified 1,326 functions across the 13 MAGs. Testing for differentially enriched functions between the five *R. bromii* MAGs from community controls and the eight *R. bromii* MAGs from dairy workers revealed no differentially enriched COG functions when controlling for the false discovery rate at 5%. However, we did observe some interesting patterns in the COG functions with the highest magnitude test statistics. Specifically, the top seven COG functions (Supplementary Table 11) with the highest magnitude test statistics were associated with dairy workers' metagenomes and came from COG categories K (Transcription), X (Mobilome: prophages, transposons), R (General function prediction only), S (Function unknown), L (Replication, recombination and repair), and V (Defense mechanisms). Functions in COG categories X (Mobilome: prophages, transposons) and V (Defense mechanisms) have been shown to be enriched in plasmids and frequently involved in horizontal gene transfer [121, 172], highlighting the potential increased adaptability of *R. bromii* species in dairy worker metagenomes

4.3.2 Fusicatenibacter saccharivorans pangenome

Across our study metagenomes, we recovered eight *F. saccharivorans* MAGs. These eight MAGs came from three community controls' metagenomes and five dairy workers' metagenomes (Figure 4.3). After manual refinement, the completion and redundancy percentages of these MAGs ranged from 42-100% and 0-7%, respectively. In total, we identified 3,477 gene clusters with 17,866 genes across the eight MAGs. The reported genome size of *F. saccharivorans* recovered from metagenomic assemblies is approximately 3.6 million base pairs in length [136, 135]. By comparison, our study *F. saccharivorans* MAGs had smaller genome lengths with a median genome length of 2.5 million base pairs (IQR 461,611) observed in our genomes.

We annotated these MAGs for AMR and virulence factor genes and found no presence of ARGs of virulence factors in these eight *F. saccharivorans* MAGs. COG function annotation, however, identified 1,381 functions across the eight genomes. We tested for differentially enriched functions between the five *F. saccharivorans* MAGs recovered from dairy workers' metagenomes and the three *F. saccharivorans* MAGs from community controls and found no differentially enriched COG functions when controlling for the false discovery rate at 5% (Supplementary Table 14).

4.3.3 Eubacterium CAG-180 sp. 000432435 pangenome

We recovered nine *Eubacterium* sp. CAG-180 MAGs from our study metagenomes and identified 2,883 gene clusters with 16,397 genes in these nine MAG (Figure 4.4). Four *Eubacterium* sp. CAG-180 MAGs were recovered from community control metagenomes and five *Eubacterium* sp. CAG-180 MAGs from dairy worker metagenomes. After manual refinement, the completion and redundancy of these *Eubacterium* sp. CAG-180 MAGs ranged from 90-94% and 0-6%, respectively. Genome sizes for *Eubacterium* sp. CAG-180 recovered from metagenomic assemblies have been approximately 1.9 million base pairs in length [136, 135]. Comparatively, our *Eubacterium* sp. CAG-180 MAGs had similar genome lengths with a median genome length of 1.9 million base pairs (IQR: 91,035).

AMR and VF gene annotation of these nine MAGs showed no evidence of ARGs or virulence factors contained in these genomes. COG function annotation identified 1,162 functions across the nine *Eubacterium* sp. CAG-180 genomes. We found no statistically significant differentially enriched COG functions at the false discovery 5% significance level. 11 COG functions with the highest magnitude test statistics (Supplementary Table 13) were all associated with dairy workers' metagenomes and came from COG categories P (Inorganic ion transport and metabolism), K (Transcription), G (Carbohydrate transport and metabolism), J (Translation, ribosomal structure and biogenesis), O (Posttranslation modification, protein turnover, chaperones), R (General function prediction only), V (Defense mechanisms), and X (Mobilome: prophages and transposons). Three (COG0732, COG2253, COG5340) COG functions were related to category V (Defense mechanisms) and one COG2801/COG2963 was related to category X (Mobilome: prophages and transposons). It is interesting to note that these functions were all more enriched in *Eubacterium* sp. CAG-180 MAGs from dairy workers' metagenomes, highlighting the potential increased adaptablity of *Eubacterium* sp. CAG-180 species in dairy worker metagenomes.

4.3.4 Ruminococcaceae UBA 1417 sp. 002305575 pangenome

We recovered a total of nine *Ruminococcaceae* UBA 1417 sp. MAGs containing 3,126 gene clusters with 18,624 genes. Four *Ruminococcaceae* UBA 1417 sp. MAGs came from community control metagenomes and five *Ruminococcaceae* UBA 1417 sp. MAGs came from dairy worker metagenomes (Figure 4.5). After manual refinement, the percent completion and redundancy ranged from 46-99% and 0-8%, respectively. The reported genome size of *Ruminococcaceae* UBA 1417 sp. 002305575 recovered from metagenome assemblies is approximately 2.3 million base pairs [136, 135]. In comparison, our *Ruminococcaceae* UBA 1417 sp. MAGs had similar genome lengths with a median genome length of 2.1 million base pairs (IQR: 300432) observed in our genomes.

AMR and virulence factor gene annotation of these MAGs showed an absence of either ARGs or virulence factors within these MAGs. Annotation of COG functions identified 1,294 functions present across the nine genomes. No COG functions were differentially enriched when controlling for the false discovery rate at 5% (Supplementary Table 12). Of the 11 COG functions with the highest magnitude test statistics, eight COG functions were differentially enriched in community controls' *Ruminococcaceae* UBA 1417 sp. MAGs and represented COG categories G (Carbohydrate transport and metabolism), M (Cell wall/membrane/envelope biogenesis), O (Posttranslational modification, protein turnover, chaperones), R (General function prediction only), T (Signal transduction mechanisms), V (Defense mechanisms), and H (Coenzyme transport and metabolism). The three COG functions that were more enriched in *Ruminococcaceae* UBA 1417 sp. MAGs from dairy workers came from COG categories K (Transcription), T (Signal transduction mechanisms), R (general function prediction only), and Q (Secondary metabolites biosynthesis, transport and catabolism). The enrichment of category V defense mechanisms in community control associated MAGs highlights the potential increased adaptability of *Ruminococcaceae* UBA 1417 sp. MAGs in community control metagenomes.

4.3.5 Phylogenomic comparisons

To assess whether there were larger patterns of differences in MAGs recovered from dairy workers and community controls' metagenome samples, we constructed phylogenomic and pangenomic trees for each species. Phylogenomic trees are useful in estimating ancestral relationships of genomes based on a set of highly conserved genes. However, estimation of phylogenies based on a set of conserved genes ignores accessory genes that may be acquired due to environmental selection or fitness [163]. Thus we additionally constructed pangenomic trees that were based on the presence and absence of all genes found across all genomes to better capture environmental similarities of our genomes [163, 32]. 21 single copy core gene clusters were used to create a phylogenomic tree for the 21 *F. prausnitzii* MAGs recovered across our 16 study metagenomes and all $6,076$ gene clusters identified across the 21 *F. prausnitzii* MAGs were used to construct a pangenomic tree. A dendrogram comparison of these two trees (Figure 4.6) showed similar genome clustering patterns between the two trees, but the organization of these *F. prausnitzii* genomes based on the occurrence of gene clusters in either tree did not appear to predict group (dairy worker compared to community control) affiliation. Similarly, we observed no clear clustering of genomes by group in the pangenomic and phylogenomic tree comparisons for *R. bromii* (103 SCGs; 4,846 total gene clusters) (Figure 4.7), *F. saccharivorans* (35 SCGs; 3,477 total gene clusters) (Figure 4.8), *Eubacterium* sp. CAG-180 (15 SCGs; 2,883 total gene clusters) (Figure 4.9), and *Ruminococcaeceae* UBA 147 sp.002305575 (22 SCGs; 3,126 total gene clusters) (Figure 4.10). These results suggest that phylogenomic similarities or clustering between these genomes are not associated with occupational exposure to livestock farming environments even when we compare evolutionary relationships using all core and accessory genes found in each genome.

4.4 Discussion

The work presented in Chapters 3 and 4 used metagenomics and pangenomics to compare dairy worker and community control microbiomes by interrogating differences in taxonomy, diversity, and genes (e.g., ARGs, CAGs, virulence factors) as well as in the functions of commensal genomes. The use of shotgun metagenomics data allowed us to circumvent some of the challenges with amplicon sequencing data by enabling detection and characterization of functions (e.g., COG functions, ARGs, virulence factor genes) and reconstruction of draft genomes. The results from our investigation lay the foundation for further research on the impact of occupational exposure to antibiotic and zoonotic pathogen rich environments on the commensal microbiome.

Previous metagenomic studies of livestock workers have found increased abundance and carriage of antibiotic resistance genes in individuals occupationally exposed to animal farming environments, raising concerns that these environments could be hotspots for antibiotic resistance and zoonotic disease emergence [184, 171, 36, 190, 179]. Two shotgun metagenomic studies conducted cross-sectional examinations of the microbiomes and resistomes of pig farmers and pig slaughterhouse workers in the Netherlands [184] and China [179] and found that the resistomes of pig farmers and slaughterhouse workers were dominated by tetracyclines, aminoglycosides, beta-lactam and macrolide resistance genes. Additionally, Van Gompel et al.'s study conducted in the Netherlands observed higher total ARG abundance and lower Shannon diversity in the fecal samples of pig slaughterhouse workers and pig farmers compared to broiler farmers and control groups [184]. Another shotgun metagenomics cross-sectional study investigated the microbiomes and resistomes of live poultry market (LPM) workers in China and found higher abundance of ARGs, lower Shannon diversity, and greater enrichment of beta-lactam and lincosamide resistance genes in LPM workers compared to controls [190]. The results from these cross-sectional studies have been further corroborated by Sun et al.'s longitudinal investigation of veterinary students in China that followed healthy students with occupational exposure to swine farms during three month internships [171]. They observed similar patterns of increased total abundance of ARGs and increased abundances of beta-lactam, aminoglycoside, and tetracycline resistance genes in veterinary students within three months of exposure to swine farm environments [171]. Furthermore, Sun et al. highlighted evidence of microbial composition shift in veterinary students' microbiota after three months of swine farm exposure that made their microbiota more similar to the swine farm workers' microbiota than a healthy control cohort of urban Chinese subjects [171].

Our study is the first, to our knowledge, shotgun metagenomics interrogation of the microbiomes and resistomes of dairy workers in the United States. We approached our investigation by looking at both metagenomes and genomes to examine differences that may occur at differing levels. Contrary to previous metagenomic studies of livestock workers in China and Europe that showed increased abundance and carriage of ARGs in livestock workers, we found no significant difference in the abundance of ARGs between dairy workers and community controls. Similarly, we found no evidence of virulence factors that were differentially enriched in one particular group. Although we did not find significant differences in ARG abundance between groups, we did observe a pattern of greater abundance of tetracycline resistant genes in dairy workers' metagenomes that was directionally consistent with findings in other farm studies [184, 171, 36, 190]. Additionally, we noted a more frequent occurrence of cephamycin (beta-lactam) resistant genes identified in the dairy worker population compared to community controls. These patterns are interesting to note since tetracyclines are commonly administered on dairy farms for treating gastrointestinal and respiratory diseases in dairy cows [78] and beta-lactam antibiotics such as ceftiofur are frequently used to treat metritis, a common post-partum uterine inflammatory disease, in dairy cows [176].

While these observed differences in carriage of ARGs between dairy workers and community controls were not statistically significant in our study, our sample size was relatively small. We therefore expect that if the patterns observed in our study are true, they would be better detected with a larger cohort of livestock workers and community controls. Additionally, we note that the community controls in our study all worked as field workers in non-animal agriculture industries. While our community controls were employed individuals, which helps to reduce healthy worker bias, there could be overlapping exposures (such as exposure to animal manure used as fertilizer) from working in agriculture that community controls shared with dairy workers, making community controls more similar to dairy workers. Additionally, in Eastern Washington state and across the United States, streptomycin and oxytetracycline are commonly sprayed during bloom time (spring) in fruit orchards and plant agriculture to control fire blight disease [38, 187]. Occupational exposure to antibiotics that have been sprayed on plants in the control group would make them more similar to the dairy workers and thereby contribute to our null findings. Furthermore, we may not have observed significant differences in ARGs between groups because of good antibiotic stewardship on farms that would limit worker exposure to these biological hazards. We did not, however, have any information on antibiotic practices and worker training conducted on the dairy farm on which the study participants worked to evaluate the possibility of good antibiotic stewardship.

Our study also highlighted the potential for commensal organisms to serve as ARG reservoirs for pathogenic bacteria. Using an assembly graph approach to reconstruct the genomic context of each antibiotic resistance gene followed by taxonomic annotation of the context, we were able to confirm the association of chromosome-mediated ARGs (i.e., *cblA*-1, *cfx*A2, *cfx*A3, *cfx*A4, *cfx*A5, *cfx*A6) with previously recognized carriers of these genes (i.e., Bacteroidetes spp.) [2]. With the same approach applied to plasmid-mediated ARGs (i.e., *tet*(B), *tet*(G), *tet*(W/N/W), *tet*(32), *tet*(M), *tet*(O), *tet*(Q), and *tet*(W)) we found that these resistance genes were associated with both commensal and pathogenic organisms. These observations suggest the potential for sharing of ARGs between commensal organisms and pathogens through conjugative plasmids. Furthermore, our results corroborate findings from a recent study that compared ARGs identified in 1,354 culture commensal strains and 45,403 pathogen strains from the human gut and found evidence of 64,188 shared ARGs that mapped to 5,931 mobile genetic elements [53]. Some of the MGEs identified by [53] had also been previously identified in data from ruminant guts , soil, and other human body sites [53].

While commensal organisms may serve as ARG reservoirs for pathogenic bacteria, they may also assist in preventing pathogenic invasion through indirect (enhancement of host immune defenses) and direct (competition of nutrients and niche) mechanisms [1, 86, 16, 82]. Further research is needed to better understand the complex dynamic that commensal organisms balance in promoting both pathogen resistance and antibiotic resistance emergence.

Our results also demonstrated evidence of lower average gene richness and genome diversity in the dairy worker metagenomes compared to community control metagenomes. Loss of microbial diversity has been associated with intestinal dysbiosis, which refers to an unbalanced microbiota that can lead to development of diseases such as inflammatory bowel disease (IBD) [118, 183, 64]. Similarly, lower gene richness has been associated with increased intestinal inflammation and metabolic disorders [25, 98, 115]. A common occupational hazard facing dairy workers is inhalation of dusts and aerosols containing endotoxins or other proinflammatory substances that can result in airway inflammation and decreased pulmonary function [126, 29, 169]. Several studies have proposed a gut-lung axis linking pulmonary inflammation to intestinal inflammation based on epidemiological and clinical observations of the co-occurrence of these diseases [188, 85, 147]. The exact mechanisms underlying organ cross-talk is still an active area of research, but one proposed mechanism is blood-mediated transport of microbial products and metabolites [147]. It is possible that the lower gene richness and genome diversity observed in dairy workers points to increased intestinal inflammation linked to possible increased airway inflammation from exposure to aerosols and endotoxins. Further investigation to explore the possibility of increased intestinal and airway inflammation of this cohort is warranted.

At the genome-level, we tested for differentially enriched COG functions between five commensal organisms found in dairy worker and community control metagenomes. Focusing on COG functions from categories X (Mobilome: prophages, transposons) and V (Defense mechanisms) revealed enrichment of these functions in *R. bromii* and *Eubacterium* CAG-180 sp000432435 MAGs associated with dairy workers and *Ruminococcaceae* UBA 1417 sp002305575 and *F. prausnitzii* MAGs associated with community controls. We hypothesized that commensal organisms in dairy workers would be enriched in functions allowing for adaptation in response to ARG rich environments; however, we found no pattern of consistent enrichment of category X and V COG functions in commensal genomes associated with dairy workers. Expansion to more commensal organisms may reveal a more consistent pattern of differential enrichment in functions related to category V and X in dairy workers' commensal bacteria.

Our study had several limitations. The first, as already mentioned was its limited sample size. A larger sample size would increase our power to reject a false null hypothesis and improve our ability to detect true effect sizes that are smaller. To address this limitation in our metagenomics investigation, we can consider future work to incorporate shotgun metagenomics data from publicly available databases (e.g., the Human Microbiome Project [124]). Selection of samples to supplement our community control group requires careful consideration to reduce selection bias. Specifically, ideal samples for our community control group would be those from participants with similar characteristics as dairy workers (e.g., age, sex, recent antibiotic use) but that differ with respect to the exposure of interest (occupational exposure to livestock farming). Similarly, to address our limited sample size in our pangenomics investigation we can consider incorporating additional genome sequences of our 5 commensal species from publicly available databases. Integrating external genomes into our pangenomics analyses would not only increase our sample size, but would also allow us to better understand the generalizability of our study genomes.

The cross-sectional design of our study was another notable limitation. While cross-sectional studies can be advantageous for conducting cost-effective comparisons of a population, they capture differences observed at a single time-point and not the changes or trends that occur over time. Substantial temporal variation has been observed within individuals' microbiomes with one recent study using densely sampled temporal microbiome data to show that single measurements of genus-level relative and quantitative abundances do not estimate a person's temporal average well [186]. The results from our cross-sectional investigation therefore may not reflect average trends or patterns of differences between dairy workers and community controls. A longitudinal study is warranted to assess longer term changes and trends to the microbiome that are induced by occupational exposure to livestock farming.

An additional limitation of our study is that we did not have detailed information available on which antibiotics were used on the dairy farms and whether samples corresponding to dairy workers were from workers who administered or directly handled antibiotics. As a result, we were unable to directly link specific antibiotic use with the ARGs found in dairy workers' metagenomes. Instead we relied on information on antibiotics approved for use and commonly reported in dairy farming throughout the United States. This limitation, however, would likely not affect the generalizability of our results to other studies of occupational exposure to livestock farming environments in the United States.

Finally, we note that shotgun metagenomics-based approaches to studying antibiotic resistance is limited to genotypic potential, which may not confer phenotypic resistance. The goals of our study were to examine differences in ARG carriage of dairy worker associated commensals and metagenomes to better understand the effect of exposure to livestock farming environments on the occurrence of these genes. Our study goals therefore did not necessitate understanding of phenotypic resistance. However, future work that would complement our study would be to consider conducting whole genome sequencing of culturable commensal species and comparing phenotypic resistance profiles between organisms found in dairy workers and community controls.

We conducted a study of the effects of occupational exposure to dairy farm environments on functional differences of gut commensal bacteria from dairy workers and community controls as well as on differences in taxonomy, diversity and genes (i.e., CAGs, ARGs, virulence factors) between dairy worker and community control metagenomes. A major strength of our study is the multi-level interrogation of this question using both genomes and metagenomes. We observed several patterns for further investigation including greater abundance of tetracycline resistance genes and higher occurrence of cephamycin resistance genes in dairy workers' metagenomes; evidence of commensal organism association with plasmid-mediated tetracycline resistance genes; and lower gene and genome diversity in dairy workers' metagenomes. This work provides a foundation for further investigations into the impact of exposure to zoonotic pathogens, antibiotic resistant organisms, and ARGs on commensal organism adaptations and resistance in livestock workers. By furthering our understanding of commensal organism adaptations and their role in AMR propagation in response to exposure to zoonotic pathogen and ARG rich environments we gain new insights to aid in the development of therapeutic interventions that utilize commensal organisms.

4.5 Methods

4.5.1 Metagenomic assembly and processing of contigs

Contigs were assembled and annotated as described in Section 3.4.4.

4.5.2 Automatic binning and genome manual refinement

We followed guidelines published for curating high-quality MAGs from metagenomes using automated and interactive visualization tools [22]. To create preliminary clusters or bins of contigs, we used `MetaBat2` [83] v2.12.1 and `MaxBin2` v2.2.6 [198] then integrated the results from these binning algorithms into a non-redundant set of bins using `metaWRAP` v1.3.2 [182] with minimum completion set at greater than 70% and less than 10% redundancy. `metaWRAP` uses `checkM` v1.0.12 [137] to calculate its completion and contamination statistics. Once preliminary bins were created from our metagenomes, we used the anvi'o [44] interactive interface for manual refinement of the bins determined by `metaWRAP`. In our refinement approach, we relied on coverage patterns across samples to flag areas of anomalous coverage that may be errors due to binning and used gene-level taxonomic annotations to guide whether a contig was included in our refined MAG. We performed one round of manual refinement for each MAG.

4.5.3 Annotation of genomes

Taxonomic annotation of our MAGs was conducted using GTDB-Tk v2.0.0 [20] toolkit and the GTDB r89 database [136]. We reported reference genome sizes for species that had no studies involving completed genomes by examining genome sizes of representative species from the GTDB database with high estimated completions. We performed functional annotation of the genes in our MAGS using the COG14 [175] database. We queried for ARGs and virulence factor genes in our MAGs using `ABRicate` v1.0.1 [160], which performs mass screening of contigs using BLAST [3] and a variety of user specified databases. We selected the Comprehensive Antibiotic Resistance Database [110] as our reference database for AMR genes and the Virulence Factor Database [21] as our reference for virulence factor genes. We considered a gene as being present in our metagenome if it met the minimum thresholds of at least 90% identity and 100% coverage of the reference database gene sequence. Versions used of each database have been referenced in Chapter 3.4.5. Gene clusters were considered as part of the "Core" genome if they occurred in at least $\approx 80\%$ of the genomes.

4.5.4 Phylogenomic tree construction

We used single-copy core genes (SCGs) that had high geometric homogeneity but were functionally diverse to construct the phylogenetic trees of each set of species genomes. We look for these characteristics in our SCGs as we want to use genes that are highly conserved across our genomes but with functional variability among aligned residues across the genomes. To do this we use the anvi'o interactive interface to filter SCGs that were found in a minimum number of genomes, occurred only once in each genome, had a minimum geometric homogeneity of 1 (perfect), and had lower than some threshold for maximum functional homogeneity. For each species we set the minimum number of genomes that the SCGs needed to be identified in as $n - 2$ where n represents the total number of MAGs in the pangenome. We used different maximum values for functional homogeneity for each species: 0.8 for *F. prausnitzii*; 0.85 for *R. bromii*; 0.95 for *F. saccharivorans*; 0.95 for *Eubacterium* CAG-180; and 0.95

for Ruminococcaeceae UBA 1417. Phylogenetic trees were then constructed using MUSCLE v3.8.1551 [41] for multiple sequence alignment and FastTree v2.1.10 [142] for tree building. Construction of the pangenomic trees consisted of using the complete set of gene clusters identified across all genomes for a given species followed by multiple sequence alignment with MUSCLE and tree building with FastTree.

4.5.5 Statistical analyses

We used **happi** v1.0 to test for differential enrichment of COG functions in MAGs recovered from dairy workers compared to MAGs recovered from community controls. We used **happi** to account for differential mean coverages of our MAGs in our functional enrichment testing followed by a false discovery rate correction using the **qvalue** v2.26.0 package [170] for q-value estimation. A COG function was considered present in a MAG if it was observed at least once and absent if it was not observed. We report patterns for COG functions that had **happi** LRT $\chi^2 < 0.05$ as these were the functions with the highest magnitude test statistics. All statistical analyses were conducted using R v4.1.2.

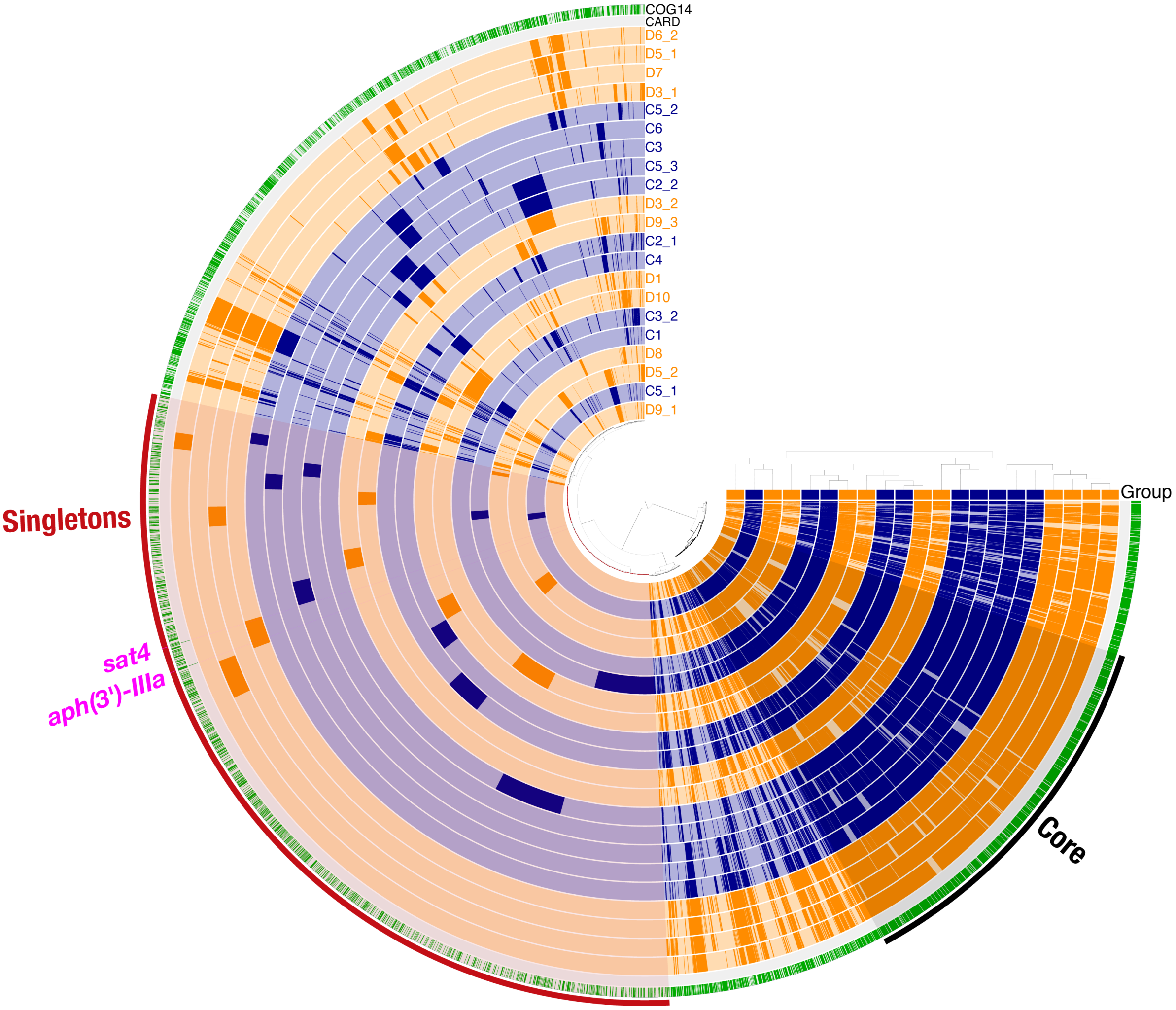


Figure 4.1: Organization of *F. prausnitzii* genomes based on gene cluster frequencies show no clear association between genome organization and group. The inner radial dendrogram corresponds to $6,076$ distinct gene clusters organized by gene presence/absence across the genomes. The inner 21 spokes represent genomes and are colored by group where orange spokes are genomes from dairy workers and blue spokes are genomes from community controls. These inner spokes corresponding to genomes are organized by their gene cluster frequencies (vertical dendrogram, top right). Cells that are filled represent gene clusters that are present. Gene clusters belonging to the "Core" genome are labeled in black and correspond to gene clusters that appear in at least $n = 18$ genomes. "Singletons" are labeled in red, corresponding to gene clusters that appear in only one genome and are part of the accessory genome. The second outermost spoke displays gene sequences from the CARD database. The outermost spoke identifies COG14 functions.

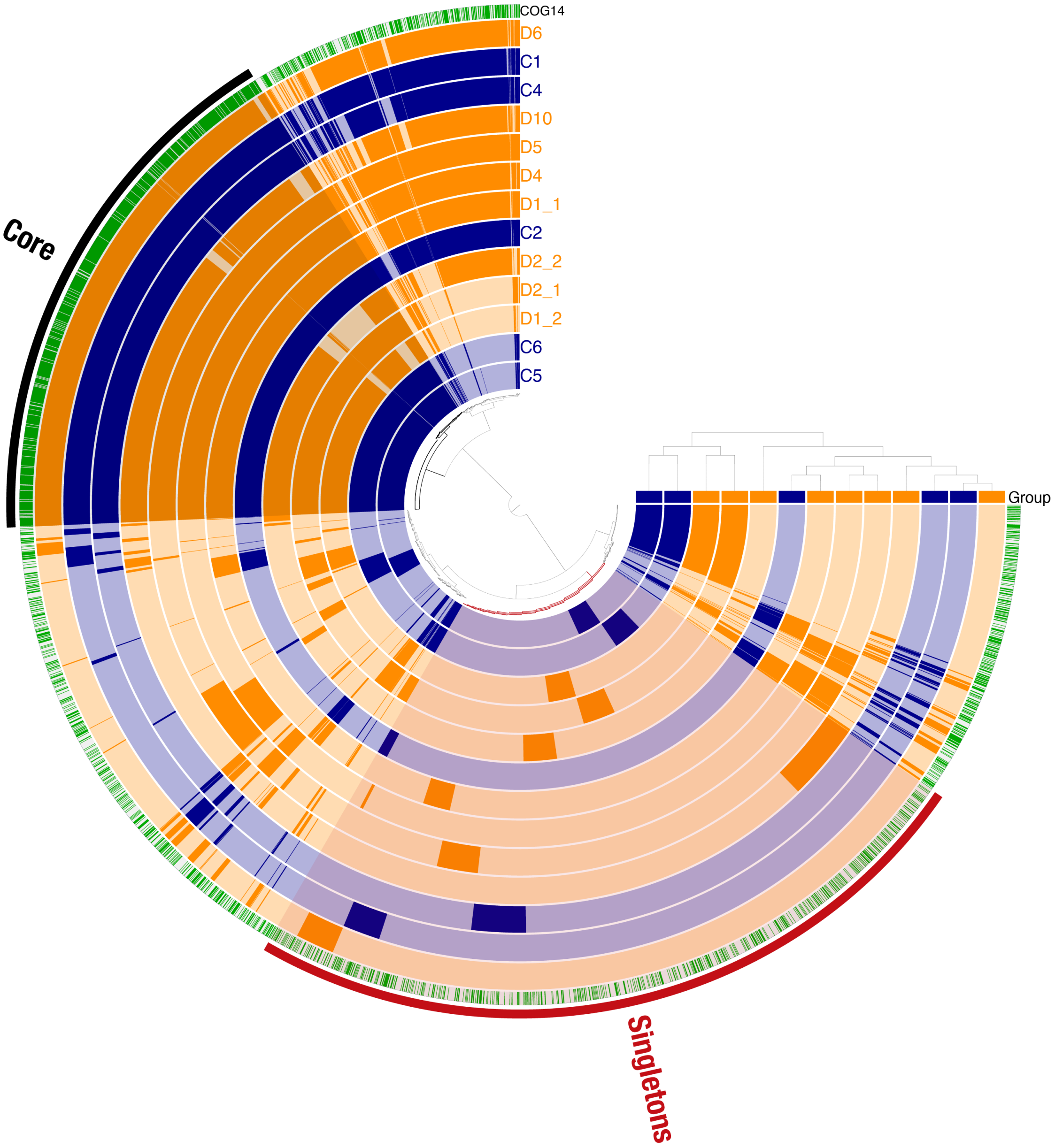


Figure 4.2: Genomes of *R. bromii* have been organized by gene content and reveal no evident association by group. The inner radial dendrogram corresponds to organization of $4,846$ gene clusters by presence/absence across genomes. The inner 13 spokes represent *R. bromii* genomes colored by affiliation with either dairy workers (orange) or community controls (blue). Groups of gene clusters corresponding to the "Core" genome are labeled in black and those corresponding to the accessory genome ("Singletons") are in red. Gene clusters are considered as part of the "Core" genome if they are present in at least $n = 11$ genomes.

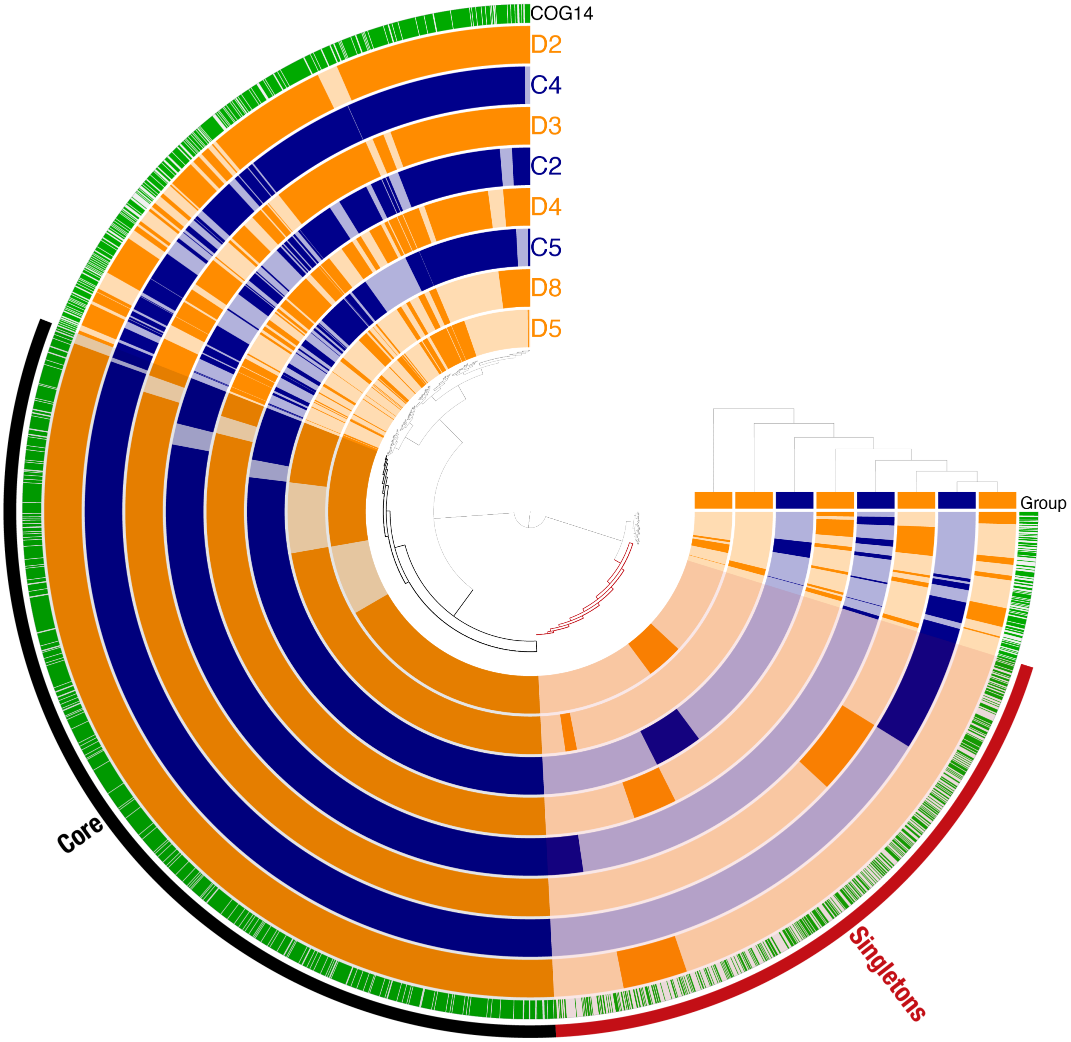


Figure 4.3: Pangenomic analysis of the species *Fusicatenibacter saccharivorans* reveals no distinct associations between genome organization based on gene cluster frequencies by group. The inner radial dendrogram corresponds to organization of 3,477 total gene clusters by presence/absence across the *F. saccharivorans* genomes. The inner eight spokes correspond to *F. saccharivorans* genomes recovered from dairy worker (orange) and community control (blue) samples. Gene clusters are considered as part of the "Core" genome if they are present in at least $n = 6$ genomes.

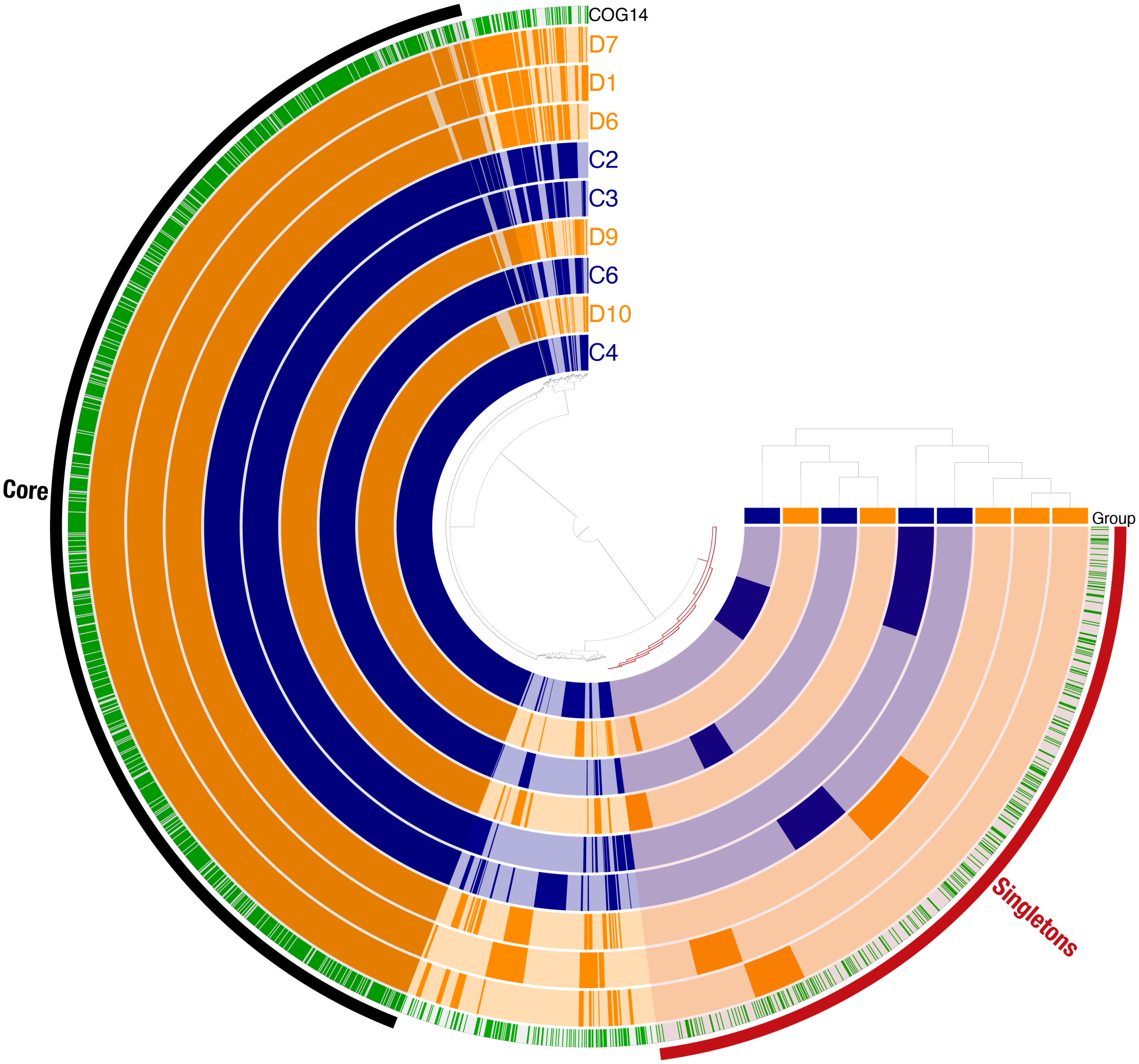


Figure 4.4: Genomes of the species *Eubacterium* CAG-180 have been organized by their gene content revealing no affiliation by group (dairy worker compared to community control). The center radial dendrogram of the phylogram organizes the $2,883$ distinct gene clusters by their presence/absence across the nine genomes. Each of the inner nine spokes corresponds to a *Eubacterium* CAG-180 genome and has been colored by association with dairy worker (orange) or community control (blue). Gene clusters are considered as part of the “Core” genome if they are present in at least $n = 7$ genomes.

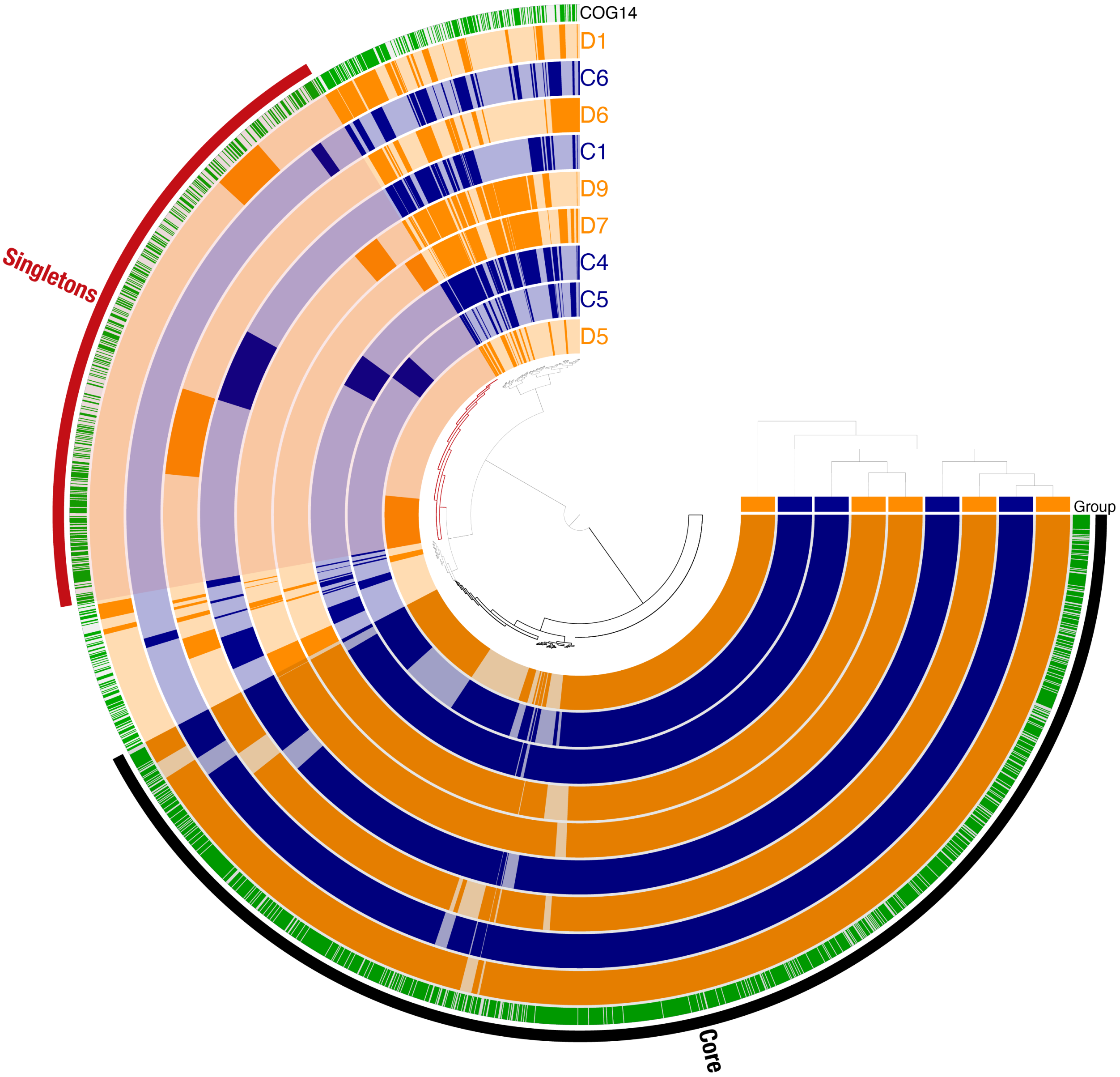


Figure 4.5: Nine *Ruminococcaceae* UBA 1417 genomes have been organized by their gene content and showed no association of organization by group. The inner nine spokes correspond to individual *Ruminococcaceae* UBA 1417 genomes that are colored by affiliation with dairy workers (orange) or community controls (blue). Gene clusters that are present are filled. The inner dendrogram organizes 3,126 gene clusters by presence/absence whereas the dendrogram at the 90° position organizes the genomes by gene frequencies. Gene clusters are considered as part of the "Core" genome if they are present in at least $n = 7$ genomes.

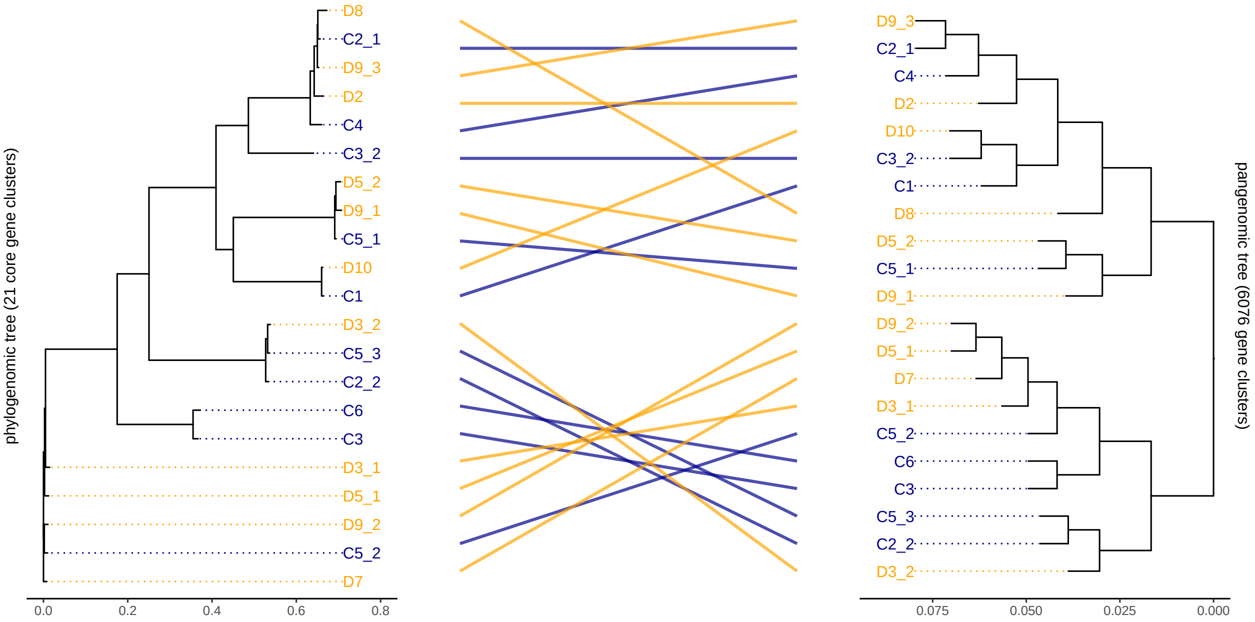


Figure 4.6: Phylogenetic organization of *Faecalibacterium prausnitzii* genomes provides no evidence of affiliation by group. The left dendrogram was constructed using 21 single copy core genes whereas the right dendrogram incorporated all 6,076 gene clusters detected in the pangenome.

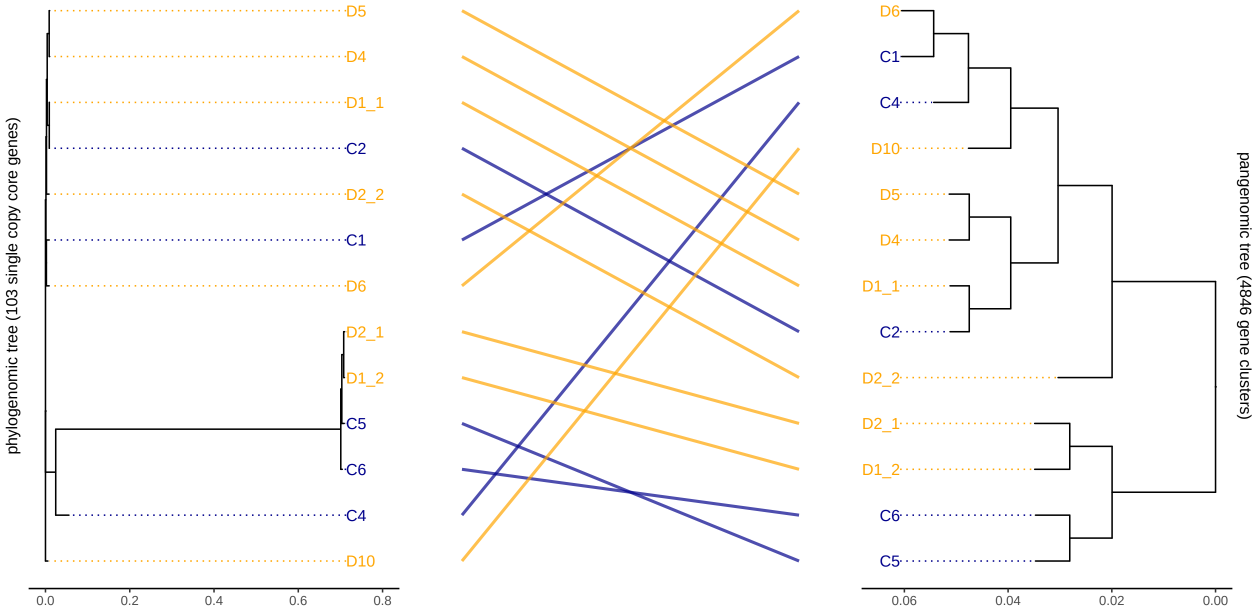


Figure 4.7: Organization of *R. bromii* genomes using phylogenomics shows no evidence of association by group. The left dendrogram was constructed using 103 single copy core genes and the right dendrogram used all 4,846 gene clusters detected in the pangenome.

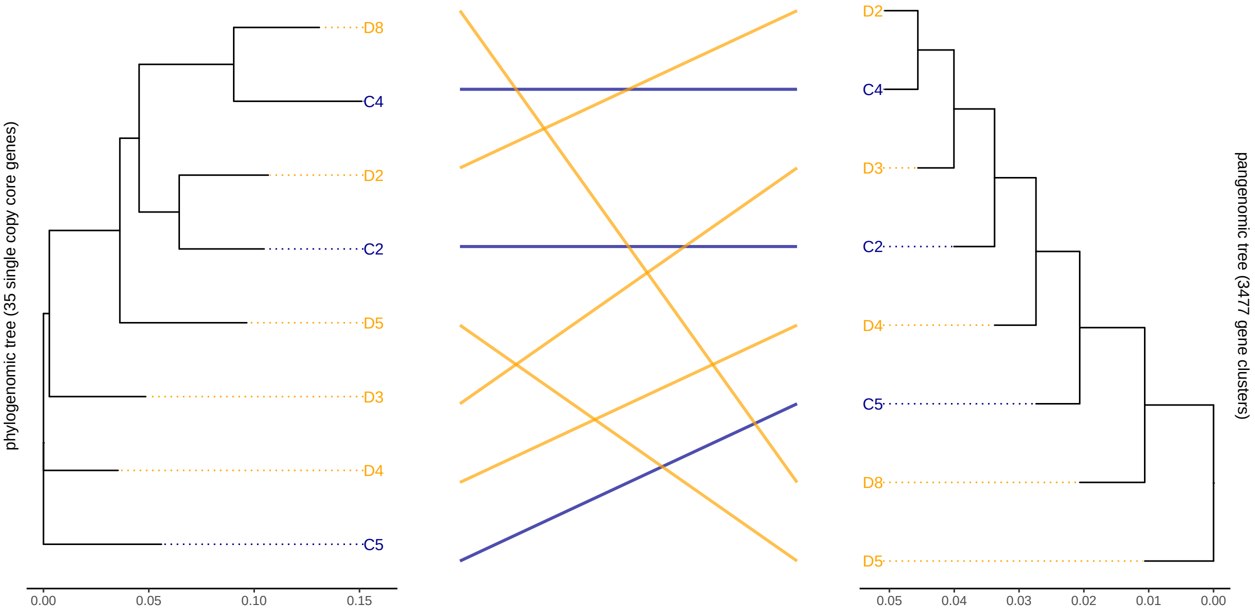

phylogenomic tree (35 single copy core genes)
D8
C4
D2
C2
D5
D3
D4
C5
0.00
0.05
0.10
0.15
D2
C4
D3
C2
D4
C5
D8
D5
0.05
0.04
0.03
0.02
0.01
0.00
pangenomic tree (3477 gene clusters)

Figure 4.8: Phylogenetic organization of *F. saccharivorans* genomes provided no evidence of affiliation by group. The left dendrogram was constructed using 35 single copy core genes whereas the right dendrogram utilized all 3,477 gene clusters detected across genomes.

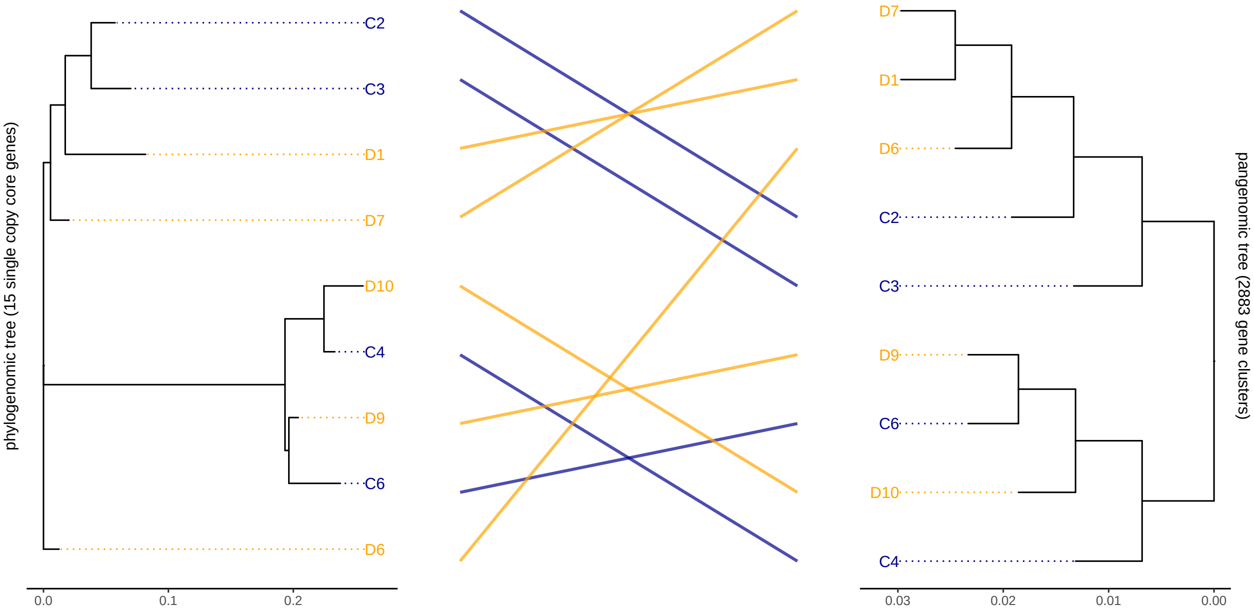


Figure 4.9: Phylogenetic organization of the *Eubacterium* CAG-180 genomes did not identify any clear associations by group. The dendrogram on the left was constructed using 15 single copy core genes whereas the dendrogram on the right utilized all 2,883 gene clusters detected in the pangenome.

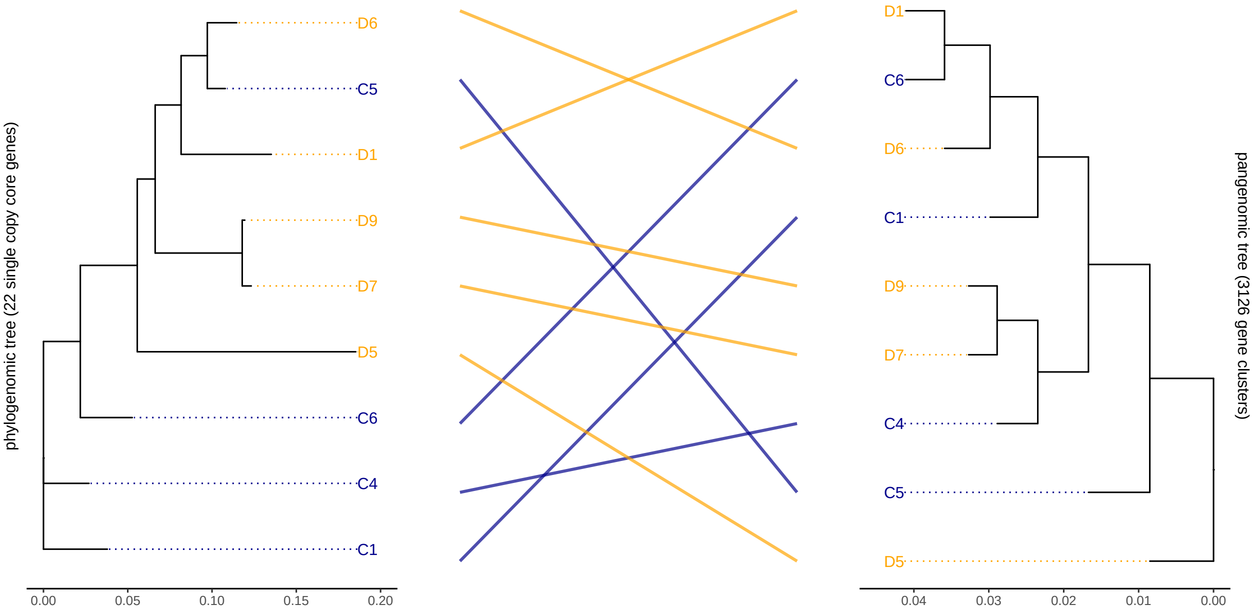

phylogenomic tree (22 single copy core genes)
D6
C5
D1
D9
D7
D5
C6
C4
C1
0.00
0.05
0.10
0.15
0.20
D1
C6
D6
C1
D9
D7
C4
C5
D5
0.04
0.03
0.02
0.01
0.00
pangenomic tree (3126 gene clusters)

Figure 4.10: Organization of these nine *Ruminococcaceae* UBA 1417 genomes using phylogenomics provided no evidence of affiliation by group. The left dendrogram was constructed using 22 single copy core genes and the right dendrogram used all 3,126 gene clusters detected in the pangenome.

Chapter 5
CONCLUDING REMARKS

Livestock farming environments are unique nexuses for the transmission and interaction of antibiotic resistance and virulence factor genes, zoonotic pathogens, and non-pathogenic bacteria [111]. Advancements in next-generation sequencing have allowed for a shift from a single pathogen to a microbial community view to understanding the transmission dynamics of zoonotic diseases and antibiotic resistant bacteria and genes. Increased knowledge of the composition, genetics, and functional capacity of commensal organisms have demonstrated the profound impact commensal organisms have on host immune homeostasis [6, 77, 5], disease development [55, 200, 122, 27], and resistance against pathogen invasion [1, 86, 16, 82]. However, there is growing concern over the potential role of commensals in facilitating antibiotic resistance emergence and propagation by serving as antibiotic resistance gene reservoirs for transmission to pathogens. It is therefore important to understand differences that may arise in the human gut microbiota from exposure to environments such as livestock farms and other situations of close human-animal contact to evaluate potential risks of these environments in propagation of zoonotic disease and antibiotic resistant bacteria and their genes.

In this dissertation I investigated differences between dairy worker and community control microbiomes and resistomes from both a metagenomics and pangenomics approach. To facilitate functional comparisons of microbial genomes between dairy workers and community controls, I addressed existing limitations with statistical methods for pangenomics by developing a statistical method for modeling gene presence that accounts for differential genome quality factors (e.g., mean coverage). Evaluation of `happi`'s performance using simulated data showed that `happi` is able to correctly control the Type 1 error rate when there is correlation between quality variables and the main covariate of interest whereas existing logistic regression methods do not. happi also exhibits greater power with larger samples sizes. However, with greater correlation between quality variables and the main covariate of interest happi has comparatively decreased power. A data analysis of Saccharibacteria genomes was also conducted and demonstrated happi's ability to produce larger p-values in situations where the lack of gene detection may be conflated with lower quality genomes, and smaller p-values when differences in detection cannot be attributed to genome quality. Taken together, the work from Chapter 2 illustrates the accuracy and robustness of happi even when genome quality is correlated with the main covariate of interest. happi can furthermore be broadly applied to functional comparisons of genomes of other microorganisms beyond bacteria and used in functional comparisons of metagenomes to adjust for differential quality of metagenomes. To facilitate usability, transparency, and reproducibility with happi, an open-source R package for happi has been made publicly available along with documentation on usage [180].

Results from the metagenome investigation of differences between dairy workers and community controls revealed no statistically significant differences at the 5% significance level in the taxonomic composition, antibiotic resistance and virulence factor gene carriage, and relative abundances of co-abundant gene groups and ARGs. The pangenome investigation similarly found no differences between functions and phylogenetic organization of commensal genomes from dairy workers and community controls. However, there were some notable patterns for further investigation. We observed greater relative abundances of tetracycline resistance genes and higher occurrence of cephamycin resistance genes in dairy workers' metagenomes. These patterns are particularly interesting given that tetracyclines and beta-lactams are two of the most commonly administered antibiotics on dairy farms in the United States [78]. A direction for future work investigating this pattern could involve shotgun metagenomic sequencing of a different set of controls. As mentioned in the Discussion section of Chapter 4, the community controls in our metagenomic study identified as field workers in non animal farming industries. This raises concerns about similarities in occupational exposure to antibiotics between the community controls and the dairy workers in our study since antibiotics have been commonly used in plant agriculture in the United States to control fire blight disease in orchards [38]. Similar occupational exposures in the dairy workers and controls would make both groups more comparable and likely contribute to null findings when examining for differences between groups. Therefore, an investigation with a different control group that might lead to greater differences between livestock workers and controls is warranted. In particular, an ideal group of controls would be those who are similar to dairy workers in age, sex, race and ethnicity, had no recent antibiotic use three months prior to sample collection, employed in a non-farming related industry such as an office worker, no co-habitation with an individual employed in a farming-related industry, live a minimum distance away from any conventional animal or plant farms, and live in a similar urban or rural environment as the dairy workers.

Our study also highlighted the potential for sharing of ARGs between commensal organisms and pathogens through conjugative plasmids and showed that taxonomic affiliation of these genes did not differentiate based on exposure to livestock farming environments. This finding emphasizes the need for further research to understand the role of commensal organisms in ARG persistence and propagation. Finally, we observed decreased gene richness and genome diversity in dairy workers compared to community controls. Lower gene richness and genome diversity have been associated with increased intestinal inflammation and dysbiosis [25, 98, 115]. Several studies have also proposed a link between intestinal inflammation and airway inflammation through the gut-lung axis [188, 85, 147]. Further investigation of intestinal and airway inflammation of this cohort is needed to contextualize these findings.

5.1 Future Directions

The work in this dissertation was limited to examining differences between microbiomes of dairy workers and controls and not changes to the microbiome that are induced by occupational exposure to livestock farming. As such, on-going complementary work is planned to understand the extent of human gut microbiota alteration in commercial livestock farming environments. This work involves a longitudinal study of individuals who are new to dairy work to investigate functional alterations and adaptations that occur to the human gut microbiome due to occupational exposure to livestock farming environments. This study of new dairy workers over time will provide insights into how and to what extent commencement of work in livestock farming environments can shape the human gut microbiota and resistome. Further directions to build on this work also include shotgun metagenomic sequencing of a different group of controls as mentioned above that do not share similar occupational exposures to antibiotics as dairy workers. Studying a control group without similar occupational exposures to antibiotics as dairy workers could provide greater contrast between the microbiomes and resistomes of dairy workers and controls. Additionally, we can also consider constructing a job exposure matrix to characterize occupational exposures to livestock using self-reported work tasks. We can then examine whether differences in the microbiome or resistome exist based on these exposures. Using more detailed characterization of occupational exposures to livestock to examine differences in the microbiome and resistome would enable identification of workers with high-risk exposures and allow for targeted interventions to be crafted to limit exposures in these individuals.

BIBLIOGRAPHY

[1] Michael C Abt and Eric G Pamer. Commensal bacteria mediated defenses against pathogens. *Current Opinion in Immunology*, 29:16–22, 2014. Host pathogens * Immune senescence.

[2] Brian P Alcock, Amogelang R Raphenya, Tammy T Y Lau, Kara K Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, Anna-Lisa V Nguyen, Annie A Cheng, Sihan Liu, Sally Y Min, Anatoly Miroshnichenko, Hiu-Ki Tran, Rafik E Werfalli, Jalees A Nasir, Martins Oloni, David J Speicher, Alexandra Florescu, Bhavya Singh, Mateusz Faltyn, Anastasia Hernandez-Koutoucheva, Arjun N Sharma, Emily Bordeleau, Andrew C Pawlowski, Haley L Zubyk, Damion Dooley, Emma Griffiths, Finlay Maguire, Geoff L Winsor, Robert G Beiko, Fiona S L Brinkman, William W L Hsiao, Gary V Domselaar, and Andrew G McArthur. Card 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic acids research*, 48(D1):D517–D525, 01 2020.

[3] S Altschul, W Gish, W Miller, E Myers, and D Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3):403–440, 1990.

[4] Yoko Arimizu, Yumi Kirino, Mitsuhiko P. Sato, Koichi Uno, Toshio Sato, Yasuhiro Gotoh, Frédéric Auvray, Hubert Brugere, Eric Oswald, Jacques G. Mainil, Kelly S. Anklam, Dörte Döpfer, Shuji Yoshino, Tadasuke Ooka, Yasuhiro Tanizawa, Yasukazu Nakamura, Atsushi Iguchi, Tomoko Morita-Ishihara, Makoto Ohnishi, Koichi Akashi, Tetsuya Hayashi, and Yoshitoshi Ogura. Large-scale genome analysis of bovine commensal Escherichia coli reveals that bovine-adapted E. Coli lineages are serving as evolutionary sources of the emergence of human intestinal pathogenic strains. *Genome Research*, 29(9):1495–1505, 2019.

[5] Marie-Claire Arrieta and Barton Finlay. The commensal microbiota drives immune homeostasis. *Frontiers in Immunology*, 3, 2012.

[6] David Artis. Epithelial-cell recognition of commensal bacteria and maintenance of immune homeostasis in the gut. *Nature Reviews Immunology*, 8(6):411–420, 2008.

[7] Satyabrata Bag, Tarini Shankar Ghosh, and Bhabatosh Das. Complete genome sequence of faecalibacterium prausnitzii isolated from the gut of a healthy indian adult. *Genome Announcements*, 5(46):e01286–17, 2017.

[8] Nielson T. Baxter, Alexander W. Schmidt, Arvind Venkataraman, Kwi S. Kim, Clive Waldron, and Thomas M. Schmidt. Dynamics of human gut microbiota and short-chain fatty acids in response to dietary interventions with three fermentable fibers. *bioRxiv*, 10(1):1–13, 2018.

[9] Francesco Beghini, Lauren J. McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, Paolo Manghi, Matthias Scholz, Andrew Maltez Thomas, Mireia Valles-Colomer, George Weingart, Yancong Zhang, Moreno Zolfo, Curtis Huttenhower, Eric A. Franzosa, and Nicola Segata. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. *eLife*, 10:1–42, 2021.

[10] Francesco Beghini, Lauren J. McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, Paolo Manghi, Matthias Scholz, Andrew Maltez Thomas, Mireia Valles-Colomer, George Weingart, Yancong Zhang, Moreno Zolfo, Curtis Huttenhower, Eric A. Franzosa, and Nicola Segata. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. *eLife*, 10:1–42, 2021.

[11] Amy Biddle, Lucy Stewart, Jeffrey Blanchard, and Susan Leschine. Untangling the genetic basis of fibrolytic specialization by lachnospiraceae and ruminococcaceae in diverse gut communities. *Diversity*, 5(3):627–640, 2013.

[12] Martin J Blaser. Antibiotic use and its consequences for the normal microbiome. *Science (New York, N.Y.)*, 352(6285):544–545, 04 2016.

[13] Ryan A. Blaustein, Alexander G. McFarland, Sarah Ben Maamar, Alberto Lopez, Sarah Castro-Wallace, and Erica M. Hartmann. Pangenomic Approach To Understanding Microbial Adaptations within a Model Built Environment, the International Space Station, Relative to Human Hosts and Soil. *mSystems*, 4(1):1–16, 2019.

[14] Ola Brynildsrud, Jon Bohlin, Lonneke Scheffer, and Vegard Eldholm. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biology*, 17(1):1–9, 2016.

[15] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using diamond. *Nature Methods*, 12(1):59–60, 2015.

[16] Charlie G. Buffie and Eric G. Pamer. Microbiota-mediated colonization resistance against intestinal pathogens. *Nature Reviews Immunology*, 13(11):790–801, 2013.

[17] B Bushnell. Bbtools software package, 2014.

[18] Brian Bushnell. Masked version of hg19. Accessed: 2022-06-26.

[19] Carlos P Cantalapiedra, Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, 38(12):5825–5829, 10 2021.

[20] Pierre-Alain Chaumeil, Aaron J Mussig, Philip Hugenholtz, and Donovan H Parks. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, 36(6):1925–1927, 11 2019.

[21] Lihong Chen, Jian Yang, Jun Yu, Zhijian Yao, Lilian Sun, Yan Shen, and Qi Jin. Vfdb: a reference database for bacterial virulence factors. *Nucleic Acids Research*, 33(1):325–328, 01 2005.

[22] Lin Xing Chen, Karthik Anantharaman, Alon Shaiber, A Murat Eren, and Jillian F Banfield. Accurate and complete genomes from metagenomes. *Genome Research*, 30(3):315–333, 2020.

[23] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890, 09 2018.

[24] Caitlin Collins and Xavier Didelot. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *bioRxiv*, pages 1–21, 2017.

[25] Aurélie Cotillard, Sean P. Kennedy, Ling Chun Kong, Edi Prifti, Nicolas Pons, Emmanuelle Le Chatelier, Mathieu Almeida, Benoit Quinquis, Florence Levenez, Nathalie Galleron, Sophie Gougis, Salwa Rizkalla, Jean-Michel Batto, Pierre Renault, Joel Doré, Jean-Daniel Zucker, Karine Clément, Stanislav Dusko Ehrlich, Hervé Blottière, Marion Leclerc, Catherine Juste, Tomas de Wouters, Patricia Lepage, Charlene Fouqueray, Arnaud Basdevant, Cornelieu Henegar, Cindy Godard, Marine Fondacci, Alili Rohia, Froogh Hajduch, Jean Weissenbach, Eric Pelletier, Denis Le Paslier, Jean-Pierre Gauchi, Jean-François Gibrat, Valentin Loux, Wilfrid Carré, Emmanuelle Maguin, Maarten van de Guchte, Alexandre Jamet, Fouad Boumezbeur, Séverine Layec, ANR MicroObes consortium, and ANR MicroObes consortium members. Dietary intervention impact on gut microbial gene richness. *Nature*, 500(7464):585–588, 2013.

[26] Emmanuelle H. Crost, Gwenaelle Le Gall, Jenny A. Laverde-Gomez, Indrani Mukhopadhya, Harry J. Flint, and Nathalie Juge. Mechanistic insights into the cross-feeding of ruminococcus gnavus and ruminococcus bromii on host and dietary carbohydrates. *Frontiers in Microbiology*, 9, 2018.

[27] A. L. Cunningham, J. W. Stephens, and D. A. Harris. Gut microbiota influence in type 2 diabetes mellitus (t2dm). *Gut Pathogens*, 13(1):50, 2021.

[28] Lawrence A. David, Corinne F. Maurice, Rachel N. Carmody, David B. Gootenberg, Julie E. Button, Benjamin E. Wolfe, Alisha V. Ling, A. Sloan Devlin, Yug Varma, Michael A. Fischbach, Sudha B. Biddinger, Rachel J. Dutton, and Peter J. Turnbaugh. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–563, 2014.

[29] Margaret E Davidson, Joshua Schaeffer, Maggie L Clark, Sheryl Magzamen, Elizabeth J Brooks, Thomas J Keefe, Mary Bradford, Noa Roman-Muniz, John Mehaffy, Gregory Dooley, Jill A Poole, Frank M Mitloehner, Sue Reed, Marc B Schenker, and Stephen J Reynolds. Personal exposure of dairy workers to dust, endotoxin, muramic acid, ergosterol, and ammonia on large-scale dairies in the high plains western united states. *Journal of occupational and environmental hygiene*, 15(3):182–193, 03 2018.

[30] Jan de Leeuw, Kurt Hornik, and Patrick Mair. Isotone optimization in R: Pool-adjacent-violators algorithm (pava) and active set methods. *Journal of Statistical Software*, 32(5):1–24, 2009.

[31] Rebekah M. Dedrick, Haley G. Aull, Deborah Jacobs-Sera, Rebecca A. Garlena, Daniel A. Russell, Bailey E. Smith, Vaishnavi Mahalingam, Lawrence Abad, Christian H. Gauthier, Graham F. Hatfull, and M. Sloan Siegrist. The prophage and plasmid mobilome as a likely driver of mycobacterium abscessus diversity. *mBio*, 12(2):e03441–20, 2021.

[32] Tom O Delmont and A Murat Eren. Linking pangenomes and metagenomes: the prochlorococcus metapangenome. *PeerJ*, 6:e4320–e4320, 01 2018.

[33] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[34] Martin Depner, Diana Hazard Taft, Pirkka V. Kirjavainen, Karen M. Kalanetra, Anne M. Karvonen, Stefanie Peschel, Elisabeth Schmausser-Hechfellner, Caroline Roduit, Remo Frei, Roger Lauener, Amandine Divaret-Chauveau, Jean Charles Dalphin, Josef Riedler, Marjut Roponen, Michael Kabesch, Harald Renz, Juha Pekkanen, Freda M. Farquharson, Petra Louis, David A. Mills, Erika von Mutius, Jon Genuneit, Anne Hyvärinen, Sabina Illi, Lucie Laurent, Petra I. Pfefferle, Bianca Schaub, Erika von Mutius, and Markus J. Ege. Maturation of the gut microbiome during the first year of life contributes to the protective farm effect on childhood asthma. *Nature Medicine*, 26(11):1766–1775, 2020.

[35] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, 2017.

[36] Ana Sofia Ribeiro Duarte, Timo Röder, Liese Van Gompel, Thomas Nordahl Petersen, Rasmus Borup Hansen, Inge Marianne Hansen, Alex Bossers, Frank M. Aarestrup, Jaap A. Wagenaar, and Tine Hald. Metagenomics-based approach to source-attribution of antimicrobial resistance determinants – identification of reservoir resistome signatures. *Frontiers in Microbiology*, 11, 2021.

[37] Carlos M Duarte, David K Ngugi, Intikhab Alam, John Pearman, Allan Kamau, Victor M Eguiluz, Takashi Gojobori, Silvia G Acinas, Josep M Gasol, Vladimir Bajic, and Xabier Irigoien. Sequencing effort dictates gene discovery in marine microbial metagenomes. *Environmental Microbiology*, 00:1–15, 2020.

[38] Brion Duffy, Eduard Holliger, and Fiona Walsh. Streptomycin use in apple orchards did not increase abundance of mobile resistance genes. *FEMS Microbiology Letters*, 350(2):180–189, 01 2014.

[39] Sarah G. Earle, Chieh Hsi Wu, Jane Charlesworth, Nicole Stoesser, N. Claire Gordon, Timothy M. Walker, Chris C.A. Spencer, Zamin Iqbal, David A. Clifton, Katie L. Hopkins, Neil Woodford, E. Grace Smith, Nazir Ismail, Martin J. Llewelyn, Tim E. Peto, Derrick W. Crook, Gil McVean, A. Sarah Walker, and Daniel J. Wilson. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nature Microbiology*, 1(5):1–8, 2016.

[40] Sean R. Eddy. Accelerated profile hmm searches. *PLOS Computational Biology*, 7(10):1–16, 10 2011.

[41] Robert C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 03 2004.

[42] Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, Erik L L Sonnhammer, Layla Hirsh, Lisanna Paladin, Damiano Piovesan, Silvio C E Tosatto, and Robert D Finn. The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1):D427–D432, 10 2018.

[43] A. Murat Eren, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3:e1319, 2015.

[44] A. Murat Eren, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3:e1319, 2015.

[45] A. Murat Eren, Joseph H. Vineis, Hilary G. Morrison, and Mitchell L. Sogin. A filtering method to generate high quality short reads using illumina paired-end technology. *PLOS ONE*, 8(6):1–6, 06 2013.

[46] FAO. *The future of food and agriculture: trends and challenges.* 2017.

[47] Carmen Veríssima Ferreira-Halder, Alessandra Valéria de Sousa Faria, and Sheila Siqueira Andrade. Action and function of faecalibacterium prausnitzii in health and disease. *Best Practice and Research Clinical Gastroenterology*, 31(6):643–648, 2017. Gut Microbiome in Health and Disease.

[48] Cormac Brian Fitzgerald, Andrey N. Shkoporov, Thomas D. S. Sutton, Andrei V. Chaplin, Vimalkumar Velayudhan, R. Paul Ross, and Colin Hill. Comparative analysis of faecalibacterium prausnitzii genomes shows a high level of genome plasticity and warrants separation into new species-level taxa. *BMC Genomics*, 19(1):931, 2018.

[49] Cormac Brian Fitzgerald, Andrey N. Shkoporov, Thomas D. S. Sutton, Andrei V. Chaplin, Vimalkumar Velayudhan, R. Paul Ross, and Colin Hill. Comparative analysis of faecalibacterium prausnitzii genomes shows a high level of genome plasticity and warrants separation into new species-level taxa. *BMC Genomics*, 19(1):931, 2018.

[50] Carla Foditsch, Thiago M. A. Santos, Andre G. V. Teixeira, Richard V. V. Pereira, Juliana M. Dias, Natália Gaeta, and Rodrigo C. Bicalho. Isolation and characterization of faecalibacterium prausnitzii from calves and piglets. *PLOS ONE*, 9(12):1–19, 12 2015.

[51] US Food. Drug administration 2015. fact sheet: veterinary feed directive final rule and next steps.

[52] US Food, Drug Administration, et al. New animal drugs and new animal drug combination products administered in or on medicated feed or drinking water of food-producing animals: recommendations for drug sponsors for voluntarily aligning product use conditions with gfi# 209. guidance for industry# 213. 2013, 2014.

[53] Samuel C. Forster, Junyan Liu, Nitin Kumar, Emily L. Gulliver, Jodee A. Gould, Alejandra Escobar-Zepeda, Tapoka Mkandawire, Lindsay J. Pike, Yan Shao, Mark D. Stares, Hilary P. Browne, B. Anne Neville, and Trevor D. Lawley. Strain-level characterization of broad host range mobile genetic elements transferring antibiotic resistance from the human microbiome. *Nature Communications*, 13(1):1445, 2022.

[54] R. Gacesa, A. Kurilshikov, A. Vich Vila, T. Sinha, M. A. Y. Klaassen, L. A. Bolte, S. Andreu-Sánchez, L. Chen, V. Collij, S. Hu, J. A. M. Dekens, V. C. Lenters, J. R. Björk, J. C. Swarte, M. A. Swertz, B. H. Jansen, J. Gelderloos-Arends, S. Jankipersadsing, M. Hofker, R. C. H. Vermeulen, S. Sanna, H. J. M. Harmsen, C. Wijmenga, J. Fu, A. Zhernakova, and R. K. Weersma. Environmental factors shaping the gut microbiome in a dutch population. *Nature*, 604(7907):732–739, 2022.

[55] Jiafeng Geng, Qingqiang Ni, Wei Sun, Liangge Li, and Xiujing Feng. The links between gut microbiota and obesity and obesity related diseases. *Biomedicine and Pharmacotherapy*, 147:112678, 2022.

[56] Dirk Gevers, Subra Kugathasan, Lee A. Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, Dan Knights, Se Jin Song, Moran Yassour, Xochitl C. Morgan, Aleksandar D. Kostic, Chengwei Luo, Antonio González, Daniel McDonald, Yael Haberman, Thomas Walters, Susan Baker, Joel Rosh, Michael Stephens, Melvin Heyman, James Markowitz, Robert Baldassano, Anne Griffiths, Francisco Sylvester, David Mack, Sandra Kim, Wallace Crandall, Jeffrey Hyams, Curtis Huttenhower, Rob Knight, and Ramnik J. Xavier. The treatment-naive microbiome in new-onset crohn's disease. *Cell Host and Microbe*, 15(3):382–392, 2014.

[57] Maarten J Gilbert, Marian E H Bos, Birgitta Duim, Bert A P Urlings, Lourens Heres, Jaap A Wagenaar, and Dick J J Heederik. Livestock-associated mrsa st398 carriage in pig slaughterhouse workers related to quantitative environmental exposure. *Occupational and Environmental Medicine*, 69(7):472–478, 2012.

[58] Brent Gilpin, Paula Scholes, Beth Robson, and Marion Savill. The transmission of thermotolerant campylobacter spp. to people living or working on dairy farms in new zealand. *Zoonoses and public health*, 55:352–60, 10 2008.

[59] Jonathan Louis Golob and Samuel Schwartz Minot. In silico benchmarking of metagenomic tools for coding sequence detection reveals the limits of sensitivity and precision. *BMC Bioinformatics*, 21(1):459, 2020.

[60] Leah Grout, Michael G Baker, Nigel French, and Simon Hales. A Review of Potential Public Health Impacts Associated With the Global Dairy Sector. *GeoHealth*, 4(2):1–46, 2020.

[61] M.V. Gryaznova, S.A. Solodskikh, A.V. Panevina, M.Y. Syromyatnikov, Yu.D. Dvoretskaya, T.N. Sviridova, E.S. Popov, and V.N. Popov. Study of microbiome changes in patients with ulcerative colitis in the central european part of russia. *Heliyon*, 7(3):e06432, 2021.

[62] Xiaoqiong Gu, Jean X. Y. Sim, Wei Lin Lee, Liang Cui, Yvonne F. Z. Chan, Ega Danu Chang, Yii Ean Teh, An-Ni Zhang, Federica Armas, Franciscus Chandra, Hongjie Chen, Shijie Zhao, Zhanyi Lee, Janelle R. Thompson, Eng Eong Ooi, Jenny G. Low, Eric J. Alm, and Shirin Kalimuddin. Gut ruminococcaceae levels at baseline correlate with risk of antibiotic-associated diarrhea. *iScience*, 25(1):103644, 2022.

[63] H Soon Gweon, Liam P Shaw, Jeremy Swann, Nicola De Maio, Manal Abuoun, Rene Niehus, Alasdair T M Hubbard, Mike J Bowes, Mark J Bailey, T E A Peto, Sarah J Hoosdally, A Sarah Walker, Robert P Sebra, Derrick W Crook, Muna F Anjum, Daniel S Read, Nicole Stoesser, M Abuoun, M Anjum, M J Bailey, L Barker, H Brett, M J Bowes, K Chau, D W Crook, N De Maio, D Gilson, H S Gweon, A T M Hubbard, S Hoosdally, J Kavanagh, H Jones, D S Read, R Sebra, L P Shaw, A E Sheppard, R Smith, E Stubberfield, J Swann, A S Walker, and N Woodford. The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples. *Environmental Microbiomes*, 14(1):1–15, 2019.

[64] Jonas Halfvarson, Colin J. Brislawn, Regina Lamendella, Yoshiki Vázquez-Baeza, William A. Walters, Lisa M. Bramer, Mauro D'Amato, Ferdinando Bonfiglio, Daniel McDonald, Antonio Gonzalez, Erin E. McClure, Mitchell F. Dunklebarger, Rob Knight, and Janet K. Jansson. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature Microbiology*, 2(5):17004, 2017.

[65] Colin Hill. Virulence or niche factors: What's in a name? *Journal of Bacteriology*, 194(21):5725–5727, 2012.

[66] Benjamin Hillmann, Gabriel A Al-Ghalith, Robin R Shields-Cutler, Qiyun Zhu, Daryl M Gohl, Kenneth B Beckman, Rob Knight, and Dan Knights. Evaluating the information content of shallow shotgun metagenomics. *MSystems*, 3(6), 2018.

[67] Matthew Holden, Lisa Crossman, Ana Cerdeño-Tárraga, and Julian Parkhill. Pathogenomics of non-pathogens. *Nature Reviews Microbiology*, 2(2):91–91, 2004.

[68] Hannes Horn, Beate M. Slaby, Martin T. Jahn, Kristina Bayer, Lucas Moitinho-Silva, Frank Förster, Usama R. Abdelmohsen, and Ute Hentschel. An enrichment of crispr and other defense-related features in marine sponge-associated microbial metagenomes. *Frontiers in Microbiology*, 7, 2016.

[69] Yongfei Hu, Xi Yang, Jing Li, Na Lv, Fei Liu, Jun Wu, Ivan Y. C. Lin, Na Wu, Bart C. Weimer, George F. Gao, Yulan Liu, and Baoli Zhu. The Bacterial Mobile Resistome Transfer Network Connecting the Animal and Human Microbiomes. *Applied and Environmental Microbiology*, 82(22):6672–6681, 2016.

[70] Shi Huang, Shuaiming Jiang, Dongxue Huo, Celeste Allaband, Mehrbod Estaki, Victor Cantu, Pedro Belda-Ferre, Yoshiki Vázquez-Baeza, Qiyun Zhu, Chenchen Ma, Congfa Li, Amir Zarrinpar, Yang-Yu Liu, Rob Knight, and Jiachao Zhang. Candidate probiotic lactiplantibacillus plantarum hnu082 rapidly and convergently evolves within human, mice, and zebrafish gut but differentially influences the resident microbiome. *Microbiome*, 9(1):151, 2021.

[71] Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, Lars J Jensen, and Others. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1):D309—-D314, 2019.

[72] Curtis Huttenhower, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H. Badger, Asif T. Chinwalla, Heather H. Creasy, Ashlee M. Earl, Michael G. FitzGerald, Robert S. Fulton, Michelle G. Giglio, Kymberlie Hallsworth-Pepin, Elizabeth A. Lobos, Ramana Madupu, Vincent Magrini, John C. Martin, Makedonka Mitreva, Donna M. Muzny, Erica J. Sodergren, James Versalovic, Aye M. Wollam, Kim C. Worley, Jennifer R. Wortman, Sarah K. Young, Qiandong Zeng, Kjersti M. Aagaard, Olukemi O. Abolude, Emma Allen-Vercoe, Eric J. Alm, Lucia Alvarado, Gary L. Andersen, Scott Anderson, Elizabeth Appelbaum, Harindra M. Arachchi, Gary Armitage, Cesar A. Arze, Tulin Ayvaz, Carl C. Baker, Lisa Begg, Tsegahiwot Belachew, Veena Bhonagiri, Monika Bihan, Martin J. Blaser, Toby Bloom, Vivien Bonazzi, J. Paul Brooks, Gregory A. Buck, Christian J. Buhay, Dana A. Busam, Joseph L. Campbell, Shane R. Canon, Brandi L. Cantarel, Patrick S. G. Chain, I-Min A. Chen, Lei Chen, Shaila Chhibba, Ken Chu, Dawn M. Ciulla, Jose C. Clemente, Sandra W. Clifton, Sean Conlan, Jonathan Crabtree, Mary A. Cutting, Noam J. Davidovics, Catherine C. Davis, Todd Z. DeSantis, Carolyn Deal, Kimberley D. Delehaunty, Floyd E. Dewhirst, Elena Deych, Yan Ding, David J. Dooling, Shannon P. Dugan, Wm Michael Dunne, A. Scott Durkin, Robert C. Edgar, Rachel L. Erlich, Candace N. Farmer, Ruth M. Farrell, Karoline Faust, Michael Feldgarden, Victor M. Felix, Sheila Fisher, Anthony A. Fodor, Larry J. Forney, Leslie Foster, Valentina Di Francesco, Jonathan Friedman, Dennis C. Friedrich, Catrina C. Fronick, Lucinda L. Fulton, Hongyu Gao, Nathalia Garcia, Georgia Giannoukos, Christina Giblin, Maria Y. Giovanni, Jonathan M. Goldberg, Johannes Goll, Antonio Gonzalez, Allison Griggs, Sharvari Gujja, Susan Kinder Haake, Brian J. Haas, Holli A. Hamilton, Emily L. Harris, Theresa A. Hepburn, Brandi Herter, Diane E. Hoffmann, Michael E. Holder, Clinton Howarth, Katherine H. Huang, Susan M. Huse, Jacques Izard, Janet K. Jansson, Huaiyang Jiang, Catherine Jordan, Vandita Joshi, James A. Katancik, Wendy A. Keitel, Scott T. Kelley, Cristyn Kells, Nicholas B. King, Dan Knights, Heidi H. Kong, Omry Koren, Sergey Koren, Karthik C. Kota, Christie L. Kovar, Nikos C. Kyrpides, Patricio S. La Rosa, Sandra L. Lee, Katherine P. Lemon, Niall Lennon, Cecil M. Lewis, Lora Lewis, Ruth E.

Ley, Kelvin Li, Konstantinos Liolios, Bo Liu, Yue Liu, Chien-Chi Lo, Catherine A. Lozupone, R. Dwayne Lunsford, Tessa Madden, Anup A. Mahurkar, Peter J. Mannon, Elaine R. Mardis, Victor M. Markowitz, Konstantinos Mavromatis, Jamison M. McCorrison, Daniel McDonald, Jean McEwen, Amy L. McGuire, Pamela McInnes, Teena Mehta, Kathie A. Mihindukulasuriya, Jason R. Miller, Patrick J. Minx, Irene Newsham, Chad Nusbaum, Michelle O'Laughlin, Joshua Orvis, Ioanna Pagani, Krishna Palaniappan, Shital M. Patel, Matthew Pearson, Jane Peterson, Mircea Podar, Craig Pohl, Katherine S. Pollard, Mihai Pop, Margaret E. Priest, Lita M. Proctor, Xiang Qin, Jeroen Raes, Jacques Ravel, Jeffrey G. Reid, Mina Rho, Rosamond Rhodes, Kevin P. Riehle, Maria C. Rivera, Beltran Rodriguez-Mueller, Yu-Hui Rogers, Matthew C. Ross, Carsten Russ, Ravi K. Sanka, Pamela Sankar, J. Fah Sathirapongsasuti, Jeffery A. Schloss, Patrick D. Schloss, Thomas M. Schmidt, Matthew Scholz, Lynn Schriml, Alyxandria M. Schubert, Nicola Segata, Julia A. Segre, William D. Shannon, Richard R. Sharp, Thomas J. Sharpton, Narmada Shenoy, Nihar U. Sheth, Gina A. Simone, Indresh Singh, Christopher S. Smillie, Jack D. Sobel, Daniel D. Sommer, Paul Spicer, Granger G. Sutton, Sean M. Sykes, Diana G. Tabbaa, Mathangi Thiagarajan, Chad M. Tomlinson, Manolito Torralba, Todd J. Treangen, Rebecca M. Truty, Tatiana A. Vishnivetskaya, Jason Walker, Lu Wang, Zhengyuan Wang, Doyle V. Ward, Wesley Warren, Mark A. Watson, Christopher Wellington, Kris A. Wetterstrand, James R. White, Katarzyna Wilczek-Boney, YuanQing Wu, Kristine M. Wylie, Todd Wylie, Chandri Yandava, Liang Ye, Yuzhen Ye, Shibu Yooseph, Bonnie P. Youmans, Lan Zhang, Yanjiao Zhou, Yiming Zhu, Laurie Zoloth, Jeremy D. Zucker, Bruce W. Birren, Richard A. Gibbs, Sarah K. Highlander, Barbara A. Methé, Karen E. Nelson, Joseph F. Petrosino, George M. Weinstock, Richard K. Wilson, Owen White, and The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.

[73] Doug Hyatt, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119, 2010.

[74] Illumina. bcl2fastq conversion software. Accessed: 2022-06-26.

[75] Illumina. Effects of index misassignment on multiplexing and downstream analysis. Technical Report 770-2017-004-D, 2018.

[76] Francesco Imperi, Luísa C.S. Antunes, Jochen Blom, Laura Villa, Michele Iacono, Paolo Visca, and Alessandra Carattoli. The genomics of Acinetobacter baumannii: Insights into genome plasticity, antimicrobial resistance and pathogenicity. *IUBMB Life*, 63(12):1068–1074, 2011.

[77] Ivaylo I Ivanov and Dan R Littman. Modulation of immune homeostasis by commensal bacteria. *Current opinion in microbiology*, 14(1):106–114, 02 2011.

[78] Saharuetai Jeamsripong, Xunde Li, Sharif Aly, Zhengchang Su, Richard Pereira, and Edward Atwill. Antibiotic resistance genes and associated phenotypes in escherichia coli and enterococcus from cattle at different production stages on a dairy farm in central california. *Antibiotics*, 10:1042, 08 2021.

[79] Xiaoqing Jiang, Xin Li, Longshu Yang, Chunhong Liu, Qi Wang, Weilai Chi, and Huaiqiu Zhu. How microbes shape their communities? a microbial community model based on functional genes. *Genomics, Proteomics, and Bioinformatics*, 17(1):91–105, 2019. Microbiome and Health.

[80] Bryony A Jones, Delia Grace, Richard Kock, Silvia Alonso, Jonathan Rushton, Mohammed Y Said, Declan McKeever, Florence Mutua, Jarrah Young, John McDermott, and Dirk Udo Pfeiffer. Zoonosis emergence linked to agricultural intensification and environmental change. *Proceedings of the National Academy of Sciences of the United States of America*, 110(21):8399–8404, 2013.

[81] Juan Jovel, Jordan Patterson, Weiwei Wang, Naomi Hotte, Sandra O'Keefe, Troy Mitchel, Troy Perry, Dina Kao, Andrew L Mason, Karen L Madsen, and Gane K S Wong. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in Microbiology*, 7(APR):1–17, 2016.

[82] Nobuhiko Kamada, Sang-Uk Seo, Grace Y. Chen, and Gabriel Núñez. Role of the gut microbiota in immunity and inflammatory disease. *Nature Reviews Immunology*, 13(5):321–335, 2013.

[83] Dongwan D. Kang, Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7(July):e7359, 2019.

[84] Ashley E. Kates, Mark Dalman, James C. Torner, and Tara C. Smith. The nasal and oropharyngeal microbiomes of healthy livestock workers. *PLOS ONE*, 14(3):1–19, 03 2019.

[85] S. Keely, N J Talley, and P M Hansbro. Pulmonary-intestinal cross-talk in mucosal inflammatory disease. *Mucosal Immunology*, 5(1):7–18, 2012.

[86] Rabia Khan, Fernanda Cristina Petersen, and Sudhanshu Shekhar. Commensal bacteria: An emerging player in defense against respiratory pathogens. *Frontiers in Immunology*, 10, 2019.

[87] Daehwan Kim, Li Song, Florian P Breitwieser, and Steven L Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome research*, 26(12):1721–1729, 12 2016.

[88] Heesoo Kim, Mincheol Kim, Sanghee Kim, Yung Mi Lee, and Seung Chul Shin. Characterization of antimicrobial resistance genes and virulence factor genes in an arctic permafrost region revealed by metagenomics. *Environmental Pollution*, 294:118634, 2022.

[89] A V Klieve, M. N. O'Leary, Lyle McMillen, and Diane Ouwerkerk. Ruminococcus bromii, identification and isolation as a dominant community member in the rumen of cattle fed a barley diet. *Journal of Applied Microbiology*, 103, 2007.

[90] Gijs Klous, Anke Huss, Dick J.J. Heederik, and Roel A. Coutinho. Human–livestock contacts and their relationship to transmission of zoonotic pathogens, a systematic review of literature. *One Health*, 2:65–76, 2016.

[91] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 08 2012.

[92] Julia G. Kraemer, Alban Ramette, Suzanne Aebi, Anne Oppliger, Markus Hilty, and Johanna Björkroth. Influence of pig farming on the human nasal microbiota: Key role of airborne microbial communities. *Applied and Environmental Microbiology*, 84(6):e02470–17, 2018.

[93] Manorama Kumari, Parul Singh, Basavaprabhu H. Nataraj, Anusha Kokkiligadda, Harshita Naithani, Syed Azmal Ali, Pradip. V. Behare, and Ravinder Nagpal. Fostering next-generation probiotics in human gut by targeted dietary modulation: An emerging perspective. *Food Research International*, 150:110716, 2021.

[94] Timothy F Landers, Bevin Cohen, Thomas E Wittum, and Elaine L Larson. A review of antibiotic use in food animals: perspective, policy, and potential. *Public health reports (Washington, D.C. : 1974)*, 127(1):4–22, Jan-Feb 2012.

[95] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[96] Ben Langmead, Christopher Wilks, Valentin Antonescu, and Rone Charles. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*, 35(3):421–432, 07 2018.

[97] Anton J.M. Larsson, Geoff Stanley, Rahul Sinha, Irving L. Weissman, and Rickard Sandberg. Computational correction of index switching in multiplexed sequencing libraries. *Nature Methods*, 15(5):305–307, 2018.

[98] Emmanuelle Le Chatelier, Trine Nielsen, Junjie Qin, Edi Prifti, Falk Hildebrand, Gwen Falony, Mathieu Almeida, Manimozhiyan Arumugam, Jean-Michel Batto, Sean Kennedy, Pierre Leonard, Junhua Li, Kristoffer Burgdorf, Niels Grarup, Torben Jørgensen, Ivan Brandslund, Henrik Bjørn Nielsen, Agnieszka S. Juncker, Marcelo Bertalan, Florence Levenez, Nicolas Pons, Simon Rasmussen, Shinichi Sunagawa, Julien Tap, Sebastian Tims, Erwin G. Zoetendal, Søren Brunak, Karine Clément, Joël Doré, Michiel Kleerebezem, Karsten Kristiansen, Pierre Renault, Thomas Sicheritz-Ponten, Willem M. de Vos, Jean-Daniel Zucker, Jeroen Raes, Torben Hansen, Eric Guedon, Christine Delorme, Séverine Layec, Ghalia Khaci, Maarten van de Guchte, Gaetana Vandemeulebrouck, Alexandre Jamet, Rozenn Dervyn, Nicolas Sanchez, Emmanuelle Maguin, Florence Haimet, Yohanan Winogradski, Antonella Cultrone, Marion Leclerc, Catherine Juste, Hervé Blottière, Eric Pelletier, Denis LePaslier, François Artiguenave, Thomas Bruls, Jean Weissenbach, Keith Turner, Julian Parkhill, Maria Antolin, Chaysavanh Manichanh, Francesc Casellas, Natalia Boruel, Encarna Varela, Antonio Torrejon, Francisco Guarner, Gérard Denariaz, Muriel Derrien, Johan E. T. van Hylckama Vlieg, Patrick Veiga, Raish Oozeer, Jan Knol, Maria Rescigno, Christian Brechot, Christine M'Rini, Alexandre Mérieux, Takuji Yamada, Peer Bork, Jun Wang, S. Dusko Ehrlich, Oluf Pedersen, and MetaHIT consortium. Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500(7464):541–546, 2013.

[99] John A. Lees, Minna Vehkala, Niko Välimäki, Simon R. Harris, Claire Chewapreecha, Nicholas J. Croucher, Pekka Marttinen, Mark R. Davies, Andrew C. Steer, Steven Y.C. Tong, Antti Honkela, Julian Parkhill, Stephen D. Bentley, and Jukka Corander. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nature Communications*, 7, 2016.

[100] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10):1674–1676, 01 2015.

[101] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 06 2009.

[102] Wenjun Li, Kathleen R O'Neill, Daniel H Haft, Michael DiCuccio, Vyacheslav Chetvernin, Azat Badretdin, George Coulouris, Farideh Chitsaz, Myra K Derbyshire, A Scott Durkin, Noreen R Gonzales, Marc Gwadz, Christopher J Lanczycki, James S

Song, Narmada Thanki, Jiyao Wang, Roxanne A Yamashita, Mingzhang Yang, Chanjuan Zheng, Aron Marchler-Bauer, and Françoise Thibaud-Nissen. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Research*, 49(D1):D1020–D1028, 12 2020.

[103] Hong Liu, Maozhen Han, Shuai Cheng Li, Guangming Tan, Shiwei Sun, Zhiqiang Hu, Pengshuo Yang, Rui Wang, Yawen Liu, Feng Chen, Jianjun Peng, Hong Peng, Hongxing Song, Yang Xia, Liqun Chu, Quan Zhou, Feng Guan, Jing Wu, Dongbo Bu, and Kang Ning. Resilience of human gut microbial communities for the long stay with multiple dietary shifts. *Gut*, 68(12):2254–2255, 2019.

[104] Mireia Lopez-Siles, Sylvia H Duncan, L Jesús Garcia-Gil, and Margarita Martinez-Medina. Faecalibacterium prausnitzii: from microbiology to diagnostics and prognostics. *The ISME Journal*, 11(4):841–852, 2017.

[105] Daniela Machado, Joana Cristina Barbosa, Melany Domingos, Diana Almeida, José Carlos Andrade, Ana Cristina Freitas, and Ana Maria Gomes. Revealing antimicrobial resistance profile of the novel probiotic candidate faecalibacterium prausnitzii dsm 17677. *International Journal of Food Microbiology*, 363:109501, 2022.

[106] Lisa Maier, Mihaela Pruteanu, Michael Kuhn, Georg Zeller, Anja Telzerow, Exene Erin Anderson, Ana Rita Brochado, Keith Conrad Fernandez, Hitomi Dose, Hirotada Mori, Kiran Raosaheb Patil, Peer Bork, and Athanasios Typas. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature*, 555(7698):623–628, 2018.

[107] Chaysavanh Manichanh, Natalia Borruel, Francesc Casellas, and Francisco Guarner. The gut microbiota in ibd. *Nature Reviews Gastroenterology & Hepatology*, 9(10):599–608, 2012.

[108] Christy Manyi-Loh, Sampson Mamphweli, Edson Meyer, and Anthony Okoh. Antibiotic use in agriculture and its consequential resistance in environmental sources: Potential public health implications. *Molecules*, 23(4), 2018.

[109] Bryan D Martin, Daniela Witten, and Amy D Willis. Modeling microbial abundances and dysbiosis with beta-binomial regression. *The annals of applied statistics*, 14(1):94–115, 03 2020.

[110] Andrew G. McArthur, Nicholas Waglechner, Fazmin Nizam, Austin Yan, Marisa A. Azad, Alison J. Baylay, Kirandeep Bhullar, Marc J. Canova, Gianfranco De Pascale, Linda Ejim, Lindsay Kalan, Andrew M. King, Kalinka Koteva, Mariya Morar, Michael R. Mulvey, Jonathan S. O'Brien, Andrew C. Pawlowski, Laura J. V. Piddock, Peter Spanogiannopoulos, Arlene D. Sutherland, Irene Tang, Patricia L. Taylor, Maulik

Thaker, Wenliang Wang, Marie Yan, Tennison Yu, and Gerard D. Wright. The Comprehensive Antibiotic Resistance Database. *Antimicrobial Agents and Chemotherapy*, 57(7):3348–3357, 2013.

[111] Clinton J. McDaniel, Diana M. Cardwell, Robert B. Moeller, and Gregory C. Gray. Humans and cattle: A review of bovine zoonoses. *Vector-Borne and Zoonotic Diseases*, 14(1):1–19, 2014.

[112] AndréE Minoche, Juliane C Dohm, and Heinz Himmelbauer. Evaluation of genomic high-throughput sequencing data generated on illumina hiseq and genome analyzer systems. *Genome biology*, 12(11):R112–R112, 11 2011.

[113] Samuel S. Minot, Kevin C. Barry, Caroline Kasman, Jonathan L. Golob, and Amy D. Willis. Geneshot: Gene-Level Metagenomics Identifies Genome Islands Associated With Immunotherapy Response. *Genome Biology*, 22(1):1–10, 2021.

[114] Samuel S. Minot and Amy D. Willis. Clustering co-abundant genes identifies components of the gut microbiome that are reproducibly associated with colorectal cancer and inflammatory bowel disease. *Microbiome*, 7(1):1–10, 2019.

[115] Samuel S. Minot and Amy D. Willis. Clustering co-abundant genes identifies components of the gut microbiome that are reproducibly associated with colorectal cancer and inflammatory bowel disease. *Microbiome*, 7(1):110, 2019.

[116] S Miquel, R Martín, O Rossi, LG Bermúdez-Humarán, JM Chatel, H Sokol, M Thomas, JM Wells, and P Langella. Faecalibacterium prausnitzii and human intestinal health. *Current Opinion in Microbiology*, 16(3):255–261, 2013. Ecology and industrial microbiology * Special Section: Innate immunity.

[117] Julia Moor, Tsering Wüthrich, Suzanne Aebi, Nadezda Mostacci, Gudrun Overesch, Anne Oppliger, and Markus Hilty. Influence of pig farming on human gut microbiota: role of airborne microbial communities. *Gut Microbes*, 13(1):1927634, 2021. PMID: 34060426.

[118] Alexis Mosca, Marion Leclerc, and Jean P Hugot. Gut microbiota diversity and human diseases: Should we reintroduce key predators in our ecosystem? *Frontiers in microbiology*, 7:455–455, 03 2016.

[119] Arghya Mukherjee, Cathy Lordan, R. Paul Ross, and Paul D. Cotter. Gut microbes from the phylogenetically diverse genus eubacterium and their various contributions to gut health. *Gut Microbes*, 12(1):1802866, 2020. PMID: 32835590.

[120] Indrani Mukhopadhya, Sarah Moraïs, Jenny Laverde-Gomez, Paul O Sheridan, Alan W Walker, William Kelly, Athol V Klieve, Diane Ouwerkerk, Sylvia H Duncan, Petra Louis, Nicole Koropatkin, Darrell Cockburn, Ryan Kibler, Philip J Cooper, Carlos Sandoval, Emmanuelle Crost, Nathalie Juge, Edward A Bayer, and Harry J Flint. Sporulation capability and amylosome conservation among diverse human colonic and rumen isolates of the keystone starch-degrader ruminococcus bromii. *Environmental microbiology*, 20(1):324–336, 01 2018.

[121] Yoji Nakamura. Prediction of horizontally and widely transferred genes in prokaryotes. *Evolutionary Bioinformatics*, 14:1176934318810785, 2018. PMID: 30546254.

[122] Y. Nakanishi, T. Sato, and T. Ohteki. Commensal gram-positive bacteria initiates colitis by inducing monocyte/macrophage mobilization. *Mucosal Immunology*, 8(1):152–160, 2015.

[123] Henrik Nielsen, Mathieu Almeida, Agnieszka Juncker, Simon Rasmussen, Junhua Li, Shinichi Sunagawa, Damian Plichta, Laurent Gautier, Anders Pedersen, Emmanuelle Le Chatelier, Eric Pelletier, Ida Bonde, Trine Nielsen, Chaysavanh Manichanh, Manimozhiyan Arumugam, Jean-Michel Batto, Marcelo Santos, Nikolaj Blom, Natalia Borruel, and Hervé Blottiere. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology*, 32:822 – 828, 08 2014.

[124] NIH Human Microbiome Portfolio Analysis Team. A review of 10 years of human microbiome research activities at the US National Institutes of Health, Fiscal Years 2007-2016. *Microbiome*, 7(1):31, 2019.

[125] Chao Niu, Dong Yu, Yuelan Wang, Hongguang Ren, Yuan Jin, Wei Zhou, Beiping Li, Yiyong Cheng, Junjie Yue, Zhixian Gao, and Long Liang. Common and pathogen-specific virulence factors are different in function and structure. *Virulence*, 4(6):473–482, 2013. PMID: 23863604.

[126] Matthew W Nonnenmann, David Gimeno Ruiz de Porras, Jeffrey Levin, David Douphrate, Vijay Boggaram, Joshua Schaffer, Michael Gallagher, Madeleine Hornick, and Stephen Reynolds. Pulmonary function and airway inflammation among dairy parlor workers after exposure to inhalable aerosols. *American journal of industrial medicine*, 60(3):255–263, 03 2017.

[127] Toru Ogata, Yo-Han Kim, Eiji Iwamoto, Tatsunori Masaki, Kentaro Ikuta, and Shigeru Sato. Comparison of ph and bacterial communities in the rumen and reticulum during fattening of japanese black beef cattle. *Animal science journal = Nihon chikusan Gakkaiho*, 91 1:e13487, 2020.

[128] Evgenii I. Olekhnovich, Artem T. Vasilyev, Vladimir I. Ulyantsev, Elena S. Kostryukova, and Alexander V. Tyakht. MetaCherchant: Analyzing genomic context of antibiotic resistance genes in gut microbiota. *Bioinformatics*, 34(3):434–444, 2018.

[129] Stephen P. Oliver, Shelton E. Murinda, and Bhushan. M. Jayarao. Impact of Antibiotic Use in Adult Dairy Cows on Antimicrobial Resistance of Veterinary and Human Pathogens: A Comprehensive Review. *Foodborne Pathogens and Disease*, 8(3):337–355, 2010.

[130] PAWIN PADUNGTOD and JOHN B. KANEENE. Campylobacter in Food Animals and Humans in Northern Thailand. *Journal of Food Protection*, 68(12):2519–2526, 12 2005.

[131] Mark J. Pallen and Brendan W. Wren. Bacterial pathogenomics. *Nature*, 449(7164):835–842, 2007.

[132] Mark J. Pallen and Brendan W. Wren. Bacterial pathogenomics. *Nature*, 449(7164):835–842, 2007.

[133] Kelli Palmer, Veronica Kos, and Michael Gilmore. Horizontal Gene Transfer and the Genomics of Enterococcal Antibiotic Resistance. 185(2):974–981, 2010.

[134] Yu Pan, Jiaxiong Zeng, Liguan Li, Jintao Yang, Ziyun Tang, Wenguang Xiong, Yafei Li, Sheng Chen, Zhenling Zeng, and Jack A. Gilbert. Coexistence of antibiotic resistance genes and virulence factors deciphered by large-scale complete genome analysis. *mSystems*, 5(3):e00821–19, 2020.

[135] Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, 50(D1):D785–D794, 09 2021.

[136] Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, 36(10):996–1004, 2018.

[137] Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7):1043–1055, 2015.

[138] Donovan H. Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz, and Gene W. Tyson. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, 2(11):1533–1542, 2017.

[139] Edoardo Pasolli, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, Paolo Manghi, Adrian Tett, Paolo Ghensi, Maria Carmen Collado, Benjamin L Rice, Casey DuLong, Xochitl C Morgan, Christopher D Golden, Christopher Quince, Curtis Huttenhower, and Nicola Segata. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*, 176(3):649—-662.e20, 2019.

[140] Alessandra Pino, Emanuela Bartolo, Cinzia Caggia, Antonio Cianci, and Cinzia L. Randazzo. Detection of vaginal lactobacilli as probiotic candidates. *Scientific Reports*, 9(1):3355, 2019.

[141] Robert A. Power, Julian Parkhill, and Tulio De Oliveira. Microbial genome-wide association studies: lessons from human GWAS. *Nature Reviews Genetics*, 18(1):41–50, 2016.

[142] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. Fasttree 2 – approximately maximum-likelihood trees for large alignments. *PLOS ONE*, 5(3):1–10, 03 2010.

[143] João T. Proença, Duarte C. Barral, and Isabel Gordo. Commensal-to-pathogen transition: One-single transposon insertion results in two pathoadaptive traits in escherichia coli -macrophage interaction. *Scientific Reports*, 7(1):4504, 2017.

[144] Youwen Qin, Aki S. Havulinna, Yang Liu, Pekka Jousilahti, Scott C. Ritchie, Alex Tokolyi, Jon G. Sanders, Liisa Valsta, Marta Brożyńska, Qiyun Zhu, Anupriya Tripathi, Yoshiki Vázquez-Baeza, Rohit Loomba, Susan Cheng, Mohit Jain, Teemu Niiranen, Leo Lahti, Rob Knight, Veikko Salomaa, Michael Inouye, and Guillaume Méric. Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort. *Nature Genetics*, 54(2):134–142, 2022.

[145] Xinyun Qiu, Xiaojing Zhao, Xiufang Cui, Xiaqiong Mao, Nana Tang, Chunhua Jiao, Di Wang, Yue Zhang, Ziping Ye, and Hongjie Zhang. Characterization of fungal and bacterial dysbiosis in young adult chinese patients with crohn's disease. *Therapeutic Advances in Gastroenterology*, 13:1756284820971202, 2020. PMID: 33240394.

[146] Christopher Quince, Alan W Walker, Jared T Simpson, Nicholas J Loman, and Nicola Segata. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, 35(9):833–844, 2017.

[147] April L Raftery, Evelyn Tsantikos, Nicola L Harris, and Margaret L Hibbs. Links between inflammatory bowel disease and chronic obstructive pulmonary disease. *Frontiers in immunology*, 11:2144–2144, 09 2020.

[148] Aathmaja Anandhi Rangarajan, Hannah E. Chia, Christopher A. Azaldegui, Monica H. Olszewski, Gabriel V. Pereira, Nicole M. Koropatkin, and Julie S. Biteen. Ruminococcus bromii enables the growth of proximal bacteroides thetaiotaomicron by releasing glucose during starch degradation. *Microbiology*, 168(4), 2022.

[149] Leah Reshef, Amir Kovacs, Amos Ofer, Lior Yahav, Nitsan Maharshak, Nirit Keren, Fred M. Konikoff, Hagit Tulchinsky, Uri Gophna, and Iris Dotan. Pouch inflammation is associated with a decrease in specific bacterial taxa. *Gastroenterology*, 149(3):718–727, 2015.

[150] Ardeshir Rineh, Michael J Kelso, Fatma Vatansever, George P Tegos, and Michael R Hamblin. Clostridium difficile infection: molecular pathogenesis and novel therapeutics. *Expert review of anti-infective therapy*, 12(1):131–150, 01 2014.

[151] Jerónimo Rodríguez-Beltrán, Javier DelaFuente, Ricardo León-Sampedro, R. Craig MacLean, and Álvaro San Millán. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nature Reviews Microbiology*, 19(6):347–359, 2021.

[152] Daphna Rothschild, Omer Weissbrod, Elad Barkan, Alexander Kurilshikov, Tal Korem, David Zeevi, Paul I. Costea, Anastasia Godneva, Iris N. Kalka, Noam Bar, Smadar Shilo, Dar Lador, Arnau Vich Vila, Niv Zmora, Meirav Pevsner-Fischer, David Israeli, Noa Kosower, Gal Malka, Bat Chen Wolf, Tali Avnit-Sagi, Maya Lotan-Pompan, Adina Weinberger, Zamir Halpern, Shai Carmi, Jingyuan Fu, Cisca Wijmenga, Alexandra Zhernakova, Eran Elinav, and Eran Segal. Environment dominates over host genetics in shaping human gut microbiota. *Nature*, 555(7695):210–215, 2018.

[153] L. Rouli, V. Merhej, P. E. Fournier, and D. Raoult. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections*, 7:72–85, 2015.

[154] Taylor M. Royalty and Andrew D. Steen. Theoretical and Simulation-Based Investigation of the Relationship between Sequencing Effort, Microbial Community Richness, and Diversity in Binning Metagenome-Assembled Genomes. *mSystems*, 4(5), 2019.

[155] James Emmanuel San, Shakuntala Baichoo, Aquillah Kanzi, Yumna Moosa, Richard Lessells, Vagner Fonseca, John Mogaka, Robert Power, and Tulio de Oliveira. Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls. *Frontiers in Microbiology*, 10(January), 2020.

[156] Keith W. Savin, Jody Zawadzki, Martin J. Auldist, Jianghui Wang, Doris Ram, Simone Rochfort, and Benjamin G. Cocks. Faecalibacterium diversity in dairy cow milk. *PLOS ONE*, 14(8):1–17, 08 2019.

[157] Matthias Scholz, Doyle V. Ward, Edoardo Pasolli, Thomas Tolio, Moreno Zolfo, Francesco Asnicar, Duy Tin Truong, Adrian Tett, Ardythe L. Morrow, and Nicola Segata. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nature Methods*, 13(5):435–438, 2016.

[158] H. Morgan Scott, Gary Acuff, Gilles Bergeron, Megan W. Bourassa, Jason Gill, David W. Graham, Laura H. Kahn, Paul S. Morley, Matthew Jude Salois, Shabbir Simjee, Randall S. Singer, Tara C. Smith, Carina Storrs, and Thomas E. Wittum. Critically important antibiotics: criteria and approaches for measuring and reducing their use in food animal agriculture. *Annals of the New York Academy of Sciences*, 1441(1):8–16, 2019.

[159] H.M. Scott, L.D. Campbell, R.B. Harvey, K.M. Bischoff, W.Q. Alali, K.S. Barling, and R.C. Anderson. Patterns of antimicrobial resistance among commensal escherichia coli isolated from integrated multi-site housing and worker cohorts of humans and swine. *Foodborne Pathogens and Disease*, 2(1):24–37, 2005. PMID: 15992296.

[160] Torsten Seemann. Abricate: mass screening of contigs for antiobiotic resistance genes. 2016.

[161] Nicola Segata, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson, and Curtis Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9(8):811–814, 2012.

[162] Corinne E. Sexton, Hayden Z. Smith, Peter D. Newell, Angela E. Douglas, and John M. Chaston. MAGNAMWAR: An R package for genome-wide association studies of bacterial orthologs. *Bioinformatics*, 34(11):1951–1952, 2018.

[163] Alon Shaiber, Amy D Willis, Tom O Delmont, Simon Roux, Lin Xing Chen, Abigail C Schmid, Mahmoud Yousef, Andrea R Watson, Karen Lolans, Özcan C Esen, Sonny T M Lee, Nora Downey, Hilary G Morrison, Floyd E Dewhirst, Jessica L Mark Welch, and A Murat Eren. Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biology*, page 21:292, 2020.

[164] Thomas J. Sharpton. An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, 5, 2014.

[165] Rachel M. Sherman and Steven L. Salzberg. Pan-genomics in the human genome era. *Nature Reviews Genetics*, 21(4):243–254, 2020.

[166] David Sims, Ian Sudbery, Nicholas E. Ilott, Andreas Heger, and Chris P. Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121–132, 2014.

[167] Harry Sokol, Bénédicte Pigneur, Laurie Watterlot, Omar Lakhdari, Luis G. Bermúdez-Humarán, Jean-Jacques Gratadoux, Sébastien Blugeon, Chantal Bridonneau, Jean-Pierre Furet, Gérard Corthier, Corinne Grangette, Nadia Vasquez, Philippe Pochart, Germain Trugnan, Ginette Thomas, Hervé M. Blottière, Joël Doré, Philippe Marteau, Philippe Seksik, and Philippe Langella. ¡i¿faecalibacterium prausnitzii¡/i¿ is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of crohn disease patients. *Proceedings of the National Academy of Sciences*, 105(43):16731–16736, 2008.

[168] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017.

[169] Saso Stoleski, Jordan Minov, Jovanka Karadzinska-Bislimovska, Dragan Mijakoski, Aneta Atanasovska, and Dragana Bislimovska. Asthma and chronic obstructive pulmonary disease associated with occupational exposure in dairy farmers - importance of job exposure matrices. *Open access Macedonian journal of medical sciences*, 7(14):2350–2359, 07 2019.

[170] John D. Storey, Andrew J. Bass, Alan Dabney, and David Robinson. *qvalue: Q-value estimation for false discovery rate control*, 2021. R package version 2.26.0.

[171] Jian Sun, Xiao-Ping Liao, Alaric W. D'Souza, Manish Boolchandani, Sheng-Hui Li, Ke Cheng, José Luis Martínez, Liang Li, You-Jun Feng, Liang-Xing Fang, Ting Huang, Jing Xia, Yang Yu, Yu-Feng Zhou, Yong-Xue Sun, Xian-Bo Deng, Zhen-Ling Zeng, Hong-Xia Jiang, Bing-Hu Fang, You-Zhi Tang, Xin-Lei Lian, Rong-Min Zhang, Zhi-Wei Fang, Qiu-Long Yan, Gautam Dantas, and Ya-Hong Liu. Environmental remodeling of human gut microbiota and antibiotic resistome in livestock farms. *Nature Communications*, 11(1):1427, 2020.

[172] Yoshihiko Suzuki, Suguru Nishijima, Yoshikazu Furuta, Jun Yoshimura, Wataru Suda, Kenshiro Oshima, Masahira Hattori, and Shinichi Morishita. Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut. *Microbiome*, 7, 08 2019.

[173] Toshihiko Takada, Takashi Kurakawa, Hirokazu Tsuji, and Koji Nomoto. Fusicatenibacter saccharivorans gen. nov., sp. nov., isolated from human faeces. *International Journal of Systematic and Evolutionary Microbiology*, 63(10):3691–3696, 2013.

[174] Kozue Takeshita, Shinta Mizuno, Yohei Mikami, Tomohisa Sujino, Keiichiro Saigusa, Katsuyoshi Matsuoka, Makoto Naganuma, Tadashi Sato, Toshihiko Takada, Hirokazu Tsuji, Akira Kushiro, Koji Nomoto, and Takanori Kanai. A Single Species of Clostridium Subcluster XIVa Decreased in Ulcerative Colitis Patients. *Inflammatory Bowel Diseases*, 22(12):2802–2810, 11 2016.

[175] Roman L. Tatusov, Natalie D. Fedorova, John D. Jackson, Aviva R. Jacobs, Boris Kiryutin, Eugene V. Koonin, Dmitri M. Krylov, Raja Mazumder, Sergei L. Mekhedov, Anastasia N. Nikolskaya, B. Sridhar Rao, Sergei Smirnov, Alexander V. Sverdlov, Sona Vasudevan, Yuri I. Wolf, Jodie J. Yin, and Darren A. Natale. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4(1):41, 2003.

[176] Ethan A Taylor, Ellen R Jordan, Jose A Garcia, Gerrit R Hagevoort, Keri N Norman, Sara D Lawhon, Juan M Piñeiro, and Harvey M Scott. Effects of two-dose ceftiofur treatment for metritis on the temporal dynamics of antimicrobial resistance among fecal escherichia coli in holstein-friesian dairy cows. *PloS one*, 14(7):e0220068–e0220068, 07 2019.

[177] Adrian Tett, Kun D. Huang, Francesco Asnicar, Hannah Fehlner-Peach, Edoardo Pasolli, Nicolai Karcher, Federica Armanini, Paolo Manghi, Kevin Bonham, Moreno Zolfo, Francesca De Filippis, Cara Magnabosco, Richard Bonneau, John Lusingu, John Amuasi, Karl Reinhard, Thomas Rattei, Fredrik Boulund, Lars Engstrand, Albert Zink, Maria Carmen Collado, Dan R. Littman, Daniel Eibach, Danilo Ercolini, Omar Rota-Stabelli, Curtis Huttenhower, Frank Maixner, and Nicola Segata. The prevotella copri complex comprises four distinct clades underrepresented in westernized populations. *Cell Host & Microbe*, 26(5):666–679.e7, 2019.

[178] Hervé Tettelin, Vega Masignani, Michael J. Cieslewicz, Claudio Donati, Duccio Medini, Naomi L. Ward, Samuel V. Angiuoli, Jonathan Crabtree, Amanda L. Jones, A. Scott Durkin, Robert T. DeBoy, Tanja M. Davidsen, Marirosa Mora, Maria Scarselli, Immaculada Margarit y Ros, Jeremy D. Peterson, Christopher R. Hauser, Jaideep P. Sundaram, William C. Nelson, Ramana Madupu, Lauren M. Brinkac, Robert J. Dodson, Mary J. Rosovitz, Steven A. Sullivan, Sean C. Daugherty, Daniel H. Haft, Jeremy Selengut, Michelle L. Gwinn, Liwei Zhou, Nikhat Zafar, Hoda Khouri, Diana Radune, George Dimitrov, Kisha Watkins, Kevin J. B. O'Connor, Shannon Smith, Teresa R. Utterback, Owen White, Craig E. Rubens, Guido Grandi, Lawrence C. Madoff, Dennis L. Kasper, John L. Telford, Michael R. Wessels, Rino Rappuoli, and Claire M. Fraser. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae:

Implications for the microbial ”pan-genome”. *Proceedings of the National Academy of Sciences*, 102(39):13950–13955, 2005.

[179] Cuihong Tong, Danyu Xiao, Longfei Xie, Jintao Yang, Ruonan Zhao, Jie Hao, Zhipeng Huo, Zhenling Zeng, and Wenguang Xiong. Swine manure facilitates the spread of antibiotic resistome including tigecycline-resistant tet(x) variants to farm workers and receiving environment. *Science of The Total Environment*, 808:152157, 2022.

[180] Pauline Trinh, David S. Clausen, and Amy D. Willis. happi: a hierarchical approach to pangenomics inference. *bioRxiv e-prints*, April 2022.

[181] Özgün C. O. Umu, Jeremy A. Frank, Jonatan U. Fangel, Marije Oostindjer, Carol Souza da Silva, Elizabeth J. Bolhuis, Guido Bosch, William G. T. Willats, Phillip B. Pope, and Dzung B. Diep. Resistant starch diet induces change in the swine microbiome and a predominance of beneficial bacterial populations. *Microbiome*, 3(1):16, 2015.

[182] Gherman V. Uritskiy, Jocelyne DiRuggiero, and James Taylor. MetaWRAP - A flexible pipeline for genome-resolved metagenomic data analysis. *bioRxiv*, pages 1–13, 2018.

[183] Ana M Valdes, Jens Walter, Eran Segal, and Tim D Spector. Role of the gut microbiota in nutrition and health. *BMJ*, 361, 2018.

[184] Liese Van Gompel, Roosmarijn E.C. Luiken, Rasmus B. Hansen, Patrick Munk, Martijn Bouwknegt, Lourens Heres, Gerdit D. Greve, Peter Scherpenisse, Betty G.M. Jongerius-Gortemaker, Monique H.G. Tersteeg-Zijderveld, Silvia García-Cobos, Wietske Dohmen, Alejandro Dorado-García, Jaap A. Wagenaar, Bert A.P. Urlings, Frank M. Aarestrup, Dik J. Mevius, Dick J.J. Heederik, Heike Schmitt, Alex Bossers, and Lidwien A.M. Smit. Description and determinants of the faecal resistome and microbiome of farmers and slaughterhouse workers: A metagenome-wide cross-sectional study. *Environment International*, 143:105939, 2020.

[185] Thea Van Rossum, Pamela Ferretti, Oleksandr M. Maistrenko, and Peer Bork. Diversity within species: interpreting strains in microbiomes. *Nature Reviews Microbiology*, 18(9):491–506, 2020.

[186] Doris Vandeputte, Lindsey De Commer, Raul Y. Tito, Gunter Kathagen, João Sabino, Séverine Vermeire, Karoline Faust, and Jeroen Raes. Temporal variability in quantitative human gut microbiome profiles and implications for clinical research. *Nature Communications*, 12(1):6740, 2021.

[187] Anne K. Vidaver. Uses of Antimicrobials in Plant Agriculture. *Clinical Infectious Diseases*, 34(Supplement 3):S107–S110, 06 2002.

[188] Hui Wang, Jing-Shi Liu, Shao-Hua Peng, Xi-Yun Deng, De-Mao Zhu, Sara Javidiparsijani, Gui-Rong Wang, Dai-Qiang Li, Long-Xuan Li, Yi-Chun Wang, and Jun-Ming Luo. Gut-lung crosstalk in pulmonary involvement with inflammatory bowel diseases. *World journal of gastroenterology*, 19(40):6794–6804, 10 2013.

[189] Wenjie Wang and Jun Yan. *splines2: Regression Spline Functions and Classes*, 2021. R package version 0.4.5.

[190] Yanan Wang, Na Lyu, Fei Liu, William J. Liu, Yuhai Bi, Zewu Zhang, Sufang Ma, Jian Cao, Xiaofeng Song, Aiping Wang, Gaiping Zhang, Yongfei Hu, Baoli Zhu, and George Fu Gao. More diversified antibiotic resistance genes in chickens and workers of the live poultry markets. *Environment International*, 153:106534, 2021.

[191] Hannah M Wexler. Bacteroides: the good, the bad, and the nitty-gritty. *Clinical microbiology reviews*, 20(4):593–621, 10 2007.

[192] Ryan R. Wick, Mark B. Schultz, Justin Zobel, and Kathryn E. Holt. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, 31(20):3350–3352, 06 2015.

[193] Amy Willis and John Bunge. Estimating diversity via frequency ratios: Estimating diversity via ratios. *Biometrics*, 71, 06 2015.

[194] Amy Willis, John Bunge, and Thea Whitman. Improved detection of changes in species richness in high diversity microbial communities. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 66(5):963–977, 2022/06/26/ 2017.

[195] Gabriela Wlasiuk and Donata Vercelli. The farm effect, or: when, what and how a farming environment protects from asthma and allergic disease. *Current Opinion in Allergy and Clinical Immunology*, 12(5), 2012.

[196] Derrick E. Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with kraken 2. *Genome Biology*, 20(1):257, 2019.

[197] S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36, 2011.

[198] Y Wu, B Simmons, and S Singer. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4):605–607, 2016.

[199] Fengzhe Xu, Yuanqing Fu, Ting-yu Sun, Zengliang Jiang, Zelei Miao, Menglei Shuai, Wanglong Gou, Chu-wen Ling, Jian Yang, Jun Wang, Yu-ming Chen, and Ju-Sheng Zheng. The interplay between host genetics and the gut microbiome reveals common and distinct microbiome features for complex human diseases. *Microbiome*, 8(1):145, 2020.

[200] J-Y Yang, Y-S Lee, Y. Kim, S-H Lee, S. Ryu, S. Fukuda, K. Hase, C-S Yang, H S Lim, M-S Kim, H-M Kim, S-H Ahn, B-E Kwon, H-J Ko, and M-N Kweon. Gut commensal bacteroides acidifaciens prevents obesity and improves insulin sensitivity in mice. *Mucosal Immunology*, 10(1):104–116, 2017.

[201] Tanya Yatsunenko, Federico E. Rey, Mark J. Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N. Baldassano, Andrey P. Anokhin, Andrew C. Heath, Barbara Warner, Jens Reeder, Justin Kuczynski, J. Gregory Caporaso, Catherine A. Lozupone, Christian Lauber, Jose Carlos Clemente, Dan Knights, Rob Knight, and Jeffrey I. Gordon. Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–227, 2012.

[202] Rahat Zaheer, Noelle Noyes, Rodrigo Ortega Polo, Shaun R. Cook, Eric Marinier, Gary Van Domselaar, Keith E. Belk, Paul S. Morley, Tim A. McAllister, Rodrigo Ortega Polo, Shaun R. Cook, Eric Marinier, Gary Van Domselaar, Keith E. Belk, Paul S. Morley, Tim A. McAllister, Rodrigo Ortega Polo, Shaun R. Cook, Eric Marinier, Gary Van Domselaar, Keith E. Belk, Paul S. Morley, and Tim A. McAllister. Impact of sequencing depth on the characterization of the microbiome and resistome. *Scientific Reports*, 8(1):1–11, 2018.

[203] Xiaolei Ze, Yonit Ben David, Jenny A. Laverde-Gomez, Bareket Dassa, Paul O. Sheridan, Sylvia H. Duncan, Petra Louis, Bernard Henrissat, Nathalie Juge, Nicole M. Koropatkin, Edward A. Bayer, Harry J. Flint, and Julian Parkhill. Unique organization of extracellular amylases into amylosomes in the resistant starch-utilizing human colonic ¡i¿firmicutes¡/i¿ bacterium ruminococcus bromii. *mBio*, 6(5):e01058–15, 2015.

[204] Xiaolei Ze, Sylvia H Duncan, Petra Louis, and Harry J Flint. Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon. *The ISME Journal*, 6(8):1535–1543, 2012.

[205] Chaofang Zhong, Chaoyun Chen, Lusheng Wang, and Kang Ning. Integrating pangenome with metagenome for microbial community profiling. *Computational and Structural Biotechnology Journal*, 19:1458–1466, 2021.

Appendix A

APPENDIX FIGURES AND TABLES

Table A.1: Antibiotic Classes and their mean abundances between dairy workers' and community controls' metagenomes. Proportion of samples from dairy workers and community controls that had identified resistance genes in each antibiotic class are displayed.

Antibiotic Class	**Dairy mean abundance (SD)**	**Dairy (N) (% detected)**	**Community mean abundance (SD)**	**Community (N) (% detected)**
tetracycline	0.00977% (0.00898%)	9 (90%)	0.00577% (0.00521%)	6 (100%)
sulfonamide	-	0	0.000337%	1 (17%)
macrolide	0.00344% (0.00402%)	7 (70%)	0.00100% (0.000656%)	5 (83%)
glycopeptide	-	0	0.000205%	1 (17%)
fluoroquinolone	0.0000793%	1 (17%)	0.00177% (0.00206%)	4 (67%)
cephamycin	0.00358% (0.00244%)	9 (90%)	0.00457% (0.00404%)	4 (67%)
cephalosporin	0.00146% (0.000978%)	3 (30%)	0.000320%	1 (17%)
aminoglycoside	0.00142% (0.00205%)	8 (80%)	0.00326% (0.00377%)	6 (100%)
multi-drug resistance	0.00177% (0.00245%)	8 (80%)	0.00717% (0.0119%)	6 (10%)

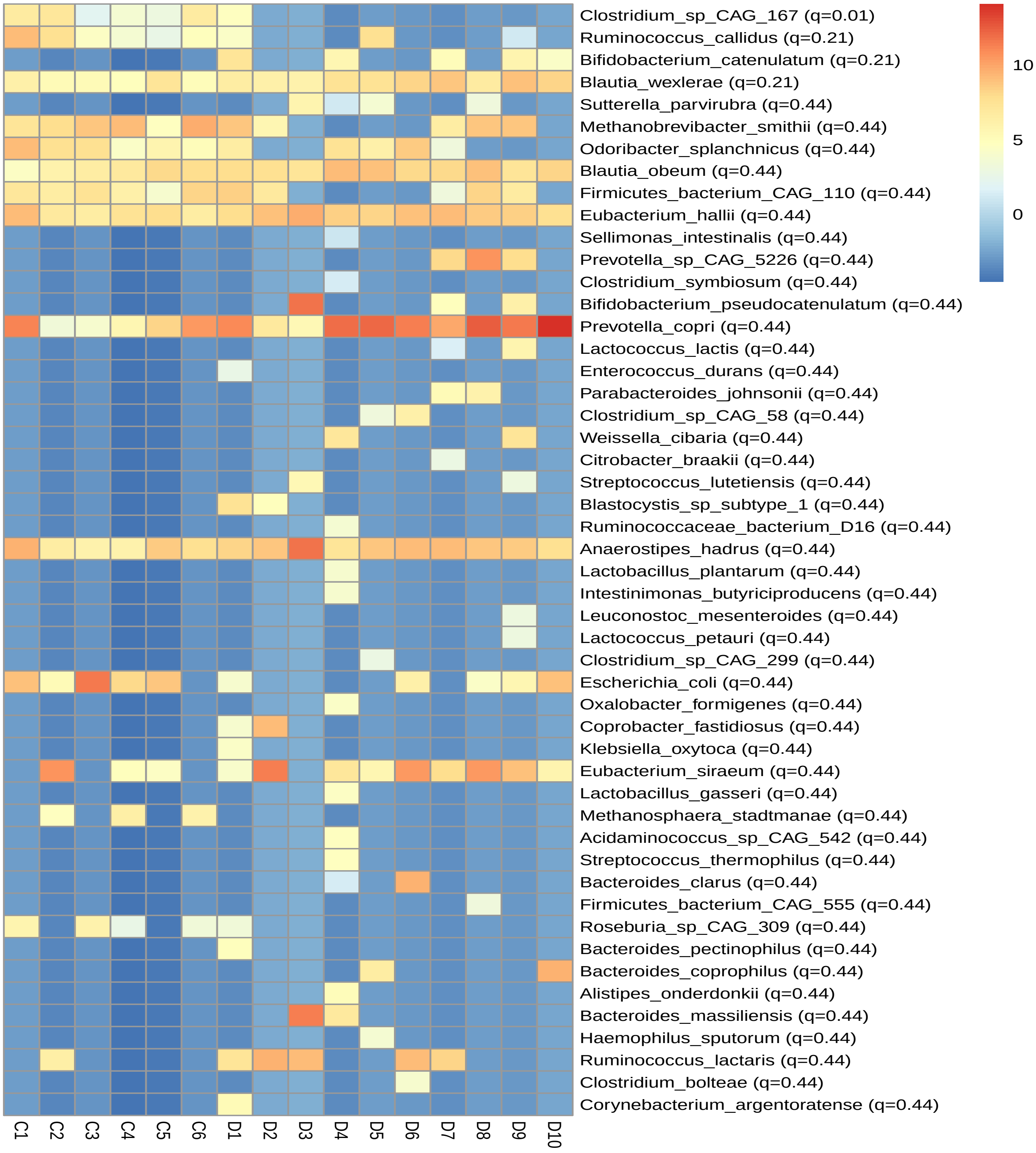


Figure A.1: To test for differential abundance of species between dairy workers and community controls, we conducted independent t-tests of centered log ratio transformed relative abundances with a Benajamini-Hochberg false discovery rate correction. The 50 species with the highest magnitude test statistics show mixed patterns of differential abundance between groups. The heatmap displays CLR transformed relative abundances from lower abundances (blue) to higher (red) abundances.

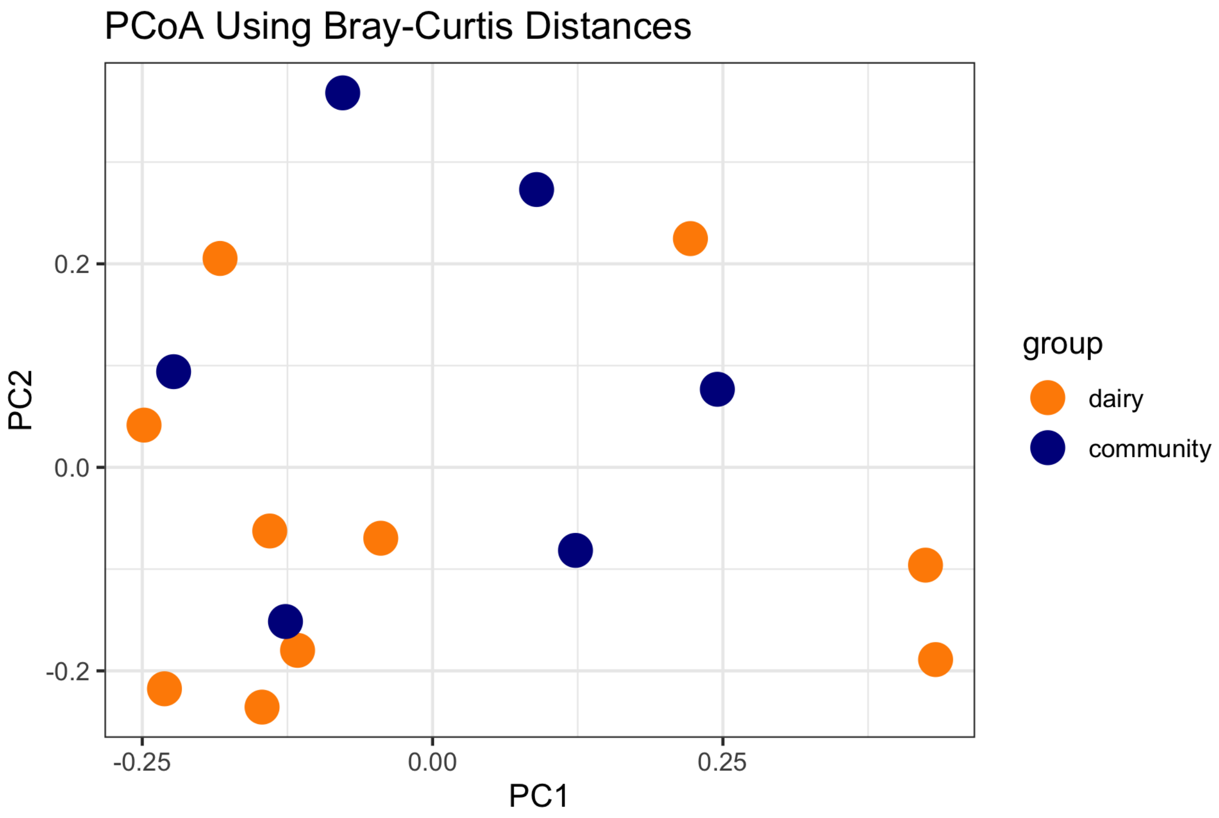


Figure A.2: Principal coordinates analysis using Bray-Curtis distances shows similarity in microbial compositions between dairy workers and community controls.

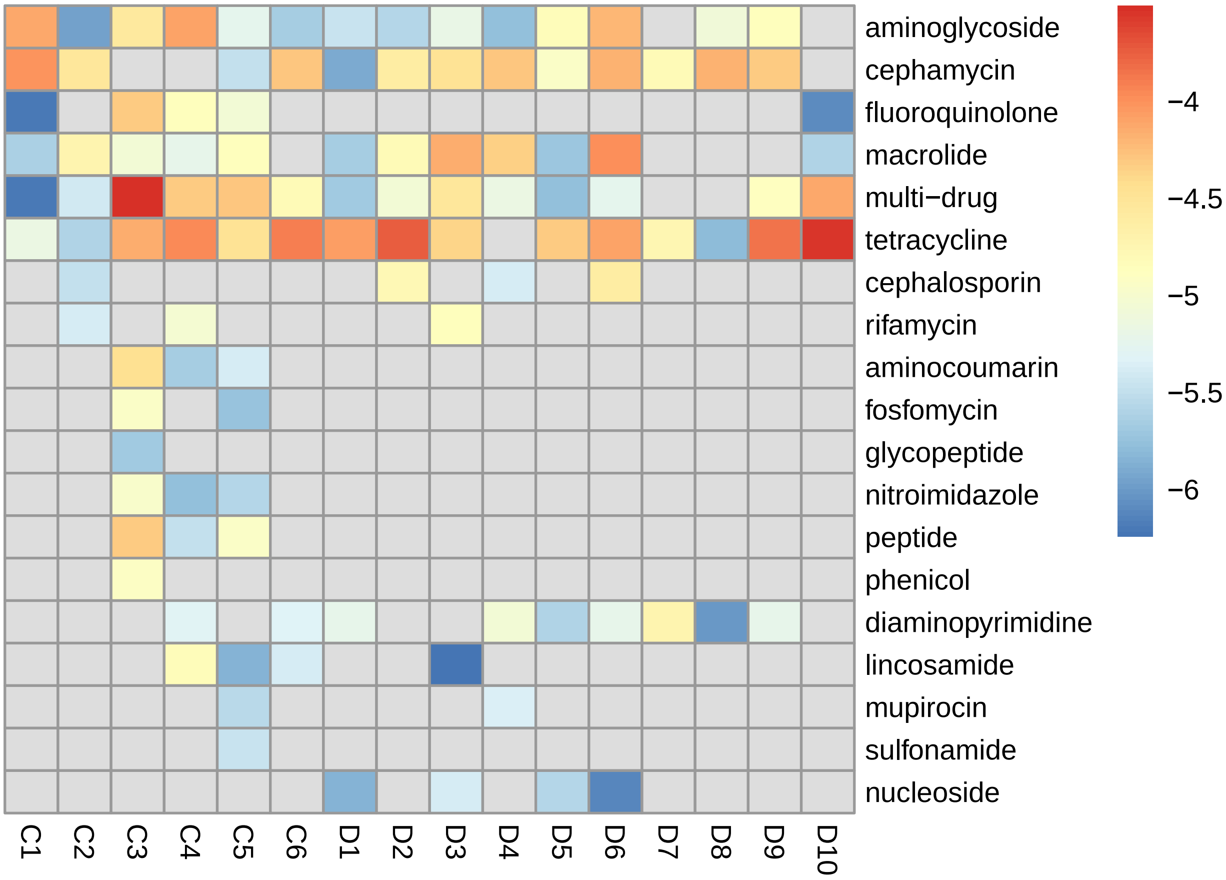


Figure A.3: log_{10} transformed relative abundances of all antibiotic resistance genes grouped by all antibiotic classes (rows) identified across the 16 metagenomes (columns). log_{10} relative abundances are colored from smaller (blue) to larger (red) with grey cells representing no identified genes corresponding to a particular sample.

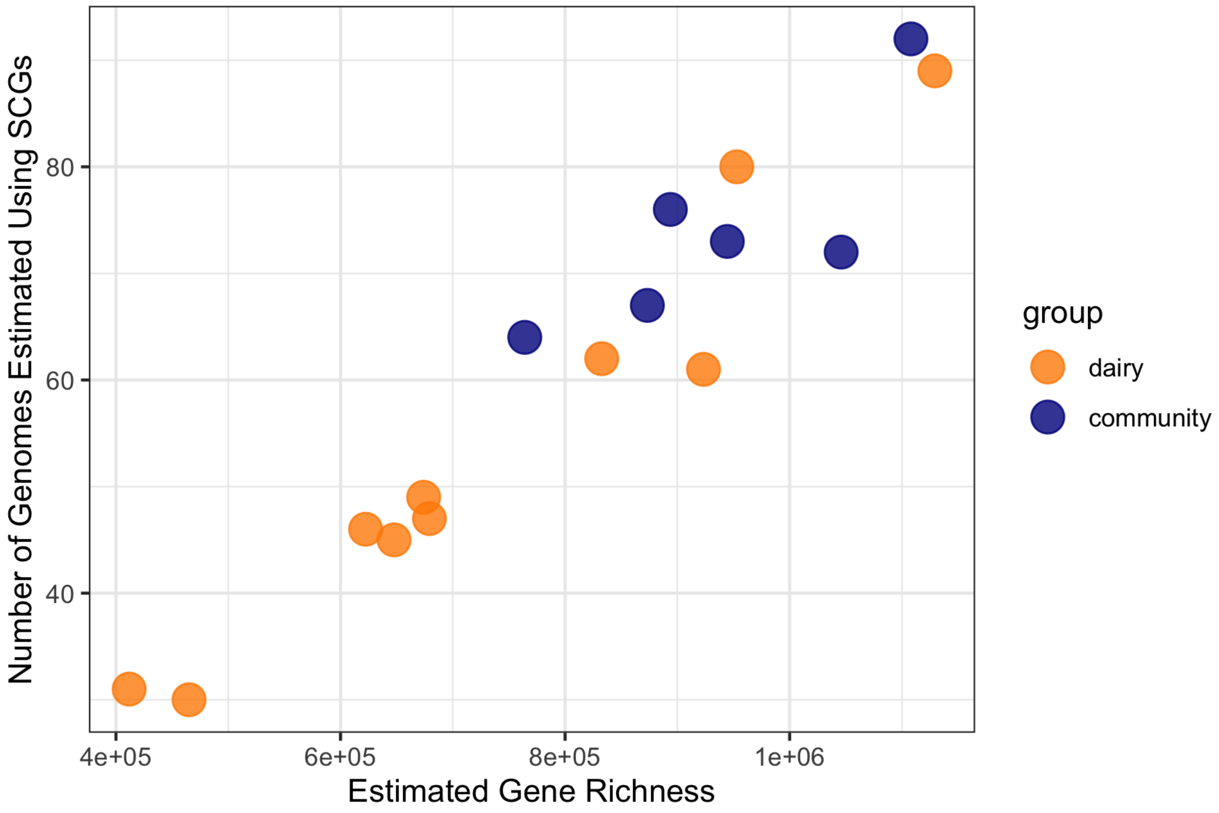


Figure A.4: Estimated gene richness using `breakaway`[193] is positively correlated with predicted numbers of genomes in a metagenome using single-copy core genes.

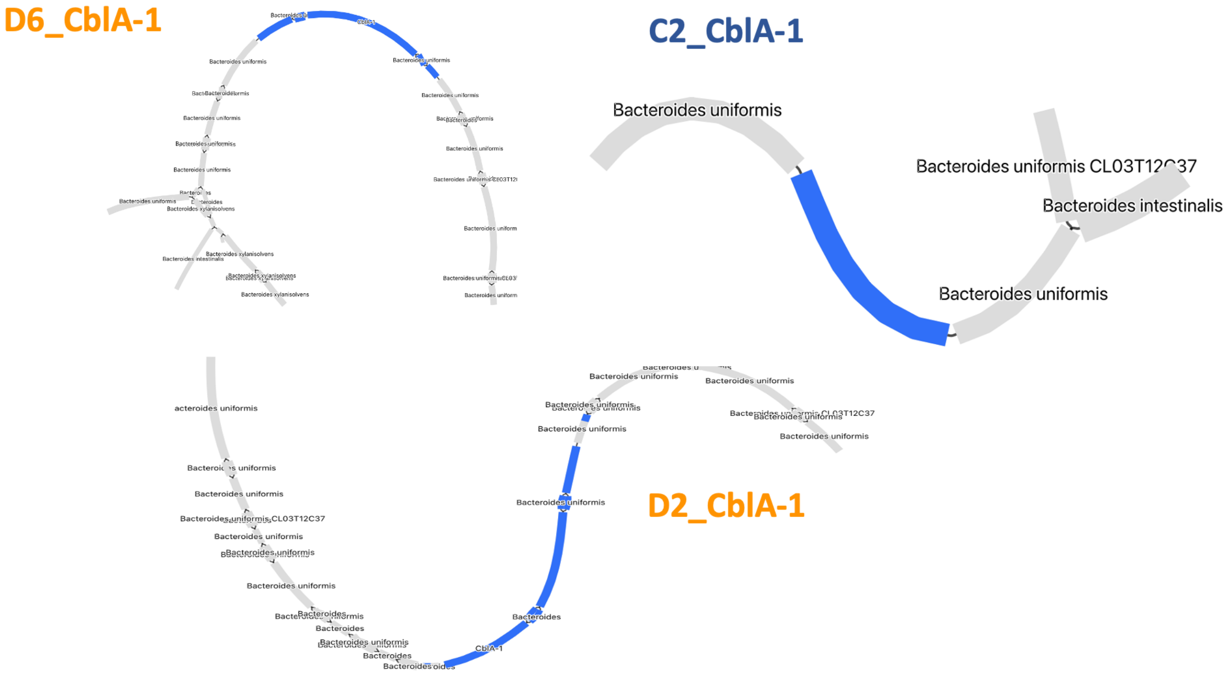


Figure A.5: Assembly graphs and taxonomic annotations for *CblA-1* resistance genes recovered from two dairy workers and one community control sample show taxonomic association of these genes with Bacteroides uniformis. Segments of the graph corresponding to the resistance gene are colored in blue and query neighborhoods are shown in grey.

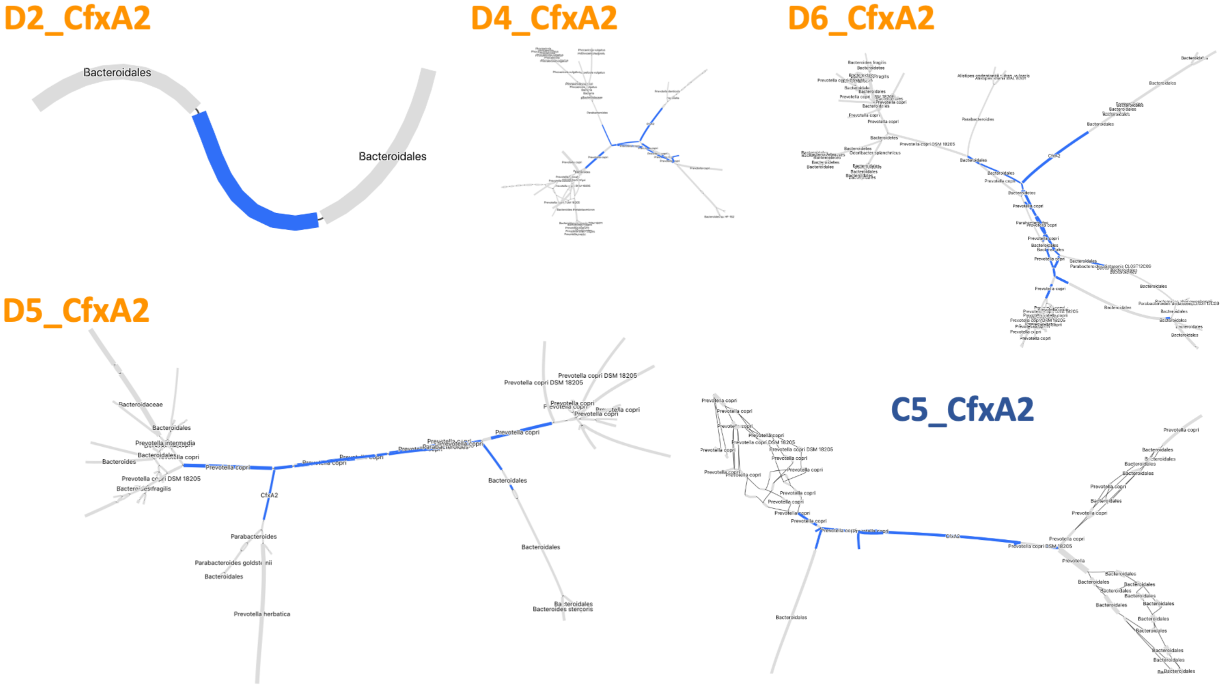


Figure A.6: Assembly graphs and taxonomic annotations for *cfxa2* resistance genes found in four dairy workers and one community control sample. Segments of the graph corresponding to the resistance gene are colored in blue and query neighborhoods are shown in grey.

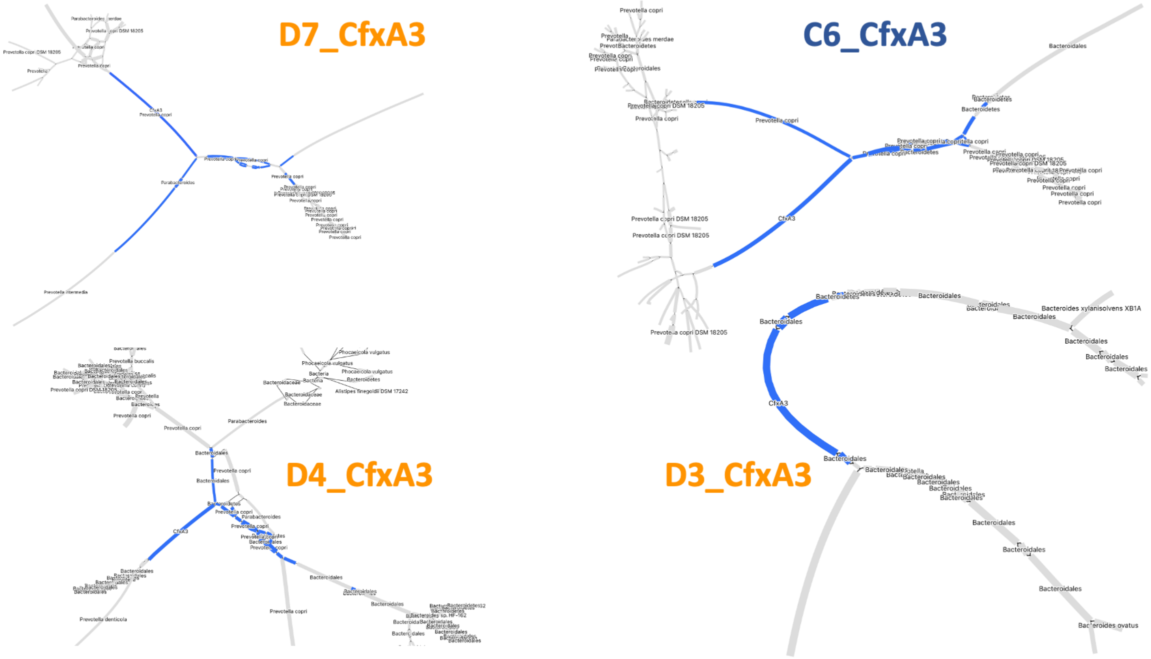


Figure A.7: Assembly graphs and taxonomic annotations for *cfxa3* resistance genes found in three dairy workers and one community control sample. Segments of the graph corresponding to the resistance gene are colored in blue and query neighborhoods are shown in grey.

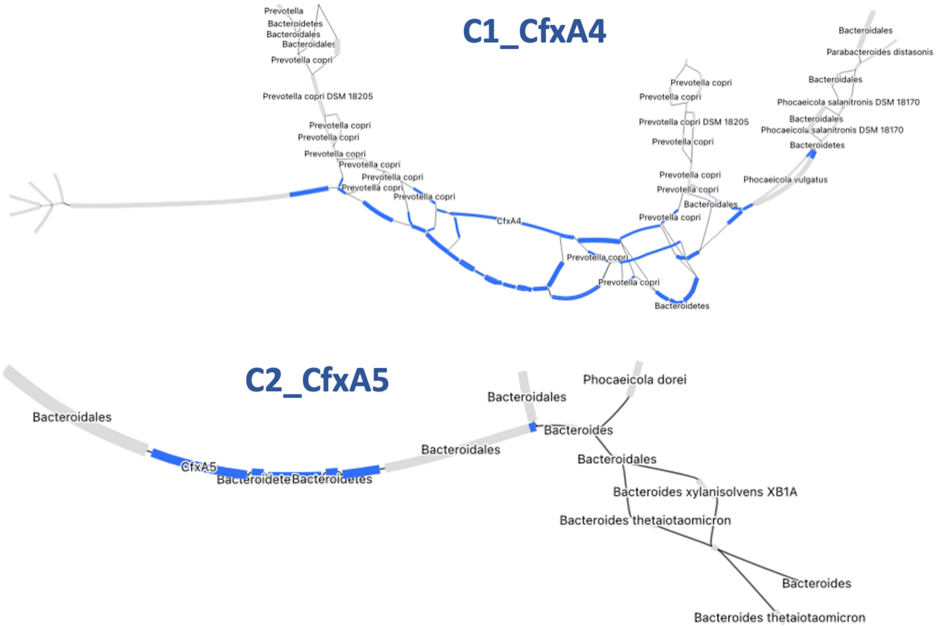


Figure A.8: Assembly graphs and taxonomic annotations for *cfxa4* and *cfxa5* resistance genes found in community control samples. Segments of the graph corresponding to the resistance gene are colored in blue and query neighborhoods are shown in grey.

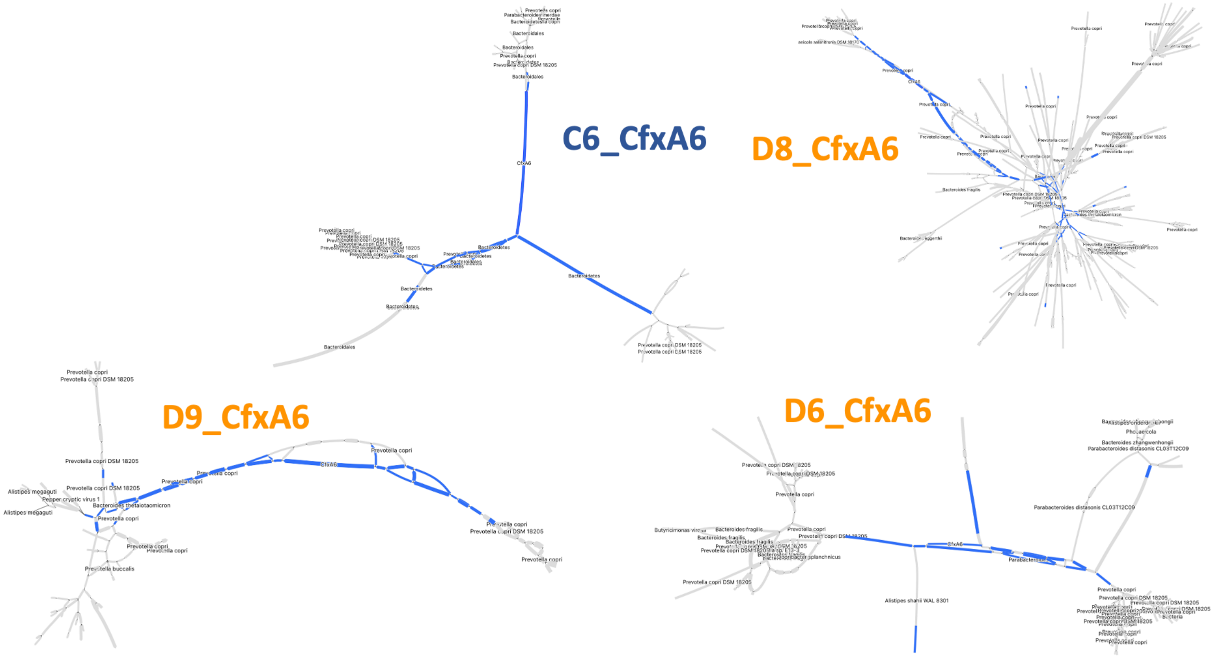


Figure A.9: Assembly graphs and taxonomic annotations for *cfxa6* and *cfxa5* resistance genes found in three dairy worker and one community control samples. Segments of the graph corresponding to the resistance gene are colored in blue and query neighborhoods are shown in grey.

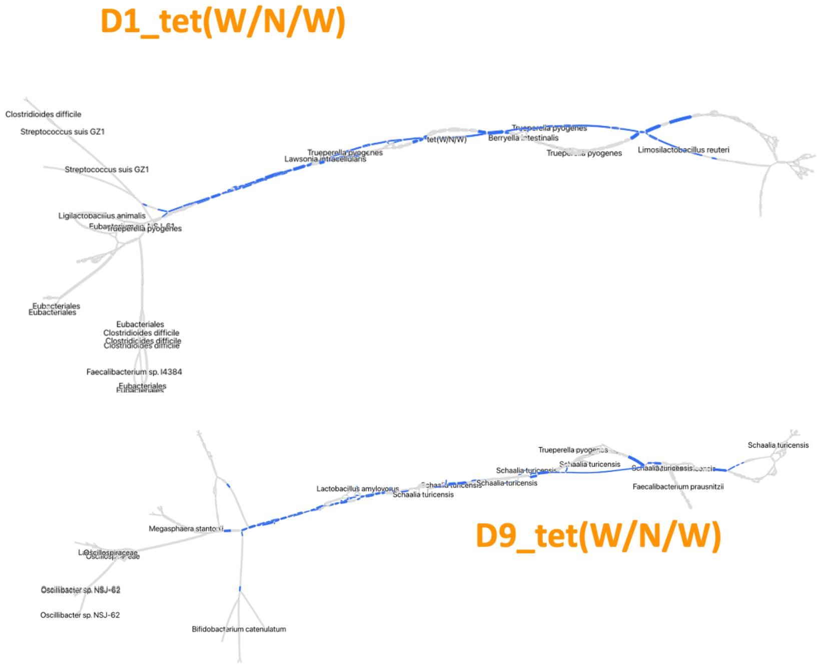


Figure A.10: Assembly graphs and taxonomic annotations for *tet W/N/W* resistance genes found in two community control samples. Segments of the graph corresponding to the resistance gene are colored in blue and query neighborhoods are shown in grey.

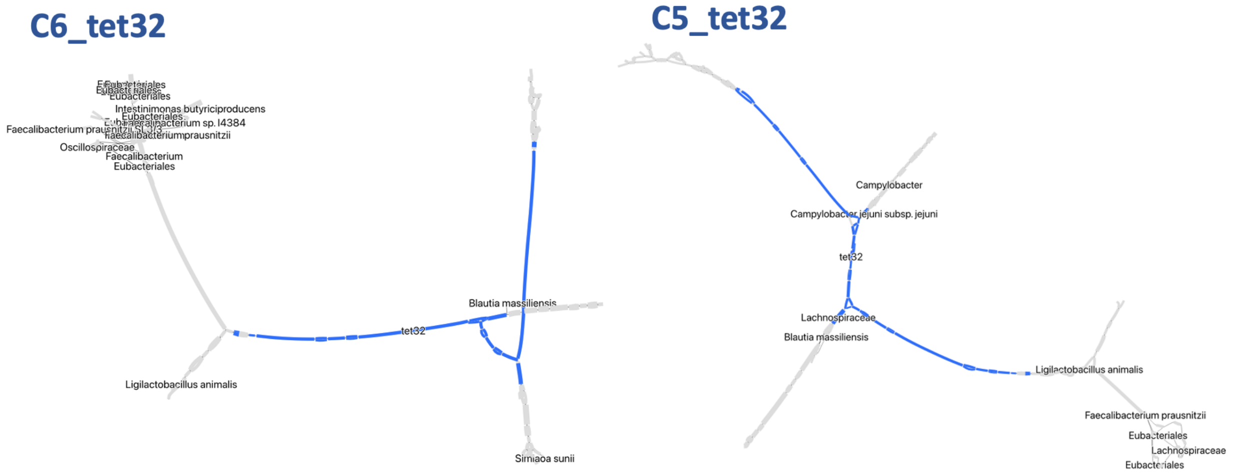


Figure A.11: Assembly graphs and taxonomic annotations for *tet32* resistance genes found in two dairy worker samples. Segments of the graph corresponding to the resistance gene are colored in blue and query neighborhoods are shown in grey.

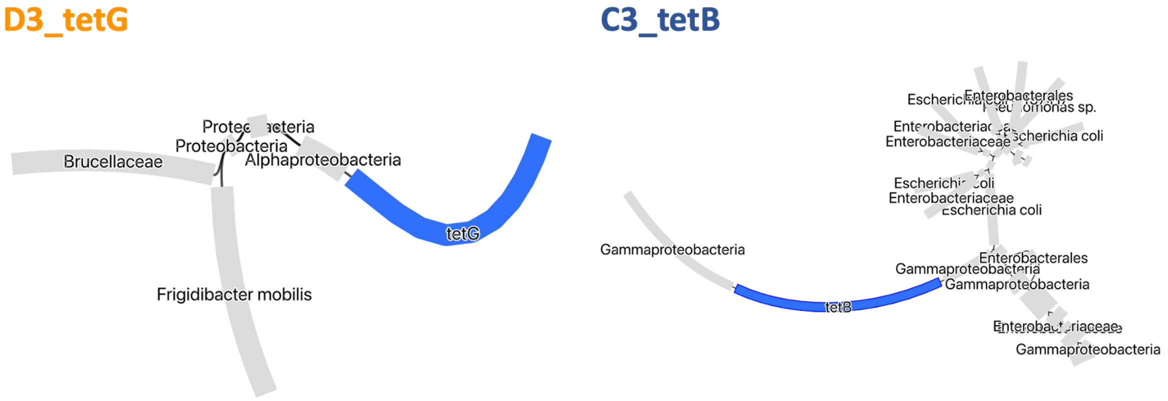


Figure A.12: Assembly graphs and taxonomic annotations for *tetG* and *tetB* resistance genes found in one dairy worker and one community control sample, respectively. Segments of the graph corresponding to the resistance gene are colored in blue and query neighborhoods are shown in grey.

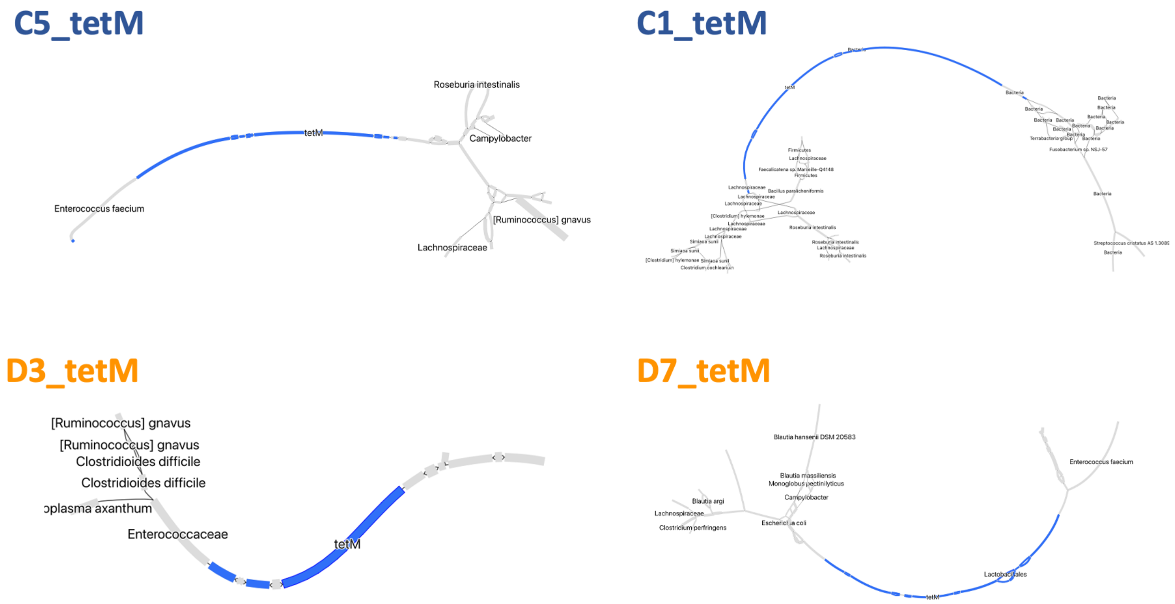


Figure A.13: Assembly graphs and taxonomic annotations for *tetM* resistance genes found in two community control and two dairy worker samples. Segments of the graph corresponding to the resistance gene are colored in blue and query neighborhoods are shown in grey.

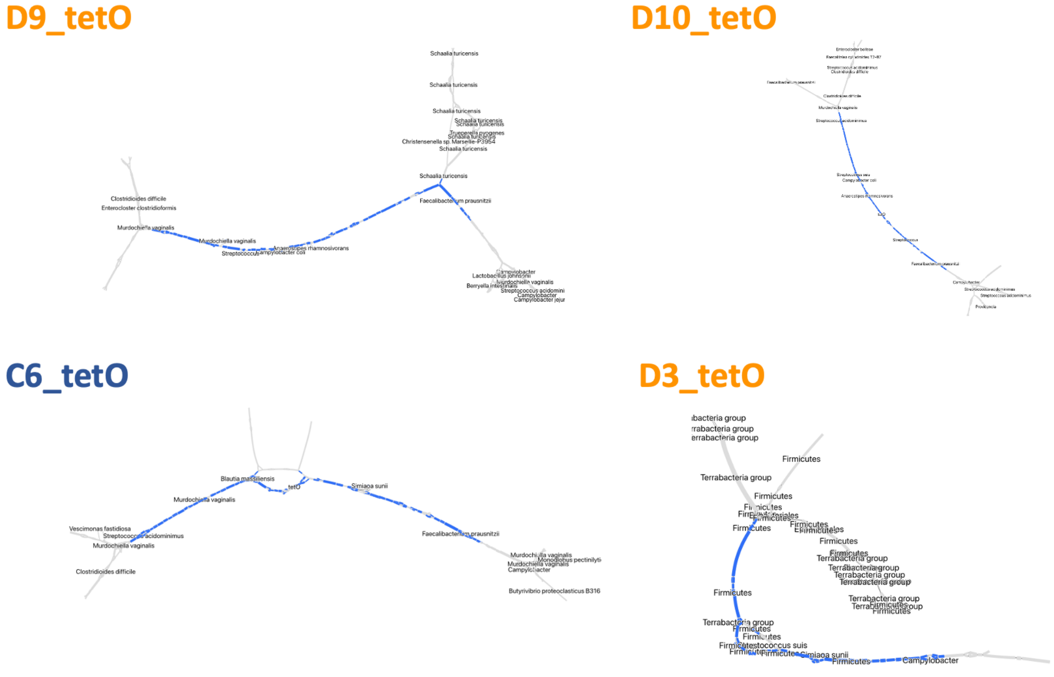


Figure A.14: Assembly graphs and taxonomic annotations for *tetO* resistance genes found in one community control and three dairy worker samples. Segments of the graph corresponding to the resistance gene are colored in blue and query neighborhoods are shown in grey.

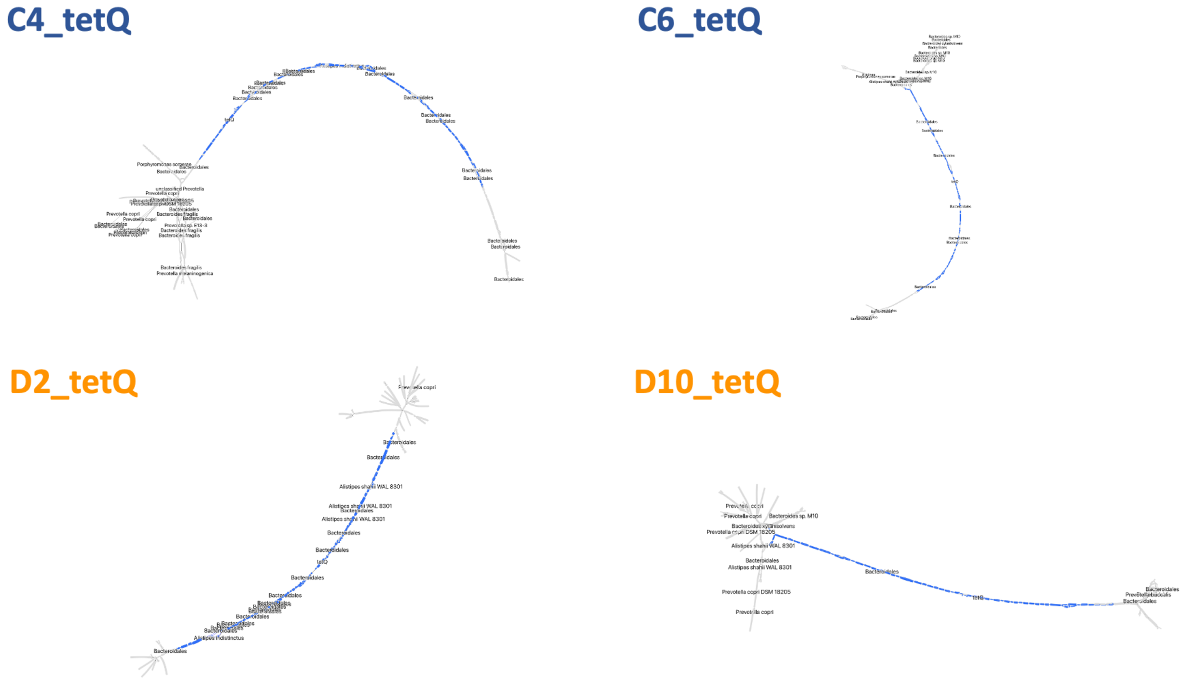


Figure A.15: Assembly graphs and taxonomic annotations for *tetQ* resistance genes identified in two community control and one dairy worker samples. Segments of the graph corresponding to the resistance gene are colored in blue and query neighborhoods are shown in grey.

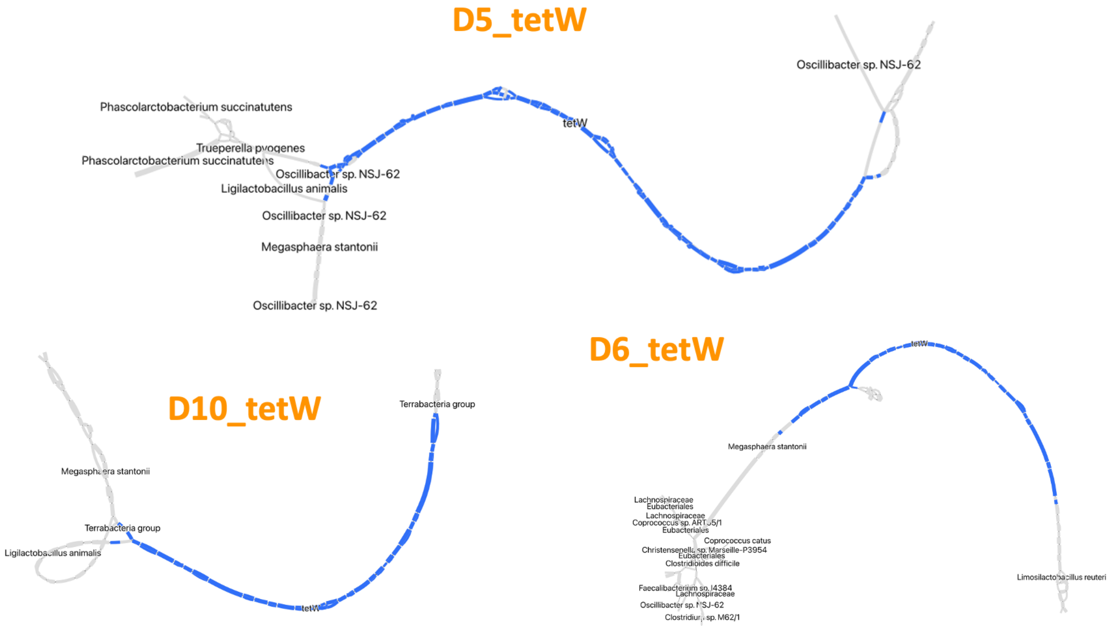


Figure A.16: Assembly graphs and taxonomic annotations for *tetW* resistance genes identified in three dairy worker samples. Segments of the graph corresponding to the resistance gene are colored in blue and query neighborhoods are shown in grey.

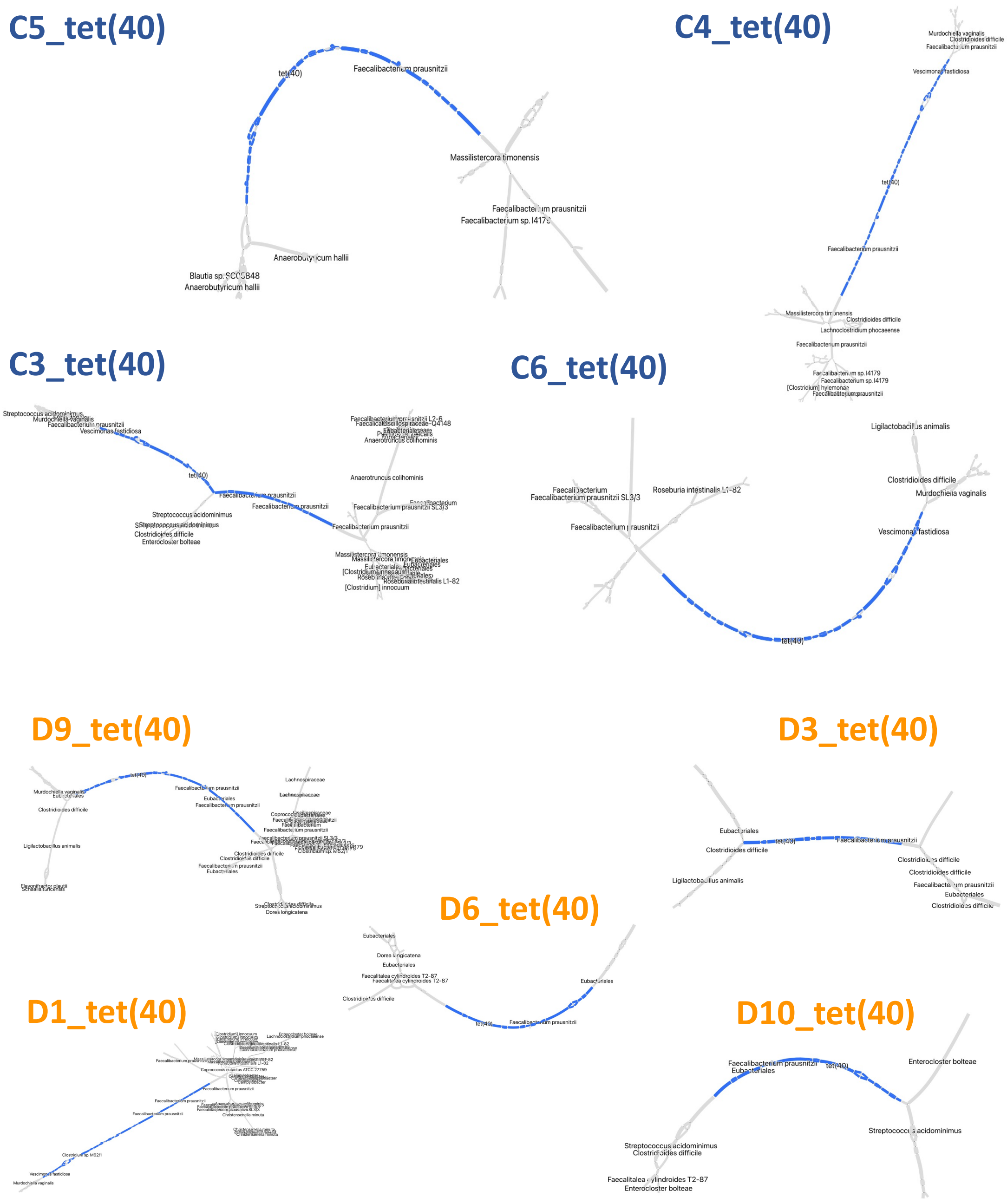

C5_tet(40)
tet(40)
Faecalibacterium prausnitzii
Massilistercora timonensis
Faecalibacterium prausnitzii
Faecalibacterium sp. I4179
Anaerobutyricum hallii
Blautia sp. SC05B48
Anaerobutyricum hallii
C4_tet(40)
Murdochiella vaginalis
Clostridioides difficile
Faecalibacterium prausnitzii
Vescimonas fastidiosa
tet(40)
Faecalibacterium prausnitzii
Massilistercora timonensis
Clostridioides difficile
Lachnoclostridium phocaeense
Faecalibacterium prausnitzii
Faecalibacterium sp. I4179
Faecalibacterium sp. I4179
[Clostridium] hylemonae
Faecalibacterium prausnitzii
C3_tet(40)
Streptococcus acidominimus
Murdochiella vaginalis
Faecalibacterium prausnitzii
Vescimonas fastidiosa
tet(40)
Faecalibacterium prausnitzii
Faecalibacterium prausnitzii
Streptococcus acidominimus
Streptococcus acidominimus
Clostridioides difficile
Enterocloster bolteae
Faecalibacterium prausnitzii L2-6
Oscillospiraceae-Q4148
Eubacteriales
Anaerotruncus colihominis
Anaerotruncus colihominis
Faecalibacterium
Faecalibacterium prausnitzii SL3/3
Faecalibacterium prausnitzii
Massilistercora timonensis
Massilistercora timonensis
Eubacteriales
Eubacteriales
[Clostridium] innocuum
Roseburia intestinalis L1-82
[Clostridium] innocuum
C6_tet(40)
Ligilactobacillus animalis
Clostridioides difficile
Murdochiella vaginalis
Vescimonas fastidiosa
Faecalibacterium
Faecalibacterium prausnitzii SL3/3
Roseburia intestinalis L1-82
Faecalibacterium prausnitzii
tet(40)
D9_tet(40)
tet(40)
Murdochiella vaginalis
Eubacteriales
Clostridioides difficile
Ligilactobacillus animalis
Flavonifractor plautii
Schaalia turicensis
Faecalibacterium prausnitzii
Eubacteriales
Faecalibacterium prausnitzii
Lachnospiraceae
Lachnospiraceae
Oscillospiraceae
Eubacteriales
Faecalibacterium
Faecalibacterium prausnitzii
Faecalibacterium prausnitzii SL3/3
Faecalibacterium sp. I4179
Clostridium sp. M62/1
Clostridioides difficile
Clostridioides difficile
Faecalibacterium prausnitzii
Eubacteriales
Clostridioides difficile
Streptococcus acidominimus
Dorea longicatena
D3_tet(40)
Eubacteriales
Clostridioides difficile
tet(40)
Faecalibacterium prausnitzii
Ligilactobacillus animalis
Clostridioides difficile
Clostridioides difficile
Faecalibacterium prausnitzii
Eubacteriales
Clostridioides difficile
D6_tet(40)
Eubacteriales
Dorea longicatena
Eubacteriales
Faecalitalea cylindroides T2-87
Faecalitalea cylindroides T2-87
Clostridioides difficile
Eubacteriales
tet(40)
Faecalibacterium prausnitzii
D1_tet(40)
[Clostridium] innocuum
[Clostridium] innocuum
Enterocloster bolteae
Lachnoclostridium phocaeense
Roseburia intestinalis L1-82
Lachnoclostridium phocaeense
Faecalibacterium prausnitzii
Massilistercora timonensis
Roseburia intestinalis L1-82
Coprococcus eutactus ATCC 27759
Campylobacter
Faecalibacterium prausnitzii
Anaerotruncus colihominis
Faecalibacterium prausnitzii SL3/3
Christensenella minuta
Christensenella minuta
Faecalibacterium prausnitzii
Faecalibacterium prausnitzii
Clostridium sp. M62/1
Vescimonas fastidiosa
Murdochiella vaginalis
D10_tet(40)
Faecalibacterium prausnitzii
Eubacteriales
tet(40)
Enterocloster bolteae
Streptococcus acidominimus
Streptococcus acidominimus
Clostridioides difficile
Faecalitalea cylindroides T2-87
Enterocloster bolteae

Figure A.17: Assembly graphs and taxonomic annotations for *tet(40)* resistance genes identified in five dairy worker and four community control samples. Segments of the graph corresponding to the resistance gene are colored in blue and query neighborhoods are shown in grey.