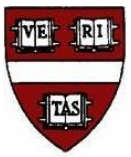
Epidemiologic Studies of the Human Microbiome and of COVID-19

HARVARD UNIVERSITY

Graduate School of Arts and Sciences



DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the
Committee on Higher Degrees in Population Health Sciences,
have examined a dissertation entitled

"Epidemiologic Studies of the Human Microbiome and of COVID-19"

presented by

Emma Accorsi

candidate for the degree of Doctor of Philosophy
and hereby certify that it is worthy of acceptance.

Dr. Marc Lipsitch, D.Phil., Committee Chair, Harvard T.H. Chan School of Public Health
Dr. William Hanage, Ph.D., Harvard T.H. Chan School of Public Health
Dr. Sebastian Haneuse, Ph.D., Harvard T.H. Chan School of Public Health
Dr. Curtis Huttenhower, Ph.D., Harvard T.H. Chan School of Public Health

In lieu of all Dissertation Advisory Committee members' signatures,
I, Tyler J. VanderWeele, Ph.D., appointed by the Ph.D. in Population Health Sciences,
confirm that the Dissertation Advisory Committee has examined the above dissertation,
presented by Emma Accorsi, and hereby certify that it is worthy of acceptance as of 15 March 2021.

Date: 15 March 2021

Epidemiologic Studies of the Human Microbiome and of COVID-19

A dissertation presented

by

Emma Accorsi

to

The Department of Epidemiology

Harvard T.H. Chan School of Public Health

&

The Department of Population Health Sciences

Harvard Graduate School of Arts and Sciences

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

In the subject of

Population Health Sciences (Field of Study: Epidemiology)

Harvard University

Cambridge, MA

March 2021

Epidemiologic Studies of the Human Microbiome and of COVID-19

Abstract

Across these three chapters, we utilize and evaluate analytical and epidemiological methods for studying how humans and microbes interact, with a focus on two pathogens, *Staphylococcus aureus* and SARS-CoV-2. We first identified predictors of *S. aureus* carriage in a paired, longitudinal, high-dimensional dataset. We determined that while mothers represented an early source for *S. aureus* transmission to the developing infant microbiome, microbiome determinants became more important later on. We also identified a gene family that was significantly anti-correlated with *S. aureus* in infants and mothers and was likely acting as a phylogenetic marker for a bacterial species not closely homologous to current reference isolates that competes with *S. aureus*. Secondly, we evaluated methods for performing statistical mediation analysis with microbiome data in realistic synthetic data. We used our findings to perform mediation analyses exploring the role of the gut microbiome as a mediator between diet and cardiometabolic disease in two datasets. We identified the top performing methods for total indirect effect estimation, hypothesis testing for the total indirect effect, and hypothesis testing for component indirect effects. However, more work is needed to improve method performance and usability; currently, an end-user may need to employ multiple methods to accomplish a full mediation analysis with microbiome data, which can give rise to conflicting estimates. Finally, with the onset of the SARS-CoV-2 pandemic, we described how epidemiological biases, such as confounding, selection bias, and measurement error, can occur across five important classes of research questions for SARS-CoV-2 and COVID-19 and provided ways to avoid these biases across different stages of the study process.

Table of Contents

Title page i

Copyright ii

Abstract iii

Table of Contents iv

Acknowledgments v

List of Tables and Figures viii

Chapter 1. Introduction 1

Chapter 2. Determinants of *Staphylococcus aureus* carriage in the developing infant nasal microbiome 3

Chapter 3. Detecting and quantifying mediation of phenotypes by microbial communities 44

Chapter 4. How to detect and reduce potential sources of biases in studies of SARS-CoV-2 and COVID-19 83

Chapter 5. Conclusion 125

References 127

Appendix 156

Acknowledgments

Firstly, I'd like to thank my PhD advisors, Curtis Huttenhower and Marc Lipsitch, for their guidance and mentorship. You have taught me how to ask big questions and answer them, you have shown me what public leadership means during a health crisis, and I aspire to continue down the paths you've opened for me. I am incredibly grateful for the opportunity to learn from you and for everything that you've taught me.

I'd like to sincerely thank my committee members, Bill Hanage, Sebastien Haneuse, Marc Lipsitch, and Curtis Huttenhower, for always taking the time to provide insightful feedback and thoughtful input on my research projects and for helping guide me towards graduation.

A very special thanks to Eric Franzosa, my science mentor, for being both a great scientist and a kind, enthusiastic teacher. Thank you for all the knowledge you've shared with me - you've taught me so much about the different dimensions of being a scientist, from the technical nuances of microbiome data to making my plots look nicer to how to best communicate research findings.

I would like to thank the members of the Huttenhower lab, especially Nicole Levesque, Tiffany Hsu, writing group, small group, and my office/desk mates, and the CCDD/Lipsitch lab, especially Tia Hira, for your advice and help.

Many thanks to the Department of Population Health Sciences and Department of Epidemiology teams, including Bruce Villineau, Eric DiGiovanni, and Matthew Boccuzzi, for making this program possible through their hard work and guiding us through any bumps along the way.

I would like to thank my family and friends for their continual love, support and encouragement, especially my parents, Sarah and Michael Accorsi, my brother, Joel Accorsi, and my partner, Scott Steinmetz.

I'm so grateful for my wonderful science mentors and teachers who inspired me to pursue a career in science and supported me on my path to pursuing a PhD, including Dr. Sherry L. Palacios, Dr. Alessandro Veneziani, Dr. Whitney Tabor, Dr. Raphael Kudela, Dr. Julia Sherman, Dr. Cheryl Granger, Dr. Emily Schaller, and Dr. Nelle Couret. Thank you for your mentorship and encouragement - I couldn't have gotten here without you.

I'd like to thank my project collaborators - it's been a true pleasure working with you and I've learned so much.

- Chapter 2: Eric A. Franzosa, Tiffany Hsu, Regina Joice Cordy, Ayala Maayan-Metzger, Hanaa Jaber, Aylana Reiss-Mandel, Madeleine Kline, Casey DuLong, Marc Lipsitch, Gili Regev-Yochay, Curtis Huttenhower (https://doi.org/10.1186/s13059-020-02209-7)

- Chapter 3: Eric A. Franzosa, Ethan K. Gough, Siyuan Ma, Amee Manges, Marc Lipsitch, Curtis Huttenhower

- Chapter 4: Xueting Qiu, Eva Rumpler, Lee Kennedy-Shaffer, Rebecca Kahn, Keya Joshi, Edward Goldstein, Mats J. Stensrud, Rene Niehus, Muge Cevik, Marc Lipsitch (https://doi.org/10.1007/s10654-021-00727-7)

Finally, I'd like to thank those who provided funding for their incredible generosity and making this journey possible, including the Charles Willinsky Award Fund at the Harvard T.H.

Chan School of Public Health; the National Institute of Allergy and Infectious Diseases of the National Institutes of Health award number T32AI007535; Ms. Nancy Glickenhaus, Ms. Sarah B. Glickenhaus and the Glickenhaus Financial Aid Fund; and the Department of Population Health Sciences, as well as the Emory Scholars Program.

List of Tables and Figures

Figure 2.1: Longitudinal shotgun metagenomics of the nasal microbiome and *Staphylococcus aureus* carriage in 36 mother-infant pairs. 9

Figure 2.2: Development of infant nasal microbiome composition over the first year of life. 12

Figure 2.3: Significant associations between nasal microbiome taxonomic and functional composition and subject phenotypes in feature-wise testing. 17

Figure 2.4: Ability of infant or maternal microbiome profiles to predict infant microbiome membership. 22

Figure 2.5: Gene-content based strain profiling for *S. aureus* and other common nasal species. 26

Figure 3.1: Overview of simulation methodology. 51

Figure 3.2: True positive and false positive rates for detection of the total indirect effect. 55

Figure 3.3: True positive and false positive rates for detection of component indirect effects. ... 57

Figure 3.4: Estimation of the total indirect effect size. 60

Figure 3.5: Estimation of mediation effects in diet-cardiometabolic disease datasets. 68

Figure 4.1: Schematic showing recruitment-based biases in a hypothetical serosurvey. 86

Figure 4.2: Biases due to misclassification by SARS-CoV-2 antibody tests. 89

Figure 4.3: The relative importance of test sensitivity and specificity depends on the underlying seroprevalence in the study population. 93

Figure 4.4: Directed acyclic graph under the alternative hypothesis showing confounding in the estimation of seroprotection. 96

Figure 4.5: Directed acyclic graph under the null hypothesis showing the possible structure of selection bias due to (a) exclusion from testing and (b) differential likelihood of testing. 101

Figure 4.6: Directed acyclic graph under the null hypothesis showing differential misclassification by (a) whether an individual is tested and (b) the timing or type of test. 104

Figure 4.7: Illustration of index case misclassification where the index and secondary cases are misclassified in a household scenario. 109

Figure 4.8: Illustration of index case misclassification when multiple index cases are present but only one is identified as the index case. 111

Figure 4.9: Illustration of misclassification of contact type and contact infection status. 113

Figure 4.10: Illustration of differential detection of infection in adults and children. 116

Table 4.1: Calculation of the SAR when there is differential detection of infection in adults and children. 118

Figure S2.1: *S. aureus* test concordance. 156

Figure S2.2: PCoA plots of maternal taxonomic and functional features show associations with infant *S. aureus* "ever" acquisition. 157

Figure S2.3: Spearman correlations of UniRef90 X5NU12 with the species-stratified abundances of other UniRef90s. 158

Figure S2.4: Variable importance plot for the prediction of infant *S. aureus* status by sequencing using infant ECs. 159

Figure S2.5: Association of EC 1.6.5.5 (NADPH:quinone reductase) with infant and mother *S. aureus* status by sequencing. 160

Figure S2.6: Variable importance plot for the prediction of infant "ever" acquisition of *S. aureus* using maternal ECs. 161

Figure S2.7: Significant associations between nasal microbiome taxonomic and subject phenotypes in a sensitivity analysis of the unmapped sample mass. 162

Table S2.1. Sample sizes for boxplots. 164

Table S2.2. Full linear model results for Fig. 2.3. 164

Table S2.3. Summary of metadata for study population and samples collected for *S. aureus* microbiome profiling. 164

Table S2.4. Summary of *S. aureus* variables for study population and samples collected for *S. aureus* microbiome profiling. 164

Table S2.5. Sample sizes for random forest models. 164

Table S2.6. Subject metadata. 164

Table S2.7: MetaPhlAn2 taxonomic profiles. 164

Table S2.8. HUMAnN2 functional profiles. 164

Figure S3.1: P-value histograms under different permutation schemes for the naive method hypothesis test. 165

Figure S3.2: Estimation of the total indirect effect by CCMM. 167

Figure S3.3: Estimation of the total indirect effect using modifications of HIMA and the naive method..... 168

Figure S3.4: Top largest component indirect effects identified in the MLVS dataset. 169

Figure S3.5: Range of spiked total indirect effect sizes across simulated datasets relative to the specified total indirect effect..... 170

Table S3.1: Simulation parameters. 172

Table S3.2: Mediation method approaches to address high-dimensionality and compositionality. 173

Figure S4.1: Illustration of index case misclassification where the index and secondary cases are misclassified in scenarios outside of the household. 175

Chapter 1. Introduction

Technological advances, including high-throughput sequencing for the study of human microbial communities and the rapid sharing of SARS-CoV-2 data and manuscripts online, have greatly expanded the complexity and number of epidemiological datasets that public health researchers have access to. In this work, I focus on methodological considerations for studies of human-microbe interaction, specifically analysis of metagenomic datasets and synthesis of the expansive SARS-CoV-2 literature.

In the first chapter "Determinants of *Staphylococcus aureus* carriage in the developing infant nasal microbiome", we use careful analytic methods to identify predictors of *S. aureus* carriage in a paired, longitudinal, high-dimensional dataset. In particular, we adjust our analyses at all stages to avoid biases due to compositional microbiome data, as well as confounding due to strong subject effects on both the microbiome and metadata variables over time. For instance, when looking for taxonomic predictors of *S. aureus* presence, the compositional nature of the data will result in any *S. aureus* abundance "squishing" all other features thereby creating an association between these features and *S. aureus* presence if not adjusted. Working on this first project made me interested in both the technical complexities of working with microbiome data and the nuances of epidemiological biases and study design, which I explored through the next two projects.

In the second project "Chapter 3. Detecting and quantifying mediation of phenotypes by microbial communities" we benchmark methods for performing statistical mediation analysis with microbiome mediators in realistic synthetic microbiome data. The process of simulating realistic microbiome data and troubleshooting a permutation-based hypothesis test for one of the methods (the naive method) showed me firsthand how properties like zero-inflation, high-dimensionality,

and compositionality interact with statistical methods, while the project as a whole will hopefully provide a resource for other microbiome researchers who want to apply statistical mediation analysis to their datasets.

In the final chapter "Chapter 4. How to detect and reduce potential sources of biases in studies of SARS-CoV-2 and COVID-19" we think about how to make sense of a rapidly expanding literature of observational studies on COVID-19 and explain how biases may be responsible for disparate results. In particular, we consider how biases, including confounding, selection bias, and measurement error, occur in the context of infectious disease epidemiology and describe best practices to avoid these biases. With access to more numerous and more complex datasets for studying human-microbe interaction, it is critical to think carefully about how biases may be introduced through both the study design and data analysis processes.

Chapter 2. Determinants of *Staphylococcus aureus* carriage in the developing infant nasal microbiome

2.1 ABSTRACT

Background

Staphylococcus aureus is a leading cause of healthcare- and community-associated infections and can be difficult to treat due to antimicrobial resistance. About 30% of individuals carry *S. aureus* asymptomatically in their nares, a risk factor for later infection, and interactions with other species in the nasal microbiome likely modulate its carriage. It is thus important to identify ecological or functional genetic elements within the maternal or infant nasal microbiomes that influence *S. aureus* acquisition and retention in early life.

Results

We recruited 36 mother-infant pairs and profiled a subset of monthly longitudinal nasal samples from the first year after birth using shotgun metagenomic sequencing. The infant nasal microbiome is highly variable, particularly within the first two months. It is weakly influenced by maternal nasal microbiome composition, but primarily shaped by developmental and external factors, such as daycare. Infants display distinctive patterns of *S. aureus* carriage, positively associated with *Acinetobacter* species, *Streptococcus parasanguinis*, *Streptococcus salivarius*, and *Veillonella* species and inversely associated with maternal *Dolosigranulum pigrum*. Furthermore, we identify a gene family, likely acting as a taxonomic marker for an unclassified species, that is significantly anti-correlated with *S. aureus* in infants and mothers. In gene-content based strain profiling, infant *S. aureus* strains are more similar to maternal strains.

Conclusions

This improved understanding of *S. aureus* colonization is an important first step toward development of novel, ecological therapies for controlling *S. aureus* carriage.

Keywords: nasal, microbiome, *S. aureus*, infant, development, carriage, shotgun metagenomics, longitudinal

2.2 INTRODUCTION

Staphylococcus aureus is a leading cause of healthcare-associated infections among infants (1) and adults (2), and it is the most common cause of hospital and community acquired skin and soft tissue infections (SSTIs) (3). Antibiotic resistance, in particular to β-lactams (e.g., methicillin-resistant *S. aureus*, MRSA), makes treatment of these infections difficult and expensive (4). *S. aureus* is responsible for approximately 119,000 bloodstream infections and 20,000 deaths per year in the United States alone (5). Nasal colonization by *S. aureus* increases the risk of SSTIs (6) and is a risk factor for bacteremia, as 80% of bloodstream infections are caused by the same strain found in the individual's nose (7–10). Additionally, colonization allows for transmission to new hosts and selection for new, possibly detrimental microbial traits, such as antibiotic resistance (11). At any one time, approximately 30% of individuals carry this species asymptomatically in their nares (12) and, over time, individuals display three carriage patterns: persistent (~20%), transient (~60%), and never (~20%) carriers (13–16). It is unknown what causes these differences between individuals (13–16), and research in infants is especially scarce (17).

Previous studies suggest that both *S. aureus* carriage and nasal microbiome composition are primarily determined by environmental, rather than host genetic factors (18,19). Co-occurring nasal species can provide resistance to *S. aureus* colonization through competition for nutrients

and binding sites (20,21), such as *Corynebacterium* strain Co304 which is believed to outcompete *S. aureus* at binding nasal mucin (21), or through the induction of a host immune response (22–24), such as the targeting of *S. aureus* by a cross-reactive antibody triggered against *Streptococcus pneumoniae* (23). Additionally, some nasal species produce antimicrobial molecules that are directly toxic to *S. aureus* (25–30), such as hydrogen peroxide produced by *S. pneumoniae* (29) or the peptide antibiotic lugdunin produced by *Staphylococcus lugdunensis* (30). While laboratory studies have identified these mechanisms, it is not clear how they translate into the observed population-level trends in *S. aureus* carriage. For instance, although lugdunin is highly effective at killing *S. aureus* and the risk of *S. aureus* colonization is sixfold-lower in those colonized with *S. lugdunensis*, the population prevalence of *S. lugdunensis* (estimated at 9% (30,31) or 26% (32)) is too low to explain why 80% of individuals only temporarily or never acquire *S. aureus* (30,33). Similarly, some *Staphylococcus epidermidis* isolates produce lantibiotics, and many produce an extracellular serine protease Esp that inhibits *S. aureus* colonization (34,35), but *S. epidermidis* also does not consistently predict *S. aureus* absence (19,32,35,36). Understanding nasal microbial community structure, ecology, metabolism, and signaling thus holds promise for elucidating patterns of *S. aureus* carriage, but many complementary and competing mechanisms likely contribute to its ultimate colonization pattern (33). Human population studies using shotgun metagenomic measures that are both detailed and comprehensive are thus needed to characterize nasal microbial communities with a high degree of phylogenetic and functional resolution.

Understanding the relationship between the nasal microbiome and *S. aureus* carriage in early life is especially important because the infant microbiome across the body undergoes dramatic shifts in the first year (37,38), and infants have also been shown to acquire and lose *S. aureus* in the nares during this time (39,40). Prior studies of *S. aureus* carriage have mainly been performed in adults

and are not necessarily generalizable to infants, as the infant microbiome differs greatly from that of adults during this early period of rapid development (41,42). Infant nasal microbiome colonization may be influenced by many of the same environmental factors as the gut microbiome (e.g., delivery mode, feeding method, antibiotic usage, and siblings (42–44)), but it has been substantially less studied than has the gut (11,43,45). Particularly with the goal of mitigating later *S. aureus* infection risk, additional longitudinal studies in this area can determine whether microbiome changes causally precede *S. aureus* acquisition or loss and shed light on the development of the infant nasal microbiome, which is also broadly involved in later asthma, allergies, and upper respiratory tract infections (44,46–48).

In this study, we aimed to identify intra-community interactions involved in the modulation of *S. aureus* carriage *in vivo* and to characterize the development of the early infant nasal microbiome. We examined nasal microbiome development in 36 mother-infant pairs over the first year after birth, combining shotgun metagenomic profiling of longitudinal samples with monthly *S. aureus* culture-based testing. Both methods generally agreed in their prevalence of *S. aureus* carriage (per sample, 43.8% by culture vs. 38% by sequencing) and were used to define subpopulations of early-, late-, and non-*S. aureus* acquirers. In gene-content based strain profiling, infants carried strains of *S. aureus* that were more similar to maternal strains compared to unrelated mothers. Furthermore, we identified a collection of species associated with patterns of *S. aureus* carriage among infants, as well as a possible uncharacterized taxon significantly anticorrelated with *S. aureus* in infants and mothers. With the CDC reporting slowed progress on methicillin-resistant and methicillin-susceptible *S. aureus* infection reduction in hospitals and communities (5), understanding such dynamics of *S. aureus* colonization in the nasal microbiome is an important

first step towards the development of safe and mechanistically-understood ecological therapies for modulating *S. aureus* carriage.

2.3 RESULTS

Patterns of *S. aureus* carriage in a mother-infant cohort

We collected monthly longitudinal nasal swab samples from a total of 36 mother-infant pairs recruited through Sheba Medical Center as part of a larger cohort study on *S. aureus* (17), as well as *S. aureus* negative mothers enrolled specifically for this study (**Methods**). A subset of samples were selected for microbial community profiling by culture testing a total of 856 samples from mothers and infants, both throughout the first year after birth. Samples were selected for shotgun metagenomics to include early time points (i.e., soon after delivery) for all subjects, later time points (as representatives) from selected mothers, and time points immediately preceding and following changes in *S. aureus* carriage status as determined by culture test in infants (as well as additional, typically later, representative infant time points in order to assess developmental changes). This resulted in a total of 284 samples profiled metagenomically, with 208 samples passing sample quality control procedures (**Fig. 2.1**, **Methods**).

During the first year of life, infants displayed distinctive *S. aureus* carriage phenotypes. Based on culture, 20 out of 36 infants (56%) acquired *S. aureus* for the first time at or before month one ("early" acquisition), and three of these infants displayed persistent colonization, defined as being positive for at least two-thirds of all time points and for at least half of time points between months six and twelve. Of those remaining, eight infants (22%) never acquired *S. aureus* over the study period ("never" acquisition), and eight infants (22%) acquired *S. aureus* after month one, but were never persistently colonized ("late" acquisition). Among retained samples, we sequenced 71, 43,

and 30 samples from early, late, and never acquirer infants, respectively, as well as 13 samples from persistently colonized infants and 64 samples from mothers (**Fig. 2.1**).

Detection of *S. aureus* by culture and sequencing showed moderate agreement; as expected, samples that were positive for *S. aureus* by culture had a higher relative abundance of *S. aureus* by sequencing compared to culture-negative samples (mean: 13.53% vs. 1.09%; median: 0.45% vs. 0%; Wilcoxon $p<10^{-7}$ across all 208 filtered samples). Since both culture and sequencing have low rates of false positives, we believe disagreement between the two tests arises from the combination of a true positive from one test and a false negative from the other, although the mechanisms causing false negatives differ between culture and sequencing. Previous work found that the sensitivity of culture is highly positively correlated with *S. aureus* absolute abundance, which suggests that culture-based methods alone may miss a significant fraction of *S. aureus* carriage (19), in part due to idiosyncratic mechanisms such as viable but non-culturable biofilm formation (49). Conversely, any sufficiently rare microbe - including *S. aureus* - can be missed by sequencing of insufficient depth (which in sites such as the nares includes human nucleotides; **Fig. S2.1**). Due to these likely false negatives from both assays, our subsequent analyses included a composite variable for *S. aureus* positivity by culture or sequencing.

Figure 2.1: Longitudinal shotgun metagenomics of the nasal microbiome and *Staphylococcus aureus* carriage in 36 mother-infant pairs.

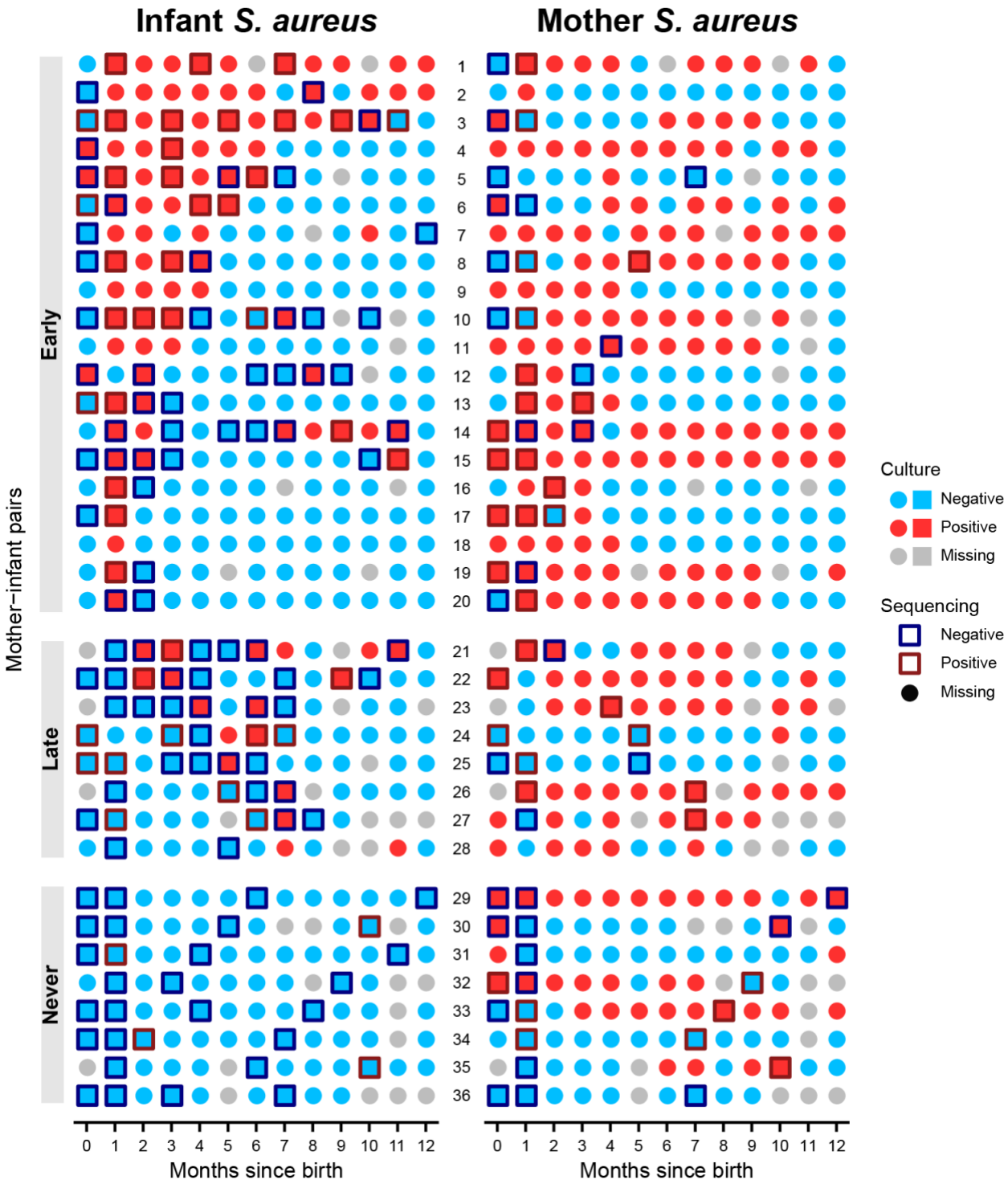


Figure 2.1 (Continued)

Nasal swabs were taken from 36 mother-infant pairs at birth and monthly for the first year after birth. Culture testing for *S. aureus* was performed on all samples to assess carriage. Culture results, red for *S. aureus* positive and blue for *S. aureus* negative, are shown as the icon fill color. A subset of 208 samples were profiled with shotgun metagenomic sequencing and retained after quality control (Methods) for further analyses. Sequencing is shown as performed only for samples that passed QC procedures. Square icons indicate samples where sequencing was performed, and the outline color, red for *S. aureus* positive and blue for *S. aureus* negative, indicates whether *S. aureus* was identified by sequencing. Infants displayed three unique patterns of *S. aureus* carriage highlighted here and used for sequenced sample prioritization: "Early" (acquisition by month one), "Late" (acquisition after month one, also typically transient), and "Never" (no acquisition during the study period).

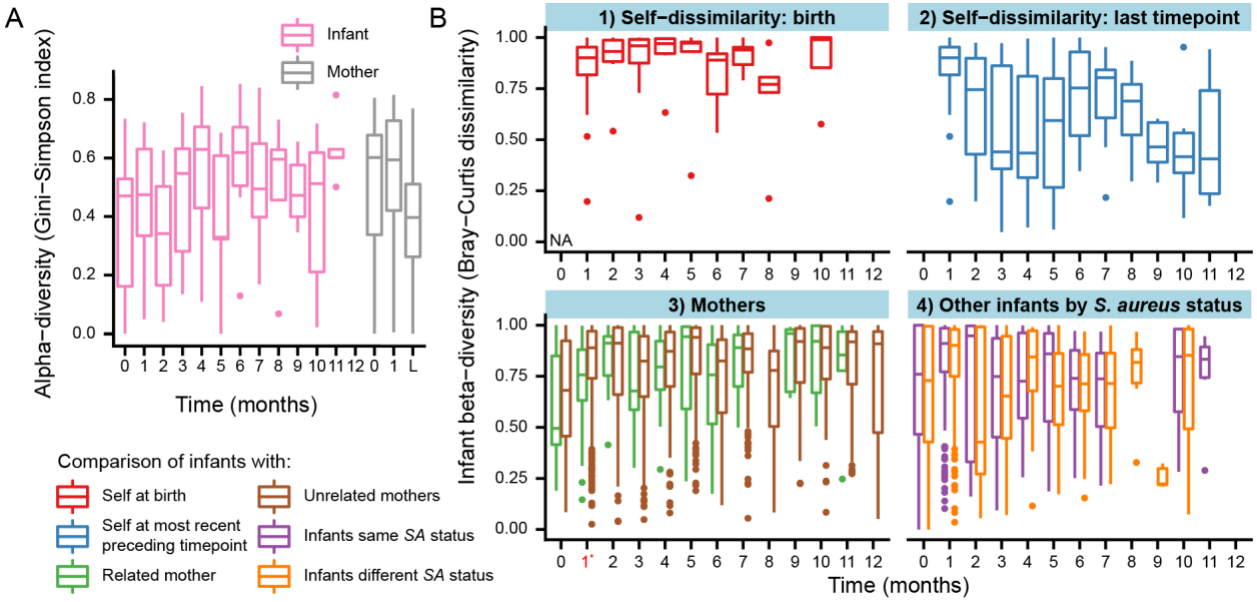
Infant nasal microbiome composition matures over the first year, but remains distinct from that of the mothers

Individual infants' species composition rapidly diverged from that observed at birth and, over time, were increasingly dissimilar to this early composition (**Fig. 2.2**). Controlling for subject, time was significantly associated with infant microbiome composition (PERMANOVA on Bray-Curtis dissimilarity, $p=0.001$). The rate of change in species composition declined over time and infants became slightly more self-conserved, consistent with stabilization toward a more mature nasal microbiome and somewhat similar to early life gut microbiome development (50) (**Fig. 2.2**). However, infant composition remained significantly different from maternal composition across all months except month eight (PERMANOVA on Bray-Curtis dissimilarity, $p<0.05$), indicating that by the end of the first year infants still did not have a fully adult nasal microbiome. These findings are consistent with those for the infant skin and gut microbiomes, which are compositionally unstable relative to adults (41,51) and increase in alpha diversity (41,51) over the first year, although reported alpha diversity trends in the nasal microbiome of healthy infants are inconsistent (44,46). In line with previous research that proposes that nasal microbiome development likely continues after year one (42), we find that infants are still significantly different from adults at the latest time points.

Infants were significantly more similar to their own mother than unrelated mothers at month one, but not at other time points (Wilcoxon-Mann-Whitney test, $p=0.0026$); this trend was generally true of early months, although it did not otherwise reach significance (**Fig. 2.2**). This supports the hypothesis that maternal influence on nasal microbiome composition mainly occurs at the earliest time points after birth. It is also consistent with studies on the infant gut microbiome, which find greater similarity between vaginally-delivered infants and their mothers shortly after birth, but that

this similarity diminishes afterwards, often quickly, based on environmental exposures (50,52–54). Lastly, the composition of the infant microbiome over time wasn't related to infant *S. aureus* status for the *S. aureus* phenotypes tested (PERMANOVA on Bray-Curtis dissimilarity, $p \geq 0.05$) (**Fig. 2.2**). This is reassuring given that omnibus testing on the full microbiome captures large differences in composition between samples, which likely would have been identified in earlier (e.g., culture-based) studies if grossly responsible for *S. aureus* phenotypes.

Figure 2.2: Development of infant nasal microbiome composition over the first year of life.



(a) Infant nasal microbiome development showed general ecological patterns similar to those of other body sites. Alpha diversity increased over time in infants, although confidence was lower at later time points due to fewer samples; it decreased slightly at later time points (L) in mothers. Mothers were sequenced at birth, month one, and one later time point (matched to a later time point for their infant, making this not the same time point for all mothers), which is labeled as "L". **(b)** Mean Bray-Curtis dissimilarity per time point comparing (1) infants to themselves at birth, (2) infants to themselves at their most recent preceding time point, (3) infants to related and unrelated mothers, (4) infants to other infants with the same or different *S. aureus* status at the same time point. In (1) the Bray-Curtis dissimilarity at birth is equal to 0 since we are comparing identical compositions and is labeled as "NA". The use of

Figure 2.2 (Continued)

the most recently recorded preceding time point for (2) is conservative because we have less frequent sampling at later time points, but still find that infants are more self-conserved at these times. *S. aureus* status as positivity by either sequencing or culture is shown here, but there were no group differences if *S. aureus* was defined using other phenotypes. In both graphs, time points with fewer than 5 data points are not displayed (see **Table S2.3** for sample sizes), and diversity measures are calculated on average subject taxonomic composition for the mothers (due to the limited sample numbers).

Species-level and functional drivers of infant *S. aureus* phenotypes

To identify if individual microbial community features correlated with *S. aureus* status, we fit linear mixed models for the association between *S. aureus* phenotype and microbial features (taxonomic and functional) for infants and mothers separately, as well as for the association between infant phenotypes and the related mother's microbial features (**Table S2.4**). Different measurements of *S. aureus* phenotype were utilized in our models to fully capture carriage patterns (**Methods**), such as the "ever" acquisition of *S. aureus* over the study period, or positivity at a single time point by sequencing, culture, or either. We found that, of the covariates tested, daycare attendance in the preceding month produced the most significant changes in the infant microbiome (**Fig. 2.3**). More specifically, daycare attendance was associated with increases in the relative abundance of the pathogens *Moraxella catarrhalis* and *Haemophilus influenzae* (**Fig. 2.3**), which are also reported in the literature (44,55,56), and linked to the development of wheezing, asthma, and respiratory infection in infants (55,57–60).

In the taxonomic subset of these linear models, abundances of infant *Acinetobacter* species, *Streptococcus parasanguinis*, *Streptococcus salivarius*, and *Veillonella* species were weakly positively associated with *S. aureus* acquisition by the next month's sample (FDR $q<0.25$, **Fig. 2.3**). These trends tended to be driven by acquisitions occurring at later time points; presence of any of the four species was more frequently observed during acquisition events occurring after the first month (Fisher's exact test, $p=0.0805$). The nasal microbiome may thus play a larger role in *S. aureus* carriage patterns as infants get older, which is concordant with a diminishing maternal and increasing environmental influence over time. However, we were limited in our cohort in the number of acquisition events occurring after the first month, causing the significance of these associations - except for *S. parasanguinis* - to be somewhat dependent on model specification (i.e.,

not always significant during robustness assessment, **Methods**, **Fig. S2.7**). Although this finding is exploratory in nature, we recommend that future studies account for early and late acquisitions events differently in order to ensure sufficient and balanced sample sizes.

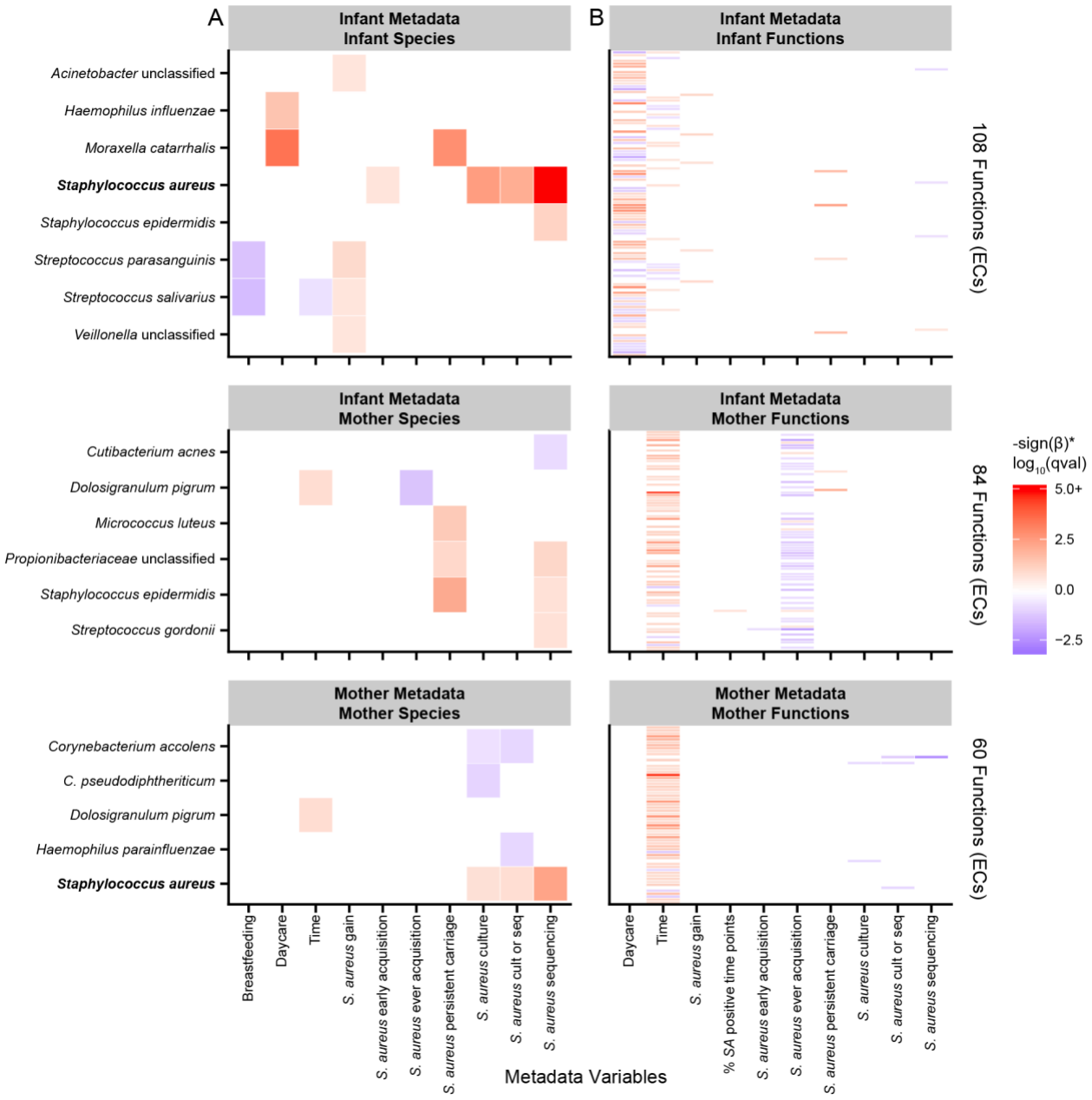
Notably, there was no overlap in the predictors of *S. aureus* in infants and mothers, which is consistent with the significant taxonomic differences between the two. However, the maternal abundance of the commensal species *Dolosigranulum pigrum* was inversely associated with infant "ever" acquisition of *S. aureus*. A number of epidemiological studies have reported an inverse association between *D. pigrum* and *S. aureus* in adults (19,36,61,62), as well as infants (63), although the potential mechanisms or causality of the relationship is unknown.

In our results, the two *Streptococcus* species positively associated with *S. aureus* gain in infants were both inversely associated with breastfeeding in the previous month (FDR $q<0.25$). These findings are consistent with previous research on the infant nasopharyngeal microbiome at six weeks of age, which reported inverse relationships between breastfeeding and *Streptococcus* OTUs (64), as well as, interestingly, inverse associations for breastfeeding with *Veillonella* species and *S. aureus* (64). Breastfeeding has been identified as a risk factor for *S. aureus* nasal carriage in infants (65) due to the presence of *S. aureus* on maternal skin and in breast milk (43), but was not associated with *S. aureus* carriage in this cohort (17). This absence of an effect may occur due to the competing pressures of increased exposure to maternal *S. aureus* during breastfeeding with protective ecological changes to the infant microbiome promoted by breastfeeding.

We next assessed microbial functions (specifically gene families grouped by Enzyme Commission number, ECs) that were contributed by diverse members of the nasal community. To avoid gene families that recapitulated individual taxa, we removed functional features for which a single

species contributed more than half the feature abundance in more than half of samples containing that feature. Among the infants, 81 ECs (of 443 total retained in infants) were associated with daycare, demonstrating the strong functional consequences of the corresponding structural rearrangement subsequent to daycare attendance. Fifty-six ECs (of 423 total in mothers) were associated with time in the mothers, indicating that the maternal nasal microbiome does change, if subtly, in the year after birth. Overall, few functions in the infant and maternal microbiomes were associated with infant *S. aureus* phenotypes: four to five ECs each (of 160 total) in the infants were associated with infant *S. aureus* positivity by sequencing, *S. aureus* gain, and persistent carriage. The only EC to be associated with a *S. aureus* phenotype in infants and mothers was EC 1.6.5.5, nominally annotated as a NADPH:quinone reductase, which was inversely associated with *S. aureus* positivity by sequencing in both populations; this feature is examined in detail below. There were 50 ECs (of 181 total) in the mothers significantly associated with infant "ever" acquisition of *S. aureus* over the study period, 44 of which were inverse associations. As was subsequently found to be linked with EC 1.6.5.5, further investigation showed that this signal was mainly driven by the maternal *D. pigrum* association with infant "ever" acquisition (**Fig. S2.2**).

Figure 2.3: Significant associations between nasal microbiome taxonomic and functional composition and subject phenotypes in feature-wise testing.



Significant associations (q<0.25) between individual taxonomic and functional features and phenotypic covariates using a MaAsLin multivariable linear model. **(a)** Of note, first attendance at daycare (during the preceding month) was an extremely strong determinant of initial microbiome colonization by several taxa (*M. catarrhalis, H. influenzae*), and the nasal microbiome exhibited a much weaker time-dependence than does the gut during infant development

Figure 2.3 (Continued)

(i.e., few taxa were consistently temporally variable between months 0 and 12). The presence of a number of oral-associated species (i.e., *Streptococcus* and *Veillonella* species) was a mild correlate of *S. aureus* gain. *S. aureus* relative abundance was included as a positive control. **(b)** After applying a species-dominance filter, we found the association of many gene families (grouped by Enzyme Commission number, ECs) in infants with daycare attendance, while many functions in mothers were associated with time and infant "ever" acquisition of *S. aureus*.

To further explore this and overall relationships between *S. aureus* and the nasal microbiome, we constructed random forest models to predict infant *S. aureus* phenotype using the taxonomic and functional composition of the infant and matched maternal microbiomes. Discriminative (e.g., random forest) rather than generative (e.g., linear) models can be used to identify a subset of features that most strongly differentiate populations with different phenotypes. In random forest models, prediction of infant *S. aureus* phenotypes using infant and maternal taxonomic profiles did not perform significantly better than chance. However, prediction accuracy improved slightly when using functional profiles; the prediction of *S. aureus* by sequencing using infant functions had an AUC of 0.748 with 95% CI [0.602,0.895], while prediction of infant "ever" acquisition of *S. aureus* using maternal functions had an AUC of 0.748 with 95% CI [0.538,0.959].

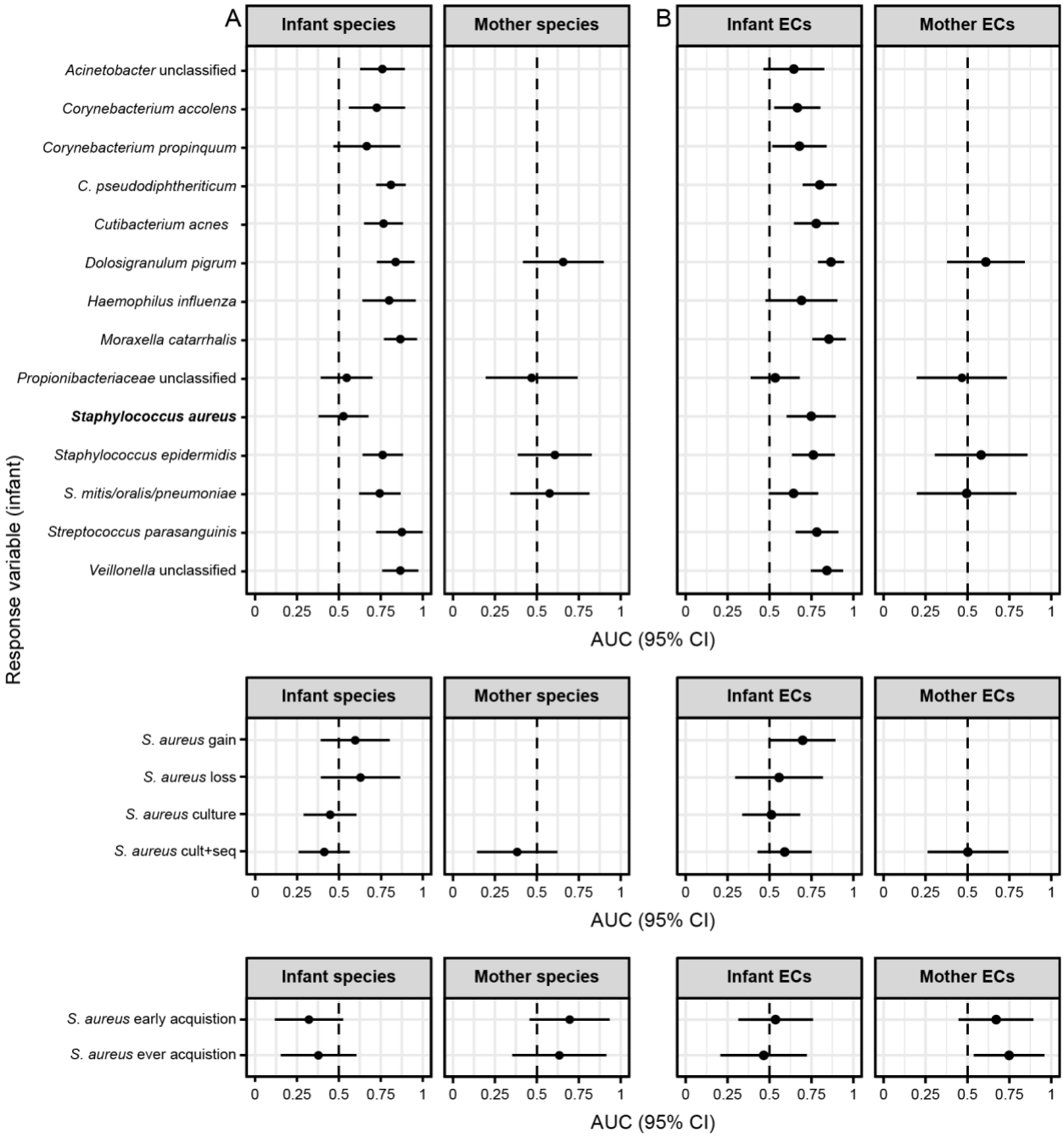
To understand why prediction improved when using functions (compared to species), we pinpointed an uncharacterized protein that proved to be associated with *D. pigrum*. The functional feature most responsible for this improvement was gene family UniRef90 X5NU12, a small 42-amino acid sequence to which taxonomy was not assigned by our analysis. However, its abundance was moderately correlated with the relative abundance of numerous other genes contributed by *D. pigrum* (1358 genes with a Spearman's rho 0.35-0.44, **Fig. S2.3**). This gene family was of major interest because it was present in about half (96 of 208) of samples and made up the majority (92%) of all sequences assigned (likely incorrectly) to EC 1.6.5.5 (NADPH:quinone reductases). Thus, the gene family summarized as EC 1.6.5.5 was the primary random forest predictor of infant *S. aureus* status by sequencing (**Fig. S2.4**), and was significantly inversely associated with *S. aureus* positivity by sequencing in both infants and mothers in linear modeling (**Fig. S2.5**). This finding was not limited to our dataset - in healthy adults in the second phase Human Microbiome Project (HMP1-II) (66), UniRef90 X5NU12 was present in about half of anterior nares samples (131 of

251), and its relative abundance was significantly inversely correlated with *S. aureus* presence by sequencing (Wilcoxon $p<10^{-5}$ across 251 samples). A BLAST (67) search of reads mapping to this gene family revealed many significant alignments with 16S rRNA gene sequences nominally assigned to species as diverse as *Streptococcus*, *Staphylococcus*, *Lactobacillus, Cutibacterium,* and *Bifidobacterium*. Given these diverse results, closer inspection of the UniRef90 X5NU12 family suggests that it may be a misannotated fragment of the highly conserved 16S rRNA gene, with sequences matching these reads from previous skin and nasal studies potentially misannotated to a variety of closely related taxa (including *D. pigrum*). Thus, UniRef90 X5NU12 may be acting as a taxonomic marker for a closely related, uncharacterized species co-occurring or sharing genetic similarity with *D. pigrum* and competing with *S. aureus*. Meanwhile, many of the maternal ECs associated with infant *S. aureus* "ever" acquisition in the linear models contributed somewhat to the predictive power of the random forest for this variable, but none were highly dominant predictors like UniRef90 X5NU12 (**Fig. S2.6**).

Random forest classifiers using infant species-level taxonomic and functional profiles were also able to successfully predict the presence/absence of a number of other taxa in the infant microbiome, including several potential pathogens and important commensals (**Fig. 2.4**). In total, 11 out of 13 of these models had AUCs that were significantly above 0.5 when using species as predictors (**Fig. 2.4**), and 9 out of 13 models had AUCs significantly above 0.5 when using functions (**Fig. 2.4**, bootstrap 95% CI > 0.5). Overall, the maternal microbiome was not predictive of the infant microbiome, with only one of the models achieving significance. This is consistent with our earlier analysis (i.e., **Fig. 2.2**), as the random forest models are built using time-matched mother-infant samples from all time points, and mothers and their infants were significantly more similar at only one of 13 time points (and even then were still quite different from each other).

Many species in the nasal microbiome appear to be interdependent, but in complex ways, thereby allowing moderate prediction by random forest models; however, *S. aureus* itself appears to be a more isolated member of the ecological community of the nose.

Figure 2.4: Ability of infant or maternal microbiome profiles to predict infant microbiome membership.



Both infant and maternal **(a)** species-level relative abundance profiles, and **(b)** functional profiles (ECs) were used as predictors in random forests to infer (1) presence/absence of other individual species by sequencing, (2) subject-

Figure 2.4 (Continued)

varying *S. aureus* variables, and (3) subject-fixed *S. aureus* culture phenotypes in infants. When predicting properties of a given species (presence/absence or any *S. aureus*-derived property), we **(a)** removed the species from the abundance table, or **(b)** removed any ECs contributed to by the species and renormalized all samples to 100% prior to model fitting to avoid circularity. The infant microbiome exhibited reasonably strong within-ecosystem cohesion (i.e., predictability), but essentially none from the maternal to infant microbiome; conversely, only certain *S. aureus* carriage phenotypes were well-predicted by this model.

Gene-level strain profiles reveal similarity between mother and infant *S. aureus*

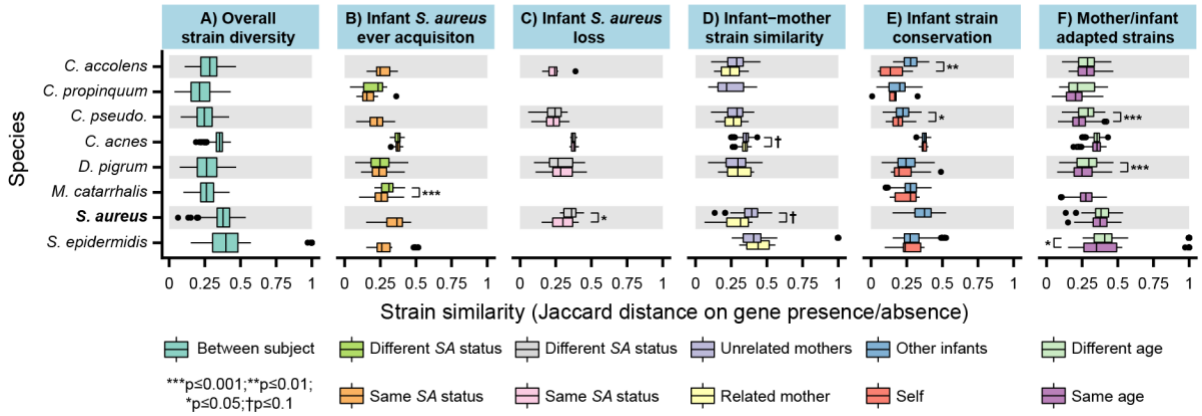
To finally explore strain-level trends in *S. aureus* acquisition and carriage and put these in context relative to other common nasal species, we compared species-level gene content (i.e., strain-specific accessory genome carriage) across samples (**Fig. 2.5**). Given the limited sequencing depth available after human nucleotide depletion, strains were characterized based on the presence/absence of gene families in each species' reference-based pan-genome (68,69) rather than by calling individual nucleotide variants. Per species, we calculated the average Jaccard distance between accessory gene profiles in different subjects to quantify genetic diversity for that species in our population. We also compared the average Jaccard distance between sample pairs matched on a metadata variable of interest (e.g., mother-infant pair) to that between discordant sample pairs. For example, one test (**Fig. 2.5B**) considered whether infants were ever positive for *S. aureus* over the study period and compared strain similarities (averaged over subject pairs) of concordant samples (i.e., both infants "ever" or "never" acquired *S. aureus*) compared to discordant samples (i.e., one infant "ever" acquired and one "never" acquired *S. aureus*).

Of the eight species with sufficient coverage to profile strains, overall population mean genetic diversity/distance was highest in *S. epidermidis* and *S. aureus*, and lowest in *Corynebacterium propinquum* and *Corynebacterium pseudodiphtheriticum* (**Fig. 2.5A**). For many of the species, gene content was more similar within an infant over time than compared to other infants (**Fig. 2.5E**), indicating stability of the strains over time (mirroring the behavior of the adult gut (66,70,71)), and this reached significance for *C. accolens* and *C. pseudodiphtheriticum.* Interestingly, individuals with the same infant/mother status (i.e., both mothers, or both infants) had more similar gene content for *C. pseudodiphtheriticum*, *D. pigrum*, and *S. epidermidis* than

individuals with differing infant/mother status (**Fig. 2.5F**), suggesting that these species may have strains adapted to microbiome maturity or be transmitted within age groups.

Infants carried strains of *S. aureus* that were somewhat more similar to those of their own mothers than to unrelated mothers (**Fig. 2.5D**) or other infants (p=0.0520). This is consistent with findings from the parent cohort study that infants often carried maternal strains (17), although we use different methods for defining similarity and, due to undersampling in our data and the high specificity of our accessory genome barcoding, infants rarely had identical strains to their mother. Interestingly, this similarity of the infant strain to the maternal strain was more pronounced in *S. aureus* compared to other species. Furthermore, infants who were positive for *S. aureus* by culture and were concordant on loss status (i.e., either both lost or did not lose *S. aureus*) had *S. aureus* with more similar gene content compared to individuals discordant on loss status (p=0.0359) (**Fig. 2.5C**), suggesting some strains may be more ecologically stable, i.e. more difficult to lose once acquired. Strain-level variation in other species may also be related to *S. aureus* status; for instance, we found that infants with the same *S. aureus* "ever" acquisition status had very significantly more similar strains of *M. catarrhalis* ($p<10^{-4}$) (**Fig. 2.5B**).

Figure 2.5: Gene-content based strain profiling for *S. aureus* and other common nasal species.



To perform gene-content based strain profiling for each species, pairwise Jaccard distances were calculated between samples on accessory gene presence/absence and averaged over same-subject pairs. Species/distance metric combinations with fewer than three subject pairs formed from six fully unique subjects are not displayed (see **Table S2.3** for sample sizes). Significance testing was performed with a one-sided Wilcoxon-Mann-Whitney test.

2.4 DISCUSSION

In this study, we performed the first shotgun metagenomic analysis of the infant nasal microbiome, with a particular focus on *S. aureus* acquisition and retention. Using multiple measures of *S. aureus* status and maternally paired, longitudinal microbiome data, we sought to identify microbiome determinants temporally preceding changes in infant *S. aureus* status, and to investigate the influence of developmental time and the maternal microbiome on the early infant microbiome. The infant nasal microbiome was highly variable over the first year and was primarily shaped by developmental and external factors (e.g., environmental exposures such as daycare). Using both linear and random forest models, as well as replication in healthy adults in the HMP1-II dataset, we found evidence of a gene family, possibly representative of an uncharacterized taxon, that was associated with *D. pigrum* and significantly anticorrelated *S. aureus*.

Generally, the taxonomic composition of our samples matched that seen in previous studies (42,72) with *Moraxella*, *Streptococcus*, *Haemophilus*, *Staphylococcus*, *Corynebacterium* and *Dolosigranulum* making up 71.78% of the total abundance of infant samples. In contrast to some previous studies, we identified *Cutibacterium* (previously *Propionibacterium*) as a major genus dominating the infant nasal microbiome in the first year of life. This difference is likely due to previous studies often sequencing hypervariable region 4 of the 16S ribosomal RNA (rRNA) gene, which has been shown to severely underestimate *Propionibacterium* abundance (73). After *Dolosigranulum pigrum* (30.60%), *Cutibacterium acnes* had the second highest mean abundance among infants (18.44%). In a linear model, birth time points were significantly ($q<0.25$) associated with *C. acnes* and *Propionibacteriaceae* unclassified (possibly *C. acnes* that could not be identified due to low sequencing read counts), which together made up 60.13% of the relative abundance of samples at birth. *D. pigrum* tended to dominate subsequently, and later (non-birth) time points were associated with significant elevations of *D. pigrum*, *C. pseudodiphtheriticum*, *M. catarrhalis*, and *S. tigurinus* ($q<0.25$). Previous research has suggested that interaction with *C. acnes* can increase *S. aureus* virulence (74), helping *S. aureus* to invade the human host (74). Similarly, the *C. acnes*-produced small molecule coproporphyrin III can promote *S. aureus* biofilm formation, potentially improving its survival in challenging environments (75). In our study, just over half of infants gained *S. aureus* at or before month one; in contrast to more complex *D. pigrum*-dominated microbiome compositions seen later in the first year, this low-diversity *C. acnes* dominated nasal microbiome at birth may fail to prevent or even promote *S. aureus* acquisition.

Our findings on infant nasal microbiome development further elucidate results from the parent cohort study, which tracked *S. aureus* phenotypes using cultured isolates and defined strains by pulsed field gel electrophoresis (PFGE) and *spa* typing (rather than sequencing) (17). In the full

cohort, it was found that 61.8% (n=76) of the infants carried *S. aureus* at month one, with the majority (72%, n=52) carrying the same strain as the mother by PFGE/*spa* (17). Our results suggest the mother is an important influence on the infant nasal microbiome at and before month one, which could allow for the sharing of both commensal nasal species, as well as *S. aureus* and other pathogens. Our strain definition via sequencing also suggests that infant strains are often genomically similar, but not identical, to the most abundant strain in the mother. In the larger cohort, we found that early acquisition (by month one) of *S. aureus* resulted in a longer length of carriage (17). Here, in the early months, full metagenomic profiling made it clear that the nasal microbiome is undeveloped with low species alpha diversity and high *C. acnes*, which may allow *S. aureus* transmitted at this time to gain a strong foothold.

Although *S. aureus* transmitted by the mother may be present very early in life, the infant nasal microbiome and immune system undergo major changes throughout the first year of life that likely act to exclude *S. aureus*, resulting in declining rates of carriage in the year after birth. Consistent with studies of the gut microbiome, in which maternal influence after birth on vaginally-delivered infants is quite strong for a short period of time, but then quickly fades (50,52–54) we found that infants were not more similar to their own mother at later time points, and that their taxonomic profiles overall rapidly diverged from the earliest time points. In contrast, environmental factors, especially daycare, strongly shaped the nasal microbiome in the first year, as it was associated with the acquisition of new species, and possible strain replacement for existing species.

Daycare attendance in particular was one of the largest effects in our study and the parent study, both on *S. aureus* specifically and the nasal microbiome generally. In the larger cohort, daycare was significantly negatively associated with persistent *S. aureus* carriage (17), and, in our subset, none of the three persistent carriers attended daycare at any point. Consistent with previous

research (44,55,56), daycare attendance was marked by a large increase in the relative abundance of *M. catarrhalis* (**Fig. 2.3**). The likely inverse relationship between *M. catarrhalis* and *S. aureus* (76–78) and the acquisition of *M. catarrhalis* at daycare then explains why daycare attendance may reduce persistent carriage. Surprisingly, we also found that the three persistent carriers had higher *M. catarrhalis* relative abundance than other daycare non-attendees (**Fig 2.3**). While this is not inherently contradictory, since none of these associations must be transitive *per se*, it may be explained by external confounding factors such as contact with siblings, since the presence of siblings is a risk factor for both *M. catarrhalis* (44,79) and *S. aureus* colonization (65), or by random chance given the small sample size. However, there may also be a more interesting biological explanation, such as combinatorial interactions with additional species like *H. influenzae*, which is also acquired with daycare attendance, or strain differences in *M. catarrhalis*, as we did find a highly significant association of *M. catarrhalis* strain type with *S. aureus* "ever" acquisition (**Fig. 2.5**). Similar to *Corynebacterium* species, in which studies have reported inconsistent associations with *S. aureus* due to the possible complexity of these interactions (11), further *in vitro* work exploring the circumstances under which *M. catarrhalis* inhibits *S. aureus* would be valuable.

One of the other most striking observations in our data was the presence of a gene family, correlated with *D. pigrum* genomic content, that was strongly co-exclusionary with *S. aureus* carriage. Due to its similarity with 16S rRNA gene sequences, this gene family likely represents a species not closely homologous to current reference isolates that competes with *S. aureus*. Its correlation with *D. pigrum* in our dataset suggests that it may either occupy a related phylogenetic space, or that these two species may act as co-colonizers. Furthermore, previous epidemiological studies have reported inverse associations between *D. pigrum* and *S. aureus* in adults (19,36,61,62)

and infants (63), and identified diverse biosynthetic gene clusters in *D. pigrum* believed to be involved in its in vitro inhibition of *S. aureus* (62). In one study of hospitalized neonates, cases (those acquiring *S. aureus*) had lower *D. pigrum* relative abundance compared to matched controls seven days prior to *S. aureus* acquisition (63), and a recent study found that 10 of 10 diverse *D. pigrum* isolates inhibited *S. aureus* growth in vitro (62). The mechanisms through which *D. pigrum* inhibited *S. aureus* in this experiment were not fully clear, but one hypothesis explored by the authors was an enrichment of biosynthetic gene clusters (BGCs) in *D. pigrum*, with strains containing both a lanthipeptide BGC and a bacteriocin inducing the greatest inhibition (62). Our finding of the inverse association between maternal *D. pigrum* and infant "ever" acquisition of *S. aureus*, however, is the first to suggest that *D. pigrum* is one pathway through which the maternal microbiome influences infant *S. aureus* carriage. From our results alone, it is not clear whether *D. pigrum* prevents the acquisition of maternal S. aureus that could be passed to the infant, or whether maternal *D. pigrum* is passed to the infant where it is protective against *S. aureus* acquisition, and in either case whether *D. pigrum* itself or an uncharacterized relative are responsible for the additional sequences co-exclusionary with *S. aureus*.

Due to the cost of metagenomic sequencing and the complexity of *S. aureus* carriage, our study has some limitations. Firstly, our data is collected monthly, which may be too coarse a temporal resolution for understanding the full relationship between *S. aureus* status and the rapidly-developing infant microbiome. Secondarily, we are limited by the size of our sample, and this is ultimately an exploratory analysis whose findings should be tested in larger cohorts. In particular, this creates challenges if early and late *S. aureus* acquisition events indeed are driven by different influences (e.g., maternal *S. aureus* carriage vs. the establishing infant nasal microbiome), as this creates a mixed signal with an already limited sample size and is worthwhile to consider for future

study design. Additionally, using metagenomics for skin habitats is challenging due to the high degree of human contamination, and we were not able to perform nucleotide-level strain analysis given relatively low sample read depths. Lastly, *S. aureus* carriage is a complex phenomenon, with ecological and microbiological variation undoubtedly contributed by unmeasured factors (e.g., immune development, contact with other caregivers and relatives, diet, etc.), and therefore microbiome contributions need to be large to be detectable. Particularly during highly dynamic infant colonization, a wide range of environmental influences (in addition to direct maternal effects) are likely to influence microbiome composition. These effects are further confounded since the mother's nasal (and other) microbiome composition may in turn be influenced by some of these same factors due to shared genetics, cohabitation, lifestyle factors, and/or transmission of strains with other children, all of which have been demonstrated for the gut microbiome (52,80–86). Through an integrative microbiome epidemiology study design, we have provided new insights into the development of the infant nasal microbiome and its role in *S. aureus* carriage that suggest further targets of investigation for eventual ecological therapies to control *S. aureus* carriage.

2.5 CONCLUSIONS

In this study, we characterized the development of the infant nasal microbiome, which was highly variable, weakly influenced by the maternal nasal microbiome composition, and strongly shaped by daycare attendance. Mothers thus represented a sporadic early source for *S. aureus* transmission to the naïve infant microbiome, but microbiome determinants became more important later on. We gained a better understanding of the role of the microbiome in *S. aureus* carriage through the identification of a specific protein family that was highly predictive of infant *S. aureus* status, significantly anticorrelated with S. *aureus* positivity in both infants and mothers, and which

ecologically interacts with the commensal species *D. pigrum*. We determined that this (misannotated) protein family was a non-protein-coding sequence acting as a phylogenetic marker of a likely novel bacterial species. An inverse relationship between *D. pigrum* and *S. aureus* has been a main result of multiple prior 16S rRNA amplicon studies; however, using metagenomic sequencing, we were able to differentiate this novel species from *D. pigrum* (which it often co-occurs with) and have shown that it is more predictive of *S. aureus* presence than is *D. pigrum* itself (with our result thus representing a likely mechanistic "driver" rather than previously identified "passengers"). Furthermore, this novel taxon was sufficiently prevalent in adults and infants to drive widespread patterns of *S. aureus* carriage. These findings were not limited to our cohort: a similar prevalence and a strong anti-correlation with *S. aureus* were also found for this novel clade in geographically distinct adults from the Human Microbiome Project. Our study provides an improved understanding of how the infant nasal microbiome develops in early life, and how it can act to promote or exclude *S. aureus* colonization.

2.6 METHODS

Cohort and experimental design

Subjects consisted of 36 mother-infant pairs recruited through Sheba Medical Center as part of a larger cohort study on *S. aureus* (17), as well as *S. aureus* negative mothers enrolled specifically for this study. More specifically, subjects were recruited from among pregnant women who attended the Sheba Medical Center obstetrics monitoring unit between 2013 to 2016, were at least 34 weeks pregnant, visited the monitoring unit during screening hours (held for three hours per week), and provided consent to participate. Exclusion criteria included delivery prior to 34 weeks of pregnancy and failure to provide consent to participate. Nasal swabs and subject covariates were

collected from mothers and infants within 48 hours of birth and then monthly for the first year after birth. At each time point, two swabs (one for *S. aureus* culture testing and one for DNA extraction and sequencing) were collected from both nostrils. The covariates included: infant sex, maternal age, delivery mode, daycare attendance in the past month, antibiotic use in the past month, and breastfeeding in the past month (**Table S2.1**). Culture testing for *S. aureus* (detailed below) was performed for all non-missing samples (n=856), and 33.2% of samples (n=284) were sequenced with shotgun metagenomic sequencing (also below, and see **Fig. 2.1**). As much as possible, shotgun sequenced samples were selected to target birth and the first month for all subjects, *S. aureus* gain and loss events in infants, and extended periods of *S. aureus* carriage or non-carriage in infants.

S. aureus culture testing

Nasal screening was performed using a cotton-tipped swab placed in Amies transport media (Copan innovation, Brescia, Italy). Swabs were streaked on CHROMagar *S. aureus* plates (HiLabs, Rehovot, Israel) within 24 hours and incubated for 24-48h at 35^{o}C. Catalase and Staphylase (PASTOREX® STAPH-PLUS, BioRad, Marnes-la-Coquette, France) were performed on suspected colonies to conclusively identify them as *S. aureus*.

DNA extraction and sequencing

DNA extraction was performed with the QIAGEN DNeasy PowerLyzer PowerSoil Kit with a bead beating protocol. After extraction, DNA concentration was measured by NanoDrop and frozen at -20°C until being shipped to the United States for sequencing. Prior to starting the project, the DNA extraction method was validated by performing extractions on double-distilled water (DDW) samples and confirming that no DNA concentration was measured by NanoDrop and that the

displayed band was a true band and not due to contamination. Negative controls were not additionally sequenced since specimen DNA concentrations were measured after extraction to be reliably high and differentiated from potential background contaminants (due to the otherwise substantial added cost).

Whole genome fragment libraries were prepared as follows at the Broad Institute (87). Metagenomic DNA samples were quantified by Quant-iT PicoGreen dsDNA Assay (Life Technologies) and normalized to a concentration of 50pg/uL. Illumina sequencing libraries were prepared from 100-250pg of DNA using the Nextera XT DNA Library Preparation kit (Illumina) according to the manufacturer's recommended protocol, with reaction volumes scaled accordingly. Prior to sequencing, libraries were pooled by collecting equal volumes (200 nl) of each library from batches of 96 samples. Insert sizes and concentrations for each pooled library were determined using an Agilent Bioanalyzer DNA 1000 kit (Agilent Technologies). Libraries were sequenced on HiSeq 2x101 to yield ~10 million paired end reads. Post-sequencing de-multiplexing and generation of BAM and Fastq files were generated using the Picard suite.

S. aureus phenotype definitions

The *S. aureus* phenotype of a sample was defined using multiple measures to capture different dimensions of *S. aureus* carriage (**Table S2.4**). Subject-fixed *S. aureus* phenotypes were defined strictly by the *S. aureus* culture data, and included *S. aureus* early acquisition, late acquisition, "ever" acquisition, persistent carriage, the percent of positive time points, and the time point of first *S. aureus* acquisition. *S. aureus* early acquisition applied only to infants, and was defined as first acquisition of *S. aureus* by or at the first month. *S. aureus* late acquisition also applied only to infants, and was defined as first acquisition of *S. aureus* after month one. If culture results for

birth were missing, early and late acquisition were calculated using month one data only. *S. aureus* "ever" acquisition for both mothers and infants was defined as ever being positive for *S. aureus* by culture over the study period. To match definitions in the larger cohort study (17), infant persistent carriage was defined as being positive for at least two-thirds of non-missing time points, and being positive for at least half of non-missing time points between months six and twelve. For mothers persistent carriage was defined as being positive for at least two-thirds of non-missing time points. For infants and mothers, the percent of positive time points was defined as the number of positive time points divided by the number of non-missing time points for the subject over the study period. For infants, the time point of first *S. aureus* acquisition was defined as the first time point for which the subject tested positive for *S. aureus* or NA for subjects who never tested positive.

Subject-varying *S. aureus* phenotypes included positivity by culture or sequencing or either, gain of *S. aureus* by the next month among culture-negative samples, and loss of *S. aureus* by the next month among culture-positive samples. Positivity by sequencing was defined as having a non-zero *S. aureus* relative abundance. Positivity by either was defined as being positive by culture or sequencing or both. Due to the false negative rate from culture, the culture data was smoothed according to the following procedure: (1) negative or missing data points that were preceded and followed by positive data points were treated as positive, and (2) missing data points that were preceded and followed by negative data points were treated as negative. Using the smoothed culture dataset, *S. aureus* gain by the next month was defined as changing from negative by culture at month t, to positive by culture at month t+1. *S. aureus* loss by the next month was defined as changing from positive by culture at month t, to negative by culture at both months t+1 and t+2. Figure 1 shows the raw, unsmoothed data.

Data processing and quality control

Three sequenced samples failed our sample tracking quality control and were excluded from further analysis. Samples were retained in the analysis if they contained more than 5×10^4 sequenced read pairs (mean 9.47×10^5 read pairs per sample), and were not taxonomically 100% unclassified (detailed below). This cut-point was selected *a priori* based on approximately 1x coverage of an *E. coli* equivalent genome size. Of 284 sequenced samples, a total of 208 samples (144 from infants and 64 from mothers) were retained for further analyses.

Taxonomic and functional profiles were generated using the bioBakery meta'omics workflow v0.9.0 (88). Briefly, reads mapping to the human genome were first filtered out using KneadData v0.7.0 with default parameters. Taxonomic profiles of shotgun metagenomes were generated using MetaPhlAn2 v2.6.0, and functional profiling was performed by HUMAnN2 v0.11.0. In taxonomic profiles, relative abundances were given at the species-level or, if unidentified, at the genus-level or family-level. Phage abundance was originally included in the taxonomic profiles, but due to the lack of any significant associations ($q<0.25$) in linear models, phage abundance was normalized out and subsequent analyses focused purely on non-viral taxa. In functional profiles, per-sample gene abundances were grouped by Enzyme Commission (EC) number using the HUMAnN2 utility script.

Due to the presence of quality-controlled DNA that didn't map, we treated the unmapped sample mass like a microbial feature (i.e., rescaled the taxonomic profiles by the percent mapped reads and included a feature for the percent unmapped reads) and assessed the robustness of our results using these new profiles. The percent unmapped reads was strongly correlated with human contamination (Spearman rho = 0.69, $p<10^{-15}$ across all 208 filtered samples) and didn't change

our main findings from linear models (**Fig. 2.3, Fig. S2.7**), other than reducing power (due to noise introduction and re-removal) and causing some weaker associations to drop out. We concluded that much of this unmapped sample mass represented mainly cryptic human contamination and that our initial approach of focusing purely on mapped reads and adjusting linear models for the proportion of human contamination was robust.

Statistical methods for assessing infant developmental trends

Alpha and beta diversity were calculated using the Simpson index, and Bray-Curtis dissimilarity, respectively, in the *vegan* R package v2.5.4. For mothers, all diversity metrics were calculated on average per-subject taxonomic composition (due to limited sample sizes). In comparisons between infants with different *S. aureus* statuses, *S. aureus* relative abundance was normalized out prior to the calculation of diversity metrics. Significance testing was performed using a one-sided Wilcoxon-Mann-Whitney test. Any time points with fewer than five data points were not included in this analysis (see **Table S2.3** for sample sizes). PERMANOVA was performed using the *adonis* function within the *vegan* R package. We tested the following variables using blocked PERMANOVA (87): *S. aureus* early acquisition, late acquisition, "ever" acquisition, and persistent carriage. Within-subject permutations were used to test the effect of time (month). We used repeated cross-sectional PERMANOVAs to test: *S. aureus* positivity by culture, sequencing or either, and infant vs. mother status.

Feature filtering

Prior to the creation of linear and random forest models, we filtered taxonomic and functional features to retain analyzable subsets of these features passing quality control. All filtering was applied separately to infant and mother samples due to the differences in microbiome composition,

and all cut-points for filtering were selected *a priori* based on best practices from similar studies. First, all feature abundances below 0.1% (taxonomy) or 0.05% (function) abundance in more than 95% of samples were aggregated to a single "other" category to avoid testing features that were too uncommon to possibly detect an effect. Next, unless otherwise noted, microbiome features trivially related to the outcome variable of interest were dropped and the remaining per-sample abundance was renormalized to 100%. For example, *S. aureus* abundance was removed from infant taxonomic profiles in this way prior to the prediction of infant *S. aureus*-related variables in the random forest models. ECs were considered related to the outcome of interest if they contained a stratification for the outcome species in any infant *or* mother sample. Lastly, we applied a filter to remove functions that were primarily contributed by a single species to avoid re-analyzing functions that essentially recapitulated the genomic abundances of a single dominant carrying species. Specifically, if a single species (excluding unclassified) contributed more than 50% of a functional feature's abundance in more than 50% of samples containing that feature, it was aggregated to an "other" category as above. Functions contributed by unclassified species were identified through the HUMAnN2 translated search (but not MetaPhlAn2) and were retained because they weren't previously captured in our taxonomic analysis.

Linear models

Linear mixed effects models to identify significant associations between microbiome features, and covariates were created in MaAsLin2 v0.2.2. Briefly, log transformed taxonomic and functional feature abundances were modeled as outcomes of a function of covariates, including a random effect for the subject and fixed effects for the *S. aureus* phenotype, time point (month), proportion of human contamination (defined as one minus the proportion of reads after vs. before human depletion), and total read count. We created models combining covariates and outcome features

for (1) infant features and infant metadata, (2) mother features and infant metadata, and (3) mother features and mother metadata. Therefore, each class of model was defined by the choice of microbiome feature type (taxonomic or functional), *S. aureus* phenotype, and the three feature/metadata configurations listed above (**Fig. 2.3**). Within each unique combination of these levels, FDR correction was applied per variable using Benjamini-Hochberg, and results with $q<0.25$ are displayed (**Fig. 2.3**).

Models were fit for both subject-fixed and subject-varying *S. aureus* phenotypes. For subject-fixed phenotypes, such as "ever" acquisition of *S. aureus*, all samples were given the subjects' phenotype value and included in the analysis. For subject-varying *S. aureus* phenotypes, such as sample positivity by sequencing, each sample with a non-missing value for the phenotype variable was included in the analysis. For *S. aureus* gain and loss, the analysis was subset to samples that were negative or positive for *S. aureus* by culture, respectively.

Other recorded exposures or environmental factors, such as antibiotic usage or delivery mode, were also evaluated as covariates potentially contributing to the relationship between microbiome features and the *S. aureus* phenotype. For mothers and infants separately, we included covariates in the model when (1) the covariate was significantly ($q<0.25$) associated with at least one species in a univariate prescreen, and (2) the covariate was at present in at least 10% of samples (including those that were missing covariate data). Furthermore, models for maternal microbiome features were not adjusted for variables considered primarily related to infants, such as breastfeeding and daycare. These criteria were decided upon *a priori*. The rationale for (2) was that these covariates were not able to be tested in our dataset because they were too rare to be individually significant. Based on these criteria, models for infant microbiome features were also adjusted for breastfeeding and daycare in the preceding month, while models for maternal microbiome features were not

adjusted for any additional covariates. The displayed results (**Fig. 2.3**) for non-*S. aureus* metadata are from a model containing all covariates except a *S. aureus* phenotype.

Random forest models

Random forest classifiers were created using the *randomForest* R package v4.6.14. Across models, the predictors consisted of filtered infant and maternal taxonomic and functional features. The binary response variable was defined as the infant *S. aureus* phenotype or the presence/absence of a species in the infant nasal microbiome by sequencing (the species tested were those that passed the prevalence/abundance filter). We did not study response variables for the mothers due to the lower sample size. As described above, features were normalized out of the predictor dataset if they were related to the outcome variable in models constructed with infant predictors/infant outcomes.

The *randomForest* function was run with default parameters, with the exception of downsampling to the size of the smaller class to handle unbalancedness in the data. To conservatively address subject-level effects, cross-validation was performed per subject: iterating through each subject, a test dataset containing all samples from the same subject was created, the model was trained on data from all other subjects, and then the outcome(s) for the excluded individual was predicted. To find 95% confidence intervals, the standard error was calculated through per-subject bootstrapping, with 1000 bootstrap samples used. Random forest models were only run if they had at least 20 samples in each class (i.e., a sufficient size for meaningful cross-validation) (see **Table S2.5** for sample sizes).

Strain-level profiling

Given the limited sequencing depth available after human nucleotide depletion, strain profiles were generated using PanPhlAn v1.2.2, which characterizes strains based on the presence/absence across gene families from across each species' reference-based pan-genome (68).

To create Figure 5, per-species gene presence/absence matrices were first filtered to remove genes present in less than 5% of samples for which PanPhlAn was run. The *vegdist* function in the *vegan* R package v2.5.4 was used to calculate pairwise-Jaccard distances between samples based on a metadata variable of interest. For each comparison in Figure 5, pairwise distances were averaged over same subject pairs to avoid undue influence by subjects with more sequenced time points. Comparisons were displayed if they had data from at least three unique pairs consisting of six unique subjects. This was to avoid bias that could occur if many pairs were created with the same subject and this subject happened to have unusual data. Significance testing was performed using a one-sided Wilcoxon-Mann-Whitney test. Due to low sample read depths, underdetection of genes is likely; however, the underdetection should not be differential with respect to the metadata variable of interest, and therefore the significance testing is unbiased.

2.7 DECLARATIONS

Ethics approval and consent to participate

Human patient research was approved by the Sheba Medical Center Institutional Review Board (protocol number: 5534-08-SMC). Each woman gave written informed consent for her and her newborn's participation and the other parent gave a non-written approval. Data analysis activities were conducted at the Harvard T.H. Chan School of Public Health, and approved by the Harvard T.H. Chan School of Public Health Institutional Review Board (protocol number: IRB13-1664). All experimental methods were in compliance with the Helsinki Declaration.

Availability of data and materials

The metagenomic sequences generated and analyzed during the current study are available in the NCBI Sequence Read Archive (SRA) repository as BioProject PRJNA610982 (89). The subject metadata and the processed taxonomic and functional datasets generated and analyzed during the current study are included in the published article (https://doi.org/10.1186/s13059-020-02209-7; Supplemental tables 6, 7 and 8, respectively). All software used in this study is free and open source.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (grants R21AI112991 to CH and T32AI007535 to EA), the Chief Scientist, Ministry of Health, Israel (grant 3-00000-5622 to GRY), and the Israel Science Foundation (grants 1590/09, and 1658/15 to GRY). The funding bodies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

EKA, EAF, TH, RJC, ML, GRY, and CH designed the research. TH, AMM, HJ, ARM, and CD generated data. EKA, EAF, and MK analyzed data. EAF, ML, GRY, and CH provided project oversight. EKA, EAF, ML, GRY, and CH wrote the paper. All authors approved the manuscript.

Acknowledgements

The authors express their gratitude to all participants for generously providing their time and samples.

Chapter 3. Detecting and quantifying mediation of phenotypes by microbial communities

3.1 ABSTRACT

Changes to the microbiome are increasingly shown to be an important intermediate step on the pathway between health exposures and the development of both chronic and infectious diseases. For human microbial communities, statistical mediation analysis aims to calculate the proportion of the effect of an exposure on an outcome that occurs through changes to the microbiome composition or function, compared to the proportion that goes through all other biological pathways. Explorations into the microbiome as a mediator can identify new biological mechanisms and suggest novel opportunities for therapeutic intervention; although, it is unknown what statistical methods perform best for complex microbiome data, which are zero-inflated, compositional, and high-dimensional. Using realistic simulated data, we compared the performance of statistical mediation methods designed for low-dimensional settings; high-dimensional, non-compositional settings; and the microbiome specifically. We make recommendations for methods to use for estimation of the total indirect effect, and hypothesis testing for total indirect and component indirect effects. However, given the respective strengths and weaknesses of the methods, an end-user may need to select multiple methods to accomplish a full mediation analysis with microbiome data. This can lead to multiple varying effect estimates that are difficult to interpret and more work is needed to continue to improve method usability and to provide tools for interpretation for end-users.

Keywords: human microbiome, benchmarking, mediation analysis, microbiome epidemiology, metagenomics

3.2 INTRODUCTION

Advances in high-throughput sequencing technologies have revolutionized the study of microbial communities. Consisting of thousands of species and their numerous enzymes, the human microbiome contains vast biochemical capabilities that continuously and bidirectionally interact with host systems. Changes to the microbiome are increasingly shown to be an important intermediate step on the pathway between health exposures and the development of both chronic and infectious diseases (90–97). Such investigations into the microbiome as a mediating variable can elucidate previously unknown biological mechanisms and suggest new avenues for interventions, as the microbiome can be modified through tools including diet, prebiotic and probiotic supplements, fecal microbiota transplantation, and antibiotics (98–100).

For human microbial communities, statistical mediation analysis aims to calculate the proportion of the effect of an exposure on an outcome that occurs through changes to microbiome composition or function (called the mediated or indirect effect). This quantity can then be compared to the proportion of the effect that goes through all other biological pathways (called the direct effect). For instance, in real world examples below we explore the effect of diet, including a Mediterranean dietary pattern, on cardiometabolic disease (CMD) risk. Diet represents one of the largest effects on the human gut microbiome (101) and alters the make-up of the gut microbiome by providing microbes in the gut with substrates (102). These alterations to the gut microbiome have downstream effects on gut permeability, the production of biologically active metabolites, and the host immune system. In this context, the indirect or mediated effect would capture - among others - the pathway through which low saturated fat in the Mediterranean diet selects against certain lipopolysaccharide(LPS)-producing bacteria in the gut microbiome. By lowering circulating levels of immunogenic LPS, the Mediterranean diet reduces the chronic subclinical inflammation characteristic of CMD (103). Indirect effects would also include pathways through which

microbiome alterations influence gut-derived microbial metabolites involved in CMD risk, including short chain fatty acids (102), trimethylamine (102), secondary bile acids (102), and imidazole propionate (101). In contrast, a direct effect (that doesn't involve microbiome mechanisms) includes reduced direct stimulation of TLR4 receptors by saturated fat, which also reduces chronic subclinical inflammation (103).

As sequencing costs fall, statistical mediation analysis will be able to be more broadly applied to microbiome datasets to identify and prioritize mediators for further, more labor-intensive, experimental manipulations; however, datasets must have certain properties for statistical mediation analysis to be advisable. First, there should be a strong total effect of the treatment on the outcome. In theory, it's possible that the direct and indirect effects cancel, such that there is still a mediated effect despite no detectable total effect of the treatment on the outcome; however, given the challenges of working with microbiome data, detecting this scenario is unlikely. It is important that the microbiome is measured after the exposure and before the outcome, or - if not - that we have reason to believe that a measurement reflects the measurement at the desired time point. Use of a randomized exposure is ideal to limit confounding, although it will not address confounding of the microbiome-outcome relationship, including by other microbiome features (in the case that mediation analysis is performed feature-by-feature).

For mediation to occur, an exposure must alter microbiome composition and/or function with these changes then affecting the outcome. Therefore, it's important to consider which scale will capture the biologically relevant microbiome alterations and to ensure that data is measured on this scale (or is convertible to this scale). The microbiome can be described by the prevalence of microbes, their relative or absolute quantities, or their functional potential (across different taxonomic and functional groupings). It can also be described by diversity metrics calculated on these values, or

- with the addition of metatranscriptomics and metabolomics - by the expression of certain genes or levels of microbial metabolites. For example, a randomized clinical trial of patients with type 2 diabetes (T2D) found that a high fiber diet increased the abundance of butyrate-producing microbes (although it didn't increase gut microbiome diversity) leading to improvements in clinical markers (101). In contrast, a healthy gut microbiome may provide colonization resistance against pathogens (104), and gut microbiome diversity specifically has been shown to be an important mediator between cancer treatment and infectious complications in patients receiving allogeneic HSCT (97). Modifications to the microbiome can take different forms and thus the proper scale for measurement must be informed by biological knowledge.

Understanding how alterations to microbiome composition and/or function link exposures to disease can improve understanding of biological mechanisms of action and suggest modification of the microbiome as a novel therapy; however, the unique characteristics of microbiome data present potential challenges for performing statistical mediation analyses. Microbiome data are compositional with the features measured as parts of a whole. Compositionality induces correlations between the features, rendering many conventional statistical analyses inappropriate (105,106). Secondly, microbiome data are sparse (i.e., zero-inflated), and may have up to 70% zero values (107). These zeros can arise due to the true biological absence of a feature, or a technical failure to detect a feature during the sequencing step (107). Lastly, microbiome data are high-dimensional with many independent variables (microbiome features) relative to the number of samples. This is beyond the scope of many standard mediation techniques, which were developed for a single mediator or a small number of mediators (108). In particular, when analyzing microbiome data, we may be interested in identifying both the total indirect effect

through all microbiome features, as well as component indirect effects through individual microbiome features (**Fig. 3.1A**).

In the last few years, multiple extensions of mediation analysis have been proposed for the microbiome specifically (109–111). However, it is not clear if they outperform non-microbiome-specific methods, and what the trade-offs are in terms of interpretability and computational complexity. In this study, we performed extensive simulations with realistic synthetic microbiome datasets to assess the performance of mediation methods designed for (a) low-dimensional settings, (b) high-dimensional, non-compositional settings (e.g., epigenetic mediators), and (c) the microbiome specifically, which is both high-dimensional and compositional. We aim to make these methods more accessible to researchers by providing guidance on method selection given the properties of a microbiome dataset.

3.3 RESULTS

Simulation methodology

To benchmark methods for statistical mediation analysis with a microbiome mediator, we evaluated method performance in realistic synthetic microbiome datasets (**Fig. 3.1B**). Briefly, SparseDOSSA (112), a tool from the Huttenhower lab that represents the null distributions of microbial features using a Bayesian hierarchical model, was used to generate realistic metagenomic count data and "spike" in a known association with a binary treatment variable. SparseDOSSA was calibrated to generate microbiome data resembling that of the healthy adult gut (113). A linear model was used to create outcome data by combining information on a binary treatment variable, the direct effect of the treatment, the absolute abundance of the mediating features (as modeled in SparseDOSSA), and the strength of the association between mediating

features and the outcome. This model assumed no treatment-mediator and no mediator-mediator interaction on the outcome. We then simulated the sequencing process (see **Methods**) to transform the absolute abundance data from SparseDOSSA into compositional data that the mediation methods were evaluated on (**Fig. 3.1B**). In addition to creating realistic metagenomic data, this approach had the advantage of testing methods on data simulated independently of their original modeling framework.

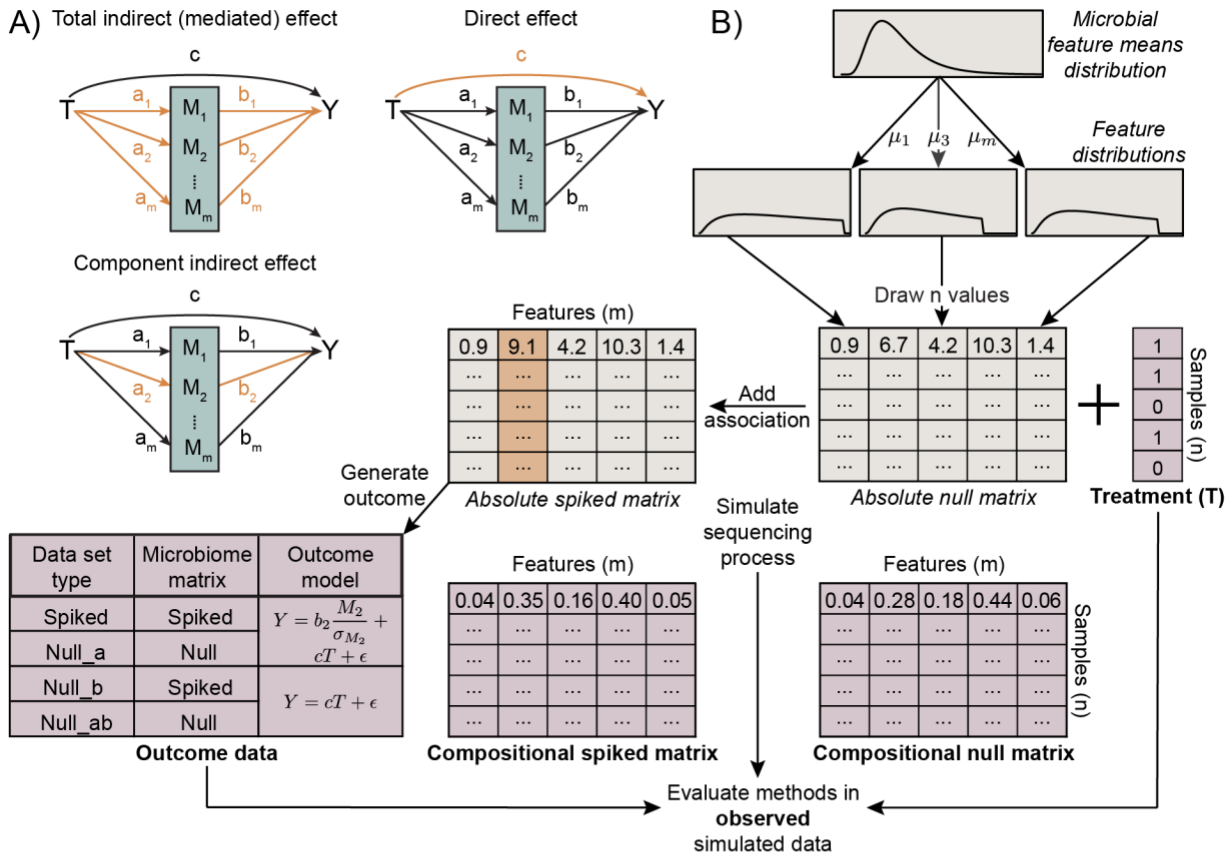
Across simulations, we varied the sample size (50, 100, 200, 400), the number of microbiome features acting as mediators (5, 10, 20), the size of the total indirect effect (0.5, 1, 1.5), the percent of the total effect that was mediated (25%, 50%, 100%) and the direction of mediation (negative or positive) across realistic values (**Table S3.1**). The average read depth (50,000) and the number of total microbiome features (100) were kept constant across simulations. For each parameter set, 300 iterations of the simulation were run. Component indirect effect p-values were FDR corrected at the 0.25 level. Prior to evaluating methods on the simulated microbiome data, low prevalence and abundance features were filtered (**Methods**). The methods were evaluated across three desired areas of functionality (**Fig. 3.1C**): hypothesis testing for the total indirect effect (**Fig. 3.2**), hypothesis testing for component indirect effects (**Fig. 3.3**), and total indirect effect estimation (**Fig. 3.4**). Methods were excluded from these evaluations if they did not offer the functionality being evaluated (**Fig. 3.1C**) or due to excessive runtimes, resulting in a different set of methods included in each evaluation.

For each simulated dataset with a mediating effect added ("spiked"), we created three null versions of the dataset using the same set of selected microbiome features: one in which the treatment had no effect on the set of microbiome features, but the features still affected the outcome ("null_a"), one in which the treatment still had an effect on the set of microbiome features, but the features

did not affect the outcome ("null_b"), and one in which no relationships were added between the set of microbiome features and the treatment or outcome ("null_ab").

We compared statistical mediation methods designed for low-dimensional settings (the "naive" method); high-dimensional, non-compositional settings (including HDMA (114), HIMA (115) and principal components regression, PCR); and the microbiome (including CCMM (109), MedTest (111), and SparseMCMM (110)). The naive method was a simplistic extension of the traditional product-of-coefficients method popularized by Baron and Kenny (116) (**Methods**). With the exception of MedTest, these methods fit two classes of models; one class captured the effect of the treatment on the microbiome features, and the other captured the effect of the microbiome features on the outcome. They used a variety of statistical tools to address compositionality (such as log ratio transformations, Dirichlet regression), and high-dimensionality (such as feature pre-selection, regularization and penalization, the transformation of microbiome features into principal components, and the use of distance metrics). The specific way each method addresses these challenges is summarized in **Table S3.2**.

Figure 3.1: Overview of simulation methodology.



Method	CCMM	MedTest	Sparse MCMM	HDMA	HIMA	PCR	Naïve
Total indirect effect hypothesis test	✓	✓	✓	✗	✗	✓	✓
Component indirect effect hypothesis test	✓	✗	✓	✓	✓	✗	✓
Total indirect effect estimation	✓	✗	✓	✓	✓	✓	✓

(a) The total indirect (mediated) effect was defined as the effect of the treatment on the outcome through all microbiome features. The component indirect (mediated) effect was defined as the effect of the treatment on the outcome through a specific microbiome feature. The direct effect was defined as the effect of the treatment on the outcome through all other non-microbiome pathways. Throughout the paper, we refer to the effect of the treatment on the microbiome feature (a's) and the effect of the microbiome features on the outcome (b's) as pathway coefficients.

(b) We used SparseDOSSA, a tool from the Huttenhower lab that represents the marginal distributions of microbial

Figure 3.1 (Continued)

features as truncated, zero-inflated, log-normal distributions, to generate realistic metagenomic count data. Simulated microbiome data representing absolute abundances was used to create simulated outcome data, while methods were evaluated on simulated microbiome data representing relative abundances. See **Methods** for a detailed description. **(c)** We compared three microbiome-specific methods (CCMM, MedTest, and SparseMCMM), three methods designed for high-dimensional mediation more generally (HDMA, HIMA, and PCR), and a naive extension of the traditional product method ("naive"). We evaluated the performance of these methods across three areas of functionality desired in a mediation model: total indirect effect estimation, hypothesis testing of the total indirect effect, and hypothesis testing of the component indirect effects. For these areas of functionality, a green check mark in the table indicates that the method performs the function. Although CCMM returns estimates of and p-values for component indirect effects, it does not provide the pathway coefficients (a's and b's).

True positive rate and false positive rates for total and component indirect effects

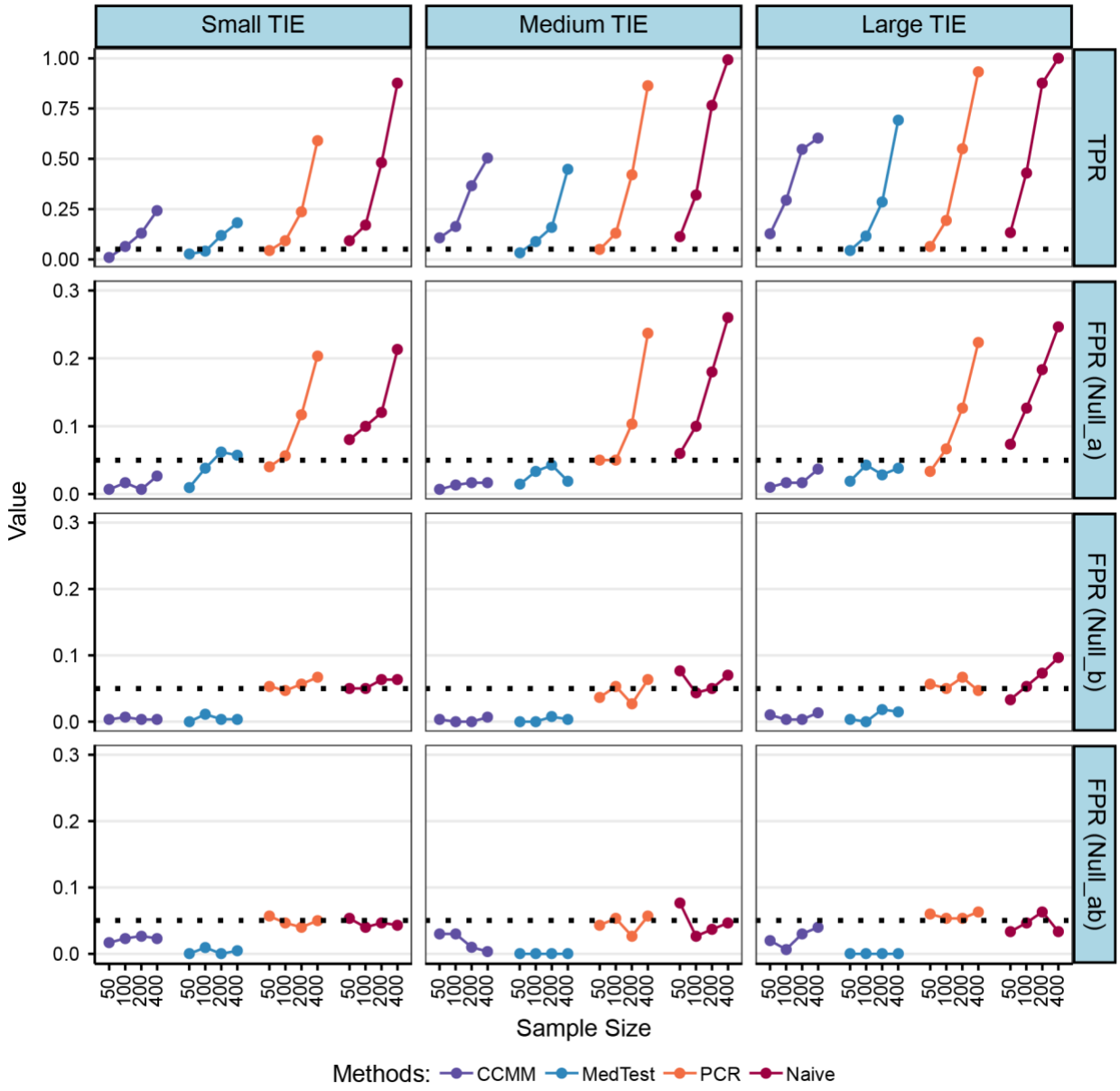
We first assessed whether the hypothesis test for each method could identify whether mediation was occurring. We calculated the true positive rate at both the level of the dataset (for the total indirect effect) and at the level of the feature (for the component indirect effects) using the "spiked" dataset. For the total indirect effect, a p-value < 0.05 was considered significant and for component indirect effects a FDR corrected $q < 0.25$ was considered significant. Since the three types of null datasets ("null_a", "null_b", "null_ab") varied in their difficulty, we present three versions of the false positive rate. A fourth false positive rate is presented for component indirect effects using the non-mediating features in datasets with a mediating effect added ("spiked").

For the total indirect effect (**Fig. 3.2**), the naive method showed the highest true positive rate at each sample size and total indirect effect size; however, it also displayed an elevated false positive rate, especially in the case of "null_a" datasets. For "null_a" datasets, this elevated false positive rate worsened as the sample size increased. Upon investigation, this occurred due to the failure of the naive method to appropriately account for compositionality in the microbiome. In particular, the addition of a spiked effect between the treatment and some microbiome features also induced a relationship between the treatment and non-spiked microbiome features due to the compositional nature of the data. In contrast, this didn't occur when an effect was added between microbiome features and the outcome because effects were added independently in the linear outcome model (see **Methods**). Due to its high true positive rate, especially compared to CCMM and MedTest, the naive method may still be of interest in an exploratory analysis where false positive findings are less problematic. We also explored different ways to perform the permutation-based hypothesis test for the naive method (**Fig. S3.1)**, which resulted in trade-offs between the true positive rate

and the false positive rate for “null_a” datasets. PCR displayed similar, although less extreme, patterns.

Among the methods without an inflated false positive rate for “null_a” datasets, CCMM tended to show a higher true positive rate than MedTest with the exception of at a large indirect effect size with 400 samples. At all combinations of sample size, total indirect effect size, and false positive rate, CCMM remained below 5% false positive findings. MedTest remained below 5% false positive findings except at a small total indirect effect size with 200 and 400 samples.

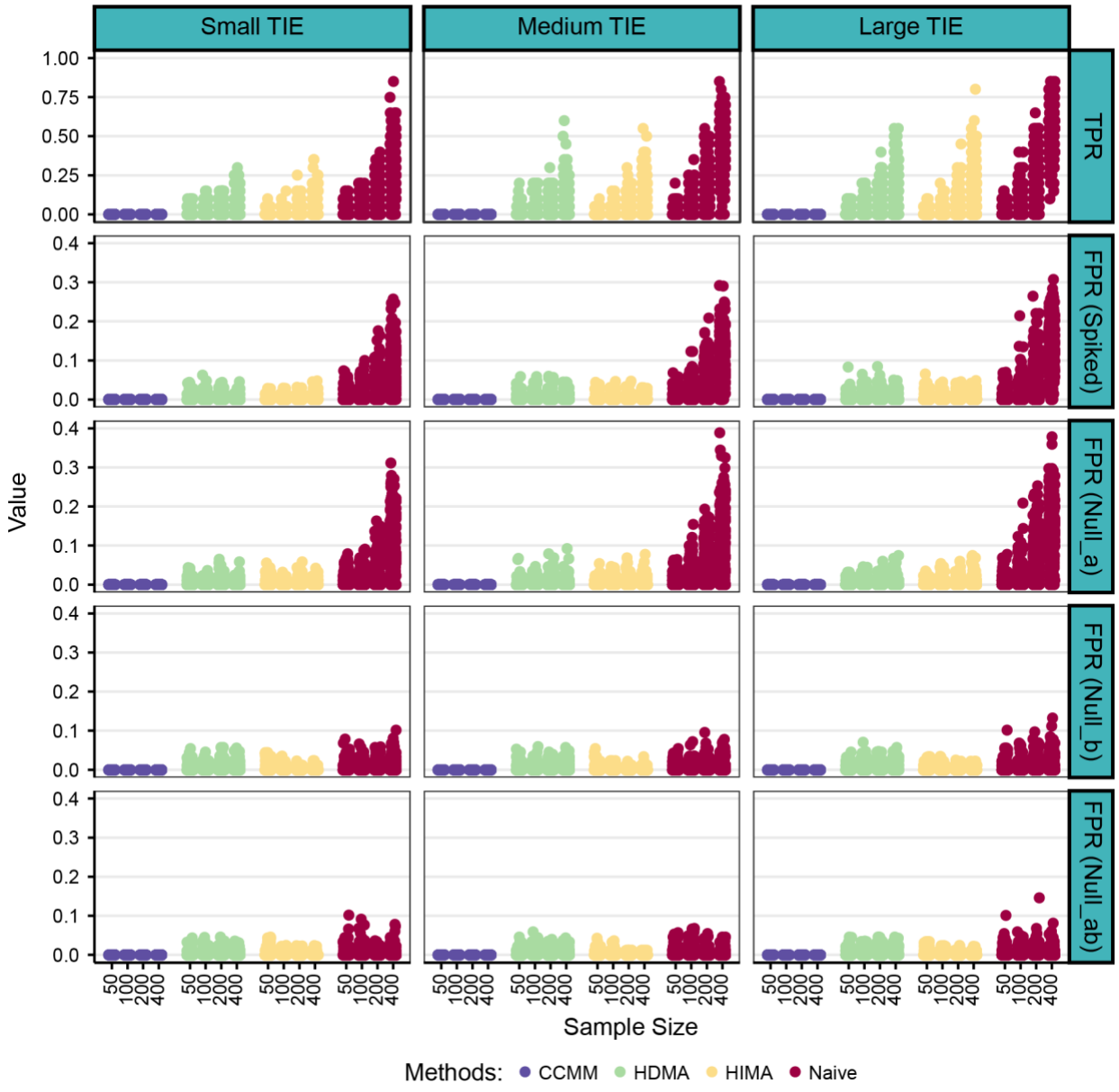
Figure 3.2: True positive and false positive rates for detection of the total indirect effect.



There are no boxplots for HDMA or HIMA because these methods do not provide a hypothesis test for the total indirect effect. SparseMCMM was excluded due to runtime issues. The simulation was run 300 times for each set of simulation parameters. The small, medium, and large total indirect effect sizes were 0.5, 1, and 1.5, respectively. Other simulation parameters besides sample size and the total indirect effect size were fixed at 20 mediating features, 100% of effect mediated, 100 total microbiome features, and a positive direction of mediation with both a positive effect of the treatment on the mediating features and a positive effect of the mediating features on the outcome ("+,+").

Similar to previous results, the naive method displayed the highest true positive rate at each combination of sample size and total indirect effect size for the detection of component indirect effects (**Fig. 3.3**). It also displayed an elevated false positive rate in the case of "null_a" and "spiked" datasets, which worsened as the sample size increased and, within sample sizes, worsened as the total indirect effect size increased. Among the methods without an overly inflated false positive rate for the "null_a" or "spiked" datasets, HDMA and HIMA performed similarly and showed the second highest true positive rate at each total indirect effect size and sample size. HDMA had a very slightly higher true positive rate than HIMA except at sample sizes of 400.

Figure 3.3: True positive and false positive rates for detection of component indirect effects.



There are no boxplots for MedTest or PCR because these methods do not provide a hypothesis test for component indirect effects. SparseMCMM was excluded due to runtime issues. The simulation was run 300 times for each set of simulation parameters. The small, medium, and large total indirect effect sizes were 0.5, 1, and 1.5, respectively. Other simulation parameters besides sample size and the total indirect effect size were fixed at 20 mediating features, 100% of effect mediated, 100 total microbiome features, and a positive direction of mediation with both a positive effect of

Figure 3.3 (Continued)

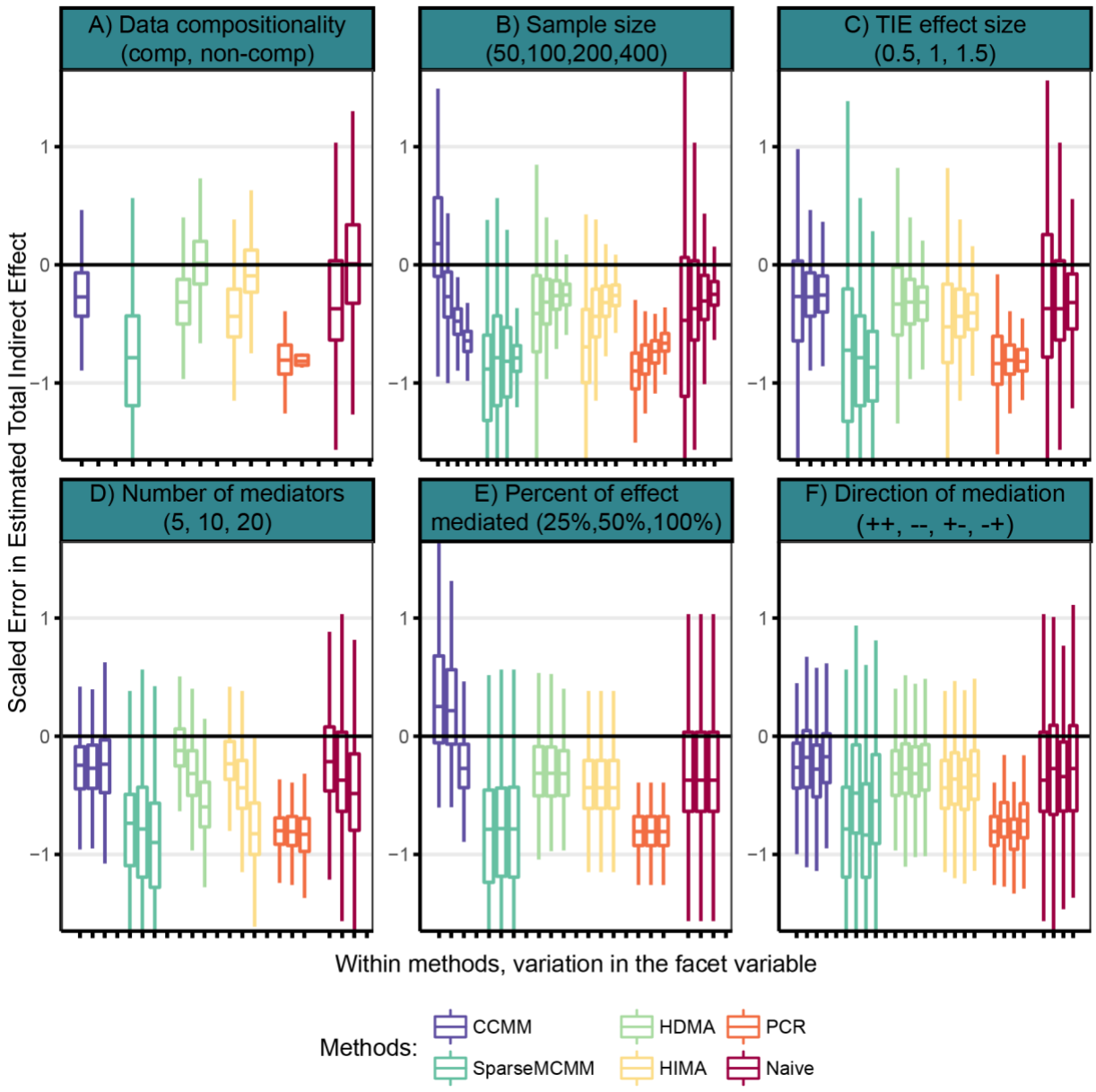
the treatment on the mediating features and a positive effect of the mediating features on the outcome ("+,+"). P-values were FDR corrected for multiple testing with $q < 0.25$ considered significant. For datasets with a mediating effect added, the true positive rate was defined as the proportion of true mediating features identified as significant by the method. For all datasets (including those with a spiked mediating effect), the false positive rate was defined as the number of non-mediating features identified as significant by the method.

Compositionality caused irrecoverable information loss

Next we explored the effect of microbiome data compositionality on effect estimation (**Fig. 3.4A**) and found that if the effect of interest was on the absolute abundance scale the full effect could not be retrieved from compositional data. To create non-compositional data, the sequencing process was not simulated and the methods were evaluated on the absolute abundance matrices from SparseDOSSA (**Fig. 3.1B**, "absolute spiked matrix" and "absolute null matrix"). Being microbiome-specific, both CCMM and SparseMCMM expected compositional data and could not be evaluated. Values are shown as scaled errors (i.e., the difference between the true and estimated effect size divided by the true effect size) to keep the same scale while allowing the total indirect effect size to change and to account for slight variation in the spiked effect size due to the complexity of the spike procedure (see **Methods**). Values above and below zero represented overestimates and underestimates of the true spiked effect, respectively.

For the non-compositional data, both HDMA and the naive method were centered on the true indirect effect, with HIMA having a scaled error slightly below zero. The scaled errors for all three methods were shifted substantially lower for compositional data indicating that the compositionality in the data caused a loss of the spiked effect that the methods couldn't correct for. Although PCR was centered around the same scaled error for both non-compositional and compositional data, it consistently underestimated the true total indirect effect in both cases. This suggests that if the effect of interest is on the absolute abundance scale even the best-designed methods couldn't retrieve the full effect from compositional data and that it may be important to use more absolute measures of microbiome data, such as qPCR, for mediation analyses to avoid underestimating the total indirect effect.

Figure 3.4: Estimation of the total indirect effect size.



Across panels, one simulation parameter was varied while the other simulation parameters were fixed at the baseline values of: compositional microbiome data, sample size of 100, medium total indirect effect size, 10 mediating features, 100% of total effect mediated, and a positive direction of mediation with both a positive effect of the treatment on the mediating features and a positive effect of the mediating features on the outcome ("+,+"). To account for slight variation in the spiked total indirect effect size (see **Methods**), values are presented as the difference between the true

Figure 3.4 (Continued)

and estimated effect size divided by the true effect size. Therefore, a value above zero represents an overestimate of the true spiked effect, while a value below zero represents an underestimate. Non-compositional data was generated by not simulating the sequencing process and evaluating the methods on the absolute abundance matrices (**Fig. 3.1B**).

Non-microbiome specific methods performed well for effect size estimation

As the sample size increased from 50 to 400, the variability in the estimates of the total indirect effect shrunk, and HDMA, HIMA, and the naive method all converged towards a value slightly below the truth, while CCMM, SparseMCMM, and PCR returned more substantial underestimates (**Fig. 3.4B**). At a sample size of 400, HDMA, HIMA and naive returned values that were closest to the true total indirect effect. At a sample size of 50 almost all methods displayed extreme variability, except PCR which consistently underestimated the total indirect effect at all sample sizes. For this evaluation the total number of microbiome features was fixed at 100 and CCMM tended to overestimate the total indirect effect at 50 samples; this was part of a larger trend where CCMM performed poorly when the number of overall microbiome features was higher than the number of samples (**Fig. S3.2**). As the total indirect effect size was varied across values representing a small, medium and large total indirect effect (**Fig. 3.4C**), the median scaled error for the methods remained approximately constant, but the variability in the scaled error decreased.

We also varied the number of mediating features while keeping the total indirect effect size fixed (i.e., comparing simulated datasets with a few strong mediating features to those with many mediating features each with only a small effect) (**Fig. 3.4D**). As the number of mediators increased, method performance worsened for most methods, especially HDMA, HIMA, and the naive method. For many weak mediators, which is a likely scenario in true biological data, CCMM was the best method to use. HDMA and HIMA showed the largest declines in performance; this occurred because they pre-selected features based on the marginal correlation of the features with the outcome (which lessened as the number of features increased) and only calculated the total indirect effect for those features.

Next the size of the total indirect effect was fixed, but the proportion of the total effect that was mediated was varied from 25% to 50% to 100% by changing the direct effect (**Fig. 3.4E**). Of the six methods, only CCMM was substantially affected by this manipulation and tended to overestimate the total indirect effect at 25% and 50% mediation of the total effect.

Lastly, the direction of mediation was varied, where a positive mediated effect meant that increasing the treatment increased the outcome via the mediating microbiome features, and vice versa for a negative direction of mediation (**Fig. 3.4F**). Each direction of mediation, positive and negative, could be achieved in two ways. For example, a positive direction of mediation occurred when the treatment had a positive effect on the mediating microbiome features, and these features in turn had a positive effect on the outcome ("+,+"), or when both these relationships were negative ("-,-"). Similarly, a negative direction of mediation occurred when the effect of the treatment on the mediating microbiome features went in the opposite direction of the effect of the mediating microbiome features on the outcome (either "+,-" or "-,+"). Overall, this manipulation did not have a large effect on method performance. The median scaled error for the methods was slightly larger when the effect of the treatment on the microbiome was positive (i.e., "+,+" or "+,-"). This difference likely occurred because in the simulation (and in real world microbiome data) microbial abundance can be increased infinitely in a positive direction, but cannot take on negative values.

We further investigated whether basic modifications to HIMA and the naive method could improve their estimation of the total indirect effect. In particular, we used the HIMA procedure to perform variable selection on the microbiome features, but estimated the regression coefficients for the effect of the microbiome features on the outcome using a non-penalized model. For the naive method, we used HIMA to pre-select the features and then performed the naive procedure. Therefore, the difference between the two modified methods was whether they fit a joint or marginal model for the outcome. Neither modification resulted in noticeable differences in total indirect effect size estimation compared to their base procedure (**Fig. S3.3**).

Despite being simplistic, the naive method performed much better than expected for total indirect effect estimation, although it tended to return more variable estimates compared to HDMA and HIMA. The naive method extends the traditional product method by fitting marginal models for the effect of the treatment on each microbiome feature, and the effect of each microbiome feature - controlling for the treatment - on the outcome. Since each outcome model for the naive method only includes a single feature, this approach will be biased if the microbiome features aren't independent conditional on the treatment variable, and even if they are independent it may be less efficient. However, because we don't use some form of feature pre-selection or penalization, a single outcome model containing all features (which would be unbiased) would be unstable because the number of features is high relative to the sample size. In the simulation, we did not specifically add correlations between the microbiome features, but these are induced due to compositionality. Despite these flaws, the naive method still performed well and, due to its simplicity, provided highly interpretable effect estimates.

In conclusion, for hypothesis testing of indirect effects, we recommend the naive method for an exploratory analysis where an inflated false positive rate may be more acceptable. Otherwise, we

recommend CCMM for testing the total indirect effect and HDMA or HIMA for testing component indirect effects. While not designed for microbiome data, HDMA, HIMA, and the naive method performed well overall at estimating the total indirect effect. When there were many weak mediations, a situation that is plausible in real world data, CCMM was the best-performing method. This suggests that an end-user who wanted to perform all three areas of functionality may need to employ multiple different methods throughout their analysis.

Vignettes: diet and cardiometabolic disease

Based on the above evaluations, we selected five methods (CCMM, MedTest, HDMA, HIMA, and naive) and applied them to two real-world datasets to explore the role of the gut microbiome in mediating the effects of diet on cardiometabolic disease (CMD) risk. This allowed us to examine the performance of the methods with real-world data.

The first dataset ("MLVS") included a subset of samples from the Men's Lifestyle and Validation Study (117). Diet was measured over up to 26 years with semi-quantitative food frequency questionnaires. Using the cumulative average diet over this time period, an index measuring adherence to the Mediterranean dietary pattern was created (117). We dichotomized the index at the median to create a binary exposure variable. The outcome was a composite score of blood biomarkers capturing the mechanisms of CVD and T2D pathogenesis (117). For the second dataset ("Blueberry RCT"), we used a subset of samples from a randomized controlled trial of blueberry polyphenol supplementation over six months (unpublished). The exposure was random assignment to daily supplementation with blueberry polyphenol powder or an isocaloric placebo powder. The outcome was change in cyclic GMP, a secondary messenger in the cardiovascular system, between baseline and month six. Samples were sequenced with shotgun sequencing and taxonomic profiles

generated with the bioBakery meta'omics workflow (88). We opted to perform mediation analysis at the level of genus. Prior to applying the methods, low prevalence and abundance features were filtered out (**Methods**). Additionally, zeros values were replaced with a small non-zero value prior to running CCMM (**Methods**).

Both datasets displayed a similar pattern where, within datasets, all methods tended to agree on the total effect, but CCMM and naive attributed most of the total effect to mediation, while HIMA and HDMA attributed it primarily to the direct effect. This difference occurred because HDMA and HIMA remove features that have a weak marginal effect on the outcome and only use the remaining features when calculating the total indirect effect. In this case, the estimates from CCMM and the naive method are likely more reliable, as HDMA and HIMA tended to substantially underestimate the total mediated effect in the case of weak mediators (**Fig. 3.4D**). Both of the estimated total effects went in the correct direction, where increased adherence to a Mediterranean dietary pattern was associated with a reduced CMD biomarker score and consumption of the blueberry polyphenol supplement was associated with increases in beneficial cGMP at six months.

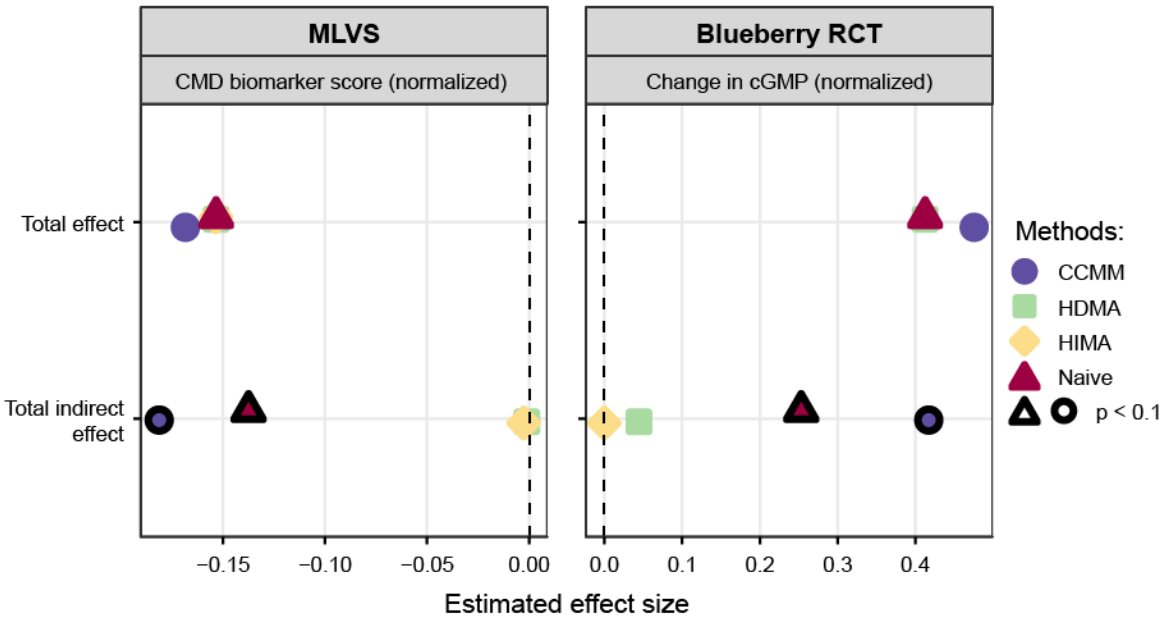
While no component indirect effects were significant in the Blueberry RCT dataset, the MLVS dataset had two component indirect effects that were significant at $q < 0.25$, *Acidaminococcus* and *Lactobacillus* (**Fig. S3.4**). The component indirect effect for *Acidaminococcus* was negative meaning that above median adherence to a Mediterranean dietary pattern was associated with a decreased CMD score via *Acidaminococcus*. In particular, we found that above median adherence to a Mediterranean dietary pattern decreased *Acidaminococcus* relative abundance, which then decreased CMD score, associations which seem biologically plausible based on the literature (118,119). As expected, *Lactobacillus* also displayed a negative component indirect effect;

however, the two associations making up the component indirect effect both went in the opposite direction of that expected (**Fig. S3.4**). The example of *Lactobacillus* shows why it's important to consider biology when interpreting such effects and how reporting the pathway coefficients (**Fig. 3.1A**), which not all methods do, can help with these interpretations.

In the original analysis of the MLVS dataset (117), the Mediterranean diet was found to interact with *Prevotella copri* carriage. In particular, the Mediterranean diet was associated with a significantly lower myocardial infarction risk in non-carriers of *P. copri*, but associated with a non-significant increase in myocardial infarction risk in carriers of *P. copri.* The epidemiological concept of interaction, in which the effect of an exposure on an outcome varies within levels of a third variable, differs from mediation; in particular, *P. copri* was not identified as a mediator because the Mediterranean diet didn't influence *P. copri* carriage.

In conclusion, these two examples show the difficulty of performing mediation analysis in real datasets. For example, in the MLVS, diet was measured over 26 years and significantly affected gut microbiome composition; however, only two significant component indirect effects were detectable. We also performed a sensitivity analysis for confounding on the MLVS results (**Methods**) and the main findings discussed above did not change.

Figure 3.5: Estimation of mediation effects in diet-cardiometabolic disease datasets.



We applied five methods (including MedTest) to two real-world datasets that looked at the effect of diet on cardiometabolic disease. MedTest did not identify a significant total indirect effect for either MLVS or the Blueberry RCT and is not shown because it doesn't return effect sizes. P-values < 0.1 for the total indirect effect were considered significant and are denoted with a black outline. The direct effect is not presented for these four methods because it is entirely determined by the total effect and total indirect effect.

3.4 DISCUSSION

We evaluated seven methods for performing mediation analysis with a microbiome mediator in realistic simulated data and considered three areas of functionality (**Fig. 3.1C**), including hypothesis testing for the total indirect effect (**Fig. 3.2**), hypothesis testing for component indirect effects (**Fig. 3.3**), and total indirect effect estimation (**Fig. 3.4**). Only three of seven methods offered all three pieces of functionality. For hypothesis testing, we recommend the naive method for exploratory analyses where an inflated false positive rate may be more acceptable. Otherwise, we recommend CCMM for testing the total indirect effect and HDMA or HIMA for testing

component indirect effects. For estimation of the total indirect effect, HDMA and HIMA performed well, but struggled in the case of many weak mediators, which is likely representative of real-world data. CCMM performed best in the case of multiple, weak mediators, but displayed strange behavior when the number of samples was less than the total number of microbiome features (**Fig. S3.2**). The naive method performed much better than expected, although showed more variability overall in estimates of the total indirect effect compared to HDMA and HIMA. Given the respective strengths and weaknesses of the methods, an end-user may need to select multiple methods in order to accomplish all three pieces of analysis; however, as seen in the real-world examples (**Fig. S3.4**) this can then make it difficult to interpret varying effect estimates. Although there has been high interest in methods development for microbiome mediators in the last few years, more work is needed to continue to improve method performance and usability, as well as provide guidance and tools for end-users to interpret the output, especially in the case of multiple conflicting estimates.

Additionally, more work on the "spiking" of microbiome data can help with future benchmarking efforts for mediation methods and statistical microbiome methods more generally. In the process of simulating realistic data, we encountered numerous challenges related to how to best create an association between the treatment variable and a microbiome feature. In particular, we saw that a spiked effect created on the absolute scale was not entirely recoverable on the relative scale (**Fig. 3.4A**). This leads to a philosophical question about whether the true effect is that on the absolute or relative scale. We also had to decide whether the spike should change the abundance or prevalence of the feature or both. In these evaluations, we assumed that the treatment only changed the abundance of microbes; more specifically, the treatment affected the microbiome by keeping the amount of bacterial biomass the same, but changing the relative proportion of one microbe.

This assumption is reasonable for environments with fixed carrying capacities, like the human gut, where the amount of biomass is constant and hard to change. When creating a negative association between the treatment and microbiome feature, we ran into another issue where values could not go below zero (since all compositions are positive), which resulted in a smaller spiked effect than what we desired. Zero inflation also reduced the maximum strength of the association that could be spiked in. In the evaluation where we fixed the total indirect effect size but varied the number of mediators (**Fig. 3.4D**), this meant that more prevalent features had to be selected as mediators in the case of few mediators to achieve the same total indirect effect. These challenges we encountered are true of real-world data as well and are not an artifact of the simulation; however, resolving them will assist in the performance of all microbiome benchmarking studies.

In our evaluations, we assessed whether statistical associations added to the synthetic data could be identified by the different methods, but multiple assumptions are needed to interpret these effects causally (116). Interpreting indirect effect estimates as causal requires that, conditional on the adjusted covariates, there is no confounding of the exposure-outcome relationship, the exposure-mediator relationship, or the mediator-outcome relationship (116). Additionally, it requires that the exposure does not affect any confounders of the mediator-outcome relationship (116), which is especially problematic for microbiome data because other microbiome features may easily fill this role when mediation analysis is performed feature-by-feature. The two diet-microbiome vignettes reveal how trade-offs regarding confounding may occur. Given that microbiome data are noisy with many features, a strong treatment effect is needed to detect mediation effects. The observational Mediterranean diet exposure had an overall stronger effect on the microbiome, in part because it occurred over many years and did not require adherence with a dietary intervention, but was more at risk of confounding. In contrast, the Blueberry RCT

randomized the treatment, thereby limiting confounding of the treatment-outcome and treatment-mediator relationships, but could only be performed over six months and required adherence with the intervention, thereby reducing effect sizes.

When selecting datasets for mediation analysis, these trade-offs between effect size and possible confounding should be considered. The timing of the measurement of the treatment, microbiome, and outcome should also be considered to ensure the temporal ordering of the variables matches that expected in mediation analysis. Lastly, the microbiome must be measured on the biologically relevant scale (or on one that can be converted to the relevant scale) to capture a mediation effect. Ideally, an absolute measure of microbial abundances, such as qPCR, would be used to limit the dilution of mediated effects via compositionality.

Limitations of this study include that real studies are likely more limited by sample size than the results presented here, which go up to a sample size of 400. Additionally, it was challenging to define the direct and indirect effect sizes for real interventions and therefore to set these simulation parameters. We did not include mediator-mediator or treatment-mediator interactions in the model used to simulate the outcome data. Including these terms may have improved the relative performance of methods like SparseMCMM that specifically take this interaction into account. In future work, we plan to include interaction terms in the outcome model, use a continuous exposure variable, and evaluate the methods on data simulated under different causal structures (i.e., when the microbiome is actually downstream of the outcome, etc.). Lastly, the performance of the methods as measured here depends on their interaction with the data generation process and therefore depends on the accuracy and assumptions of our data generation process.

3.5 CONCLUSIONS

We evaluated seven methods for performing mediation analysis with a microbiome mediator in realistic simulated data and considered three areas of functionality (**Fig. 3.1C**), including hypothesis testing for the total indirect effect (**Fig. 3.2**), hypothesis testing for component indirect effects (**Fig. 3.3**), and total indirect effect estimation (**Fig. 3.4**). Only three of seven methods offered all three pieces of functionality. For hypothesis testing, we recommend the naive method for exploratory analyses where an inflated false positive rate may be more acceptable. Otherwise, we recommend CCMM for testing the total indirect effect and HDMA or HIMA for testing component indirect effects. For estimation of the total indirect effect, HDMA and HIMA performed well, but struggled in the case of many weak mediators, which is likely representative of real-world data. CCMM performed best in the case of multiple, weak mediators, but displayed strange behavior when the number of samples was less than the total number of microbiome features (**Fig. S3.2**). The naive method performed much better than expected, although showed more variability overall in estimates of the total indirect effect compared to HDMA and HIMA. Given the respective strengths and weaknesses of the methods, an end-user may need to select multiple methods in order to accomplish all three pieces of analysis; however, as seen in the real-world examples (**Fig. S3.4**) this can then make it difficult to interpret varying effect estimates. Although there has been high interest in methods development for microbiome mediators in the last few years, more work is needed to continue to improve method performance and usability, as well as provide guidance and tools for end-users to interpret the output, especially in the case of multiple conflicting estimates.

3.6 METHODS

Simulation parameters

We explored how method performance changed as important dataset properties were varied over ranges of realistic values. We varied multiple parameters during simulation, including sample size, number of microbiome features acting as mediators, strength of mediation, proportion of the total effect mediated, and direction of mediation (where a positive mediated effect meant that increasing the treatment increased the outcome via a given mediator, and vice versa for a negative direction of mediation) (**Table S3.1**). The average read depth (50,000) and the number of total microbiome features (100) were kept constant across simulations. For each parameter set, 300 iterations of the simulation were run (**Table S3.1**).

Assumptions were made to reduce simulation complexity and limit the number of parameters. We assumed that the linear effect of the treatment variable on each mediating feature was the same, and that the linear effect of each mediating feature on the outcome was the same (**Fig. 3.1A**). Furthermore, we assumed consistent absolute values of these two sets of coefficients. Therefore, after selecting the number of microbiome features acting as mediators, the strength of mediation, and the mediation direction, the values for both sets of coefficients could be solved for under these assumptions.

Synthetic data generation

Overview: Simulated data was created in three steps. First, we randomly generated a binary treatment variable using ~Bern(0.5). Secondly, SparseDOSSA version 1 (112), a tool from the Huttenhower lab that generates realistic metagenomic count datasets with properties similar to a calibration dataset, was used to simulate microbiome count data containing a specified association with the treatment variable. Lastly, a linear model was used to combine information on the

treatment and microbiome features to generate simulated outcomes. As described above, for each simulated dataset with a mediating effect added ("spiked"), we created three null versions of the dataset using the same set of selected microbiome features for comparison purposes: one with no effect of the treatment on the set of microbiome features ("null_a"), one with no effect of the set of microbiome features on the outcome ("null_b"), and one in which no relationships were spiked between the set of microbiome features and the treatment or outcome ("null_ab").

Microbiome data generation with SparseDOSSA: During the SparseDOSSA simulation process, a parent lognormal distribution was fit based on the calibration dataset (**Fig. 3.1B**, "microbial feature means distribution"), in this case using gut microbiome samples from healthy adults (113). The marginal distribution of each microbiome feature was represented by a truncated lognormal distribution with a mean drawn from the parent distribution (**Fig. 3.1B**, "feature distributions"). To create the null matrix (**Fig. 3.1B,** "absolute null matrix") - a matrix representing the underlying absolute abundance of each feature with no added association with the treatment variable - we drew a number of values from each feature's distribution equal to the sample size. To represent the zero-inflation in microbiome data, each feature was assigned a probability of zeros (p_z) proportional to its mean, such that features with lower mean abundances had a higher probability of zero values. These zeros represented true biological zeros (i.e., the microbe is truly absent rather than undetectable by sequencing). After generating the null matrix, each value for a feature was replaced with a zero value according to a Bernoulli coin flip with probability p_z.

Effect of the treatment on the microbiome:

In creating an association between the treatment and a selected microbiome feature, we assumed that the treatment affected the microbiome by keeping the amount of bacterial biomass the same,

but changing the relative proportion of one microbe; this assumption is reasonable for environments with fixed carrying capacities, like the human gut, where the amount of biomass is constant and hard to change. Therefore, the aim of the "spike" procedure was to keep the mean and standard deviation of the new "spiked" feature similar to that of the original microbiome feature, while not increasing overall abundance. First, we rescaled the treatment variable to have the same mean and standard deviation as the microbiome feature. We then combined one part microbiome feature with δ parts rescaled treatment variable (where δ was the SparseDOSSA "spike strength" input parameter) and divided by (δ+1) to keep the overall feature abundance the same. Additionally, after spiking in the association, the degree of zero-inflatedness in the data was kept the same by setting any values that were originally zero back to zero; the biological explanation for this being that we assumed the treatment changes the growth of microbes (i.e., changes abundance), but doesn't introduce any new microbes that weren't present in the sample (i.e., change prevalence). Since this data is later converted to compositions, any negative values were recoded to zero - this created a small amount of truncation error (depending on the strength and direction of the "spike strength" parameter, δ) and therefore we calculated the mediated effect size for each simulation iteration empirically (see the discussion on the challenges of spiking complex data in **Discussion**).

Selection of mediating microbiome features: To create a matrix containing the absolute abundances of microbiome features, including spiked associations with the treatment variable for the specified number of mediating features (**Fig. 3.1B**, "absolute spiked matrix"), we selected microbiome features to spike based on their percent zero values. Given that zero-inflation dilutes the spiked association, we proactively increased the spike strength based on the proportion of zeros in each spiked feature such that each component indirect effect was the same (except for truncation

error). No effects of microbiome features on other microbiome features were spiked in; however, it is possible that some features had weak correlations with each other randomly.

Generation of outcome data: To create simulated outcome data, we used a linear model to combine data on the treatment and the microbiome features. This model assumed no treatment-mediator interaction, no mediator-mediator interaction on outcome, and a standard normal error. Inclusion of treatment-mediator and mediator-mediator interactions would complicate estimation and likely reduce model performance below that found above - therefore our results may represent an upper bound on model performance. Depending on the effect type of the simulation (i.e., "null_a", "null_b", "null_ab", "spiked"), data from either the absolute null matrix or absolute spiked matrix was used in the outcome model.

Conversion of microbiome data from absolute to relative abundance: Lastly, we simulated the sampling process on both the null and spiked absolute matrices to produce compositional matrices that the methods were evaluated on (**Fig. 3.1B,** "compositional spiked matrix", "compositional null matrix"). Each sample was assigned a total read count, and a multinomial distribution was used to draw those counts based on the proportion of each feature in each sample. As with sequencing, since the total read count per sample is arbitrary and the samples have different total read counts, we divided each sample by its total read count to get relative abundances for comparisons. This process captured technical zeros that occur during sequencing (i.e., low abundance species that are missed by sampling). To represent realistic data analysis practices, a quality control step was run prior to method evaluation where all feature abundances below 0.1% abundance in more than 95% of samples were aggregated into a single "other" category. Additionally, CCMM and SparseMCMM couldn't handle zero values so prior to evaluating these

two methods the data was adjusted by replacing each zero with half of the lowest non-zero value for that feature (i.e., Laplace smoothing).

Calculation of the empirical mediated effect size: To calculate the true mediated effect size in each simulated dataset, we used an empirical calculation due to the addition of a small amount of error via truncation during the process of inducing an association between the treatment variable and mediating microbiome features. For a binary treatment variable, the total mediated effect was calculated as E[Y|T=1]-E[Y|T=0] - c, where c was the specified direct effect and E[Y|T=1]-E[Y|T=0] was the difference in mean outcomes among simulated samples with and without the binary treatment.

Implementation of mediation methods

The selected statistical mediation methods are as follows:

- Causal Compositional Mediation Model (CCMM) (109): we used the default implementation of CCMM in the R package *ccmm* with the "normal" method used to estimate the variance of the indirect effects. This method performs an additive log-ratio transformation on the microbiome features prior to fitting marginal models for the effect of the treatment on the microbiome features and a penalized linear log-contrast model using the L1 norm for the outcome.

- MedTest (111): we used the implementation of MedTest from GitHub: https://github.com/jchen1981/MedTest and ran the method with three distance matrices (Bray-Curtis, Jaccard, and Euclidean) computed in the R *vegan* package. MedTest performs multidimensional scaling on the supplied distance matrices to calculate a test statistic and then runs a permutation-based hypothesis test for the total indirect effect.

- Sparse Microbial Causal Mediation Model (SparseMCMM) (110): we used the default implementation of SparseMCMM in the R package *SparseMCMM*. This method uses penalized Dirichlet regression using the L1 norm to handle compositionality when modeling the effect of the treatment on the microbiome features, and fits a penalized linear log-contrast model of the outcome using an original penalty that ensures that microbiome feature-treatment interaction terms are only included in the outcome model if the main feature effect is also included.

- High dimensional mediation analysis (HDMA) (114): we used the default implementation of HDMA from GitHub: https://github.com/YuzhaoGao/High-dimensional-mediation-analysis-R/blob/master/HDMA.R, with the specification of "gaussian" for family and "lasso" for method. HDMA uses a similar procedure to HIMA, including performing feature pre-selection with sure independence screening, but instead fits a penalized linear outcome model using the L1 norm with a debias procedure.

- HIMA (115): we used the default implementation of HIMA in the R package *hima*, including the default minimax concave penalty, with the specification of "gaussian" for family. HIMA utilizes a sure independence screening procedure that filters out microbiome features with a weak marginal correlation with the outcome variable and selects a number of features less than the sample size to evaluate. HIMA fits multiple marginal mediator models and uses a linear outcome model with a minimax concave penalty, which compared to the L1 norm applies less shrinkage to non-zero parameters.

- PCR (principal component regression): PCR was implemented in R. First, the microbiome data were arcsine square root transformed. Principal components were calculated using the

prcomp function in R with centering and scaling (i.e., the features were centered at zero and scaled to have a variance of one before calculating the PCs). We selected the top PCs that together explained 90% of the variation in the data. The total indirect effect was calculated by fitting marginal linear regression models for the effect of the treatment on each PC and the effect of each PC - controlling for the treatment - on the outcome and summing the products of the coefficients from the regression models. While the naive method implemented a similar procedure, the use of PCs in the models (instead of microbiome features directly) both reduced dimensionality and eliminated bias because the PC's are independent; however, it also prevented estimation of the component indirect effects.

- Naive (simplistic extension of the traditional product-of-coefficients method): The naive method was implemented in R. First, the relative abundance of each feature was normalized by subtracting the mean feature relative abundance and dividing by the standard deviation of the feature relative abundance. We calculated the total indirect effect and the component mediated effects by fitting marginal linear regression models for effect of the treatment on each feature and the effect of each feature - controlling for the treatment - on the outcome. The component indirect effect for each feature was calculated as the product of the coefficients from the two regressions. The total indirect effect was calculated as the sum of all the component indirect effects.

We used a permutation-based hypothesis test for the naive and PCR methods. To get the distribution of the test statistic under the null hypothesis of no mediated effect, microbiome data were reassigned across samples (i.e., shuffling the rows without breaking them) and the test statistic recalculated. The permutation p-value was calculated as the proportion of test statistics

that were farther from 0 than the original, non-permuted test statistic with adjustment to prevent the underestimation of p-values (120). For each p-value calculation, the data were permuted 1000 times. We visually examined the p-value distributions from simulated datasets with and without mediation to assess the performance of the permutation-based hypothesis test (**Fig. S3.1**).

Vignette methods

MLVS: For full methods, including the collection of dietary information via semi-quantitative food frequency questionnaires, calculation of the MedDiet index capturing adherence to a Mediterranean dietary pattern, and calculation of the composite biomarker score for CVD and T2D pathogenesis see (117). In the MLVS, subjects had two sets of paired stool samples (i.e., up to four samples) collected six months apart between the years 2011 to 2013. Only microbiome samples from the second study visit were used. For subjects with more than one sample at the second study visit, a random sample was selected. We performed a sensitivity analysis for confounding using a residualized outcome. To create the residualized outcome, a linear model for the normalized CVD risk score was fit with terms for total energy intake, physical activity, Probiotic use, Bristol stool scale, age, use of proton pump inhibitors, use of metformin, and use of antibiotics and the residuals were taken. The main results for the total effect, total indirect effect, and two significant component indirect effects did not change when using the residualized outcome. The final microbiome data contained samples from 248 distinct subjects and 58 genus-level microbiome features.

Blueberry RCT: We used data from a randomized, double-blind, placebo controlled trial of blueberry polyphenol supplementation (unpublished). The trial enrolled 115 subjects who met three or more criteria for metabolic syndrome at baseline. Subjects were randomly assigned to consume 26g freeze-dried blueberry powder (corresponding to one cup of blueberries per day),

26g isocaloric placebo powder, or 13 of each powder (corresponding to one half cup blueberries per day) every day for six months. Fecal samples, fasting blood samples, and 24 hour urine samples were collected at baseline and after six months. Only microbiome samples from the six month time point were used. The final microbiome data contained samples from 112 distinct subjects and 45 genus-level microbiome features.

Data processing: Taxonomic and functional profiles were generated using the bioBakery meta'omics workflow (88). Briefly, reads mapping to the human genome were first filtered out using KneadData v0.7.0 with default parameters. Taxonomic profiles of shotgun metagenomes were generated using MetaPhlAn2 v2.6.0. In taxonomic profiles, relative abundances were given at the genus-level or, if unidentified, at the family-level. Phage abundance was normalized out and analyses focused purely on non-viral taxa. A prevalence-abundance filter was applied and all feature abundances below 0.1% abundance in more than 95% of samples were aggregated into a single "other" category. Prior to running CCMM each zero value was replaced with half of the lowest non-zero value in that feature (i.e., Laplace smoothing) since CCMM cannot run with zero values in the microbiome dataset.

3.7 DECLARATIONS

Availability of data and materials

All software used in this study is free and open source. SparseDOSSA version 1 used in this study is available at: https://github.com/biobakery/sparseDOSSA. The data for the MLVS vignette is available at: https://www.nature.com/articles/s41591-020-01223-3#data-availability. The data for the blueberry RCT are not yet publicly available.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (grant T32AI007535 to EA). The funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

EKA, EAF, EKG, AM, and CH designed the research. EKA, and EAF analyzed data. EKG, SM, AM, and ML provided conceptual feedback. EAF, ML, and CH provided project oversight. EKA, EAF and CH wrote the paper. All authors approved the manuscript.

Acknowledgements

The authors express their gratitude to William Hanage, Sebastien Haneuse, and Eric Rimm for their insightful feedback on the project; Kelsey Thompson and Daniel (Dong) Wang for their input on the vignette datasets; Aedin Cassidy for the use of the blueberry (poly)phenol RCT dataset; all participants of the MLVS and blueberry (poly)phenol RCT for generously providing their time and samples; and the National Institute of Allergy and Infecstious Diseases of the National Institutes of Health (award number T32AI007535) for funding.

Chapter 4. How to detect and reduce potential sources of biases in studies of SARS-CoV-2 and COVID-19

4.1 ABSTRACT

In response to the coronavirus disease (COVID-19) pandemic, public health scientists have produced a large and rapidly expanding body of literature that aims to answer critical questions, such as the proportion of the population in a geographic area that has been infected; the transmissibility of the virus and factors associated with high infectiousness or susceptibility to infection; which groups are the most at risk of infection, morbidity and mortality; and the degree to which antibodies confer protection to re-infection. Observational studies are subject to a number of different biases, including confounding, selection bias, and measurement error, that may threaten their validity or influence the interpretation of their results. To assist in the critical evaluation of a vast body of literature and contribute to future study design, we outline and propose solutions to biases that can occur across different categories of observational studies of COVID-19. We consider potential biases that could occur in five categories of studies: (1) cross-sectional seroprevalence, (2) longitudinal seroprotection, (3) risk factor studies to inform interventions, (4) studies to estimate the secondary attack rate, and (5) studies that use secondary attack rates to make inferences about infectiousness and susceptibility.

Keywords: epidemiological biases, selection bias, misclassification, measurement error, COVID-19, observational data

4.2 INTRODUCTION

Since the onset of the coronavirus disease (COVID-19) pandemic, public health scientists have worked tirelessly to provide the knowledge needed to address this new, global crisis. The pandemic has spurred an exceptional number and breadth of scientific studies (121,122), including epidemiologic ones, with the pace making it both important and challenging for researchers to design and analyze studies in the most robust way possible, and for reviewers and users to accurately evaluate the strength of evidence such studies provide. Using relevant examples from the literature, we discuss potential epidemiological biases arising in various phases of observational studies of COVID-19 and outline possible solutions.

We consider biases arising across five classes of research questions: (1) estimates of seroprevalence (123), (2) estimates of seroprotection (124,125), (3) studies of risk factors for becoming infected (126), (4) estimates of the secondary attack rate (127), and (5) comparisons of secondary attack rates to make inferences about susceptibility and infectiousness (127).

4.3 SEROPREVALENCE MEASUREMENT TO ESTIMATE CUMULATIVE INCIDENCE

Serological surveillance studies detect SARS-CoV-2 specific antibodies in the population and can provide an estimate of the seroprevalence -- the proportion of various groups (e.g., age groups) harboring such antibodies at a single time point or, if repeated, over time. If antibodies are a marker of protection, seroprevalence may provide a direct estimate of the fraction of individuals immune to the virus, although, as we note in the next section, the protective role of antibodies against future infection remains uncertain and may wane over time (128). If antibodies are a reliable measure of prior infection, then seroprevalence can also be used as a proxy for the cumulative incidence of

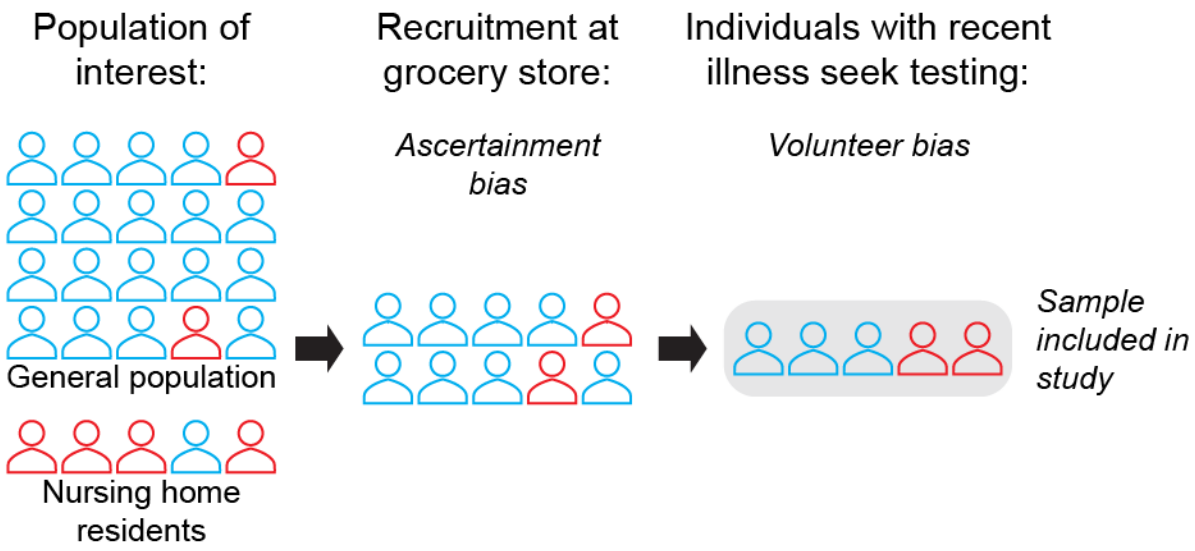
infection until that time point (more precisely, until an earlier time point, because antibodies take time to rise to detectable levels after infection (128,129)).

As measures of cumulative incidence, seroprevalence studies can be more accurate than direct counting of case reports, especially for an infection that is often not detected due to limited testing and/or lack of symptoms and where case ascertainment rates have varied drastically over time. Yet to achieve accurate estimates from seroprevalence studies it is important to recruit a representative sample of the population of interest, and to consider measurement error, which can still create bias in a perfectly collected sample. Bias is always in reference to the specific variable which is being estimated in a statistical analysis (i.e., the estimand) so we will suppose we want to estimate the cumulative incidence of infection using seroprevalence among all people living in a particular city. We note the following considerations regarding potential biases in studies where seroprevalence is used to estimate cumulative incidence and use the terms interchangeably below.

Seroprevalence estimates may be unrepresentative of the target population when the individuals enrolled in the study are not representative of that population. The direction and magnitude of the resulting bias depend on the population for which inference of seroprevalence is being attempted, for example, all residents of a county, and the degree to which the individuals tested diverge from a random sample of that population. Depending on the sampling location and time, people who are present to be sampled may be at higher or lower risk of COVID-19 than average (**Fig. 4.1**). Important populations, including those in congregate settings (e.g., nursing homes, prisons), are often excluded. For example, persons residing in long term care facilities (LTCFs) may be over- or underrepresented in serosurveys (130,131), which depending on their seroprevalence can produce an over or underestimate of true seroprevalence in the population. If persons in LTCFs are not sampled, this can result in a lower seroprevalence estimate for older individuals (131). For

example, the authors of the New York State serosurvey, which recruited a convenience sample at grocery stores across the state, acknowledge that enrollment disproportionately excluded persons from vulnerable groups who may be more likely to self-isolate at home; individuals who died from or were hospitalized or housebound with COVID-19 infection; and individuals living in LTCFs (132). If researchers hope to generalize inferences to specific risk groups, they must ensure these groups are included in the survey. Once the population of interest has been clearly defined, they should endeavor to randomly recruit individuals from the population of interest and upweight under-sampled groups. Standardization or inverse-probability-of-sampling weighting can mitigate this type of bias, but only if all relevant predictors of seropositivity are included in the correction (133).

Figure 4.1: Schematic showing recruitment-based biases in a hypothetical serosurvey.



This figure shows a hypothetical serosurvey that aims to measure the underlying seroprevalence in the entire population of a geographic region and performs recruitment among shoppers at a grocery store. Outline color represents prior SARS-CoV-2 infection status (red for prior infection, blue for no prior infection). Ascertainment bias occurs because (1) individuals recruited at the grocery store are likely at slightly higher risk of COVID-19 than average

Figure 4.1 (Continued)

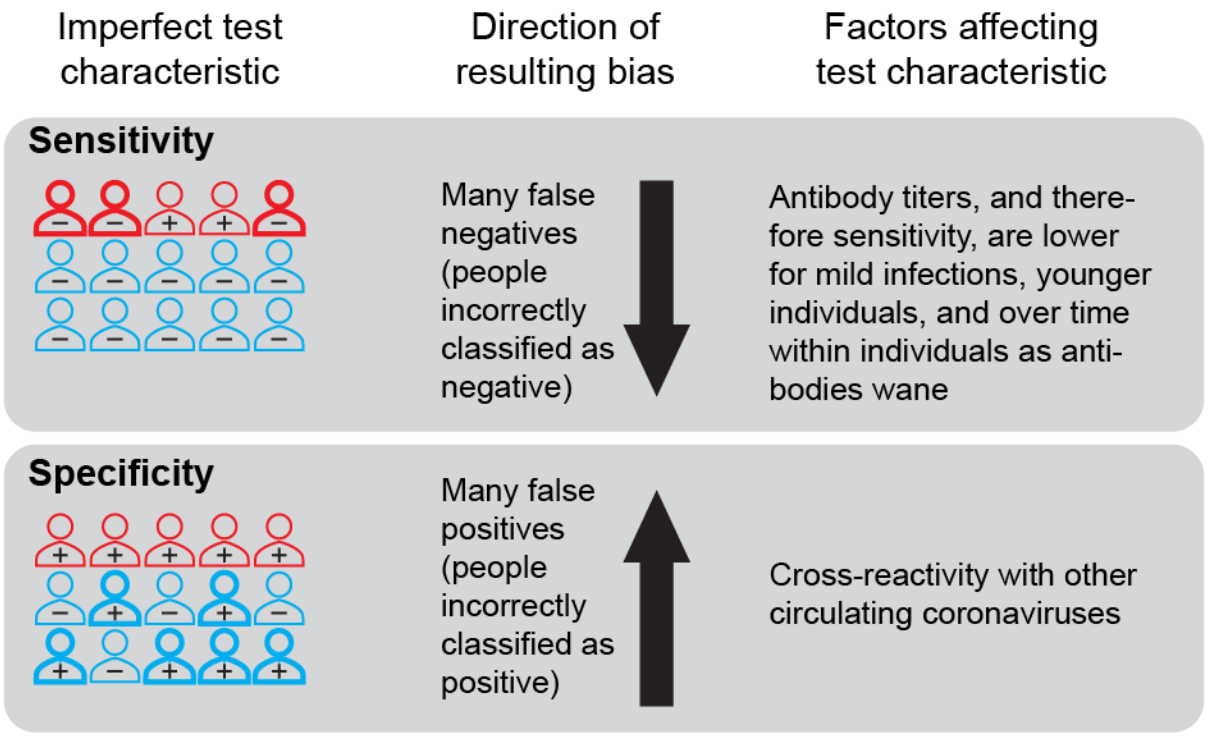
(since individuals who are isolated at home and rarely grocery shop are less likely to be sampled), and (2) nursing home residents and other populations in congregate settings are excluded from the sample. Among individuals present at the grocery store, volunteer bias occurs when individuals who believe they have been infected participate in order to receive testing. Meanwhile, individuals who do not think they have been infected may avoid testing.

Biases may arise from nonrandom willingness to participate in a survey, even if a random sample of the population is approached to participate. A serosurvey that successfully reaches the population of interest may still suffer from volunteer bias in who participates in testing. This can bias estimates in either direction. Estimates of seroprevalence will be too high if individuals are more likely to accept testing because they think they have been exposed to SARS-CoV-2. On the other hand, a downward bias will occur if individuals accept testing because they are overcautious or if exposed individuals avoid testing because they do not want a positive test result (**Fig. 4.1**). In the study design, volunteer bias can be reduced by sampling from a pre-established cohort with high rates of participation. At least some demographic information should be collected on those who do and do not consent to testing, in order to assess aspects of how representative the consenting population is of those approached. If predictors of non-response are collected, estimates can be corrected in the analysis stage, for example through inverse probability weighting. Serosurveys may want to ask subjects "Do you think you've had COVID-19 previously?", or collect data on symptoms to assess whether volunteer bias is occurring in their sample. Lastly, one innovative approach (134) proposes splitting the survey group into subsets and giving each subset an increasing incentive (for example, money) for participation, enabling researchers to construct a statistical model to predict how seroprevalence would change if everyone participated.

False negative serologic tests, if not properly accounted for, can underestimate seroprevalence, while false positive tests, if not properly accounted for, can overestimate it -- the latter problem being most serious near the start of the epidemic. Tests for SARS-CoV-2 antibodies are imperfect. Test performance is described by the test's sensitivity, which is the ability to identify those who have SARS-CoV-2-specific antibodies, and specificity, which is the ability to identify those who do not have such antibodies. Unless adjusted for in the analysis, the use of imperfectly sensitive

tests will underestimate the cumulative incidence of past infections due to the presence of infections not detected by the test (**Fig. 4.2**). Conversely, a test with imperfect specificity will incorrectly classify individuals without antibodies as positive, resulting in an overestimate of cumulative incidence if not adjusted for in the analysis (**Fig. 4.2**). When a disease is rare, such as COVID-19 early in the pandemic or in areas with low transmission, high test specificity is needed to accurately measure the seroprevalence. For example, the Santa Clara study, which claimed that there were 50–85 times more COVID-19 cases in Santa Clara than previously identified, found 50 individuals positive for antibodies out of 3330 tested (135); however, the specificity of the test used in that study was uncertain, and a test with 98.5% specificity would be expected to generate 50 false positives on average in that sample if no one had antibodies.

Figure 4.2: Biases due to misclassification by SARS-CoV-2 antibody tests.



The sensitivity of a SARS-CoV-2 antibody test is the probability the test is positive given an individual has been

Figure 4.2 (Continued)

infected with the virus, while the specificity is the probability of a negative test given an individual has not been infected with SARS-CoV-2. Test performance is imperfect; low sensitivity can result in an estimate of cumulative incidence that is too low (as individuals with prior infection are misclassified as negative), and low specificity can result in an estimate of cumulative incidence that is too high (as individuals without prior infection are misclassified as positive). Outline color represents prior SARS-CoV-2 infection status (red for prior infection, blue for no prior infection). The annotation ("+" or "-") indicates the result of a test for SARS-CoV-2 antibodies. Bold outlines indicate individuals who are misclassified by the test.

Adjustments for test sensitivity and specificity should be done with care, accounting for the often small numbers of validation samples and possible differences between the populations in which the tests were validated and the study population. While most serologic studies do adjust for the test sensitivity and specificity using available estimates for each test, the values of sensitivity and specificity in the study population may be different than in the population used for evaluating test performance, which is commonly made up of hospitalized patients (123). In particular, sensitivity is often lower for individuals with lower antibody titers. Therefore rates of false negatives are expected to be higher among individuals with less severe disease (128,136), such as younger individuals (137,138), individuals who were recently infected and have not yet mounted an antibody response (128,129), or individuals who were infected long before testing, as antibody titers wane over the weeks and months after infection (128,129). For instance, if a seroprevalence estimate among a college student population (where test sensitivity is likely lower) is corrected using a measurement of test sensitivity from older, sicker individuals in validation data then the adjusted estimate will be too low, and vice versa (see (123) for illustrations of this bias through simulation).

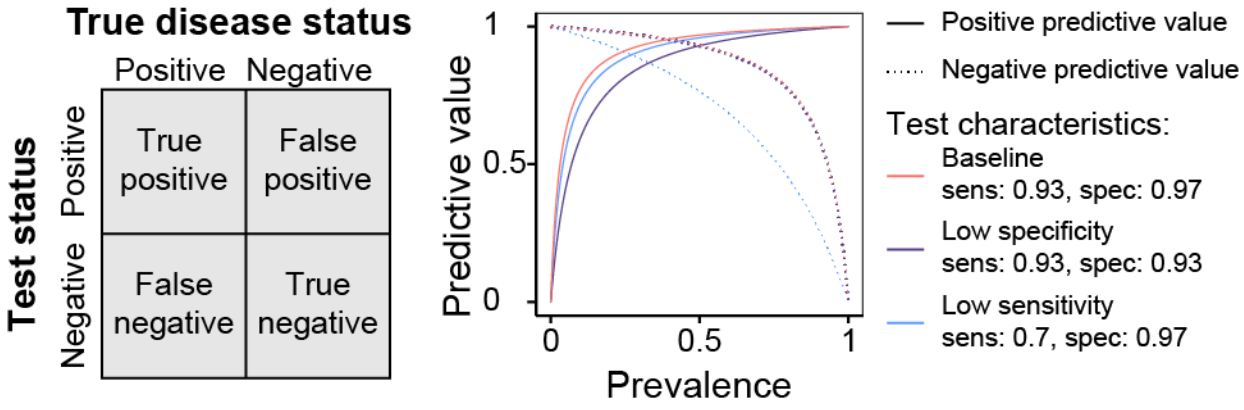
As mentioned above, seroprevalence may underestimate cumulative incidence if some individuals who initially have antibody levels sufficient to test positive on a serologic test have waning levels that drop below the threshold for positivity, a phenomenon sometimes called "seroreversion". Low antibody values occur as antibodies are increasing and as they are declining; however, the increase is fast compared to the decline (129,139), so most individuals with low titers will be those on the decline, except perhaps in a very rapidly growing epidemic, where there will be many very recent infections (e.g., (140) but with antibody titers instead of viral load). Antibodies to seasonal coronaviruses have been shown to decline substantially within a period of a few months to a year

(141). Recent evidence points to the similar disappearance of antibodies to some components of SARS-CoV-2 when data is presented as the percent above a threshold defined as positive (142). Low sensitivity due to waning antibodies is problematic when using seropositivity as a proxy for the cumulative incidence of infection in a population; for example, the observed temporal decline in seroprevalence in several studies (130,143), if more than is explainable by sampling variation, likely indicates waning of antibody titers, as the cumulative incidence of infection cannot decrease with time in a closed population. Much remains to be learned (from seroprotection studies) about the nature and duration of protection following infection, so we do not take a position here on whether cumulative incidence and immunity are the same, but we note that it is biologically possible for an individual to be at least partially immune to infection and/or disease due to T cell and B cell memory despite low antibody titers (144–147). By presenting full distributions of quantitative (e.g., ELISA) values, instead of reporting the percent positive above a threshold, seroprevalence studies can preserve the data for reanalysis as our understanding of antibody kinetics improves.

Solutions to misclassification include prioritizing high specificity when seroprevalence is low, and high sensitivity when seroprevalence is high (**Fig. 4.3**) either through test selection or by using multiple independent tests (e.g., (131)). While estimates of seroprevalence at the population level can potentially be corrected for imperfect test characteristics, this may not remove all sources of bias. As discussed above, bias can remain if, for example, the estimates of test characteristics are obtained from "gold standard" positives and negatives with a different distribution of antibody levels than the true positives and negatives, respectively, in the study population (123). Additionally, there is uncertainty in the measurement of test specificity and sensitivity, especially for newly developed SARS-CoV-2 antibody tests for which validation datasets may be small.

Adjustment using point estimates instead of the full ranges of plausible sensitivity and specificity values underestimates the true uncertainty in seroprevalence estimates. A way to rectify these biases is through the use of a Bayesian approach to adjust seroprevalence estimates for ranges of values for test characteristics (148).

Figure 4.3: The relative importance of test sensitivity and specificity depends on the underlying seroprevalence in the study population.



The value of a test can be described through the positive predictive value (PPV), which is defined as the probability that an individual truly has been infected with the virus given that they test positive and is calculated as the number of true positives divided by the total number of positive tests. Similarly, the negative predictive value (NPV) is defined as the probability that an individual truly has not been infected with the virus given that they test negative and is calculated as the number of true negatives divided by the total number of negative tests. When the underlying seroprevalence is low, test performance is largely a function of specificity, as the majority of individuals in the population have not been infected, while sensitivity is more important as seroprevalence increases. Note that the negative predictive values for the baseline and low specificity tests are very similar so the curves nearly overlap in the figure.

SUMMARY: An ideal study design for SARS-CoV-2 seroprevalence would:

- Use a sample that is representative of the target population. In particular:

    - Recruit participants in a way that does not systematically favor those with unusually high or low levels of exposure.
    - Ask participants and nonparticipants whether they believe they have been infected to detect volunteer bias in a sample.
    - Consider demographics and other information about participants (and ideally also nonparticipants) to facilitate adjustment of results.

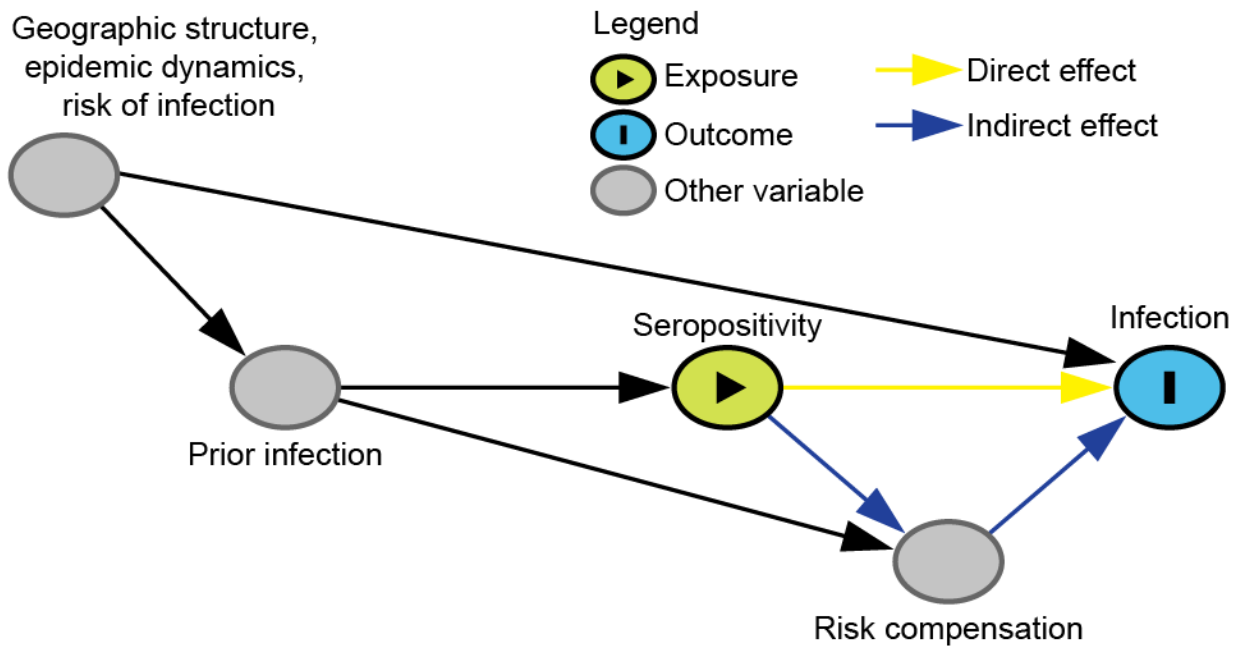
- Use an assay for which sensitivity and specificity estimates are available from a population similar to that being studied in terms of disease severity and timing of infections.
- Report the distribution of quantitative (e.g., ELISA) values and not just the percent positive above a threshold to allow analysis of possible seroreversions.
- Adjust seroprevalence estimates for test characteristics, including uncertainty in the measurements of these characteristics.

4.4 SEROPROTECTION

There is evidence to suggest that prior infection with a coronavirus, including SARS-CoV-2, confers some level of immunity and protection against reinfection with the same viral species (141,149–151). However, the extent and duration of this protection is unknown, and studies are needed to better characterize immunity to this novel virus. While seroprevalence studies focus on one point in time, seroprotection studies are longitudinal, following people over time to evaluate whether seropositivity confers protection against infection compared to seronegativity. In this case, the causal effect of interest is the direct (biological) effect of seropositivity on future infection. A number of biases can arise in these observational studies.

Estimates of the (total) effect of prior infection on (re)-infection may be biased toward the null (no protection) if seropositive individuals remain more exposed to infection than seronegative ones (confounding by risk of infection). This scenario occurs when there is a confounder that persists through time, such as residing in a crowded household or being an essential worker, that may predict both the exposure (seropositivity) and the outcome (future infection). For example, people in higher-risk occupations are more likely to become infected at each point in time, meaning they are more likely to be seropositive and also more likely to be reinfected (**Fig. 4.4**). These positive associations of the persistent confounder with both the exposure and the outcome create a downward bias, causing seropositivity to appear less protective against (or even harmful for) future infection. Limiting the study to groups with high rates of infection and risk of exposure can mitigate this bias while improving power (125), otherwise this bias can be addressed by adjusting for occupation or other factors associated with risk of infection.

Figure 4.4: Directed acyclic graph under the alternative hypothesis showing confounding in the estimation of seroprotection.



This figure shows the causal relationship between important variables that influence the infection status of an individual. To analyze the effect of seropositivity on the risk of infection, we would need to adjust for geographic structure, epidemic dynamics, the risk of infection and any other variables that are confounders of this exposure-outcome relationship. The effect of seropositivity on infection risk may be mediated by behavior change (induced by knowledge of serostatus) that affects the risk of infection. Disentangling direct (biological) effects of seropositivity and indirect effects through risk compensation is not straight-forward. Geographic structure, epidemic dynamics, and risk of infection are likely or guaranteed confounders of the relationship between seropositivity and future infection. For the purposes of illustrating this particular bias, the directed acyclic graph is drawn under the strong assumption of no additional unmeasured confounding; however, a study of seroprotection, like any observational study, may have other common causes of the exposure (seropositivity) and the outcome (future infection) and it is important to think carefully about additional confounders given unique study settings and designs.

Seroprotection estimates may be biased in either direction if individuals are enrolled at varying phases of their local epidemics or from communities with differently sized outbreaks. People who

are enrolled into a seroprotection study in, for example, the early phase of an epidemic are less likely to be seropositive and have a lower daily hazard of infection than those enrolled during the peak of an epidemic. Similarly, study participants enrolled in communities with lower population infection rates are less likely to be seropositive at enrollment and less likely to become infected after enrollment than study participants from communities with higher population infection rates. Adjustments for day of enrollment and community can reduce this bias (125).

Imperfect sensitivity or specificity of serologic tests may result in bias toward the null due to misclassification of exposure status (seropositivity) in seroprotection studies. As noted in the above section on seroprevalence, seropositivity at the population level may imperfectly represent cumulative incidence due to limited sensitivity and specificity and possible changes in these over the course of the antibody response (i.e., declining sensitivity as titers decline). Analogously, at the individual level, which matters for seroprotection studies, seropositivity may be an imperfect representation of an individual's prior infection status. Some of the corrections which are effective at the population level for seroprevalence estimates are not effective at the individual level (152). Misclassification at the individual level may reduce power as well as cause bias (153).

Increases in risky behavior by those who are seropositive (risk compensation) may increase the risk of reinfection for such individuals, thereby reducing the magnitude of seroprotection by creating an indirect effect through which prior infection/seropositivity increases the risk of infection. While this will not bias estimates of the total effect of seroprotection, it will make estimating the direct effect of seropositivity on infection (the effect of interest, which does not include indirect effects mediated by changes in behavior) more challenging; without explicit consideration of behavioral changes, the effects that are evaluated in seroprotection studies will be a combination of the direct and indirect effects (**Fig. 4.4**). One potential study design to better

isolate the direct effect includes restricting the study to seropositive individuals and comparing high vs. low antibody levels (since people do not usually know this value it should not affect behavior). On the other hand, if antibody levels are a function of previous disease severity (128,136) and disease severity in turn affects future behavior, this could create a different bias. Moreover, negative control outcomes (154) (e.g., risk of other respiratory infections such as respiratory syncytial virus) could be considered to assess the magnitude of the effect due to behavior differences between seropositive and seronegative individuals. Assuming that SARS-CoV-2 antibodies do not protect against influenza, differences in the number of cases of influenza between SARS-CoV-2 seropositives and seronegatives may indicate behavioral differences between the two groups. Another option may be to perform a formal mediation analysis, but the interventions defining this analysis must be explicit and plausible (155–157).

SUMMARY: An ideal study design for SARS-CoV-2 seroprotection would:

- Explicitly define a causal effect (estimand) of interest, e.g. with respect to a target trial (133).
- Adjust for factors associated with the risk of infection to reduce confounding.
- Control for (in the analysis) or match on (in the study design) time of enrollment and geographic location to mitigate confounding by epidemic dynamics.
- Think carefully about additional confounders given the unique study setting and design.
- Account for, or at least acknowledge, possible bias and/or loss of power due to imperfect sensitivity and specificity of the serologic assay.

- Give thought to the impact of risk compensation among seropositives on the effects estimated in the study.

- Consider the generalizability of the results given the dynamics of the epidemic during the trial.

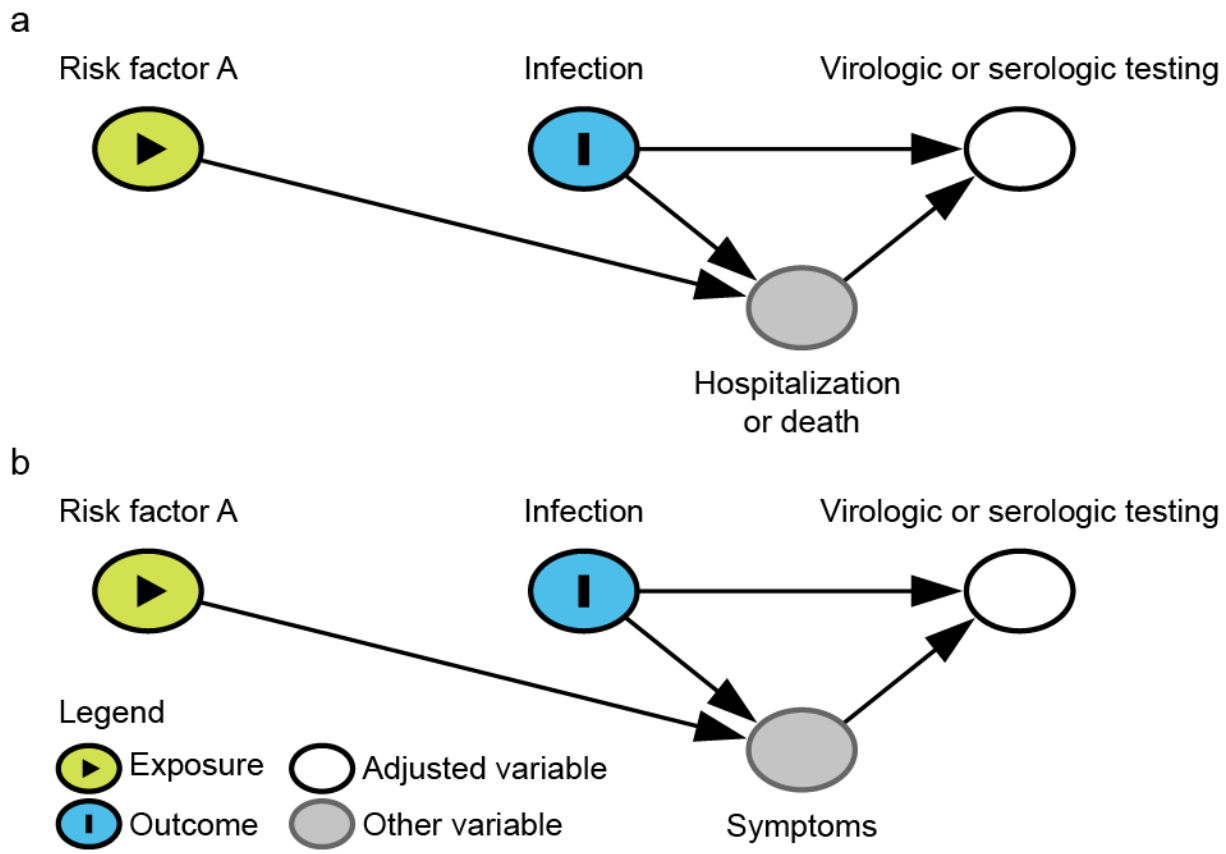
4.5 INFECTION RISK FACTORS

When a new infectious disease epidemic arises, some of the most important questions are who is most at risk of acquiring infection and, among those most vulnerable to infection, which groups are more likely to face severe illness or death. These questions are especially relevant for COVID-19, as the pandemic has disproportionately affected communities of color, people living in poverty, and other marginalized groups in the United States and internationally (158–161). Epidemiological studies are crucial for identifying demographic factors (e.g., age, gender, race/ethnicity, disability status, socio-economic status, job type), as well as structural factors (e.g., living and working conditions, literacy, racism, gender inequity) that are associated with the risk of infection in order to inform the allocation of resources and optimize the impact of prevention and treatment interventions (162). We refer to these as "risk factors" whether they are true causal factors or statistical predictors of infection (163). Even if we are only looking to identify statistical predictors of infection risk, these studies can still suffer from selection bias due to excluding certain populations or differential testing rates among populations, and from differential misclassification bias due to assuming non-tested individuals are uninfected or by combining test results across test types, timing of tests, and reasons for testing.

If a study considers a selected group of individuals who are tested to ascertain infection risk factors, selection bias can play a role (in either direction) if the risk factor of interest is related to

the likelihood of hospitalization/death, and hospitalization/death also affects the likelihood of being tested. For example, if serologic testing is performed in the community (i.e., among non-hospitalized individuals at some point in time during the outbreak), individuals who are currently hospitalized for severe infection or have died because of COVID-19 will not be included in this cohort. If infected individuals with a certain risk factor, such as a comorbidity, are more likely to experience severe disease and become hospitalized or die, there may be a negative correlation between that risk factor and infection in the study because those individuals are underrepresented in the study (**Fig. 4.5A**). This will lead to a spurious correlation under the null hypothesis and an attenuated effect estimate if the factor of interest is a true risk factor for infection. This bias can be limited by including individuals who are hospitalized at the time of testing in the sample or accounting separately for their exclusion, as could also be done for individuals who have died due to COVID-19.

Figure 4.5: Directed acyclic graph under the null hypothesis showing the possible structure of selection bias due to (a) exclusion from testing and (b) differential likelihood of testing.



Under the null hypothesis (of no effect of Risk Factor A on COVID-19 infection) selection bias can be in either direction depending on whether Risk Factor A increases or decreases the likelihood of **(a)** severe disease or **(b)** symptoms among infected individuals. The figures are simplified to illustrate these particular biases so make the strong assumption of no additional unmeasured confounding (i.e., no common causes of any two variables in the figure).

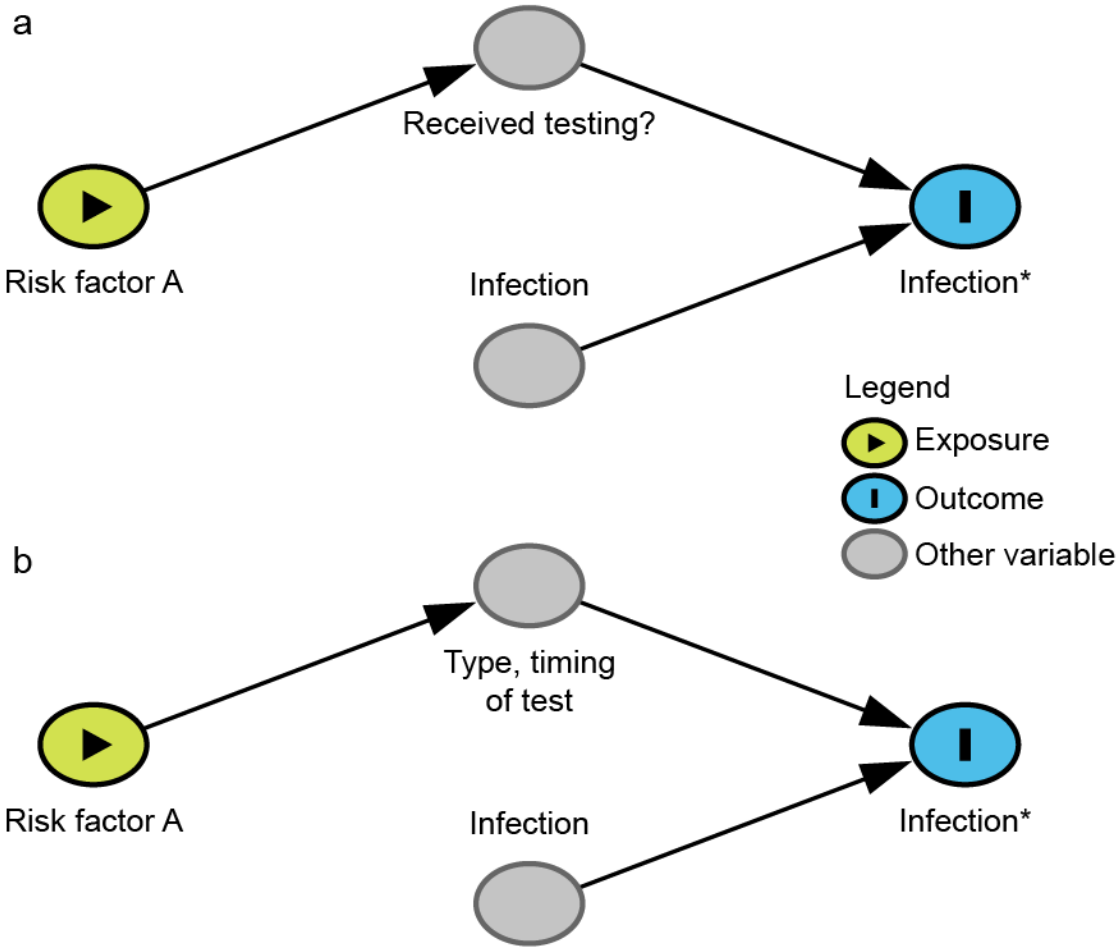
Studies of infection risk factors that include selected groups of individuals who are tested may also suffer from another selection bias if there are differential testing rates or differences in testing criteria between the groups. For instance, testing for SARS-CoV-2 in many parts of the world has often been limited to suspected cases with symptoms. Thus, symptomatic people are more likely to seek testing or may be more likely to be identified as a contact and tested compared to

individuals with atypical or no symptoms. If the risk factor of interest affects the likelihood of such symptoms, then selection bias may occur (**Fig. 4.5B**). As another example, suppose a study aimed to investigate the relationship between gender and COVID-19 among tested individuals; women, especially younger women, are more likely to seek medical care than men and therefore are more likely to be tested (164). Men who receive a test may primarily be those with more severe symptoms or with known exposure to SARS-CoV-2. This selection bias will cause a spurious negative correlation between female gender and infection among those who are tested and enrolled in the study. The bias can be avoided by 1) testing all members of a cohort regardless of symptoms or 2) stratifying on the reason for testing, such as respiratory symptoms (165), although this can limit the generalizability of the results.

In studies that use confirmed cases among the population as the outcome of interest, including both tested and untested individuals, differential misclassification bias can occur if untested individuals are assumed to be uninfected. To avoid the selection bias described in the preceding paragraph, studies may include individuals who were not tested for infection and assume they were never infected. Due to this assumption, misclassification of the outcome (infection status) will be much higher among individuals who were not tested (because of lack of symptoms and/or not seeking medical care) compared to tested individuals. This design is unlikely in a formal epidemiological study, but could occur in an ad hoc analysis of case counts. However, this approach can cause differential misclassification bias, in which risk factors for testing appear as risk factors for COVID-19 (166,167). Differential misclassification bias will occur if the risk factor increases (or decreases) the chance of testing through a causal pathway unrelated to the probability of infection (**Fig. 4.6A**). For instance, suppose a study aims to investigate whether pregnancy is a risk factor for SARS-CoV-2 infection. In some healthcare facilities and municipalities, pregnant

women are routinely screened for SARS-CoV-2 infection upon admission for delivery (168). In this case, pregnancy will be associated with a higher likelihood of being tested and, therefore, a lower likelihood of being misclassified as uninfected due to not receiving testing. This will induce a spurious positive correlation between pregnancy (or factors correlated with pregnancy, such as gender and age) and infection. Similar differential misclassification by whether an individual is tested may also occur with age, as younger individuals are tested less frequently than older individuals, especially in the early phases of the pandemic (169,170).

Figure 4.6: Directed acyclic graph under the null hypothesis showing differential misclassification by (a) whether an individual is tested and (b) the timing or type of test.



A study is trying to determine the relationship between Risk Factor A and observed infection status (Infection*), where observed infection status is a proxy for the variable of interest, true infection status (Infection). If **(a)** Risk Factor A influences whether someone is tested and all non-tested individuals are assumed to be uninfected or **(b)** Risk Factor A affects the type of test and timing of testing conducted then under the null hypothesis of no effect of Risk Factor A on COVID-19 rates misclassification can cause upward or downward bias. The figures are simplified to illustrate these particular biases, and therefore make the strong assumption of no additional unmeasured confounding (i.e., no common causes of any two variables in the figure).

Differential misclassification bias can also arise if testing is performed at different time points, both virologic and serologic testing are used in the outcome measure, or the presence of symptoms affects test performance (171) (**Fig. 4.6B**). For example, suppose a study aims to investigate the relationship between age and SARS-CoV-2 infection risk. Virologic (e.g., PCR) testing of children early in the outbreak was less frequent compared to adults because disease in children is relatively mild (169,170). Therefore, children may seem less affected in studies that include only virologic testing. On the other hand, given that serologic testing only became available later (172), children may be more likely to be tested with serologic tests compared to adults in a study that combines both serologic and virologic testing. If the total positivity rates for children and adults were compared, combining both test types, children would appear to have a higher risk of infection under the null because they were more likely to undergo serologic testing (a cumulative vs. point-in-time measure) and it occurred at a later time point. Even in studies using only serologic testing, if children were on average tested later than adults, they would have had more opportunities to become infected, which may still induce a correlation. To prevent this bias, studies of infection risk factors should avoid comparison of serology results from one group with virologic testing from another group, or comparison of combined serologic and virologic testing between groups. Additionally, this bias can be reduced in analysis by adjusting for (e.g., through stratification, matching, or control for) the type and timing (with respect to epidemic time) of the test.

SUMMARY: An ideal study design of SARS-CoV-2 infection risk factors would:

- Test all enrollees using the same test type at a fixed time point or set of fixed time points; and

- Include individuals who were hospitalized or died due to COVID-19 in the enrolled population or account for their exclusion; and
- Enroll individuals who have been randomly selected for testing through an infection surveillance program or some other mechanism; or
- Include only tested individuals (limiting the generalizability) and stratify on or otherwise adjust for the type, timing, and reason for receiving the test, as well as account for individuals not tested due to hospitalization or death.

4.6 SECONDARY ATTACK RATE ESTIMATION

We first define and differentiate the terms "infectiousness" and "secondary attack rate" (SAR). We define the term infectiousness as the probability that an infected host transmits the infection to a susceptible person during some well-defined type of contact or interaction. Infectiousness depends on factors associated with the pathogen (e.g., quantity shed, transmission-favoring mutations), host factors in the infected person (e.g., age, symptoms, severity of illness, aerosol generation), and host factors in the susceptible individual (e.g., age, health status). With the accumulation of viral genomes, recent studies have reported evidence of mutations that increase transmissibility. For example, the famous D614G mutation is associated with higher viral load and infection of younger hosts (173), and the SARS-CoV-2 VOC-202012/01 strain with the N501Y mutation has been estimated to have a R_0 1.75 times higher than the 501N strain (174). However, without long-term observation and rich sequencing data we cannot know the relative transmissibility of the virus strains observed in different studies. Therefore, we will limit these discussions to factors relevant for control measures – that is, we assume that the biological features of the virus do not change when comparing different studies. Under this assumption, we expect

that infectiousness will differ for different kinds of contact (e.g., being in proximity vs. touching vs. kissing), with different precautions (e.g., with or without mask wearing), and in different environments (e.g., indoor vs. outdoor, degree of ventilation). In an ideal world, infectiousness would be measured by the "susceptible-exposure attack rate", which is the proportion of exposures per susceptible contact leading to a transmission event -- that is, infectiousness per contact, with contact being precisely defined (175). Since exposures themselves are rarely observed, the SAR is often used as a proxy measure for infectiousness or susceptibility.

The SAR is the proportion of susceptible individuals who become infected within a group of susceptible contacts of a primary (index) case within a given time period (176). The denominator is the total number of individuals contacted in a particular setting (a particular study may but often does not assess the susceptibility of these individuals at baseline). The numerator is the number of infected secondary cases among those contacts. Contact tracing studies identify and collect information about the index case(s) (usually defined as the first identified infected individual(s)) as well as the close contacts, who are followed to observe the outcome of the exposure. SAR estimation in a particular setting, such as a household or school, can help identify the role of different social interactions, environmental factors, characteristics of the index case, and susceptibility of contacts, which can inform effective strategies to prevent onward transmission. However, biases in study design or data analysis can give rise to inaccurate SAR estimation leading to the mischaracterization of infectiousness or susceptibility. Here we summarize biases that result in inaccurate estimation of the SAR, provide some examples from current studies, and outline recommendations that should be considered to provide accurate estimates and interpretations of infectiousness and susceptibility.

Biases in estimating the SAR can be introduced through two key mechanisms: misclassification of the index case(s) (section 4.6.1, **Fig. 4.7, Fig. 4.8, Fig. S4.1**) and misclassification of close contacts (section 4.6.2, **Fig. 4.9**).

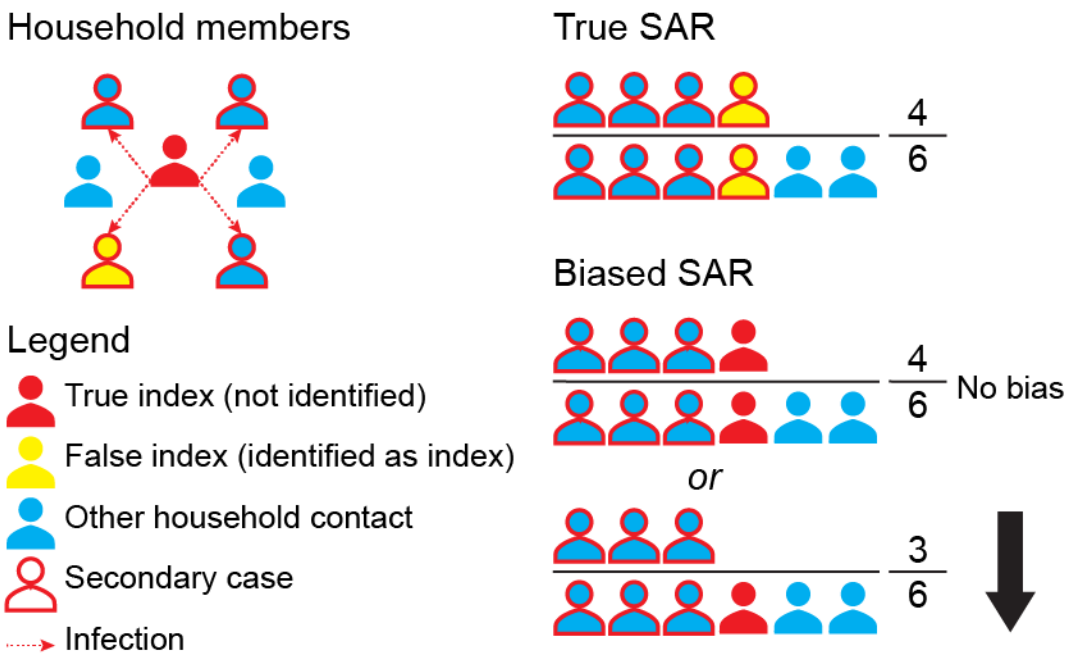
4.6.1 Secondary attack rate estimation: misclassification of the index case(s)

Misclassification of the index case(s) can happen in two ways: a non-primary (i.e., secondary or tertiary) case is falsely identified as the index case (**Fig. 4.7, Fig. S4.1**) or only one index case is identified when in fact multiple index cases are present (**Fig. 4.8**).

Identifying a non-primary case as the index case may bias SAR estimates up, down, or create no bias in some household settings. In an outbreak investigation or a household contact tracing study, a secondary case that has more obvious clinical symptoms or epidemiological characteristics (i.e., imported vs. local cases) relative to the true index case may falsely be classified as the index case. For COVID-19, it is possible for an index case to develop symptoms later than the secondary cases (177); therefore, studies that define the index as the first identified case in a cluster, especially in household context (178,179), are prone to this error, which could bias the SAR downward or cause no bias (**Fig. 4.7**). For scenarios outside the household, the direction of the bias depends on the relative number of contacts from the true index case and the misclassified one (**Fig. S4.1**). Epidemiological history can cause an analogous bias; imported cases may be more likely to be identified as index cases than local community cases, especially at a relatively early stage of an outbreak when importation of cases from the epidemic center and local transmissions are both happening simultaneously. For example, identified asymptomatic cases during the third wave in Hong Kong in July and August 2020 were more likely to be imported, which may indicate different screening and testing practices for travelers (180). This tends to bias the SAR of imported cases

upwards and the SAR of local index cases downwards, which may lead to underestimation of existing community transmission or delay the detection of local transmission chains.

Figure 4.7: Illustration of index case misclassification where the index and secondary cases are misclassified in a household scenario.

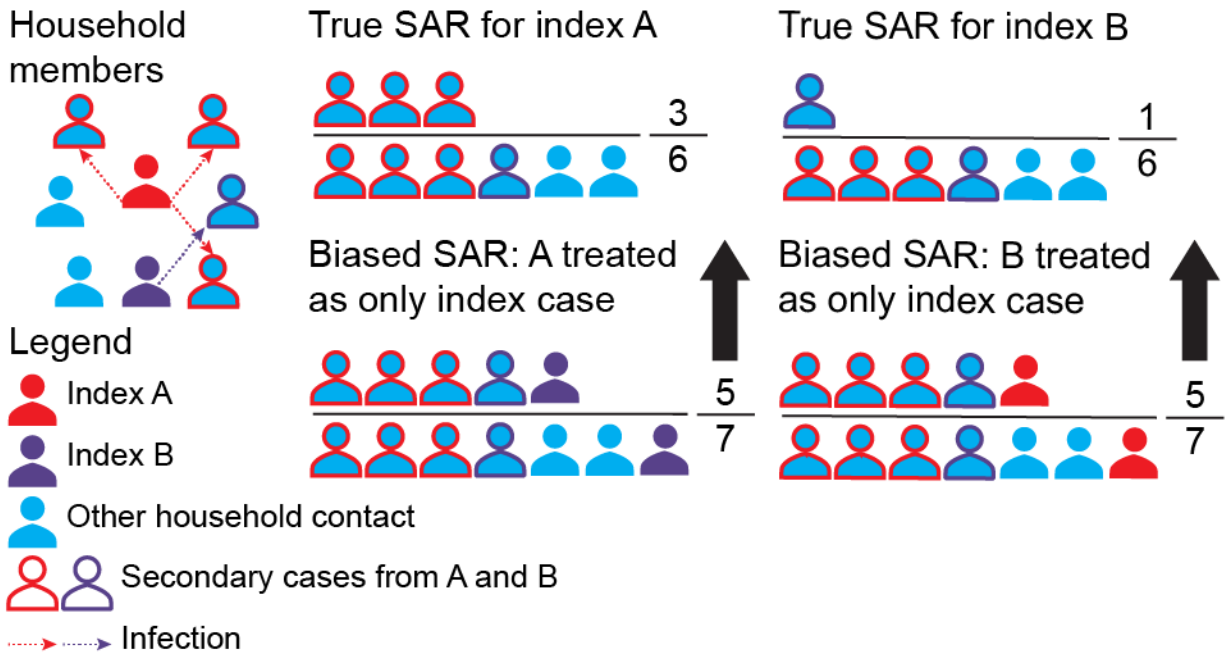


In this scenario (top left), each individual has close contacts with every other household member, and the red arrows indicate infections transmitted by the true index case (red individual) to other household members. The true SAR is shown in the top right; the infected contacts of the true index are in the numerator and all contacts are in the denominator. Index case misclassification can happen if one of the secondary cases of the index is falsely identified as the index case (yellow individual). This may cause no bias in the estimation of the SAR value; however, the interpretation of this SAR may be incorrect because we mistakenly attribute the SAR to the false index case, who may have different characteristics, such as age, from the true index case. It can also introduce downward bias if the true index is no longer detected by PCR by the time they are tested (bottom right).

Failure to identify the existence of multiple index cases in a cluster can bias SAR estimates upward (**Fig. 4.8**). If two (or more) cases - A and B - are index cases, one of them may be classified as the

only index case and all the secondary cases detected will be attributed to this individual. This causes the SAR to be biased upwards as the secondary cases from B and perhaps even B themselves will be attributed as secondary cases of A. This is most likely in settings with dynamic populations where the source of infection is unclear (i.e., gyms, bars, nursing homes, or gathering events) and multiple index cases may be infecting people at the same time (181,182). One example of this bias can be seen in a study from South Korea (178) in which the initial report suggested that the SAR of index cases aged 10-19 years was significantly higher than for index cases aged 20-29, 30-39 and 40-49 years. However, there was a great deal of uncertainty about the true index case in the 10-19 year group due to common sources of exposure with other family members. In a re-analysis (183) of this data, the authors removed household members who potentially shared a common source of exposure with the pediatric cases, resulting in a much lower SAR for the 10-19 year group. Another example is a nursing home outbreak investigation from the Netherlands, where a church service was initially thought to be the source of the outbreak; however, genome sequencing showed multiple clusters in the viral genomes, suggesting multiple introductions to the nursing home (182).

Figure 4.8: Illustration of index case misclassification when multiple index cases are present but only one is identified as the index case.



As shown in the top left, two index cases (red and purple individuals) acquired the infection and transmitted it (red and purple arrows) to other household members. As members of a household are often considered to all be in contact with one another, we cannot distinguish who truly infected whom. The true SARs for each index case are shown in the top. The numerator consists of infected contacts and the denominator consists of all contacts. Upward bias in the SAR can be introduced by falsely attributing all infections, including the other index case, to one of the two index cases.

Index case misclassification can be minimized through rigorous follow-up of the selected study population, using frequent and standardized testing, symptom monitoring, and daily contact diaries to track potential infections and index cases. Furthermore, viral genomic analyses, combined with information on contacts, have the potential to more accurately reconstruct transmission pairs or chains of transmission, and further reduce bias in the identification of index cases (184,185). In addition, the chain-binomial model can also be used to avoid the misclassification of secondary or

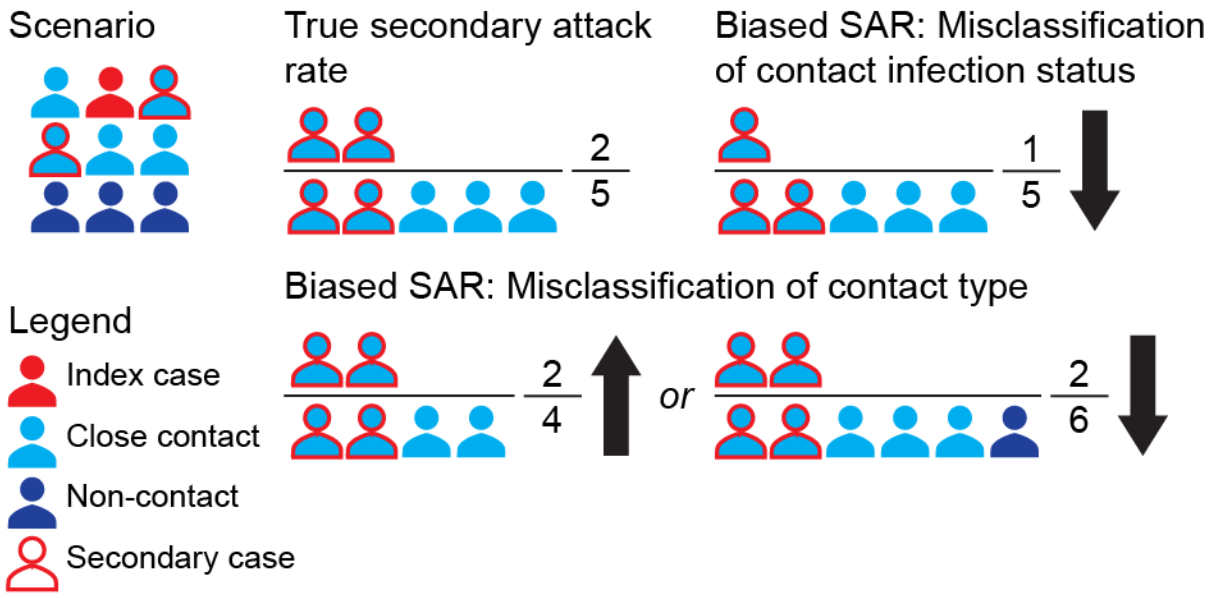
tertiary cases as index cases, especially in household settings (186). An ideal but less practical solution, particularly early in an outbreak, is to conduct a prospective cohort study, either a household study or another population-based study, in which subjects are enrolled before infection and followed up over time.

3.6.2 Secondary attack rate estimation: misclassification of close contacts

Imprecise definitions of "close contact" can complicate estimation and interpretation of the SAR. Ambiguity or bias due to misclassification of close contacts is a common problem in contact tracing studies.

Misclassification of close-contact identification is often related to how the study defines and recruits close contacts, which directly impacts the denominator of the SAR estimate (**Fig. 4.9**). If the total number of contacts is not fully documented due to an unclear definition of close contact, it could bias the SAR upward because the closest contacts are more likely to be documented. For example, one study investigated how mask wearing in pre-symptomatic patients can prevent SARS-CoV-2 transmission (187). In this study, a maximum of ten close contacts of each index case were selected even though the index case may have had more than ten close contacts. This artificial limit would bias the SAR upward for index cases with over ten close contacts. On the other hand, if all contacts including non-close contacts are counted as close contacts, that is, all individuals in the setting rather than well-defined close contacts, it would bias the SAR downward. For example, one study (188) included all employees, family members, and clients of a supermarket over a certain time period, although not all individuals had close contact with the index case. While this study reported separate SARs for different contact groups, the overall SAR reported in this study is not comparable to that of other studies that include only close contacts.

Figure 4.9: Illustration of misclassification of contact type and contact infection status.



As shown in the top left, an infected individual infects some of their close contacts. The true SAR is represented in the top middle; the infected contacts are in the numerator and all close contacts are in the denominator. Bias due to misclassification of contact type can go in both directions. Bias is in the upward direction if some close contacts are missed during contact-tracing (bottom middle), and in the downward direction if non-close contacts are falsely considered as close contacts (bottom right). Misclassification of contact infection status can happen when close contacts are not appropriately tested or followed-up and creates downward bias (top right).

Misclassification of infection status in the close contacts directly affects the numerator of the SAR (**Fig. 4.9**). If some secondary infections are missed because contacts are not followed for an adequate duration (188,189) or different outcome ascertainment procedures are used for different groups (190), it would bias the SAR downward. For example, if only contacts with respiratory symptoms are tested then this could bias the SAR downward as seen in a study that recruited 445 close contacts, but only tested the 54 who developed new or worsening symptoms during active symptom monitoring (190). This means untested contacts were counted as uninfected. A similar but less severe bias occurs when different tests with imperfect sensitivity are used for contacts,

which could also bias the SAR downward. Misclassification may also exist when a study uses real-time polymerase chain reaction (RT-PCR) but tests contacts too early or too late after exposure, resulting in a low yield in test positivity (an example under **Fig. 4.10**). For example, RT-PCR missed 36% (95% CI: 28%, 44%) of infected close-contacts, especially among those who were tested in the first few days after exposure (191).

Proposed solutions to avoid misclassification of the identification and infection status of close contacts are to refine study protocols beforehand. For example, a clear close contact definition and a standardized protocol to define and identify all potential close contacts throughout the study can help eliminate misclassification. A standardized and high sensitivity testing plan for all close contacts (ideally, regardless of symptom status), and reasonable follow-up (ideally, at least for the duration of the incubation period, for example, 14 days for COVID-19) to observe the outcome of exposure can provide complete ascertainment of the secondary cases.

4.6.3 Interpreting comparisons of the secondary attack rate to make inference about relative susceptibility or infectiousness

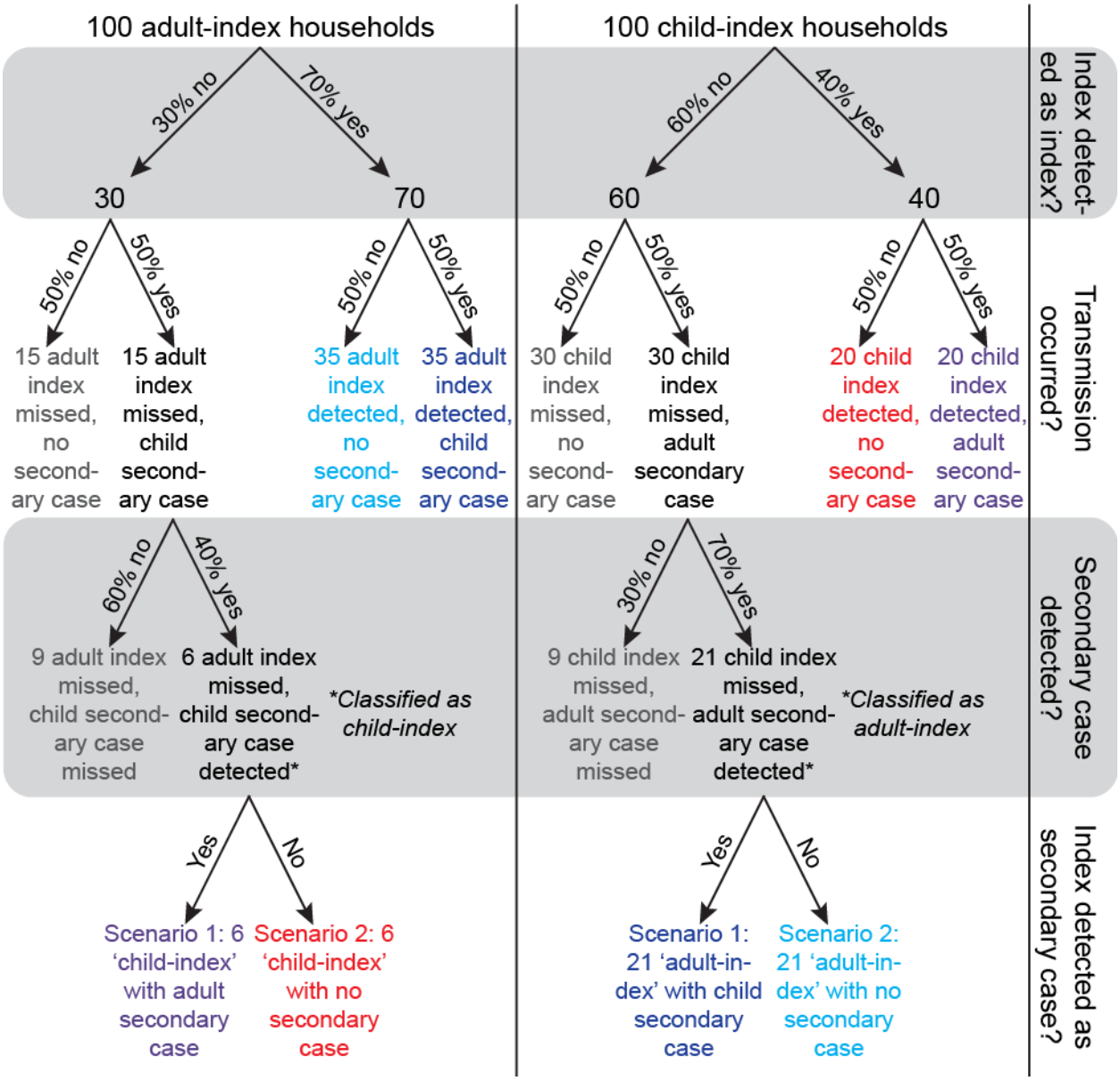
The previous section cataloged the ways in which various biases could affect SAR estimation. If the SAR can be used to estimate infectiousness, then comparing the SAR between index cases can provide information about the association of index case characteristics (e.g., age) with infectiousness. Likewise, SAR comparisons between contacts of different demographic types can be used to infer the relative susceptibility of different types of individuals. However, there are several factors that need to be considered when using the SAR to infer relative susceptibility or infectiousness. Here we discuss factors that influence the interpretation of the SAR, such as comparing the SAR by characteristics of the index cases or close contacts, contact patterns, and

across environmental settings. For demographics, we use adults compared to children as the running example.

If identical biases were present in estimating the SAR for adult and child index cases, we might expect the comparison of infectiousness and/or susceptibility to be unaffected. However, many of the biases depend on, for example, the probability of identifying someone correctly as an index case (as discussed previously in sub-section 4.6.1), and this may differ between adults and children because children with SARS-CoV-2 infection often have milder or no symptoms (192), and thus may be less likely to be classified as the index case for the cluster than adults. Similarly, children may also be less likely to be identified as secondary cases. Thus, these biases in SAR estimation can differentially affect child and adult index cases and/or contacts, and may lead to biases in comparative infectiousness or susceptibility.

One hypothetical scenario we illustrate here shows how this bias is introduced and the corresponding results, where the bias can go in either direction even under the null hypothesis of equal infectiousness (**Fig. 4.10, Table 4.1**). In this example, where all households have only one child and one adult, the infectiousness of children will be underestimated (and their susceptibility will be overestimated) if missed index cases are misclassified as secondary cases (scenario 1), while the opposite bias will occur if missed index cases are not identified as secondary cases (scenario 2), perhaps because they are no longer positive by the time they are tested. As shown in detail in **Fig. 4.10** and **Table 4.1**, differential misclassification between adults and children can lead to biases in either direction in the estimation of their relative infectiousness or susceptibility, even when no such differences exist (i.e., the null hypothesis is true, **Fig. 4.10, Table 4.1**). We use this simple example for illustration; actual studies will typically have a mixture of household structures and may also face the issues described in the next section.

Figure 4.10: Illustration of differential detection of infection in adults and children.



Flowcharts of infection and detection are presented in the diagram. Households shown in the same colors represent the same results, no matter whether they are true or misclassified. Households in grey remain completely undetected.

We consider a simplified example for intergenerational household transmission; all households are composed of one adult and one child, so that the only transmission opportunity is to one individual of the other age category. We use 100 households with an adult index case (left column), and 100 households with a child index (right column). This scenario is drawn under the null hypothesis of equal infectiousness of adults and children, and both age groups transmit

Figure 4.10 (Continued)

the infection half (50%) of the time. The only difference between infected adults and children is the probability that they are detected, reflecting differential symptom presentation. We assume that 70% of adults and 40% of children are detected (the numbers chosen are illustrative and the key information is that adults are more likely to be detected). Additionally, we assume that testing works perfectly (i.e., all contacts are tested and identified accurately) and no testing is triggered by contacts outside of the household. We consider two scenarios when an index case is missed and their secondary case is detected. In both scenarios the secondary case is falsely considered to be the index case and the true index is tested as a potential secondary case. In scenario 1, the true index can still be detected and will falsely be considered a secondary case of the false index. In scenario 2, the true index can no longer be detected and the false index will be considered to have not infected anyone. The SARs under both scenarios are calculated in **Table 4.1**, which shows that the differential detection of infections in adults and children creates a bias that can go in either direction. Under scenario 1, the SAR is higher for adult indices than for child indices, while for scenario 2 the SAR is higher for child indices than for adult indices.

Table 4.1: Calculation of the SAR when there is differential detection of infection in adults and children.

Scenario 1: The true index can still be detected and is falsely considered a secondary case of the false index			
	Number of households with transmission	Total number of households	**SAR**
Households classified as **adult index** with child at risk	**35 + 21 = 56**	**56 + 35 = 91**	**56/91 = 62%**
Households classified as **child index** with adult at risk	**20 + 6 = 26**	**26 + 20 = 46**	**26/46 = 57%**

Scenario 2: The true index can no longer be detected and the false index is not considered to have any secondary cases			
	Number of households with transmission	Total number of households	**SAR**
Households classified as **adult index** with child at risk	**35**	**35 + 56 = 91**	**35/91 = 38%**
Households classified as **child index** with adult at risk	**20**	**20 + 26 = 46**	**20/46 = 43%**

The tables show the SAR calculation under the two scenarios displayed in **Fig. 4.10**. The differential detection of infections in adults and children creates a bias, which can go in either direction. In scenario 1, adult indices have a higher SAR than child indices. In scenario 2, adult indices have a lower SAR than child indices.

Besides factors related to the identification of index cases and close contacts, attention must be paid to contact patterns, including the duration of contact, contact frequency, and the setting where contact occurs, when using the SAR to make inferences about infectiousness. This includes

heterogeneity in contact behaviors, exposure settings, and contact populations. Individuals may infect more secondary cases simply because they have prolonged and closer contacts or riskier behaviors. Higher risks can come from the gathering pattern, such as living together, sleeping in the same room, dining together, or activities such as singing/shouting, and playing board games (193). For example, spouses have higher attack rates compared to other household contacts with odds ratios of 2.27 (95% CI: 1.22, 4.22) (194) and 3.66 (95% CI: 1.28, 10.5) (195), indicating that the marital status of household contacts should be accounted for in household studies to better disentangle biological from behavioral factors in the infectiousness of household index cases. The setting of the exposure is also important. Different environmental settings, such as indoor vs. outdoor, household vs. non-household (196), environments that tend to generate aerosols (some clinical treatment processes (197)), or poor ventilated settings, may lead to differences in transmission. Suppose there are two types of individuals - type A and type B - who are identical other than the factor being considered. If, all else equal, individuals of type A have prolonged contact with others in a poorly ventilated setting, then type A indices will have more secondary cases than type B indices. Or if type A indices have more contacts in these settings and engage in higher risk activities when the viral load is highest (i.e., -2 to 5 days after symptom onset (198)) and they are more likely to be highly infectious then they may generate more secondary cases (i.e., superspreading events) than type B indices. The third component in contact heterogeneity is the age structure or demographic characteristics of the close contacts. For example, if type A individuals have more contacts with older individuals, then type A indices may end up infecting more people than type B indices, as emerging evidence suggests susceptibility to infection increases somewhat with age (127). However, studies rarely report the age structure or underlying

conditions of close contacts. While this issue can influence studies about infectiousness or susceptibility, it can be easily avoided by collecting relevant information.

Proposed solutions to accurately infer and compare infectiousness include testing stool samples in children because the duration of viral RNA shedding is longer so they are less likely to be misclassified (199,200); reporting differences in contact patterns, including activities/behaviors, duration of contact, contact frequency, and contact setting; and collecting detailed epidemiological characteristics of close contacts, such as age, gender, and underlying conditions.

SUMMARY: An ideal study design for SARS-CoV-2 secondary attack rate, infectiousness, and susceptibility estimation would:

- Ensure rigorous follow-up of the study population to minimize misclassification of the index case. This could integrate whole-genome sequencing and phylogenetic analysis to improve the identification of the index case(s), introduction of multiple index cases, transmission directions, chains of transmission, and network interactions. Repeat testing that gives information about viral load may also inform inference of the relative probability that different individuals are index cases (201).

- Use a prospective cohort design, such as a household study, in which subjects are enrolled before infection and followed over time. This could involve frequent serial testing, symptom monitoring, and the use of daily contact diaries. In contact tracing studies, stool samples might be an option for testing children to reduce misclassification of child cases and contacts.

- Clearly define “close contact” and use a standardized protocol for identifying all potential close contacts. Have a standardized and highly sensitive testing plan for assessing infection in all close contacts.

- Use a sufficiently long length of follow-up to observe the outcome of exposure (ideally, at least for the duration of the incubation period, for example, 14 days for COVID-19) among all close contacts.

- Consider hypotheses about shared exposures and collect information on them (for example, travel together to an infected area) (202) or stratify results in main or sensitivity analyses to exclude possible shared exposures (203).

- Simulate *in silico* outcome data for contact tracing studies under different plausible infection scenarios to understand the impacts of potential misclassifications of index cases or close contacts.

- Be aware of differences in contact patterns, including the duration of contact, contact frequency, contact setting, and epidemiological characteristics of close contacts, when using the SAR to infer that certain groups, such as adults, have higher biological, per-contact infectiousness or susceptibility than other groups, such as children.

4.7 CONCLUSIONS

To assist in the evaluation of a continually expanding body of literature on COVID-19, we have outlined and proposed solutions to common biases that can occur across different types of observational studies of COVID-19, including cross-sectional seroprevalence, longitudinal seroprotection, risk factor studies to inform interventions, studies to estimate the secondary attack

rate, and studies that use the secondary attack rate to make inferences about relative infectiousness or susceptibility. Across study designs, we identified issues of interpretation, as well as possible biases due to measurement error, selection bias, confounding, and recruitment of non-representative samples. In particular, we highlighted how studies of seroprevalence are subject to misclassification by antibody tests and the possible recruitment of non-representative samples, while studies of seroprotection may suffer from confounding by geographic structure, epidemic dynamics and risk of infection, and their interpretation may be complicated by risk compensation. Studies of infection risk factors may be prone to biased selection of subjects, resulting from the presence of symptoms or hospitalization/death status, and differential misclassification of infection status due to testing factors. Lastly studies of the secondary attack rate can be biased due to misclassification of the index case(s), and failure to correctly identify close contacts and determine their infection statuses, while the use of secondary attack rates to make inferences about infectiousness and susceptibility must be performed carefully with awareness of contact patterns. Although each bias is discussed separately in each study design, multiple biases may coexist and need to be examined carefully in real settings. We hope these thorough descriptions of biases can provide a map or checklist of potential biases to assist with both future study design and the critical interpretation of existing study results.

4.8 DECLARATIONS

Figure creation

Figures were created using Adobe Illustrator, the *ggplot2* package in R, and the DAGitty online tool.

Availability of data and material

Not applicable.

Conflicts of interest/Competing interests

ML reports grants from NIH/NIGMS, during the conduct of the study; personal fees from Affinivax, personal fees from Merck, grants and personal fees from Pfizer, grants from PATH Vaccine Solutions, outside the submitted work.

Funding

EKA was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number T32AI007535. XQ, EG, and RN were supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number U54GM088558. ER and LKS were supported by the Morris-Singer Fund. ML was supported by the National Institutes of Health under cooperative agreement U01 CA261277, and the Morris-Singer Fund. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Morris-Singer Fund.

Authors' contributions

EKA, LKS, RK, EG, and ML conceptualized the project. EKA, XQ, ER, LKS, RK, KJ, MC, and ML wrote the manuscript. EKA, ER, LKS, RK, and KJ created manuscript figures. EG, MJS, RN, MC, and ML provided conceptual feedback. MC and ML supervised the research. All authors edited the manuscript and approved the final manuscript.

Acknowledgements

The authors express their gratitude to Ali Ihsan Nergiz for his help identifying and extracting data from studies cited in the "Secondary attack rate (SAR) estimation" section; Shae Gantt, Yonatan Grad, Siyang Xia, and James Hay for their insightful feedback on the project; and the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (award number T32AI007535), the National Institute of General Medical Sciences of the National Institutes of Health (award number U54GM088558), the Morris-Singer Fund, and the National Institutes of Health (cooperative agreement U01 CA261277) for funding.

Chapter 5. Conclusion

We first examined the nasal carriage of *Staphylococcus aureus* bacteria, a leading cause of healthcare- and community-associated infections, in a longitudinal cohort of mother-infant pairs profiled with shotgun metagenomic sequencing. We characterized the development of the infant nasal microbiome, and identified ecological and functional genetic elements within the infant and maternal nasal microbiomes that influenced *S. aureus* acquisition and retention in early life. In particular, we found a gene family, likely acting as a taxonomic marker for an unclassified species, that was significantly anti-correlated with *S. aureus* in infants and mothers. Additionally, we determined that mothers were an important early influence of the naive infant microbiome, likely acting as a source of both beneficial species and *S. aureus* for infants, while microbiome determinants of *S. aureus* acquisition became more important later on.

Secondly, we considered mediation analysis for complex microbiome data, which calculates the proportion of the effect of an exposure on an outcome that occurs through changes in microbiome composition and function. Using realistic synthetic data, we performed benchmarking for statistical methods for mediation analysis and explored the properties under which mediation analyses performed well by varying the number of mediating microbiome features, the strength and direction of mediation, and other properties of the datasets. We used our findings to inform mediation analyses exploring the role of the gut microbiome as a mediator between diet and cardiometabolic disease in two datasets. We made recommendations for the methods to use for estimation of the total indirect effect, and hypothesis testing for total indirect and component indirect effects. However, given the respective strengths and weaknesses of the methods, an end-user may need to select multiple methods to accomplish a full mediation analysis with microbiome data. This can lead to multiple varying effect estimates that are difficult to

interpret and more work is needed to continue to improve method usability and to provide tools for interpretation for end-users.

Lastly with the onset of the SARS-CoV-2 pandemic and accompanying rapid expansion of the literature, we reviewed epidemiological biases that can occur across five important classes of research questions for SARS-CoV-2 and COVID-19, including estimates of seroprevalence, estimates of seroprotection, studies of risk factors for becoming infected, estimates of the secondary attack rate, and comparisons of secondary attack rates to make inferences about susceptibility and infectiousness. To aid in interpretation of the literature and future study design, we outlined how these biases can be corrected at different stages of the study process and described the features of an ideal study design for each research question.

References

1. Tong SYC, Davis JS, Eichenberger E, Holland TL, Fowler VG Jr. Staphylococcus aureus infections: epidemiology, pathophysiology, clinical manifestations, and management. Clin Microbiol Rev. 2015 Jul;28(3):603–61.

2. Magill SS, O'Leary E, Janelle SJ, Thompson DL, Dumyati G, Nadle J, et al. Changes in Prevalence of Health Care-Associated Infections in U.S. Hospitals. N Engl J Med. 2018 Nov 1;379(18):1732–44.

3. Olaniyi R, Pozzi C, Grimaldi L, Bagnoli F. Staphylococcus aureus-Associated Skin and Soft Tissue Infections: Anatomical Localization, Epidemiology, Therapy and Potential Prophylaxis. Curr Top Microbiol Immunol. 2017;409:199–227.

4. Klevens RM, Edwards JR, Gaynes RP, National Nosocomial Infections Surveillance System. The impact of antimicrobial-resistant, health care-associated infections on mortality in the United States. Clin Infect Dis. 2008 Oct 1;47(7):927–30.

5. Kourtis AP, Hatfield K, Baggs J, Mu Y, See I, Epson E, et al. Vital Signs: Epidemiology and Recent Trends in Methicillin-Resistant and in Methicillin-Susceptible Staphylococcus aureus Bloodstream Infections - United States. MMWR Morb Mortal Wkly Rep. 2019 Mar 8;68(9):214–9.

6. Johnson RC, Ellis MW, Lanier JB, Schlett CD, Cui T, Merrell DS. Correlation between nasal microbiome composition and remote purulent skin and soft tissue infections. Infect Immun. 2015 Feb;83(2):802–11.

7. Luzar MA, Coles GA, Faller B, Slingeneyer A, Dah GD, Briat C, et al. Staphylococcus aureus

nasal carriage and infection in patients on continuous ambulatory peritoneal dialysis. N Engl J Med. 1990 Feb 22;322(8):505–9.

8. Kluytmans JA, Mouton JW, Ijzerman EP, Vandenbroucke-Grauls CM, Maat AW, Wagenvoort JH, et al. Nasal carriage of Staphylococcus aureus as a major risk factor for wound infections after cardiac surgery. J Infect Dis. 1995 Jan;171(1):216–9.

9. von Eiff C, Becker K, Machka K, Stammer H, Peters G. Nasal carriage as a source of Staphylococcus aureus bacteremia. Study Group. N Engl J Med. 2001 Jan 4;344(1):11–6.

10. Wertheim HFL, Vos MC, Ott A, van Belkum A, Voss A, Kluytmans JAJW, et al. Risk and outcome of nosocomial Staphylococcus aureus bacteraemia in nasal carriers versus non-carriers. Lancet. 2004;364(9435):703–5.

11. Brugger SD, Bomar L, Lemon KP. Commensal-Pathogen Interactions along the Human Nasal Passages. PLoS Pathog. 2016 Jul;12(7):e1005633.

12. Gorwitz RJ, Kruszon-Moran D, McAllister SK, McQuillan G, McDougal LK, Fosheim GE, et al. Changes in the prevalence of nasal colonization with Staphylococcus aureus in the United States, 2001-2004. J Infect Dis. 2008 May 1;197(9):1226–34.

13. Williams RE. Healthy carriage of Staphylococcus aureus: its prevalence and importance. Bacteriol Rev. 1963 Mar;27:56–71.

14. Eriksen NH, Espersen F, Rosdahl VT, Jensen K. Carriage of Staphylococcus aureus among 104 healthy persons during a 19-month period. Epidemiol Infect. 1995 Aug;115(1):51–60.

15. Hu L, Umeda A, Kondo S, Amako K. Typing of Staphylococcus aureus colonising human nasal carriers by pulsed-field gel electrophoresis. J Med Microbiol. 1995 Feb;42(2):127–32.

16. Kluytmans J, van Belkum A, Verbrugh H. Nasal carriage of Staphylococcus aureus: epidemiology, underlying mechanisms, and associated risks. Clin Microbiol Rev. 1997 Jul;10(3):505–20.

17. Reiss-Mandel A, Rubin C, Maayan-Mezger A, Novikov I, Jaber H, Dolitzky M, et al. Patterns and Predictors of Staphylococcus aureus Carriage during the First Year of Life: a Longitudinal Study. J Clin Microbiol [Internet]. 2019 Sep;57(9). Available from: http://dx.doi.org/10.1128/JCM.00282-19

18. Andersen PS, Pedersen JK, Fode P, Skov RL, Fowler VG Jr, Stegger M, et al. Influence of host genetics and environment on nasal carriage of staphylococcus aureus in danish middle-aged and elderly twins. J Infect Dis. 2012 Oct;206(8):1178–84.

19. Liu CM, Price LB, Hungate BA, Abraham AG, Larsen LA, Christensen K, et al. Staphylococcus aureus and the ecology of the nasal microbiome. Sci Adv. 2015 Jun;1(5):e1400216.

20. Prince T, McBain AJ, O'Neill CA. Lactobacillus reuteri protects epidermal keratinocytes from Staphylococcus aureus-induced cell death by competitive exclusion. Appl Environ Microbiol. 2012 Aug;78(15):5119–26.

21. Uehara Y, Nakama H, Agematsu K, Uchida M, Kawakami Y, Abdul Fattah AS, et al. Bacterial interference among nasal inhabitants: eradication of Staphylococcus aureus from nasal cavities by artificial implantation of Corynebacterium sp. J Hosp Infect. 2000 Feb;44(2):127–33.

22. Lai Y, Cogen AL, Radek KA, Park HJ, Macleod DT, Leichtle A, et al. Activation of TLR2

by a small molecule produced by Staphylococcus epidermidis increases antimicrobial defense against bacterial skin infections. J Invest Dermatol. 2010 Sep;130(9):2211–21.

23. Lijek RS, Luque SL, Liu Q, Parker D, Bae T, Weiser JN. Protection from the acquisition of Staphylococcus aureus nasal carriage by cross-reactive antibody to a pneumococcal dehydrogenase. Proc Natl Acad Sci U S A. 2012 Aug 21;109(34):13823–8.

24. Wanke I, Steffen H, Christ C, Krismer B, Götz F, Peschel A, et al. Skin commensals amplify the innate immune response to pathogens by activation of distinct signaling pathways. J Invest Dermatol. 2011 Feb;131(2):382–90.

25. Cogen AL, Yamasaki K, Sanchez KM, Dorschner RA, Lai Y, MacLeod DT, et al. Selective antimicrobial action is provided by phenol-soluble modulins derived from Staphylococcus epidermidis, a normal resident of the skin. J Invest Dermatol. 2010 Jan;130(1):192–200.

26. Gonzalez DJ, Haste NM, Hollands A, Fleming TC, Hamby M, Pogliano K, et al. Microbial competition between Bacillus subtilis and Staphylococcus aureus monitored by imaging mass spectrometry. Microbiology. 2011 Sep;157(Pt 9):2485–92.

27. Iwase T, Uehara Y, Shinji H, Tajima A, Seo H, Takada K, et al. Staphylococcus epidermidis Esp inhibits Staphylococcus aureus biofilm formation and nasal colonization. Nature. 2010 May 20;465(7296):346–9.

28. Shu M, Wang Y, Yu J, Kuo S, Coda A, Jiang Y, et al. Fermentation of Propionibacterium acnes, a commensal bacterium in the human skin microbiome, as skin probiotics against methicillin-resistant Staphylococcus aureus. PLoS One. 2013 Feb 6;8(2):e55380.

29. Regev-Yochay G, Trzcinski K, Thompson CM, Malley R, Lipsitch M. Interference between

Streptococcus pneumoniae and Staphylococcus aureus: In vitro hydrogen peroxide-mediated killing by Streptococcus pneumoniae. J Bacteriol. 2006 Jul;188(13):4996–5001.

30. Zipperer A, Konnerth MC, Laux C, Berscheid A, Janek D, Weidenmaier C, et al. Human commensals producing a novel antibiotic impair pathogen colonization. Nature. 2016 Jul 28;535(7613):511–6.

31. Bieber L, Kahlmeter G. Staphylococcus lugdunensis in several niches of the normal skin flora. Clin Microbiol Infect. 2010 Apr;16(4):385–8.

32. Kaspar U, Kriegeskorte A, Schubert T, Peters G, Rudack C, Pieper DH, et al. The culturome of the human nose habitats reveals individual bacterial fingerprint patterns. Environ Microbiol. 2016 Jul;18(7):2130–42.

33. Krismer B, Weidenmaier C, Zipperer A, Peschel A. The commensal lifestyle of Staphylococcus aureus and its interactions with the nasal microbiota. Nat Rev Microbiol. 2017 Oct 12;15(11):675–87.

34. Laux C, Peschel A, Krismer B. Staphylococcus aureus Colonization of the Human Nose and Interaction with Other Microbiome Members. Microbiol Spectr [Internet]. 2019 Mar;7(2). Available from: http://dx.doi.org/10.1128/microbiolspec.GPP3-0029-2018

35. Fredheim EGA, Flægstad T, Askarian F, Klingenberg C. Colonisation and interaction between S. epidermidis and S. aureus in the nose and throat of healthy adolescents. Eur J Clin Microbiol Infect Dis. 2015 Jan;34(1):123–9.

36. Yan M, Pamp SJ, Fukuyama J, Hwang PH, Cho D-Y, Holmes S, et al. Nasal microenvironments and interspecific interactions influence nasal microbiota complexity and

S. aureus carriage. Cell Host Microbe. 2013 Dec 11;14(6):631–40.

37. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, et al. Succession of microbial consortia in the developing infant gut microbiome. Proc Natl Acad Sci U S A. 2011 Mar 15;108 Suppl 1:4578–85.

38. Maynard CL, Elson CO, Hatton RD, Weaver CT. Reciprocal interactions of the intestinal microbiota and immune system. Nature. 2012 Sep 13;489(7415):231–41.

39. Leshem E, Maayan-Metzger A, Rahav G, Dolitzki M, Kuint J, Roytman Y, et al. Transmission of Staphylococcus aureus from mothers to newborns. Pediatr Infect Dis J. 2012 Apr;31(4):360–3.

40. Lebon A, Labout JAM, Verbrugh HA, Jaddoe VWV, Hofman A, van Wamel W, et al. Dynamics and determinants of Staphylococcus aureus carriage in infancy: the Generation R Study. J Clin Microbiol. 2008 Oct;46(10):3517–21.

41. Capone KA, Dowd SE, Stamatas GN, Nikolovski J. Diversity of the human skin microbiome early in life. J Invest Dermatol. 2011 Oct;131(10):2026–32.

42. Bomar L, Brugger SD, Lemon KP. Bacterial microbiota of the nasal passages across the span of human life. Curr Opin Microbiol. 2018 Feb;41:8–14.

43. Rawls M, Ellis AK. The microbiome of the nose. Ann Allergy Asthma Immunol. 2019 Jan;122(1):17–24.

44. Ta LDH, Yap GC, Tay CJX, Lim ASM, Huang C-H, Chu CW, et al. Establishment of the nasal microbiota in the first 18 months of life: Correlation with early-onset rhinitis and wheezing. J Allergy Clin Immunol. 2018 Jul;142(1):86–95.

45. Peterson SW, Knox NC, Golding GR, Tyler SD, Tyler AD, Mabon P, et al. A Study of the Infant Nasal Microbiome Development over the First Year of Life and in Relation to Their Primary Adult Caregivers Using cpn60 Universal Target (UT) as a Phylogenetic Marker. PLoS One. 2016 Mar 28;11(3):e0152493.

46. Mika M, Mack I, Korten I, Qi W, Aebi S, Frey U, et al. Dynamics of the nasal microbiota in infancy: a prospective cohort study. J Allergy Clin Immunol. 2015 Apr;135(4):905–12.e11.

47. Proctor DM, Relman DA. The Landscape Ecology and Microbiota of the Human Nose, Mouth, and Throat. Cell Host Microbe. 2017 Apr 12;21(4):421–32.

48. Shilts MH, Rosas-Salazar C, Tovchigrechko A, Larkin EK, Torralba M, Akopov A, et al. Minimally Invasive Sampling Method Identifies Differences in Taxonomic Richness of Nasal Microbiomes in Young Infants Associated with Mode of Delivery. Microb Ecol. 2016 Jan;71(1):233–42.

49. Li L, Mendis N, Trigui H, Oliver JD, Faucher SP. The importance of the viable but non-culturable state in human bacterial pathogens. Front Microbiol. 2014 Jun 2;5:258.

50. Vatanen T, Franzosa EA, Schwager R, Tripathi S, Arthur TD, Vehik K, et al. The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. Nature. 2018 Oct;562(7728):589–94.

51. Milani C, Duranti S, Bottacini F, Casey E, Turroni F, Mahony J, et al. The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota. Microbiol Mol Biol Rev [Internet]. 2017 Dec;81(4). Available from: http://dx.doi.org/10.1128/MMBR.00036-17

52. Ferretti P, Pasolli E, Tett A, Asnicar F, Gorfer V, Fedi S, et al. Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. Cell Host Microbe. 2018 Jul 11;24(1):133–45.e5.

53. Vatanen T, Plichta DR, Somani J, Münch PC, Arthur TD, Hall AB, et al. Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. Nat Microbiol. 2019 Mar;4(3):470–9.

54. Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N, et al. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc Natl Acad Sci U S A. 2010 Jun 29;107(26):11971–5.

55. Teo SM, Mok D, Pham K, Kusel M, Serralha M, Troy N, et al. The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. Cell Host Microbe. 2015 May 13;17(5):704–15.

56. Sá-Leão R, Nunes S, Brito-Avô A, Alves CR, Carriço JA, Saldanha J, et al. High rates of transmission of and colonization by Streptococcus pneumoniae and Haemophilus influenzae within a day care center revealed in a longitudinal study. J Clin Microbiol. 2008 Jan;46(1):225–34.

57. Mansbach JM, Hasegawa K, Henke DM, Ajami NJ, Petrosino JF, Shaw CA, et al. Respiratory syncytial virus and rhinovirus severe bronchiolitis are associated with distinct nasopharyngeal microbiota. J Allergy Clin Immunol. 2016 Jun;137(6):1909–13.e4.

58. Bisgaard H, Hermansen MN, Buchvald F, Loland L, Halkjaer LB, Bønnelykke K, et al. Childhood asthma after bacterial colonization of the airway in neonates. N Engl J Med. 2007

Oct 11;357(15):1487–95.

59. von Linstow M-L, Schønning K, Hoegh AM, Sevelsted A, Vissing NH, Bisgaard H. Neonatal airway colonization is associated with troublesome lung symptoms in infants. Am J Respir Crit Care Med. 2013 Oct 15;188(8):1041–2.

60. Ederveen THA, Ferwerda G, Ahout IM, Vissers M, de Groot R, Boekhorst J, et al. Haemophilus is overrepresented in the nasopharynx of infants hospitalized with RSV infection and associated with increased viral load and enhanced mucosal CXCL8 responses. Microbiome. 2018 Jan 11;6(1):10.

61. Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE, Lemon KP. New Insights into Human Nostril Microbiome from the Expanded Human Oral Microbiome Database (eHOMD): a Resource for the Microbiome of the Human Aerodigestive Tract. mSystems [Internet]. 2018 Nov;3(6). Available from: http://dx.doi.org/10.1128/mSystems.00187-18

62. Brugger SD, Eslami SM, Pettigrew MM, Escapa IF, Henke MM, Kong Y, et al. Dolosigranulum pigrum cooperation and competition in human nasal microbiota [Internet]. bioRxiv. 2019 [cited 2019 Aug 29]. p. 678698. Available from: https://www.biorxiv.org/content/10.1101/678698v1

63. Khamash DF, Mongodin EF, White JR, Voskertchian A, Hittle L, Colantuoni E, et al. The Association between the Developing Nasal Microbiota of Hospitalized Neonates and Staphylococcus aureus Colonization. Open Forum Infect Dis [Internet]. 2019 Feb 11 [cited 2019 Feb 11]; Available from: https://academic.oup.com/ofid/advance-article/doi/10.1093/ofid/ofz062/5315771?searchresult=1

64. Biesbroek G, Bosch AATM, Wang X, Keijser BJF, Veenhoven RH, Sanders EAM, et al. The impact of breastfeeding on nasopharyngeal microbial communities in infants. Am J Respir Crit Care Med. 2014 Aug 1;190(3):298–308.

65. Peacock SJ, Justice A, Griffiths D, de Silva GDI, Kantzanou MN, Crook D, et al. Determinants of acquisition and carriage of Staphylococcus aureus in infancy. J Clin Microbiol. 2003 Dec;41(12):5718–25.

66. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature. 2017 Oct 5;550(7674):61–6.

67. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol. 2000 Feb;7(1-2):203–14.

68. Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods. 2016 May;13(5):435–8.

69. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods. 2018 Nov;15(11):962–8.

70. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. 2017 Apr;27(4):626–38.

71. Human Microbiome Project Consortium. A framework for human microbiome research. Nature. 2012 Jun 13;486(7402):215–21.

72. Kumpitsch C, Koskinen K, Schöpf V, Moissl-Eichinger C. The microbiome of the upper respiratory tract in health and disease. BMC Biol. 2019 Nov 7;17(1):87.

73. Meisel JS, Hannigan GD, Tyldsley AS, SanMiguel AJ, Hodkinson BP, Zheng Q, et al. Skin Microbiome Surveys Are Strongly Influenced by Experimental Design. J Invest Dermatol. 2016 May;136(5):947–56.

74. Lo C-W, Lai Y-K, Liu Y-T, Gallo RL, Huang C-M. Staphylococcus aureus hijacks a skin commensal to intensify its virulence: immunization targeting β-hemolysin and CAMP factor. J Invest Dermatol. 2011 Feb;131(2):401–9.

75. Wollenberg MS, Claesen J, Escapa IF, Aldridge KL, Fischbach MA, Lemon KP. Propionibacterium-produced coproporphyrin III induces Staphylococcus aureus aggregation and biofilm formation. MBio. 2014 Jul 22;5(4):e01286–14.

76. Pan H, Cui B, Huang Y, Yang J, Ba-Thein W. Nasal carriage of common bacterial pathogens among healthy kindergarten children in Chaoshan region, southern China: a cross-sectional study. BMC Pediatr. 2016 Sep 30;16(1):161.

77. Dunne EM, Murad C, Sudigdoadi S, Fadlyana E, Tarigan R, Indriyani SAK, et al. Carriage of Streptococcus pneumoniae, Haemophilus influenzae, Moraxella catarrhalis, and Staphylococcus aureus in Indonesian children: A cross-sectional study. PLoS One. 2018 Apr 12;13(4):e0195098.

78. Bae S, Yu J-Y, Lee K, Lee S, Park B, Kang Y. Nasal colonization by four potential respiratory bacteria in healthy children attending kindergarten or elementary school in Seoul, Korea. J Med Microbiol. 2012 May;61(Pt 5):678–85.

79. Verhaegh SJC, Lebon A, Saarloos JA, Verbrugh HA, Jaddoe VWV, Hofman A, et al. Determinants of Moraxella catarrhalis colonization in healthy Dutch children during the first 14 months of life. Clin Microbiol Infect. 2010 Jul;16(7):992–7.

80. Song SJ, Lauber C, Costello EK, Lozupone CA, Humphrey G, Berg-Lyons D, et al. Cohabiting family members share microbiota with one another and with their dogs. Elife. 2013 Apr 16;2:e00458.

81. Brito IL, Gurry T, Zhao S, Huang K, Young SK, Shea TP, et al. Transmission of human-associated microbiota along family and social networks. Nat Microbiol. 2019 Jun;4(6):964–71.

82. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, et al. Human genetics shape the gut microbiome. Cell. 2014 Nov 6;159(4):789–99.

83. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature. 2012 May 9;486(7402):222–7.

84. Browne HP, Neville BA, Forster SC, Lawley TD. Transmission of the gut microbiota: spreading of health. Nat Rev Microbiol. 2017 Sep;15(9):531–43.

85. Yassour M, Jason E, Hogstrom LJ, Arthur TD, Tripathi S, Siljander H, et al. Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. Cell Host Microbe. 2018 Jul 11;24(1):146–54.e4.

86. Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, et al. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe.

2015 Jun 10;17(6):852.

87. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019 May;569(7758):655–62.

88. McIver LJ, Abu-Ali G, Franzosa EA, Schwager R, Morgan XC, Waldron L, et al. bioBakery: a meta'omic analysis environment. Bioinformatics. 2018 Apr 1;34(7):1235–7.

89. Accorsi EK, Franzosa EA, Hsu T, Cordy RJ, Maayan-Metzger A, Jaber H, et al. Determinants of S. aureus carriage in the developing infant nasal microbiome. BioProject PRJNA610982. NCBI Sequence Read Archive [Internet]. Available from: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA610982/ (2020).

90. Forslund K, Hildebrand F, Nielsen T, Falony G, Le Chatelier E, Sunagawa S, et al. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. Nature. 2015 Dec 10;528(7581):262–6.

91. Latorre M, Krishnareddy S, Freedberg DE. Microbiome as mediator: Do systemic infections start in the gut? World J Gastroenterol. 2015 Oct 7;21(37):10487–92.

92. Mehta RS, Nishihara R, Cao Y, Song M, Mima K, Qian ZR, et al. Association of Dietary Patterns With Risk of Colorectal Cancer Subtypes Classified by Fusobacterium nucleatum in Tumor Tissue. JAMA Oncol. 2017 Jul 1;3(7):921–7.

93. Penders J, Gerhold K, Thijs C, Zimmermann K, Wahn U, Lau S, et al. New insights into the hygiene hypothesis in allergic diseases: mediation of sibling and birth mode effects by the gut microbiota. Gut Microbes. 2014 Mar;5(2):239–44.

94. Tun HM, Bridgman SL, Chari R, Field CJ, Guttman DS, Becker AB, et al. Roles of Birth Mode and Infant Gut Microbiota in Intergenerational Transmission of Overweight and Obesity From Mother to Offspring. JAMA Pediatr. 2018 Apr 1;172(4):368–77.

95. Yang H, Duan Z. The Local Defender and Functional Mediator: Gut Microbiome. Digestion. 2018 Jan 8;97(2):137–45.

96. Zhao L, Zhang F, Ding X, Wu G, Lam YY, Wang X, et al. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. Science. 2018 Mar 9;359(6380):1151–6.

97. Taur Y, Pamer EG. Microbiome mediation of infections in the cancer setting. Genome Med. 2016 Apr 18;8(1):40.

98. Cammarota G, Ianiro G, Bibbò S, Gasbarrini A. Gut microbiota modulation: probiotics, antibiotics or fecal microbiota transplantation? Intern Emerg Med. 2014 Jun;9(4):365–73.

99. Gallo A, Passaro G, Gasbarrini A, Landolfi R, Montalto M. Modulation of microbiota as treatment for intestinal inflammatory disorders: An uptodate. World J Gastroenterol. 2016 Aug 28;22(32):7186–202.

100. Singh RK, Chang H-W, Yan D, Lee KM, Ucmak D, Wong K, et al. Influence of diet on the gut microbiome and implications for human health. J Transl Med. 2017 Apr 8;15(1):73.

101. Attaye I, Pinto-Sietsma S-J, Herrema H, Nieuwdorp M. A Crucial Role for Diet in the Relationship Between Gut Microbiota and Cardiometabolic Disease. Annu Rev Med. 2020 Jan 27;71:149–61.

102. Tindall AM, Petersen KS, Kris-Etherton PM. Dietary Patterns Affect the Gut Microbiome-The Link to Risk of Cardiometabolic Diseases. J Nutr. 2018 Sep 1;148(9):1402–7.

103. Bailey MA, Holscher HD. Microbiome-Mediated Effects of the Mediterranean Diet on Inflammation. Adv Nutr. 2018 May 1;9(3):193–206.

104. Ducarmon QR, Zwittink RD, Hornung BVH, van Schaik W, Young VB, Kuijper EJ. Gut Microbiota and Colonization Resistance against Bacterial Enteric Infection. Microbiol Mol Biol Rev [Internet]. 2019 Aug 21;83(3). Available from: http://dx.doi.org/10.1128/MMBR.00007-19

105. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional [Internet]. Vol. 8, Frontiers in Microbiology. 2017. Available from: http://dx.doi.org/10.3389/fmicb.2017.02224

106. Quinn TP, Erb I, Richardson MF, Crowley TM. Understanding sequencing data as compositions: an outlook and review. Bioinformatics. 2018 Aug 15;34(16):2870–8.

107. Silverman JD, Roche K, Mukherjee S, David LA. Naught all zeros in sequence count data are the same. Comput Struct Biotechnol J. 2020 Sep 28;18:2789–98.

108. Chén OY, Crainiceanu C, Ogburn EL, Caffo BS, Wager TD, Lindquist MA. High-dimensional multivariate mediation with application to neuroimaging data. Biostatistics. 2018 Apr 1;19(2):121–36.

109. Sohn MB, Li H. Compositional mediation analysis for microbiome studies. Ann Appl Stat. 2019 Mar;13(1):661–81.

110. Wang C, Hu J, Blaser MJ, Li H. Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. Bioinformatics [Internet]. 2019 Jul 22; Available from: http://dx.doi.org/10.1093/bioinformatics/btz565

111. Zhang J, Wei Z, Chen J. A distance-based approach for testing the mediation effect of the human microbiome. Bioinformatics. 2018 Jun 1;34(11):1875–83.

112. Ren B, Schwager E, Tickle TL, Huttenhower C. SparseDOSSA: Sparse data observations for simulating synthetic abundance [Internet]. Available from: https://github.com/biobakery/sparseDOSSA

113. Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser HJ, Reinker S, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat Microbiol. 2019 Feb;4(2):293–305.

114. Gao Y, Yang H, Fang R, Zhang Y, Goode EL, Cui Y. Testing Mediation Effects in High-Dimensional Epigenetic Studies. Front Genet. 2019 Nov 22;10:1195.

115. Zhang H, Zheng Y, Zhang Z, Gao T, Joyce B, Yoon G, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. Bioinformatics. 2016 Oct 15;32(20):3150–4.

116. VanderWeele TJ. Mediation Analysis: A Practitioner's Guide. Annu Rev Public Health. 2016;37:17–32.

117. Wang DD, Nguyen LH, Li Y, Yan Y, Ma W, Rinott E, et al. The gut microbiome modulates the protective association between a Mediterranean diet and cardiometabolic disease risk. Nat Med [Internet]. 2021 Feb 11; Available from: https://doi.org/10.1038/s41591-020-01223-3

118. Gaike AH, Paul D, Bhute S, Dhotre DP, Pande P, Upadhyaya S, et al. The Gut Microbial Diversity of Newly Diagnosed Diabetics but Not of Prediabetics Is Significantly Different from That of Healthy Nondiabetics. mSystems [Internet]. 2020 Mar 31;5(2). Available from:

http://dx.doi.org/10.1128/mSystems.00578-19

119. Zheng J, Hoffman KL, Chen J-S, Shivappa N, Sood A, Browman GJ, et al. Dietary inflammatory potential in relation to the gut microbiome: results from a cross-sectional study. Br J Nutr. 2020 Nov 14;124(9):931–42.

120. Phipson B, Smyth GK. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. Stat Appl Genet Mol Biol. 2010 Oct 31;9:Article39.

121. Teixeira da Silva JA, Tsigaris P, Erfanmanesh M. Publishing volumes in major databases related to Covid-19. Scientometrics. 2020 Aug 28;1–12.

122. Else H. How a torrent of COVID science changed research publishing - in seven charts. Nature. 2020 Dec;588(7839):553.

123. Takahashi S, Greenhouse B, Rodríguez-Barraquer I. Are SARS-CoV-2 seroprevalence estimates biased? J Infect Dis [Internet]. 2020 Aug 28; Available from: http://dx.doi.org/10.1093/infdis/jiaa523

124. Addetia A, Crawford KH, Dingens A, Zhu H, Roychoudhury P, Huang M-L, et al. Neutralizing antibodies correlate with protection from SARS-CoV-2 in humans during a fishery vessel outbreak with high attack rate. medRxiv [Internet]. 2020 Aug 14; Available from: http://dx.doi.org/10.1101/2020.08.13.20173161

125. Kahn R, Kennedy-Shaffer L, Grad YH, Robins JM, Lipsitch M. Potential Biases Arising from Epidemic Dynamics in Observational Seroprotection Studies. Am J Epidemiol [Internet]. 2020 Sep 1; Available from: http://dx.doi.org/10.1093/aje/kwaa188

126. Griffith G, Morris TT, Tudball M, Herbert A, Mancano G, Pike L, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity [Internet]. Epidemiology. medRxiv; 2020. Available from: https://www.medrxiv.org/content/10.1101/2020.05.04.20090506v2.abstract

127. Goldstein E, Lipsitch M, Cevik M. On the effect of age on the transmission of SARS-CoV-2 in households, schools and the community. J Infect Dis [Internet]. 2020 Oct 29; Available from: http://dx.doi.org/10.1093/infdis/jiaa691

128. Poland GA, Ovsyannikova IG, Kennedy RB. SARS-CoV-2 immunity: review and applications to phase 3 vaccine candidates. Lancet. 2020 Nov 14;396(10262):1595–606.

129. Sethuraman N, Jeremiah SS, Ryo A. Interpreting Diagnostic Tests for SARS-CoV-2. JAMA [Internet]. 2020 May 6; Available from: http://dx.doi.org/10.1001/jama.2020.8259

130. Herzog S, De Bie J, Abrams S, Wouters I, Ekinci E, Patteet L, et al. Seroprevalence of IgG antibodies against SARS coronavirus 2 in Belgium: a prospective cross-sectional nationwide study of residual samples [Internet]. Epidemiology. medRxiv; 2020. Available from: https://www.medrxiv.org/content/10.1101/2020.06.08.20125179v2.abstract

131. Pollán M, Pérez-Gómez B, Pastor-Barriuso R, Oteo J, Hernán MA, Pérez-Olmeda M, et al. Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. Lancet [Internet]. 2020 Jul 6; Available from: https://doi.org/10.1016/S0140-6736(20)31483-5

132. Rosenberg ES, Tesoriero JM, Rosenthal EM, Chung R, Barranco MA, Styer LM, et al. Cumulative incidence and diagnosis of SARS-CoV-2 infection in New York. Ann Epidemiol.

2020 Aug;48:23–9.e4.

133. Hernán MA, Robins JM. Causal Inference: What If. CRC Boca Raton, FL; 2020.

134. Porter E. Don't Think You Need a Coronavirus Test? What if I Paid You? The New York Times [Internet]. 2020 Apr 21 [cited 2020 Sep 11]; Available from: https://www.nytimes.com/2020/04/21/business/economy/coronavirus-tests-infections.html

135. Bendavid E, Mulaney B, Sood N, Shah S, Ling E, Bromley-Dulfano R, et al. COVID-19 Antibody Seroprevalence in Santa Clara County, California [Internet]. Epidemiology. medRxiv; 2020. Available from: http://dx.doi.org/10.1101/2020.04.14.20062463

136. Lynch KL, Whitman JD, Lacanienta NP, Beckerdite EW, Kastner SA, Shy BR, et al. Magnitude and kinetics of anti-SARS-CoV-2 antibody responses and their relationship to disease severity [Internet]. Infectious Diseases (except HIV/AIDS). medRxiv; 2020. Available from: https://www.medrxiv.org/content/10.1101/2020.06.03.20121525v1.abstract

137. Lumley SF, Wei J, O'Donnell D, Stoesser NE, Matthews PC, Howarth A, et al. The duration, dynamics and determinants of SARS-CoV-2 antibody responses in individual healthcare workers [Internet]. Infectious Diseases (except HIV/AIDS). medRxiv; 2020. p. CD013652. Available from: http://dx.doi.org/10.1101/2020.11.02.20224824

138. Liu Y, Mao B, Liang S, Yang J-W, Lu H-W, Chai Y-H, et al. Association between age and clinical characteristics and outcomes of COVID-19. Eur Respir J [Internet]. 2020 May;55(5). Available from: http://dx.doi.org/10.1183/13993003.01112-2020

139. Iyer AS, Jones FK, Nodoushani A, Kelly M, Becker M, Slater D, et al. Dynamics and significance of the antibody response to SARS-CoV-2 infection. medRxiv [Internet]. 2020 Jul

20; Available from: http://dx.doi.org/10.1101/2020.07.18.20155374

140. Hay JA, Kennedy-Shaffer L, Kanjilal S, Lipsitch M, Mina MJ. Estimating epidemiologic dynamics from single cross-sectional viral load distributions [Internet]. bioRxiv. medRxiv; 2020. Available from: http://medrxiv.org/lookup/doi/10.1101/2020.10.08.20204222

141. Edridge AWD, Kaczorowska J, Hoste ACR, Bakker M, Klein M, Loens K, et al. Seasonal coronavirus protective immunity is short-lasting. Nat Med [Internet]. 2020 Sep 14; Available from: http://dx.doi.org/10.1038/s41591-020-1083-1

142. Iyer AS, Jones FK, Nodoushani A, Kelly M, Becker M, Slater D, et al. Persistence and decay of human antibody responses to the receptor binding domain of SARS-CoV-2 spike protein in COVID-19 patients. Sci Immunol [Internet]. 2020 Oct 8;5(52). Available from: http://dx.doi.org/10.1126/sciimmunol.abe0367

143. Emmenegger M, De Cecco E, Lamparter D, Jacquat RPB, Ebner D, Schneider MM, et al. Early peak and rapid decline of SARS-CoV-2 seroprevalence in a Swiss metropolitan region [Internet]. Epidemiology. medRxiv; 2020. Available from: https://www.medrxiv.org/content/10.1101/2020.05.31.20118554v4.abstract

144. Dan JM, Mateus J, Kato Y, Hastie KM, Yu ED, Faliti CE, et al. Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. Science [Internet]. 2021 Jan 6; Available from: http://dx.doi.org/10.1126/science.abf4063

145. Rodda LB, Netland J, Shehata L, Pruner KB, Morawski PA, Thouvenel CD, et al. Functional SARS-CoV-2-Specific Immune Memory Persists after Mild COVID-19. Cell. 2021 Jan 7;184(1):169–83.e17.

146. Zuo J, Dowell A, Pearce H, Verma K, Long HM, Begum J, et al. Robust SARS-CoV-2-specific T-cell immunity is maintained at 6 months following primary infection [Internet]. Cold Spring Harbor Laboratory. 2020 [cited 2021 Jan 21]. p. 2020.11.01.362319. Available from: https://www.biorxiv.org/content/10.1101/2020.11.01.362319v1?ijkey=032ae87a45b76ab5cdefeae919d1aec84c893222&keytype2=tf_ipsecsha

147. Gaebler C, Wang Z, Lorenzi JCC, Muecksch F, Finkin S, Tokuyama M, et al. Evolution of antibody immunity to SARS-CoV-2. Nature [Internet]. 2021 Jan 18; Available from: http://dx.doi.org/10.1038/s41586-021-03207-w

148. Larremore DB, Fosdick BK, Zhang S, Grad YH. Jointly modeling prevalence, sensitivity and specificity for optimal sample allocation [Internet]. bioRxiv. 2020 [cited 2020 Oct 20]. p. 2020.05.23.112649. Available from: https://www.biorxiv.org/content/10.1101/2020.05.23.112649v1

149. Callow KA, Parry HF, Sergeant M, Tyrrell DA. The time course of the immune response to experimental coronavirus infection of man. Epidemiol Infect. 1990 Oct;105(2):435.

150. Reed SE. The behaviour of recent isolates of human respiratory coronavirus in vitro and in volunteers: evidence of heterogeneity among 229E-related strains. J Med Virol [Internet]. 1984 [cited 2020 Aug 28];13(2). Available from: https://pubmed.ncbi.nlm.nih.gov/6319590/

151. Bao L, Deng W, Gao H, Xiao C, Liu J, Xue J, et al. Lack of Reinfection in Rhesus Macaques Infected with SARS-CoV-2. bioRxiv.(2020) [Internet]. bioRxiv. 2020. p. 13–990226. Available from: https://www.biorxiv.org/content/10.1101/2020.03.13.990226v2

152. Cauchemez S, Horby P, Fox A, Le Quynh M, Thanh LT, Thai PQ, et al. Influenza Infection Rates, Measurement Errors and the Interpretation of Paired Serology. PLoS Pathog. 2012 Dec 13;8(12):e1003061.

153. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. American Journal of Epidemiology. 1977;105(5):488–95.

154. Lipsitch M, Tchetgen ET, Cohen T. Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. Epidemiology. 2010 May;21(3):383.

155. Stensrud MJ, Robins JM, Sarvet A, Tchetgen Tchetgen EJ, Young JG. Conditional separable effects [Internet]. arXiv. 2020. Available from: http://arxiv.org/abs/2006.15681

156. Robins JM, Richardson TS, Shpitser I. An Interventionist Approach to Mediation Analysis [Internet]. arXiv. 2020. Available from: http://arxiv.org/abs/2008.06019

157. Robins, James M., and Thomas S. Richardson. Alternative graphical causal models and the identification of direct effects. In: Patrick Shrout Katherine Keyes, editor. Causality and psychopathology: Finding the determinants of disorders and their cures. Oxford University Press; 2010. p. 103–58.

158. Kim SJ, Bostwick W. Social Vulnerability and Racial Inequality in COVID-19 Deaths in Chicago. Health Educ Behav. 2020 Aug 1;47(4):509–13.

159. Shadmi E, Chen Y, Dourado I, Faran-Perach I, Furler J, Hangoma P, et al. Health equity and COVID-19: global perspectives. Int J Equity Health. 2020 Jun 26;19(1):104.

160. Patel JA, Nielsen FBH, Badiani AA, Assi S, Unadkat VA, Patel B, et al. Poverty, inequality and COVID-19: the forgotten vulnerable. Public Health. 2020 Jun 1;183:110–1.

161. Hooper MW, Nápoles AM, Pérez-Stable EJ. COVID-19 and Racial/Ethnic Disparities. JAMA. 2020 Jun 23;323(24):2466–7.

162. Cevik M, Marcus JL, Buckee C, Smith TC. SARS-CoV-2 transmission dynamics should inform policy. Clin Infect Dis [Internet]. 2020 Sep 23; Available from: http://dx.doi.org/10.1093/cid/ciaa1442

163. Huitfeldt A. Is caviar a risk factor for being a millionaire? BMJ [Internet]. 2016 Dec 9;355(6536). Available from: https://www.bmj.com/bmj/section-pdf/935674?path=/bmj/355/8086/Food_for_thought.full.pdf

164. Number of coronavirus (COVID-19) cases in England as of July 30, 2020, by age and gender [Internet]. Statista. [cited 2020 Aug 25]. Available from: https://www.statista.com/statistics/1115083/coronavirus-cases-in-england-by-age-and-gender/

165. de Lusignan S, Dorward J, Correa A, Jones N, Akinyemi O, Amirthalingam G, et al. Risk factors for SARS-CoV-2 among patients in the Oxford Royal College of General Practitioners Research and Surveillance Centre primary care network: a cross-sectional study. Lancet Infect Dis [Internet]. 2020 May 15; Available from: http://dx.doi.org/10.1016/S1473-3099(20)30371-6

166. VanderWeele TJ, Hernán MA. Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. Am J Epidemiol. 2012 Jun 15;175(12):1303–10.

167. Hernán MA, Cole SR. Invited Commentary: Causal diagrams and measurement bias. Am

J Epidemiol. 2009 Oct 15;170(8):959–62; discussion 963–4.

168. Sutton D, Fuchs K, D'Alton M, Goffman D. Universal Screening for SARS-CoV-2 in Women Admitted for Delivery. N Engl J Med [Internet]. 2020 Apr 13; Available from: http://dx.doi.org/10.1056/NEJMc2009316

169. Ludvigsson JF. Systematic review of COVID-19 in children shows milder cases and a better prognosis than adults. Acta Paediatr. 2020 Jun;109(6):1088–95.

170. Cruz AT, Zeichner SL. COVID-19 in Children: Initial Characterization of the Pediatric Disease. Pediatrics [Internet]. 2020 Jun;145(6). Available from: http://dx.doi.org/10.1542/peds.2020-0834

171. Eyre DW, Lumley SF, O'Donnell D, Campbell M, Sims E, Lawson E, et al. Differential occupational risks to healthcare workers from SARS-CoV-2 observed during a prospective observational study. Elife [Internet]. 2020 Aug 21;9. Available from: http://dx.doi.org/10.7554/eLife.60675

172. Krammer F, Simon V. Serology assays to manage COVID-19. Science. 2020 Jun 5;368(6495):1060–1.

173. Volz E, Hill V, McCrone JT, Price A, Jorgensen D, O'Toole Á, et al. Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. Cell. 2021 Jan 7;184(1):64–75.e11.

174. Leung K, Shum MH, Leung GM, Lam TT, Wu JT. Early transmissibility assessment of the N501Y mutant strains of SARS-CoV-2 in the United Kingdom, October to November 2020. Euro Surveill [Internet]. 2021 Jan;26(1). Available from: http://dx.doi.org/10.2807/1560-

7917.ES.2020.26.1.2002106

175. Hope Simpson RE. Infectiousness of communicable diseases in the household (measles, chickenpox, and mumps). Lancet. 1952 Sep 20;260(6734):549–54.

176. Liu Y, Eggo RM, Kucharski AJ. Secondary attack rate and superspreading events for SARS-CoV-2. Lancet. 2020 Mar 14;395(10227):e47.

177. Du Z, Xu X, Wu Y, Wang L, Cowling BJ, Meyers LA. Serial Interval of COVID-19 among Publicly Reported Confirmed Cases. Emerg Infect Dis. 2020 Jun;26(6):1341–3.

178. Park YJ, Choe YJ, Park O, Park SY, Kim Y-M, Kim J, et al. Contact Tracing during Coronavirus Disease Outbreak, South Korea, 2020. Emerg Infect Dis [Internet]. 2020 Jul 16;26(10). Available from: http://dx.doi.org/10.3201/eid2610.201315

179. Zhu Y, Bloxham CJ, Hulme KD, Sinclair JE, Tong ZWM, Steele LE, et al. Children Are Unlikely to Have Been the Primary Source of Household SARS-CoV-2 Infections [Internet]. The Lancet Infectious Diseases. 2020 [cited 2020 Sep 22]. Available from: https://papers.ssrn.com/abstract=3564428

180. Tao J, Zhang X, Zhang X, Zhao S, Yang L, He D, et al. The time serial distribution and influencing factors of asymptomatic COVID-19 cases in Hong Kong. One Health. 2020 Dec;10:100166.

181. Sukbin Jang, Si Hyun Han, Ji-Young Rhee. Cluster of Coronavirus Disease Associated with Fitness Dance Classes, South Korea. Emerging Infectious Disease journal [Internet]. 2020;26(8). Available from: https://wwwnc.cdc.gov/eid/article/26/8/20-0633_article

182. Voeten HA, Sikkema RS, Damen M, Oude Munnink BB, Arends C, Stobberingh E, et al.

Unravelling the modes of transmission of SARS-CoV-2 during a nursing home outbreak: looking beyond the church super-spread event. 2020 Sep 9 [cited 2020 Sep 11]; Available from: https://www.researchsquare.com/article/rs-63027/v1.pdf

183. Kim J, Choe YJ, Lee J, Park YJ, Park O, Han MS, et al. Role of children in household transmission of COVID-19. Arch Dis Child [Internet]. 2020 Aug 7; Available from: http://dx.doi.org/10.1136/archdischild-2020-319910

184. Campbell F, Didelot X, Fitzjohn R, Ferguson N, Cori A, Jombart T. outbreaker2: a modular platform for outbreak reconstruction. BMC Bioinformatics. 2018 Oct 22;19(Suppl 11):363.

185. Worby CJ, Lipsitch M, Hanage WP. Shared Genomic Variants: Identification of Transmission Routes Using Pathogen Deep-Sequence Data. Am J Epidemiol. 2017 Nov 15;186(10):1209–16.

186. Becker N. A general chain binomial model for infectious diseases. Biometrics. 1981 Jun;37(2):251–8.

187. Hong L-X, Lin A, He Z-B, Zhao H-H, Zhang J-G, Zhang C, et al. Mask wearing in pre-symptomatic patients prevents SARS-CoV-2 transmission: An epidemiological analysis. Travel Med Infect Dis. 2020 Jun 24;101803.

188. Zhang JZ, Zhou P, Han DB, Wang WC, Cui C, Zhou R, et al. Investigation on a cluster epidemic of COVID-19 in a supermarket in Liaocheng, Shandong province. Zhonghua Liu Xing Bing Xue Za Zhi. 2020 Apr 27;41(0):E055.

189. Kwok KO, Wong VWY, Wei WI, Wong SYS, Tang JW-T. Epidemiological characteristics of the first 53 laboratory-confirmed cases of COVID-19 epidemic in Hong Kong, 13 February

2020. Euro Surveill [Internet]. 2020 Apr;25(16). Available from: http://dx.doi.org/10.2807/1560-7917.ES.2020.25.16.2000155

190. Burke RM, Midgley CM, Dratch A, Fenstersheib M, Haupt T, Holshue M, et al. Active Monitoring of Persons Exposed to Patients with Confirmed COVID-19 - United States, January-February 2020. MMWR Morb Mortal Wkly Rep. 2020 Mar 6;69(9):245–6.

191. Zhang Z, Bi Q, Fang S, Wei L, Wang X, He J, et al. Insights into the practical effectiveness of RT-PCR testing for SARS-CoV-2 from serologic data, a cohort study [Internet]. Epidemiology. medRxiv; 2020. Available from: https://www.medrxiv.org/content/10.1101/2020.09.01.20182469v2.abstract

192. Waterfield T, Watson C, Moore R, Ferris K, Tonry C, Watt AP, et al. Seroprevalence of SARS-CoV-2 antibodies in children - A prospective multicentre cohort study [Internet]. Pediatrics. medRxiv; 2020. Available from: https://www.medrxiv.org/content/10.1101/2020.08.31.20183095v1.abstract

193. Qiu X, Nergiz AI, Maraolo AE, Bogoch II, Low N, Cevik M. Defining the role of asymptomatic SARS-CoV-2 transmission: a living systematic review [Internet]. Epidemiology. medRxiv; 2020. Available from: https://www.medrxiv.org/content/10.1101/2020.09.01.20135194v1.abstract

194. Li W, Zhang B, Lu J, Liu S, Chang Z, Cao P, et al. The characteristics of household transmission of COVID-19. Clin Infect Dis [Internet]. 2020 Apr 17; Available from: http://dx.doi.org/10.1093/cid/ciaa450

195. Wu J, Huang Y, Tu C, Bi C, Chen Z, Luo L, et al. Household Transmission of SARS-CoV-

2, Zhuhai, China, 2020. Clin Infect Dis [Internet]. 2020 May 11; Available from: http://dx.doi.org/10.1093/cid/ciaa557

196. Hu S, Wang W, Wang Y, Litvinova M, Luo K, Ren L, et al. Infectivity, susceptibility, and risk factors associated with SARS-CoV-2 transmission under intensive contact tracing in Hunan, China [Internet]. Infectious Diseases (except HIV/AIDS). medRxiv; 2020. Available from: http://dx.doi.org/10.1101/2020.07.23.20160317

197. Epstein JB, Chow K, Mathias R. Dental procedure aerosols and COVID-19. Lancet Infect Dis [Internet]. 2020 Aug 10; Available from: http://dx.doi.org/10.1016/S1473-3099(20)30636-8

198. Cevik M, Tate M, Lloyd O, Maraolo AE, Schafers J, Ho A. SARS-CoV-2, SARS-CoV-1 and MERS-CoV viral load dynamics, duration of viral shedding and infectiousness: a living systematic review and meta-analysis [Internet]. Infectious Diseases (except HIV/AIDS). medRxiv; 2020. Available from: https://www.medrxiv.org/content/10.1101/2020.07.25.20162107v2.abstract

199. Xu Y, Li X, Zhu B, Liang H, Fang C, Gong Y, et al. Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding. Nat Med. 2020 Apr;26(4):502–5.

200. Hua C-Z, Miao Z-P, Zheng J-S, Huang Q, Sun Q-F, Lu H-P, et al. Epidemiological features and viral shedding in children with SARS-CoV-2 infection. J Med Virol [Internet]. 2020 Jun 15; Available from: http://dx.doi.org/10.1002/jmv.26180

201. Kissler SM, Fauver JR, Mack C, Tai C, Shiue KY, Kalinich CC, et al. Viral dynamics of

SARS-CoV-2 infection and the predictive value of repeat testing [Internet]. Epidemiology. medRxiv; 2020. Available from: https://www.medrxiv.org/content/10.1101/2020.10.21.20217042v1.abstract

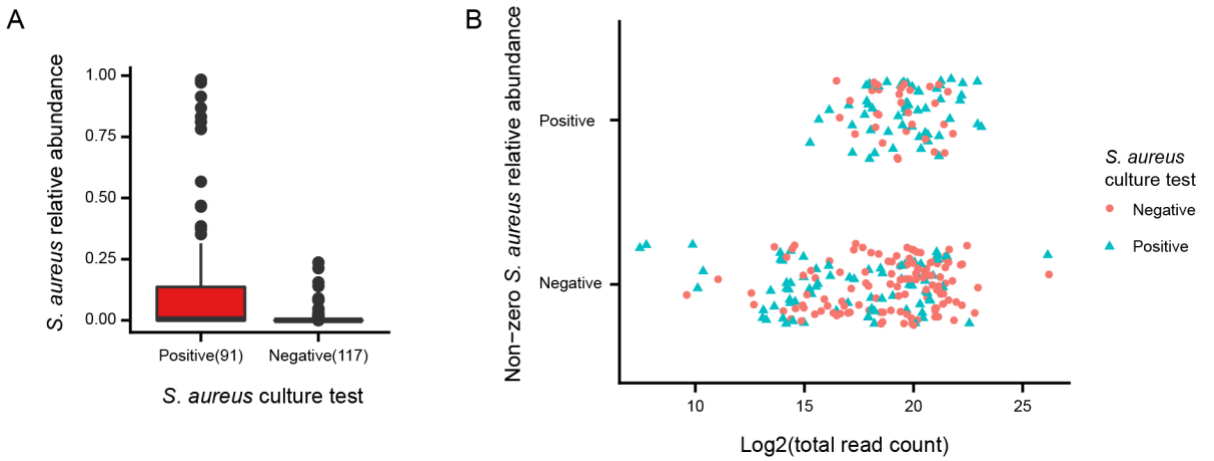
202. Bi Q, Wu Y, Mei S, Ye C, Zou X, Zhang Z, et al. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. Lancet Infect Dis [Internet]. 2020 Apr 27; Available from: http://dx.doi.org/10.1016/S1473-3099(20)30287-5

203. Hu M, Lin H, Wang J, Xu C, Tatem AJ, Meng B, et al. The risk of COVID-19 transmission in train passengers: an epidemiological and modelling study. Clin Infect Dis [Internet]. 2020 Jul 29; Available from: http://dx.doi.org/10.1093/cid/ciaa1057

Appendix

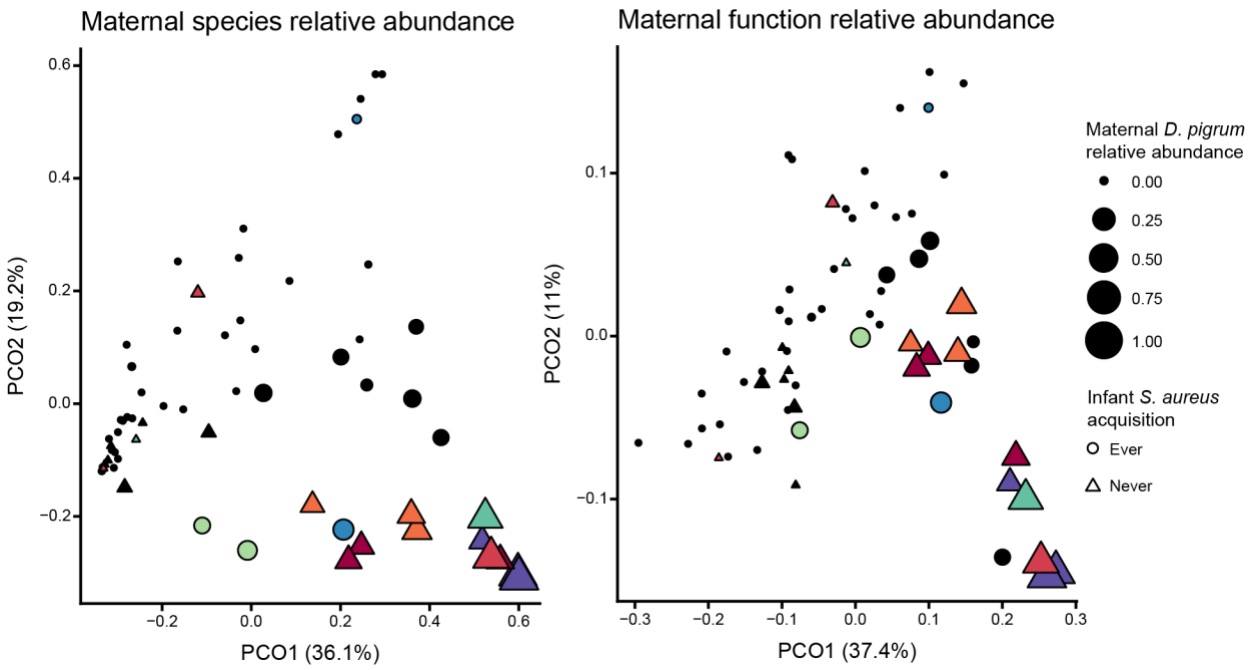
Chapter 2

Figure S2.1: *S. aureus* test concordance.



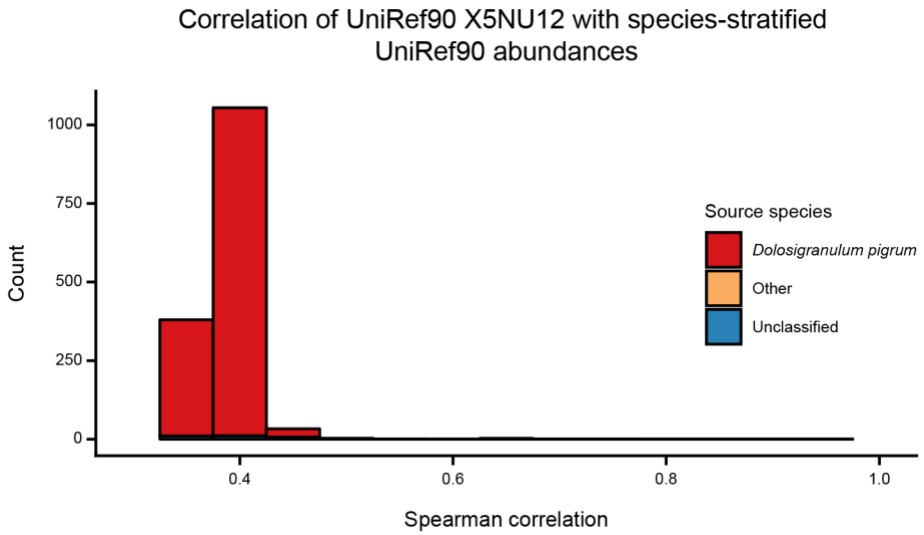
(a) Detection of *S. aureus* by culture and sequencing showed strong, but not complete, concordance, due to high rates of false negatives. **(b)** Sequencing false negatives (i.e., samples positive for *S. aureus* by culture, but with zero *S. aureus* relative abundance by sequencing) tended to have overall lower sample read depth compared to sequencing true negatives.

Figure S2.2: PCoA plots of maternal taxonomic and functional features show associations with infant *S. aureus* "ever" acquisition.



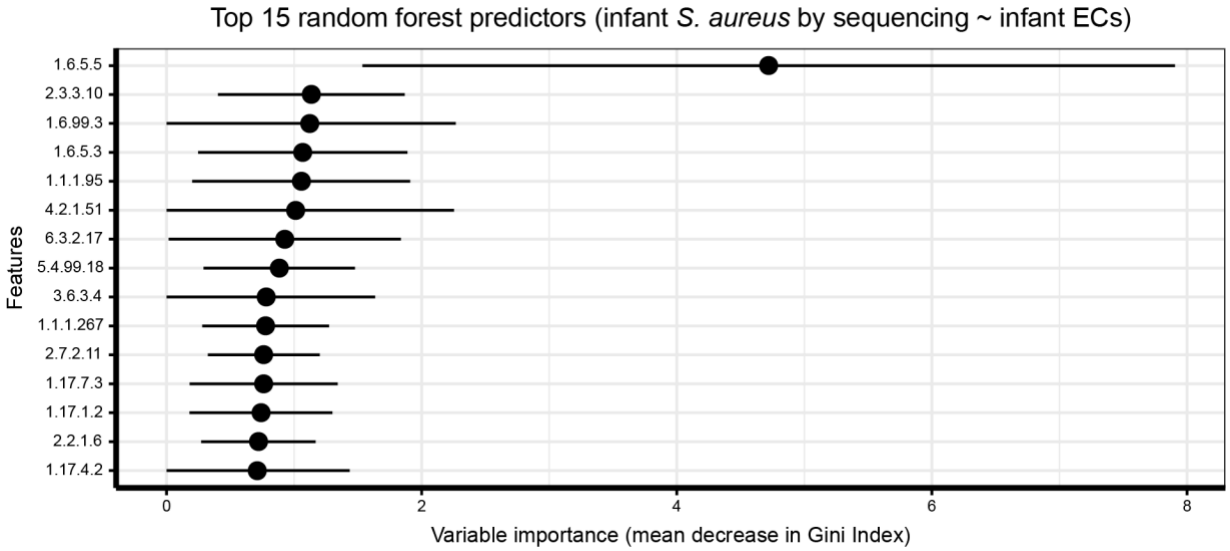
Principal coordinate analysis (PCoA) plots were created using Bray-Curtis dissimilarity. Subjects with at least one sample with more than 25% *D. pigrum* relative abundance are displayed with a unique color. **(a)** As identified in the linear modeling results (**Fig. 2.3**), maternal relative abundance of *D. pigrum* was inversely associated with their infant "ever" acquiring *S. aureus* over the study period. **(b)** This association created a pronounced signal in the functional data (as also seen in **Fig. 2.3**).

Figure S2.3: Spearman correlations of UniRef90 X5NU12 with the species-stratified abundances of other UniRef90s.



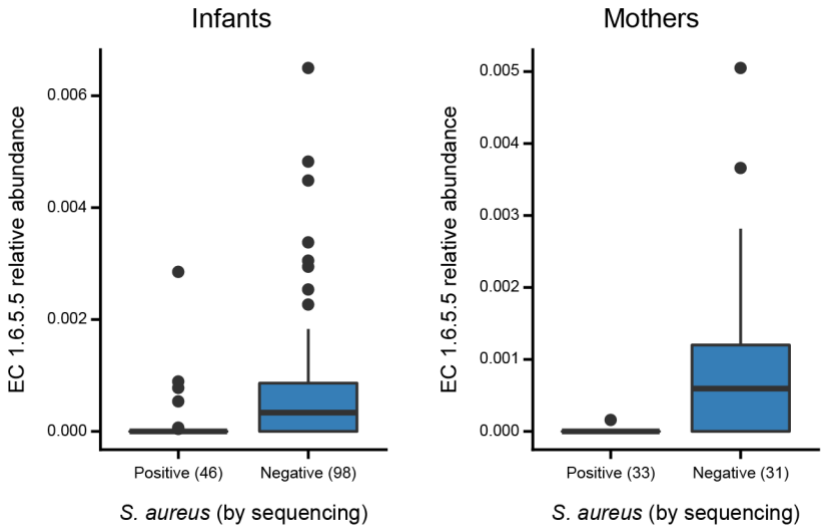
The histogram shows all correlations greater than 0.35, and is colored to show the species of origin. *D. pigrum* contributed the large majority of the abundance of species-stratified UniRef90s moderately positively correlated with UniRef90 X5NU12.

Figure S2.4: Variable importance plot for the prediction of infant *S. aureus* status by sequencing using infant ECs.



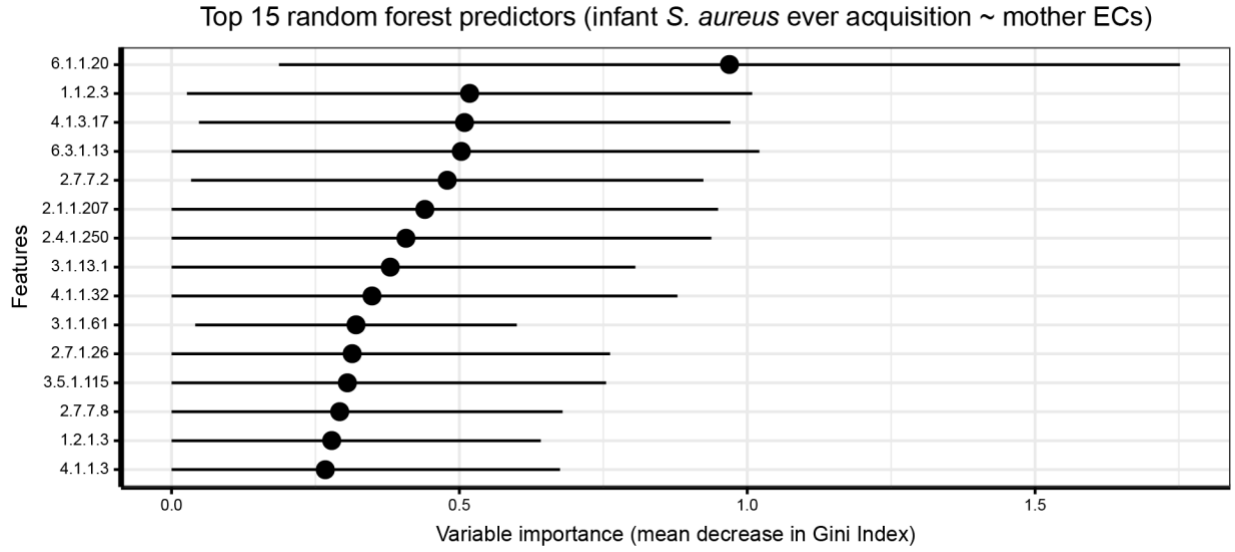
In random forest models, EC 1.6.5.5 (NADPH:quinone reductase) in the infant nasal microbiome was the dominant predictor of infant *S. aureus* status by sequencing.

Figure S2.5: Association of EC 1.6.5.5 (NADPH:quinone reductase) with infant and mother *S. aureus* status by sequencing.



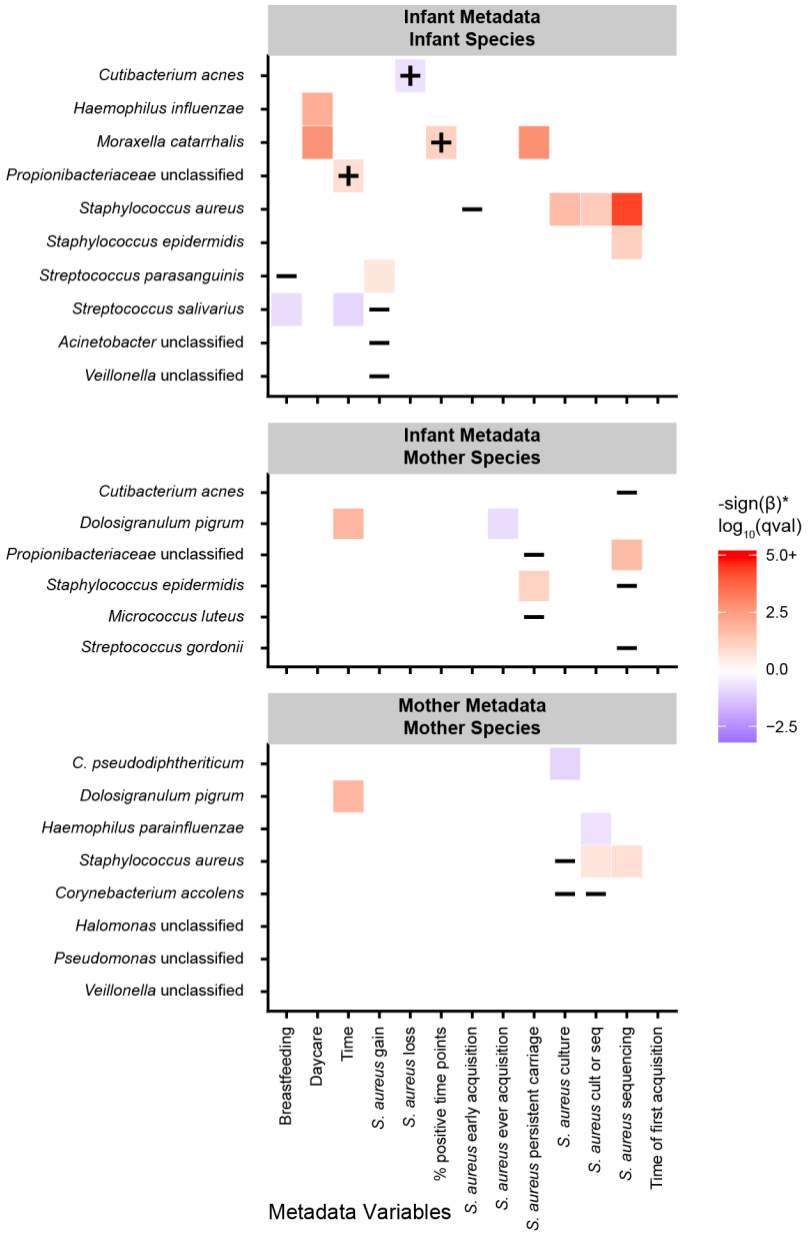
EC 1.6.5.5 was significantly inversely associated with *S. aureus* positivity by sequencing in both infants and mothers.

Figure S2.6: Variable importance plot for the prediction of infant "ever" acquisition of *S. aureus* using maternal ECs.



In random forest models, a number of maternal ECs were important for the prediction of infant "ever" acquisition, but none were as dominant as EC 1.6.5.5 (**Fig. S2.4**).

Figure S2.7: Significant associations between nasal microbiome taxonomic and subject phenotypes in a sensitivity analysis of the unmapped sample mass.



The unmapped sample mass was treated like a taxonomic feature (i.e., taxonomic profiles were rescaled by the percent mapped reads, and the percent unmapped reads was included as a feature), and linear models were rerun. Significant associations ($q<0.25$) between individual taxonomic features and phenotypic covariates using a MaAsLin multivariable linear model for the rescaled profiles are shown. An annotation of "+" indicates that the association was

Figure S2.7 (Continued)

not seen in the original (**Fig. 2.3**) linear models, and an annotation of "-" indicates that the association was only seen in the original (**Fig. 2.3**) linear models. Overall, the percent unmapped sample mass was mostly representative of human contamination. As a result, model power was reduced by inducing a relationship between taxonomic composition and human contamination, and then also subsequently re-adjusting for human contamination as a linear model covariate (**Methods**), causing weaker associations to drop over the threshold of significance.

The following supplementary tables are attached as a Zip file and are also available at: https://doi.org/10.1186/s13059-020-02209-7

Table S2.1. Sample sizes for boxplots.

Table S2.2. Full linear model results for Fig. 2.3.

Table S2.3. Summary of metadata for study population and samples collected for *S. aureus* microbiome profiling.

Table S2.4. Summary of *S. aureus* variables for study population and samples collected for *S. aureus* microbiome profiling.

Table S2.5. Sample sizes for random forest models.

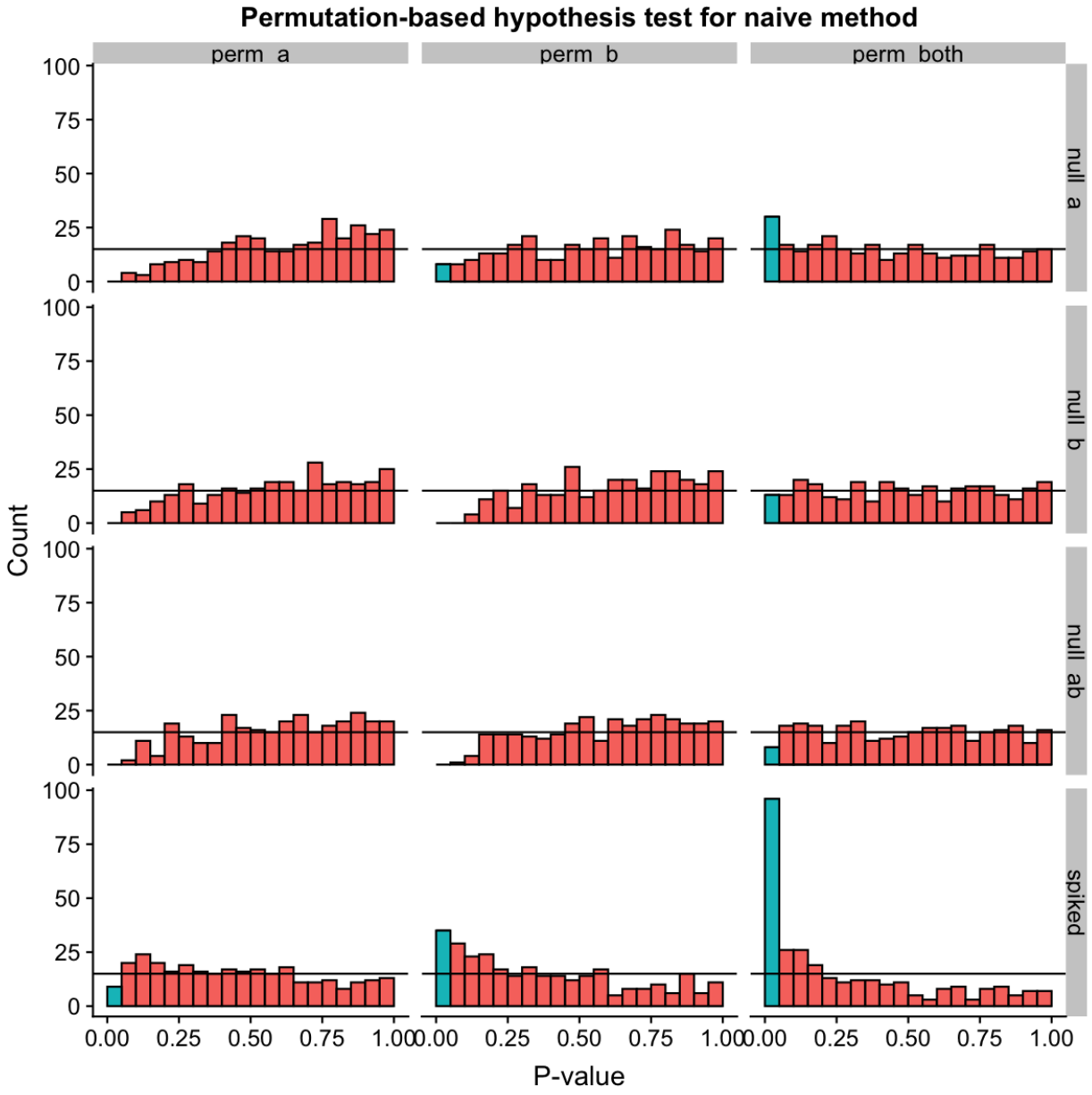
Table S2.6. Subject metadata.

Table S2.7: MetaPhlAn2 taxonomic profiles.

Table S2.8. HUMAnN2 functional profiles.

Chapter 3

Figure S3.1: P-value histograms under different permutation schemes for the naive method hypothesis test.

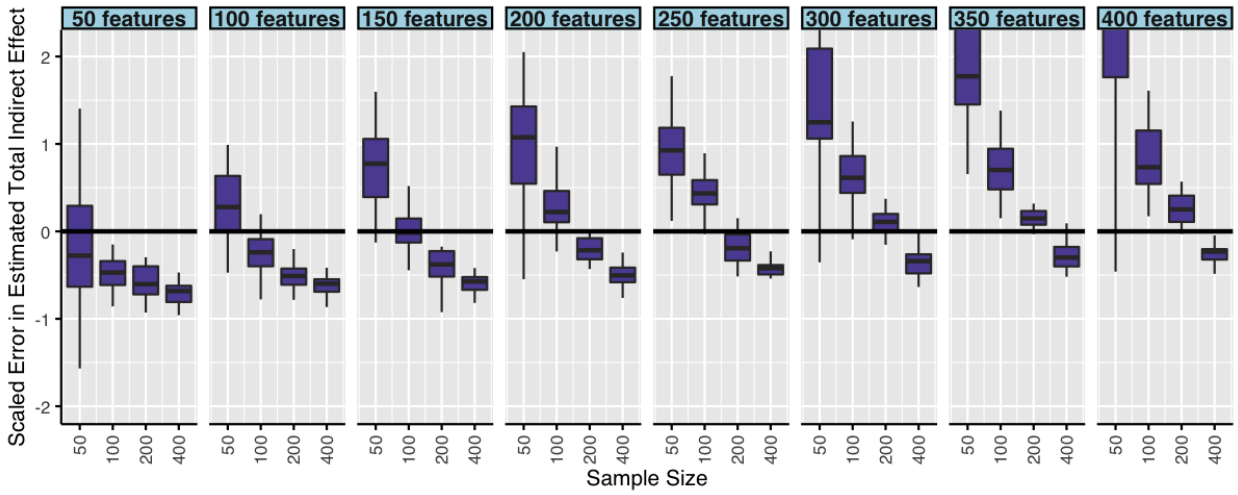


For the different types of simulated datasets (rows, "null_a", "null_b", "null_ab", and "spiked"), we compared three different ways of performing the permutations (columns) for the permutation-based hypothesis test for the naive

Figure S3.1 (Continued)

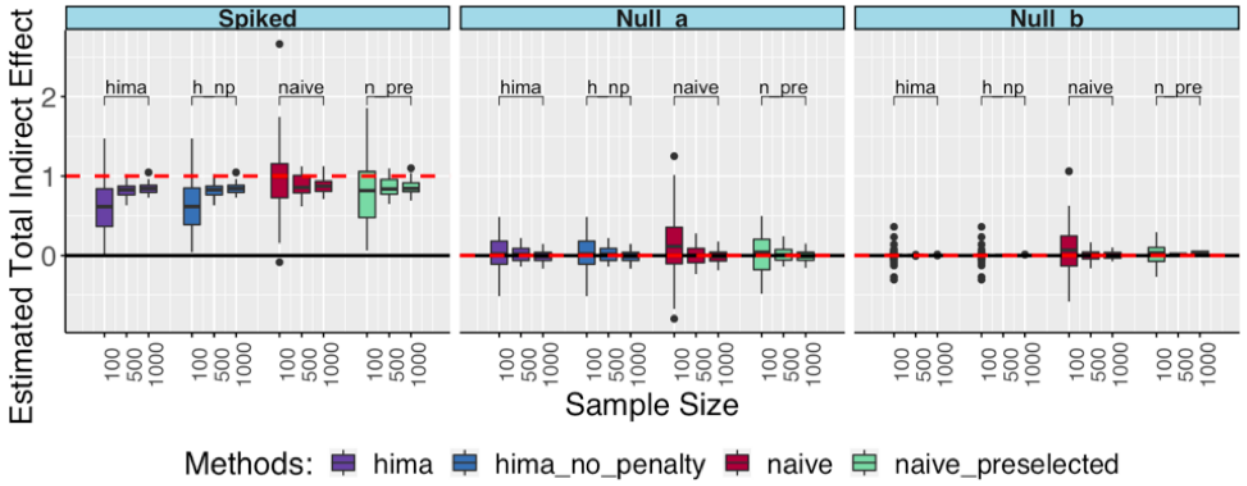
method. The third column ("perm_both") is the permutation scheme used for the naive method in **Fig. 3.2** and **Fig. 3.3**.

Figure S3.2: Estimation of the total indirect effect by CCMM.



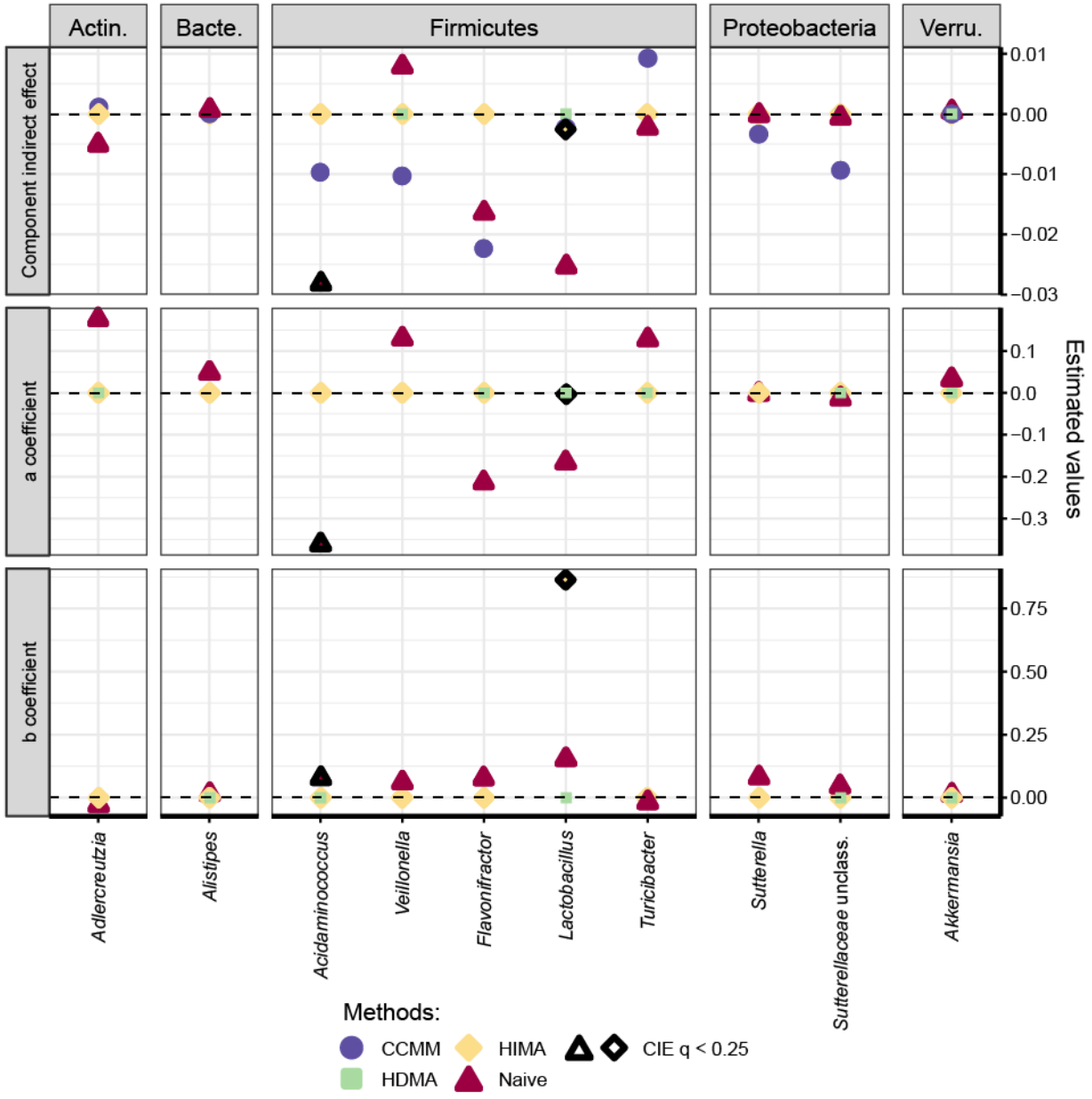
Although CCMM was developed specifically for microbiome mediators and was shown to be a consistently top-performing method for hypothesis testing and effect size estimation (when tested above with 100 microbiome features), it displayed some strange behavior. When the number of microbiome features exceeded the number of samples, CCMM overestimated the total indirect effect. Furthermore, the error in CCMM's estimate of the total indirect scaled with the number of microbiome features. To account for slight variation in the spiked total indirect effect size (see **Methods**), values are presented as the difference between the true and estimated effect size divided by the true effect size.

Figure S3.3: Estimation of the total indirect effect using modifications of HIMA and the naive method.



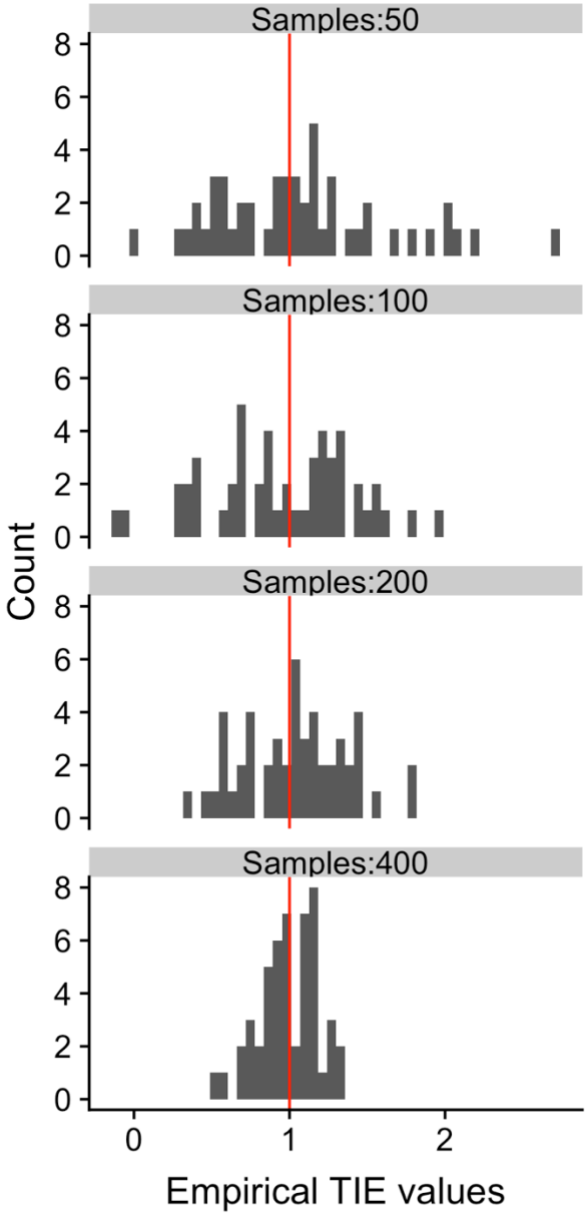
Neither of the modified versions of naive and HIMA that we compared resulted in large differences in effect size estimation for the total indirect effect.

Figure S3.4: Top largest component indirect effects identified in the MLVS dataset.



We identified the top five and bottom five component indirect effects by value for each method. Those that were identified as a top component indirect effect by at least two methods are shown above. The top panel ("est_CIE") displays the component indirect effect for each microbiome, while the bottom two panels ("est_a") and ("est_b") give the pathway coefficients that make up each component indirect effect.

Figure S3.5: Range of spiked total indirect effect sizes across simulated datasets relative to the specified total indirect effect.



Due to the complexity of the simulation, the true spiked effect size in each simulated dataset varied around the specified spiked effect size (red line) with the mean of the true spiked effect sizes falling at the desired effect size. For simulation runs with smaller sample sizes (i.e., top vs. bottom rows), the spread of the true spiked effect sizes around the mean

Figure S3.5 (Continued)

was larger, as expected. To account for this variation in the spiked total indirect effect size, values are presented as the difference between the true and estimated effect size divided by the true effect size (i.e., in **Fig. 3.4**).

Table S3.1: Simulation parameters.

Simulation parameters	Baseline value	Range of values
Iterations per parameter set	300	300
Total microbiome features	100	100
Average read count	50,000	50,000
Direct effect	0	0, 1, 3
Total indirect effect	1	0.5 (small), 1 (medium), 1.5 (large)
Number of samples	100	50, 100, 200, 400
Number of mediators	20 (**Fig. 3.2**, **Fig. 3.3**), 10 (**Fig. 3.4**)	5, 10, 20
Direction of mediation (a's)	pos (+)	pos (+), neg (-)
Direction of mediation (b's)	pos (+)	pos (+), neg (-)
Effect type	NA	spiked, null_a, null_b, null_ab

Table S3.2: Mediation method approaches to address high-dimensionality and compositionality.

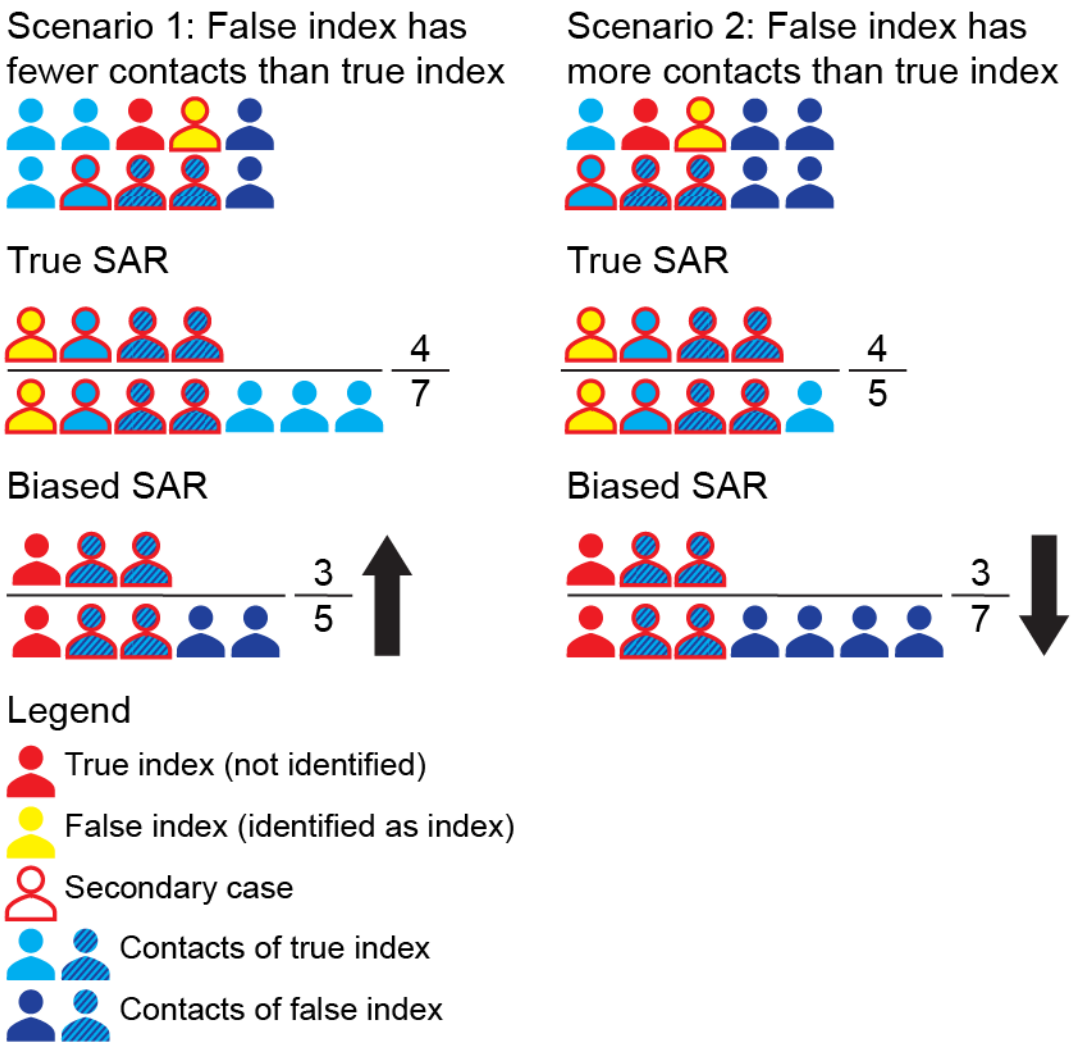
Method	Feature preselection	Mediator model: form	Mediator model: feature transforms	Mediator model: penalty	Outcome model: form	Outcome model: feature transforms	Outcome model: penalty
CCMM	-	Multiple, linear regressions	Additive log-ratio	-	Linear log-contrast	Additive log-ratio	L_1 norm with debias procedure
MedTest	-	-	-	-	-	-	-
Sparse MCMM	-	Dirichlet regression	-	L_1 norm	Linear log-contrast with interactions	Additive log-ratio	Original penalty
HDMA	Sure independence screening	Multiple linear regressions	-	-	Linear	-	L_1 norm with debias procedure

Table S3.2 (Continued)

Method	Feature preselection	Mediator model: form	Mediator model: feature transforms	Mediator model: penalty	Outcome model: form	Outcome model: feature transforms	Outcome model: penalty
HIMA	Sure independence screening	Multiple, linear regressions	-	-	Linear	-	Minimax concave penalty
PCR	-	Multiple, linear regressions	Principal component analysis	-	Multiple, linear regressions	Principal component analysis	-
Naïve	-	Multiple, linear regressions	-	-	Multiple, linear regressions	-	-

Chapter 4

Figure S4.1: Illustration of index case misclassification where the index and secondary cases are misclassified in scenarios outside of the household.



In these scenarios, one of the true index's (red individual) secondary cases is falsely identified as the index case (yellow individual). The true index and false index share some contacts (stripped individuals) but, as this is not a household context, they also have their own contacts (light and dark blue individuals, respectively). Depending on their respective number of contacts, the bias introduced by index case misclassification can go in either direction (Scenarios 1 and 2).